Module : Analyse et fouille de données	
Responsables du Cours: Bouaziz Souhir, Abbes Amal **Enseignants TP**: Barhoumi Chawki, Rekik Amal, Njeh Maissa	**Auditoire:** D-LSI-ADBD **A-U:** 2023-2024

Projet

Analyse et Fouille de données d'une base de données médicales

1) Motivations du projet

Selon l'Organisation Mondiale de la Santé (OMS), le cancer du poumon est l'une des principales causes de décès dans le monde avec une incidence estimée à 2 206 771 en 2020 (voir figure 1). L'ampleur croissante de ce type de cancer constitue un problème majeur de santé publique, d'où la nécessité d'une solution à la fois automatique et performante, permettant une identification précoce à encourage plusieurs développeurs à proposer des systèmes permettant de répondre à ce défi.

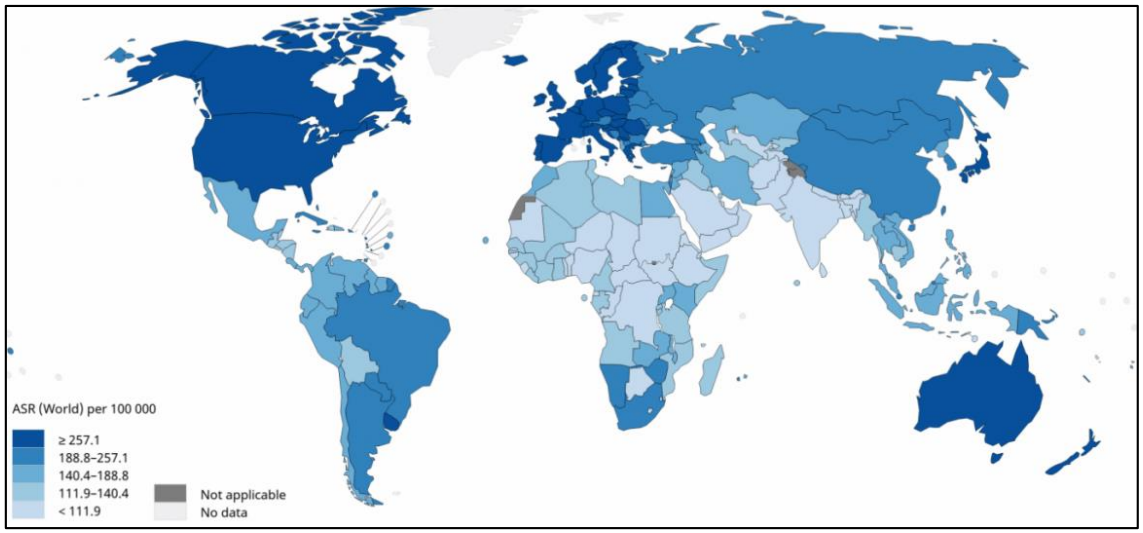

Fig1. Statistiques mondiales du cancer en 2020 : le cancer du poumon toujours le plus mortel

2) Objectives du projet et protocole de réalisation

Vous disposez, en annexe de cette description, une base de données (appelée ***cancer_des_poumons.csv***) portant des patients ainsi que des caractéristiques permettant d'aider à détecter cette anomalie.

L'objective de ce projet est d'appliquer les méthodes d'analyse et de fouille de données étudiées et d'interpréter les résultats trouvés afin d'aider les médecins dans leur prise de décision.

Vous êtes demandés de déposer vos avancements du travail sur Google Classroom (comme pour les exercices de TPs). Le nom du dossier à déposer porte vos noms prénoms et vos groupes. Ce dossier doit se composer de :

- ✓ Dossier contenant les codes sources bien commentés de la partie réalisée en Python ;
- ✓ Un exposé (qui va être représenté dans la dernière séance du TP) présentant les résultats trouvés ainsi que les interprétations effectuées.

3) Description

Il s'agit de développer un système de reconnaissance de la maladie des poumons. Le système proposé sera composé trois phases : phase de préparation comportant le prétraitement et la transformation de données, phase d'extraction des caractéristiques, et phase de data Mining.

Afin de faciliter aux décideurs la prise de décision, ces phases peuvent être représentées graphiquement à travers une interface graphe. **Cette dernière sera considérée comme un travail supplémentaire.**

3.1. La phase de préparation de données

Cette phase consiste à appliquer des méthodes permettent le prétraitement et la transformation des données brutes.

Prétraitement :

- ✓ Télécharger et lire la base de données ***cancer_des_poumons.csv*** existent dans votre Classroom.
- ✓ Interpréter le jeu de données : indiquer le nombre des observations dans la base ainsi que le nombre des caractéristiques.
- ✓ Vérifier s'il existe des observations qui sont manquantes ou NaN, si c'est le cas alors remplacer les valeurs manquantes dans chaque colonne par la moyenne de la variable.

Transformations :

- ✓ Appliquer le codage nécessaire pour transformer les caractéristiques dont les valeurs sont de type chaine de caractères en entier.
- ✓ Vérifier si la base est normalisée ou non (centrée-réduite), effectuer les transformations nécessaires.
- ✓ Afficher la matrice de corrélation puis analyser les dépendances des variables. Quels sont les couples de variables les plus corrélées.

3.2. La phase d'extraction des caractéristiques

Au cours de cette phase, vous êtes demandés d'appliquer l'algorithme d'Analyse en Composantes Principales (ACP) pour extraire des nouveaux facteurs permettant de réduire la dimensionnalité par une meilleure représentation des données.

- ✓ Appliquer sur la base une ACP normée. Interpréter les valeurs propres.
- ✓ Déterminer le pourcentage d'inertie à partir de l'éboulis des valeurs propres. Quelles sont les composantes principales à tirées ?
- ✓ Afficher la saturation des variables et tracer le cercle de corrélation. Interpréter.

3.3. La phase de data Mining

Cette phase consiste à appliquer des méthodes de regroupement pour distinguer entre les observations correspondantes aux sujets malades des sains.

- ✓ Appliquer l'algorithme des **K-means** pour diviser les données en deux classes.
- ✓ Afficher dans un graphe les centroïdes et les données appartenant à chaque classe.
- ✓ Appliquer l'algorithme **Classification Ascendante Hiérarchique** (CAH) pour diviser les données en deux classes.
- ✓ Comparer les résultats des deux algorithmes.

3.4. Représentation graphique du système proposé

Vous êtes demandés dans cette phase, de créer une interface graphique en Python qui permet aux médecins (décideurs) d'interagir et d'appliquer toutes les étapes effectuées.

En effet, le médecin peut importer la base de données puis l'afficher. Par la suite, il peut effectuer tous les prétraitements et les transformations indiqués dans le projet (représentées par des boutons dans l'interface). L'interface permet, ensuite, au médecin de visualiser les résultats de l'ACP ainsi que les méthodes de regroupement appliqués.