**DESCRIPTION DE NOTRE DATASET ‘BREAST CANCER PREDICTION’**

1. **Introduction**

Le jeu de données "Breast Cancer Prediction" est une collection de données qui vise à prédire si une tumeur mammaire est bénigne ou maligne en fonction de diverses caractéristiques. Le jeu de données contient des informations sur 569 tumeurs mammaires, chacune décrite par 32 variables. Les variables incluent des mesures de la taille de la tumeur, de sa forme, de son adhérence, de son uniformité de la cellule, de la taille du noyau, de la chromatin et du nucléole, ainsi que le nombre de mitoses observées.

Chargement des données et bibliothèques

Commençons par importer les bibliothèques nécessaires et charger les données dans notre environnement Python.

1. **Description des données**

Le jeu de données contient 569 observations et 32 variables. Les variables sont les suivantes :

id: identifiant unique de la tumeur

diagnosis: diagnostic de la tumeur (M = maligne, B = bénigne)

radius\_mean: moyenne des distances du centre à chaque point sur le périmètre de la tumeur

texture\_mean: écart type des valeurs de niveaux de gris de l'image

perimeter\_mean: périmètre de la tumeur

area\_mean: surface de la tumeur

smoothness\_mean: variation locale des longueurs de rayon

compactness\_mean: périphérie^2 / aire - 1.0

concavity\_mean: gravité des parties concaves du contour

concave points\_mean: nombre de parties concaves du contour

symmetry\_mean: symétrie

fractal\_dimension\_mean: côte fractale - "environ 1"

radius\_se: écart-type des distances du centre aux points du périmètre

texture\_se: écart type des valeurs de niveaux de gris de l'image

perimeter\_se: écart type du périmètre

area\_se: écart type de la surface

smoothness\_se: écart type des variations locales des longueurs de rayon

compactness\_se: écart type de la périphérie^2 / aire - 1.0

concavity\_se: écart type de la gravité des parties concaves du contour

concave points\_se: écart type du nombre de parties concaves du contour

symmetry\_se: écart type de la symétrie

fractal\_dimension\_se: écart type de la côte fractale - "environ 1"

radius\_worst: plus grand des moyennes des distances du centre à chaque point sur le périmètre de la tumeur

texture\_worst: écart type des valeurs de niveaux de gris de l'image

perimeter\_worst: plus grand périmètre de la tumeur

area\_worst: plus grande surface de la tumeur

smoothness\_worst: plus grande variation locale des longueurs de rayon

compactness\_worst: plus grande périphérie^2

1. **Visualisation des données**

Nous pouvons utiliser des graphiques pour visualiser les caractéristiques de notre jeu de données. Commençons par visualiser la distribution des valeurs dans chaque colonne pour les tumeurs malignes (M) et bénignes (B) séparément.

Nous pouvons voir que certaines caractéristiques, comme texture\_mean et area\_worst, ont des distributions très différentes pour les tumeurs malignes et bénignes. D'autres caractéristiques, comme fractal\_dimension\_mean, ont des distributions similaires pour les deux types de tumeurs.

1. **Corrélation entre les caractéristiques**

Nous pouvons également examiner la corrélation entre les caractéristiques pour voir s'il y a des relations importantes à noter.

Nous pouvons voir que certaines caractéristiques sont fortement corrélées entre elles, comme radius\_mean, perimeter\_mean et area\_mean. Nous pouvons également voir que certaines caractéristiques sont fortement corrélées avec la variable cible, comme concavity\_mean, concave points\_mean et perimeter\_worst.

1. **Relation entre les caractéristiques et la cible**

Enfin, nous pouvons examiner la relation entre chaque caractéristique et la variable cible (diagnosis).

Nous pouvons voir que certaines caractéristiques, comme texture\_mean et concavity\_mean, ont des distributions très différentes pour les tumeurs malignes et bénignes. D'autres caractéristiques, comme fractal\_dimension\_mean, ont des distributions similaires pour les deux types de tumeurs.

1. **Conclusion**

Dans cette EDA, nous avons examiné le jeu de données "Breast Cancer Prediction" et effectué plusieurs analyses exploratoires pour mieux comprendre les caractéristiques des tumeurs mammaires et leur relation avec la variable cible. Nous avons constaté que certaines caractéristiques sont fortement corrélées entre elles et avec la variable cible, et que certaines caractéristiques ont des distributions très différentes pour les tumeurs malignes et bénignes.