**TEXTE ORAL MATHS**

***DIAPO 1***

Le sujet sur lequel nous avons travaillé cette période pour le module « Modélisation et Intelligence Artificielle » s’intitule « PREDICTION DU CANCER DU SEIN ».

***DIAPO 2***

Pour commencer, je vais poser et définir le problème auquel nous avons cherché à répondre, puis je vais vous décrire la base de données utilisée. Clara prendra ensuite le relais pour expliquer comment nous avons rendu ces données utilisables, puis nous expliquerons comment nous avons appliqué les 3 différents modèles aux données existantes. Et nous finirons par conclure sur quel modèle est le plus efficace pour répondre à notre sujet.

***DIAPO 3***

Dans un premier temps, nous avons dû extraire le problème du sujet.

A partir d’une base de données contenant des images échantillons, nous voulons implémenter du Machine Learning pour détecter la présence de cellules cancéreuses.

A partir de ces informations, nous avons retenu le problème suivant : **Pouvons-nous détecter des cellules cancéreuses sur une coupe d’échantillon de tumeur afin de localiser précisément les zones cancéreuses et ainsi évaluer la gravité du cancer ?**

Pour répondre à cette question, le modèle va prendre comme entrée une image provenant de la base de données, et en sortie il dira s’il y a cancer ou non. Il s’agit donc d’une classification.

***DIAPO 4***

Notre première tâche a été de trouver une base de données exploitable, avec un nombre assez important d’image. La base de données que nous avons choisie est constituée de 277 524 (deux cent soixante-dix-sept mille cinq cent vingt-quatre) photos dont 198 738 où il n’y a pas de cellules cancéreuses et 78 786 où il y en a. Ces données sont représentées par les diagrammes en barre qui s’affiche à l’écran.

Cette base de données a été créé avec les photos issues de 279 patientes, pour chaque patiente il y a un dossier avec deux sous dossiers : le dossier 0 qui contient les photos non cancéreuses et le dossier 1 les photos où des cellules cancéreuses ont été identifiées. Le nom de chaque image contient le numéro de la patiente, les coordonnées de l’image ainsi que le diagnostic associé (1 pour positif, 0 pour négatif).

Avant de pouvoir utiliser ces données, nous avons dû vérifier s’il n’y avait pas de données manquantes ou non exploitables :

Nous avons rencontré des images qui n’étaient pas toutes de la même taille, on a donc redimensionné celles trop grande à la taille 50x48 et nous avons supprimé les images trop petites.

Après avoir fait ce traitement, notre base finale et exploitable contient 275 246 images.

***DIAPO 8***

Le premier modèle que nous avons appliqué à notre petite base de données est la classification avec les plus proches voisins.

C’est une classification par apprentissage supervisé, c’est à dire que le modèle sait à quelle classe appartiennent les données. Pour chaque image, le programme regarde k voisins les plus proches (k une constante prédéfinie). Il classe ensuite l’image étudiée dans la catégorie dans laquelle se trouve la majorité de ses k voisins les plus proches.

Comme on peut le voir à l’écran, cette méthode à mit environ 8 heures à s’exécuter et sa précision est de 79.5%

***DIAPO 9 (MOITIE !!)***

Le second modèle que nous avons appliqué est la classification avec arbres de décisions. Il s’agit ici aussi d’un apprentissage supervisé. Le programme crée un arbre de décision où les nœuds sont des critères et les feuilles des valeurs discrètes. Suivant les valeurs des feuilles, l’image sera classée dans une catégorie ou dans l’autre.

Ce modèle à mit 11heures pour s’exécuter et à obtenue une précision de 80.6%. Cette précision est détaillée dans la matrice de confusion affichée à l’écran et que Clara va maintenant présenter.