**PREDICTION CANCER DU SEIN**

**I – Identifier et définir le problème**

Le problème est : 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 ?

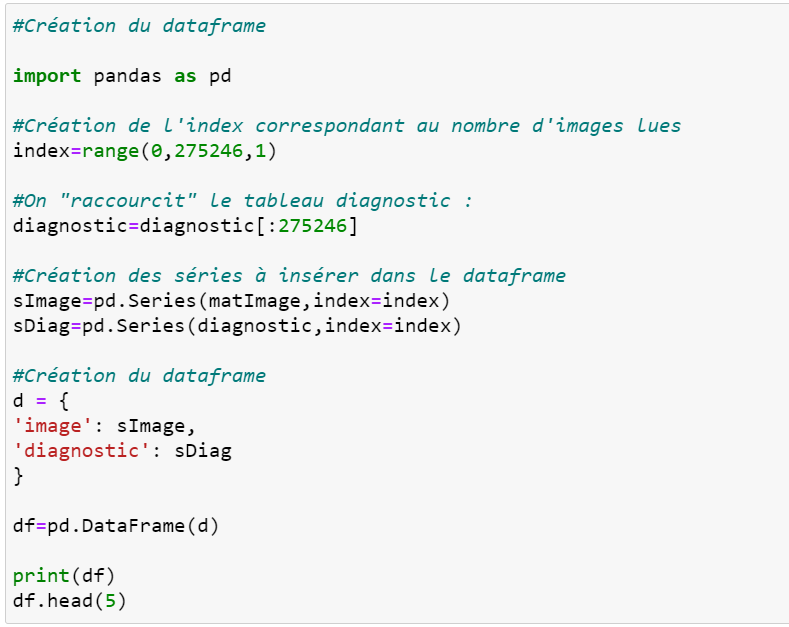
Pour répondre à cette question, le modèle va prendre comme entrée une image provenant de la base de données, et en sortie il dira s’il y a cancer ou non. Il s’agit donc d’une classification.

**II – Comprendre les données**

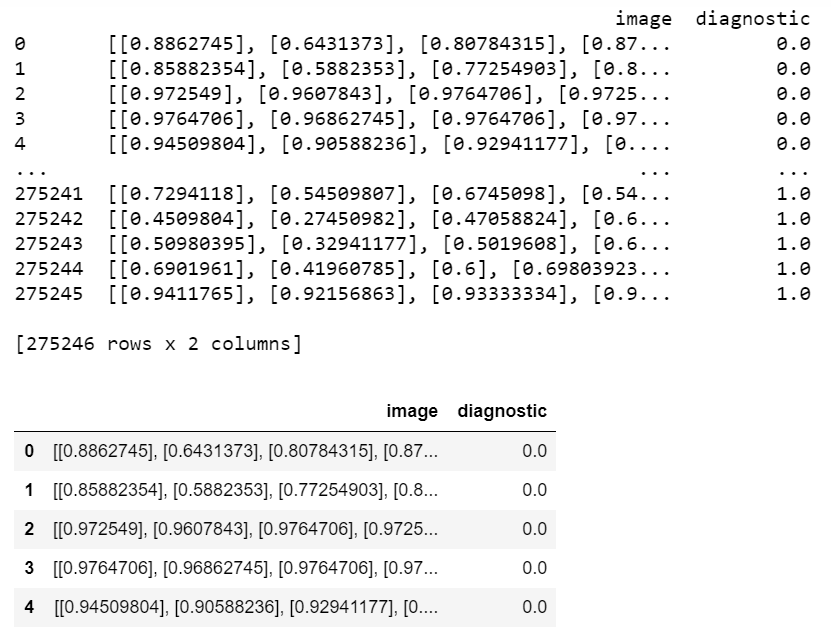
* Charger la base de données dans un tableau

La base de données que nous avons choisie est constituée de 277 524 photos dont 198 738 où il n’y a pas de cellules cancéreuses et 78 786 où il y en a. Cette base de données a été créé avec les photos 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.

Pour pouvoir utiliser les données, nous les chargeons dans un dataframe. Ce dataframe est composé d’une colonne avec l’image (c’est à dire un tableau à 7200 lignes correspondant au nombre de pixels multiplié par 3, car l’image est en couleur et comporte donc trois valeurs pour chaque pixel) et d’une deuxième colonne avec le diagnostic, c’est à dire 1 si la photo présente des cellules cancéreuses et 0 sinon. Pour cela, nous avons utilisé le code suivant :



Nous obtenions ainsi le dataframe suivant :



Par la suite, nous avons rencontré des problèmes causés par la manière dont nous avions sauvegardé les données. Nous avons donc décidé de recharger toutes les images sous une autre forme. Nous avons stocké les images dans un tableau de dimension 275246\*7200. La première dimension correspond au nombre d’images sélectionnées (cf. point suivant). Dans un deuxième tableau à une seule dimension, nous avons inscrit le diagnostic des images (0 si négative, 1 si positive). Nous avons ensuite sauvegardé ces tableaux dans deux fichiers .npy.

* Sélectionner les photos

Lors de la création de notre fichier, nous nous sommes rendues compte que toutes les photos n’avaient pas la même taille. En effet, la plupart des photos sont de la taille 50x50 mais certaines mesurent 50x48 ou 48x50.

Pour résoudre ce problème, nous avons choisi de réduire toutes les images à la même taille : 50x48 (en pratique, on enlève les deux dernières colonnes à toutes les images de taille 50x50). Après avoir fait ça, nous nous sommes rendu compte qu’il restait une infime partie des images qui étaient encore plus petites, nous avons donc décidé de les ignorer et de ne pas les garder afin de ne pas perdre davantage de données.

**III – Choisir une méthode**

**IV – choisir un modèle et comprendre les grandes lignes**

(Être capable en 2-3 phrases de faire comprendre à vos collègues) le fonctionnement du modèle et le mécanisme d'apprentissage sous-jacent

**V – Appliquer le modèle et commenter les résultats**