DATAFRAME

-lecture des images en parcourant les dossiers,

->on obtient alors une matrice d'images sous forme d'un tableau d'une ligne et 7200 colonnes

-> un tableau avec les diagnostics (0 ou 1)

-pour créer le dataframe : on crée l'index, on raccourcit le tableau diagnostic pour qu'il ait le bon nombre de lignes

-on crée deux séries, chacune avec l'index

-on compile ces deux séries en un dataframe

-on obtient donc un dataframe qui comporte 275246 lignes et deux colonnes, une pour les images et une pour le diagnostic qui lui est associé

==>> Lorsque nous avons voulu utiliser ce dataframe pour faire la classification, nous avons rencontré des problèmes d'incompatibilité. Nous avons donc dû revoir notre manière de gérer les données.

TABLEAU

-on lit toujours de la même manière les images, et on est reparties de notre matrice d'images et de notre tableau de diag

-tout d'abord on transforme cette matrice en un tableau numpy, duquel on change la taille pour qu'il comporte le même nombre de lignes que d'images et 7200 colonnes

-ensuite on raccourcit le tableau des diagnostics et on concatène les deux de sorte à obtenir un tableau de 7201 colonnes, dont la dernière représente les diagnostics

==>On a ainsi un format adapté à la suite de notre projet

PREPARATION

Puisque notre sujet appelle à une classification par apprentissage supervisé, nous devons diviser notre base en 2 parties. D'une part, les données qui serviront à l'entrainement, avec lesquelles le programme va déterminer la classification à effectuer, et d'autre part les données qui serviront au programme à tester la précision de son entrainement. Nous avons décidé d'attribuer 75% des données à l'entrainement et 25% au test.

Comme notre dataset de départ est très important, nous avons dû réduire la taille de notre base de données avant de pouvoir effectuer les calculs. Nous avons conservé 102 592 images, dont 24504 positives et 78088 négatives. Notre base d'entrainement est donc composée de 76944 images et la partie « test » en contient 25648.

MATRICE DE CONFUSION

Horizontalement, on lit les diagnostics réels des images. Verticalement, on lit les résultats donnés par la machine après son entrainement. On peut donc voir que la machine a détecté 2619+3788=6407 images positives mais seulement 3788 d'entre elles étaient réellement négatives. La machine a effectué une mauvaise prédiction pour 2367 images qu'elle a considérées négatives alors d'elles étaient positives (case en bas à gauche)

SVM

La classification Support Vector Machine est un apprentissage supervisé qui cherche à séparer les deux classes d'images par un plan. Elle consiste en la recherche du plan pour lequel l'écart entre le plan et les échantillons les plus proches, appelé marge, est le plus élevé.

Ce modèle est le plus efficace puisqu'il permet à la machine d'effectuer une bonne prédiction dans plus de 86% des cas.
Pour obtenir d'encore meilleurs résultats, nous pourrions envisager de traiter davantage de données (plusieurs paquets contenant peu d'images mis en commun par la suite).