Projet de Programmation Python
M1 MIAGE IDA – 2021-2022

Contexte

Dans ce projet, nous vous proposons d'analyser un dataset sur les stades de maturité dentaire. Pour chaque dent d'un secteur de l'arcade dentaire (incisives, cuspides, prémolaires, molaires), le stade de maturité sur des radiographies dentaires a été défini.
Nous avons donc 8 dents et leur stade de maturité (VAL_I1, VAL_I2, VAL_C1, VAL_P1, VAL_P2, VAL_M1, VAL_M2, VAL_M3). La figure 1 représente un panoramique dentaire vous permettant d'imaginer le positionnement des dents. Vous noterez ici l'absence de la dent M3 (dent de sagesse).

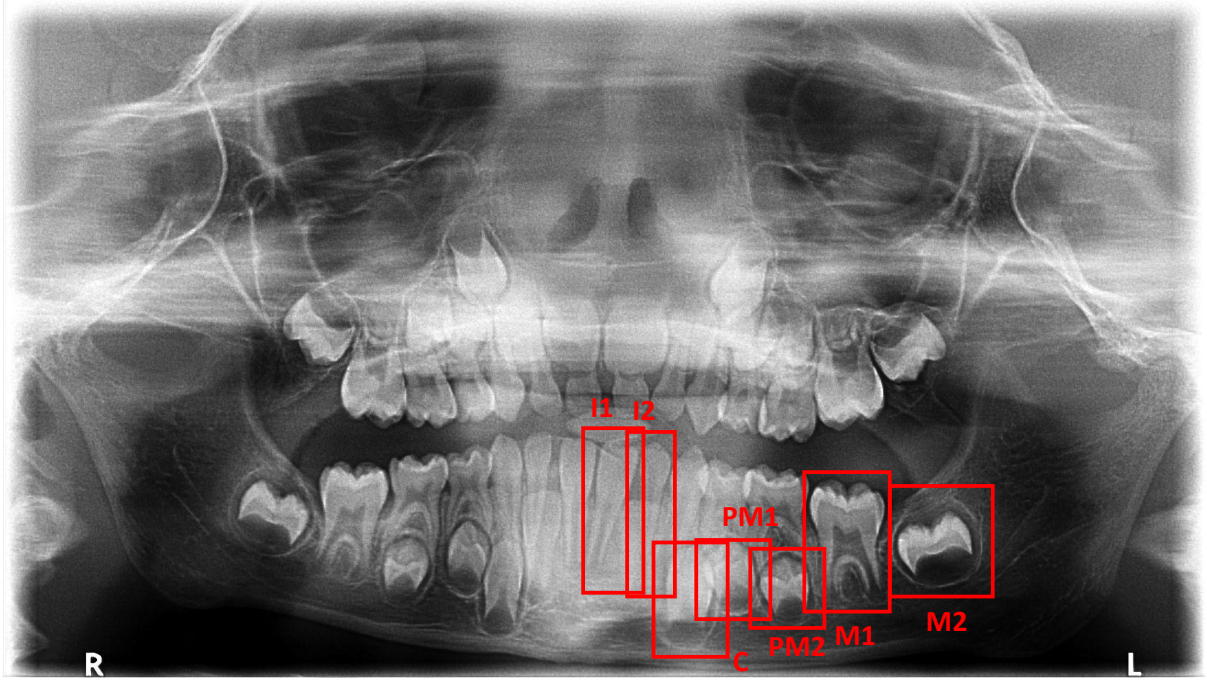

Figure 1 : Exemple de panoramique dentaire avec l'indentification des différentes dents.

Les stades ont été définis de 0 à H (0, 1, A, B, C, D, E, F, G, H). 0 signifie une maturité très faible et H signifie la maturité la plus élevée.
Le sexe du patient est aussi fourni (PAT_SEX).

A long terme, l'objectif de ce dataset est d'étudier les possibilités de retrouver l'âge du patient à partir du stade de développement des différentes dents. C'est un projet d'intérêt en particulier pour la médecine légale par exemple. Dans ce cadre, des seuils d'intérêts ont aussi été définis : 13 - 16 - 18 - 21.

Un dataset contenant 2847 patients vous est fourni (voir sur l'espace de cours).

Livrable

Vous devez construire un Notebook Jupyter (en Python) permettant de répondre aux besoins exprimés ci-dessous. Ce notebook doit être parfaitement annoté et proposer une courte analyse des statistiques demandées. Vous devrez rendre en le déposant sur l'espace de dépôt dans l'espace de cours correspondant. Celui-ci devra être livré avant **Mardi 23 Novembre 2021 à 23h59**. Le projet doit être réalisé par **groupe de 2 étudiants maximum**.

Besoins

1/ Charger le dataset dans un dataframe Pandas

2/ Pour chacune des caractéristiques (hors sexe), calculer la moyenne et l'écart-type des valeurs. Proposer une visualisation de ces statistiques.

3/ Pour chacune des caractéristiques, afficher la distribution des valeurs. Que pensez-vous de ces distributions ?

4/ Vous devriez avoir remarqué qu'un certain nombre de données sont marquées *NULL* car non renseignées. Proposer une ou plusieurs méthodes permettant de remplacer ces données par une valeur cohérente vis-à-vis du dataset. Ces méthodes peuvent varier d'une caractéristique à l'autre.

5/ En réutilisant les fonctions développées dans les besoins 2 et 3, évaluer l'impact de ces méthodes sur le dataset.

6/ Développer un algorithme permettant de diviser le dataset en 2 dataset : 1 pour l'entrainement des futurs algorithmes de machine learning et l'autre pour leur évaluation sur des données jamais vu au précédent. Pour cela, l'algorithme :

- Prend en entrée : le dataset, la répartition de la division (par exemple 70% pour l'entrainement et 30% pour l'évaluation)
- Retourne : 2 datasets (entrainement et évaluation)
- Contraintes :
    - Une ligne de données ne peut être présente que dans un et un seul dataset de sortie,
    - La répartition doit être complètement aléatoire et différente à chaque appel de la fonction,
    - L'algorithme doit être parfaitement utilisable quel que soit le dataset.

7/ En utilisant les algorithmes précédents, évaluer l'impact de la répartition entrainement/évaluation sur le dataset.