HACKATHON SUR UNE CAMPAGNE DE DON DE SANG

CONCEPTION DE LA CARTE

Nous débutons avec l'importation des bases nécessaires

-Par suite, on poursuit avec le nettoyage de la base notamment des variables "Arrondissement de résidence" et "Quartier de

En réalité, lors de la collecte de données, plusieurs noms de quartier ont été mal écrits, le quartier de certains individus n'a pas été renseigné. Certains

arrondissements aussi ont été mal renseigné. On écrit une fonction "corriger_quartier" pour corriger les quartiers. Cette fonction est donc utilisée pour corriger

les quartiers.

Normalisation des arrondissements

Les individus dont l'arrondissement a été écrit DOUALA (non précisé sont transformés en DOUALA. En le faisant, l'arrondissement DOUALA qu'on vient de créer contient les quartiers de DOUALA 1; DOUALA 2; DOUALA 3.... Pour attribuer leur véritable arrondissement, on écrira une boucle for qui va identifier les quartiers qui se répètent à la fois à DOUALA et dans d'autres arrondissements. Des qu'un quartier de ce genre est identifié, on substituera DOUALA par son vrai arrondissement, mettant ainsi chaque quartier dans son vrai arrondissement. Ensuite vient la phase du géocodage des adresses pour la conception de la carte.

Pour la conception de la cartes, nous avons besoin de la position géographique de chaque quartier. En réalité la position exacte de chaque individu ne peut être connue car les données géographiques dont nous disposons sont l'arrondissement et le quartier: on se servira donc de la bibliothèque geopy de python pour le faire. On se sert essentiellement des variables Arrondissement et Quartier pour géocoder les adresses de quartier du maximum d'individus. On obtient donc la longitude et la latitude qui seront ajoutées comme nouvelles colonnes dans

la base et comme le géocodage étant une longue opération, on enregistre le fichier en le renommant "LE_FICHIER_AVEC_LES_MEILLEURS_GEOCODES.xlsx". Ce fichier est ensuite importé pour l'établissement de la carte. Mais juste avant, APRES LE GEOCODAGE 264 individus n'ont pas pu être géocodés à cause du fait que le quartier a été mal écrit(il n'existe donc pas), c'est à dire que la position exacte du quartier de 264 individus n'est pas connue. Par contre on connaît leur Arrondissement. On va donc les placer à la moyenne des positions de tous les individus de l'arrondissement. Pour bien identifier chaque individu sur la carte le marqueur de position de chaque personne va indiquer des informations sur le Genre, le sexe, la date de naissance, le poids, la taille…. On conçoit donc la carte avec folium.

CONDITIONS DE SANTE ET ELIGIBILITE

La base contient des variables telles que la Religion, Profession qui ne sont pas correctement écrits et normalisés. On écrit un dictionnaire pour corriger les noms des différents métiers et Religions. Celà nous permet de concevoir certains graphes.

-Conditions de santé. J'utilise les variables comme "Raison d'indisponibilité totale", "l'âge", "le taux d'hémoglobines"… pour la conception des graphiques avec plotly.

ANALYSE DES CLUSTERS DE DONNEURS DE SANG

Malgré les données incomplètes, l'analyse a révélé 4 profils distincts de donneurs, le score de silhouette étant de 0,63. Les résultats suggèrent une forte homogénéité démographiques, mais mettent en lumière des opportunités pour des campagnes ciblées

ANALYSE DE LA RECURRENCE DES DONS

La récurrence des dons est de 43% et les facteurs qui influent sont le genre et la profession(Khi 2 et p value). En réalité, il nous a fallu segmenter chaque variable en plusieurs sous partie et étudier la fidélité variable après variable.

API

Grâce à la bibliothèque joblib, j'encapsule le modèle et l'encodage des variables utilisées lors de la conception du modèle. Globalement, les variables utilisées pour la prédiction sont "Niveau_d_etude", "Genre","Situation_Matrimoniale", "Déjà_donné_le_sang", "Religion", "Profession", "Âge", "Taux_d_hémoglobine".

OBSERVATIONS SUR LA BASE

La base contient beaucoup de variables inutilisables, qui ne contiennent pas assez de données. Les variables importantes comme les noms des quartiers sont mal écrits.

BIEN VOULOIR INTRODUIRE LES FICHIERS: "Updated Challenge dataset.xlsx", "LE_FICHIER_AVEC_LES_MEILLEURS_GEOCODES.xlsx" dans le répertoire courant avant exécution du code.

NB: L'exécution du code peut rencontrer un soucis à cause l'incompatibilité internes de certains modules. Bien veuillez à s'assurer que les modules ont des versions compatibles entre elles.