Atelier Composants IA

21 septembre 2023

Présenté par : Lina EL JERROUDI / Marine BRIOUDE / PAUL RIMBAUD

Contexte

Nous avons un jeu de données qui donne une liste d'inspections sanitaires chez des établissements avec le résultat de cette évaluation et toutes les informations liées aux établissements.

Problématique

Ce que nous voulons déduire à partir de ce jeu de données est si l'évaluation sanitaire est liée au type d'établissement, et par les communes.

Traitement

Afin de pouvoir répondre à cette problématique, nous avons commencé par le nettoyage des données.

Nous avons remarqué que certaines données comme les 3 colonnes 'Agrement', 'geores' et 'Numero_inspection' étaient inutiles pour notre problématique.

Nous avons ainsi remarqué que les colonnes 'filtre', 'ods_type_activite' et 'APP_Libelle_activite_etablissement' contient des informations redondante et dans le but de fusionner les valeur nous avons pris la décision de modifier le fichier brut en remplissant les valeur **Autres** dans le champ 'ods_type_activite' par les valeurs existant dans 'filtre'.

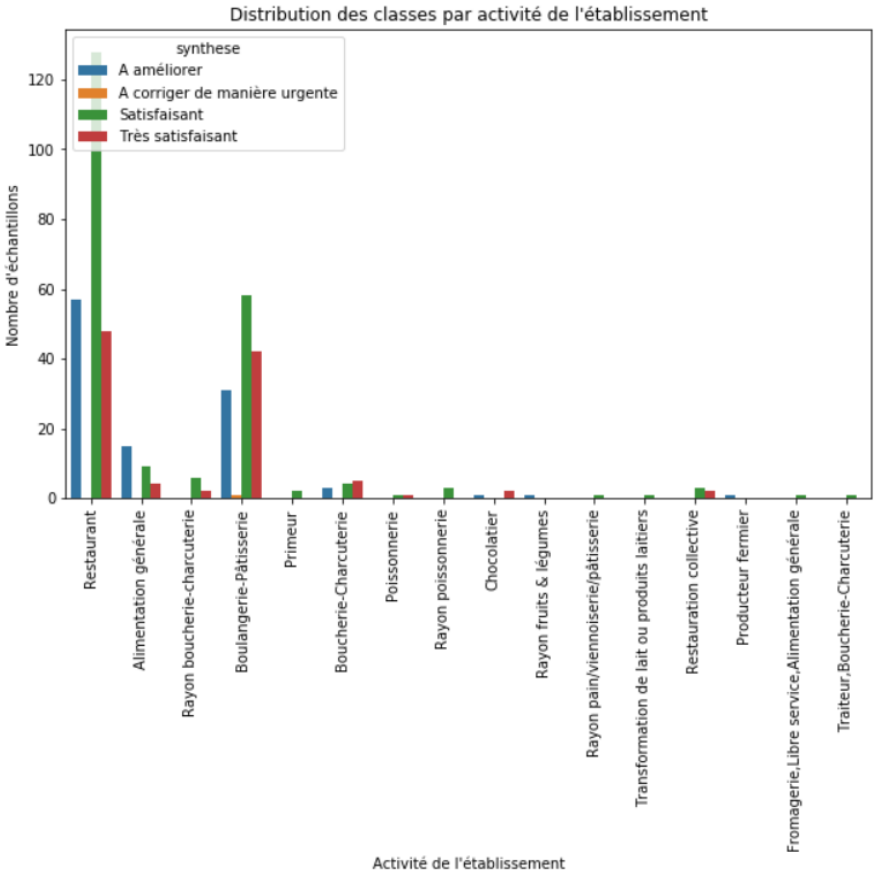
Concernant la colonnes 'Code_postal' nous avons remarqué que quelques valeurs contenaient moins que 5 chiffres et afin de résoudre cette erreur nous avons modifier le type de champ de float à string et en passant une condition si la valeur contient 4 chiffres on rajoute un 0 au début.

Choix du modèle

Pour notre modèle et puisque les caractéristiques et la variable cible choisies sont des catégories données qualitative et quantitative exemple **synthèse** : ‘Satisfaisant’, ‘À corriger de manière urgente’ , ‘À améliorer’ et ‘Très Satisfaisant’ nous avons choisi de mettre un modèle de régression logistique.

Résultat

Pour les résultats obtenus sur les différents modèles fait, nous avons eu des résultat satisfaisant par rapport à la qualité des données dans le fichier brut, avec une précision entre 0.52 et 0.62 qui est plutôt pas mal.



Les deux plus gros secteurs évalués sont les boulangeries et les restaurants, ils sont tous satisfaisants.

Cependant, la prédiction des autres types de secteurs risque d'être moins précise du fait du manque de données cohérentes.