TP - Atelier composant métier

Table des matières

[**1. Profilage des données 2**](#_nwh9d2w5cuca)

[**2. Problématique 2**](#_12azzjmfbxzm)

[**3. Formulation des hypothèses 2**](#_pikatetuwhsc)

[**4. Analyse du jeu de données 2**](#_knn0sf6xh618)

# 

# Profilage des données

Colonne agrément : taux de satisfaction (84% vide)

# Problématique

Nous disposons de plusieurs informations sur différents établissements. Pour une date donnée, une note de satisfaction est attribuée à un établissement donné. Nous cherchons à savoir quels types d’établissements ont les meilleures notes de satisfaction par rapport à leur code d’hygiène.

# Formulation des hypothèses

Le type d’établissement influe sur la note de propreté.

La période de l’année (saisonnalité) influe sur la note de propreté.

L’heure de l'inspection peut influer sur la note.

Le département influe sur la note.

# Analyse du jeu de données

Notre jeu de données est composé de 32720 lignes et de 13 colonnes. Nous remarquons que la colonne ‘Agrement’, représentant le score de l’évaluation sanitaire, comporte plusieurs valeurs manquantes (plus de 80% de valeurs vides). Nous n’allons donc pas l’inclure dans notre analyse.

La colonne ‘APP\_Libelle\_activite\_etablissement’ représente l’activité de l'établissement (ce que nous appellerons ‘type de l'établissement’) et cette dernière comporte de multiples valeurs dont certaines sont répétées une centaine de fois alors d’autres ne dépassent pas les deux ou trois occurrences. Afin que notre futur modèle ne soit pas biaisé, nous proposerons de regrouper certaines valeurs sous un seul libellé, ex : ‘Restaurant’ et ‘Restaurant, Boucherie-Charcuterie’ afin d’avoir un jeu de données plus équilibré.

La colonne agrements est inutilisable car vide a plus de 80%.

La plupart des données disponibles sont des données qualitatives donc notre modèle a de grandes chances d'être un modèle de classification.

# Modèles

## Premier modèle

Le premier modèle que nosu avons réalisé est un modèle de classification multiclass. Dans ce dernier, nous nous basons sur la région, la temporalité et le type d’établissement.

La précisionobtenus par ce derneir est de 0,54. Ce n’est pas encore satisfaisant et ne nous permet pas de répondre à nos hypothèses.

## Second modèle