TP - Atelier composant métier

Table des matières

1. Profilage des données

Colonne agrément : taux de satisfaction (84% vide)

2. Problématique

Nous disposons de plusieurs informations sur différents établissements. Pour une date donnée, une note d'évaluation sanitaire est attribuée à un établissement donné. Nous cherchons à savoir les paramètres qui influencent la note de d'évaluation sanitaire pour les établissements.

3. Formulation des hypothèses

De part la problématique du sujet, trois hypothèses ont alors émergées et nous avons cherché afin d'y répondre :

- Le type d'établissement influe sur la note de propreté.
- La période de l'année (saisonnalité) influe sur la note de propreté.
- L'emplacement géographique de l'établissement influe sur la note.

4. 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 que 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.

Les données que nous cherchons à prédire sont qualitatives. Nous opterons alors pour un modèle de classification.

Nous avons réalisé diverses visualisations de ces données afin d'avoir un meilleur aperçu de ces dernières.

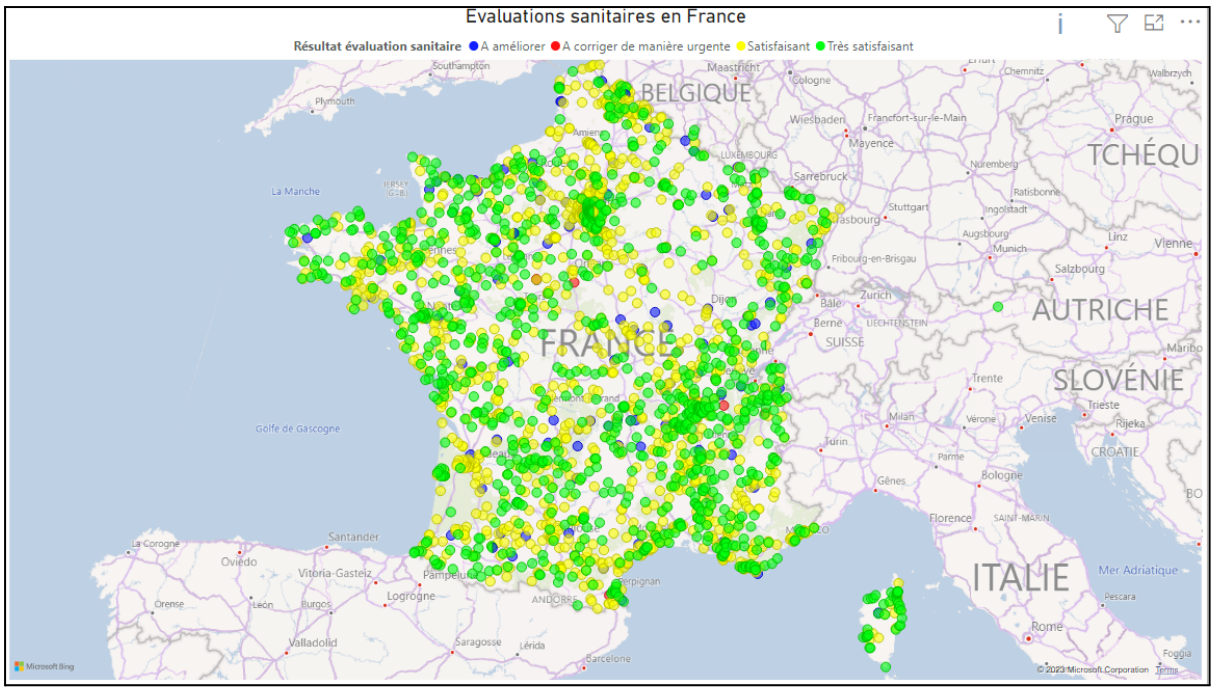

Figure 1 : Répartition des établissements en fonction de l'emplacement géographique et de la note d'évaluation sanitaire

Le graphe ci-dessus pourrait nous amener à penser que Paris contient plus d'établissements dont les conditons laisse à désirer mais il faut prendre en compte que c'est la plus grande ville de France avec donc le plus grand nombre d'établissmeents.

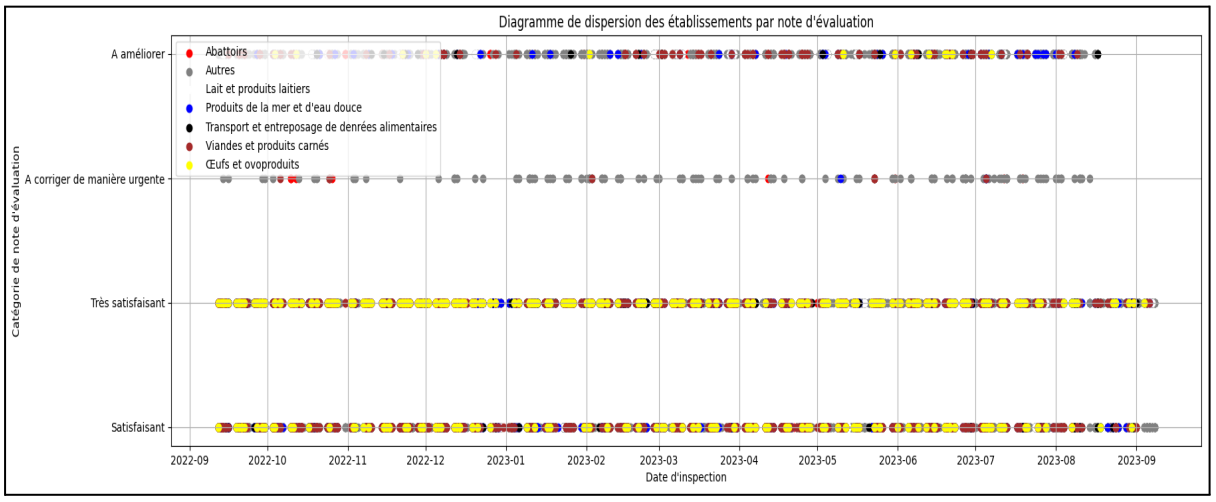

Figure 2 : Répartition des établissements en fonction leur type, de la date et de la note d'évaluation sanitaire

À travers le graphe ci-dessus, nous cherchons, en premier temps, à identifier la note la plus attribuée à un établissement en fonction de son type, et en deuxième temps, à observer l'impact qu'a le facteur temporel sur la note d'évaluation sanitaire pour les établissements. Nous cherchons par exemple à voir si un type d'établissement reçoit de meilleures notes au fil du temps.

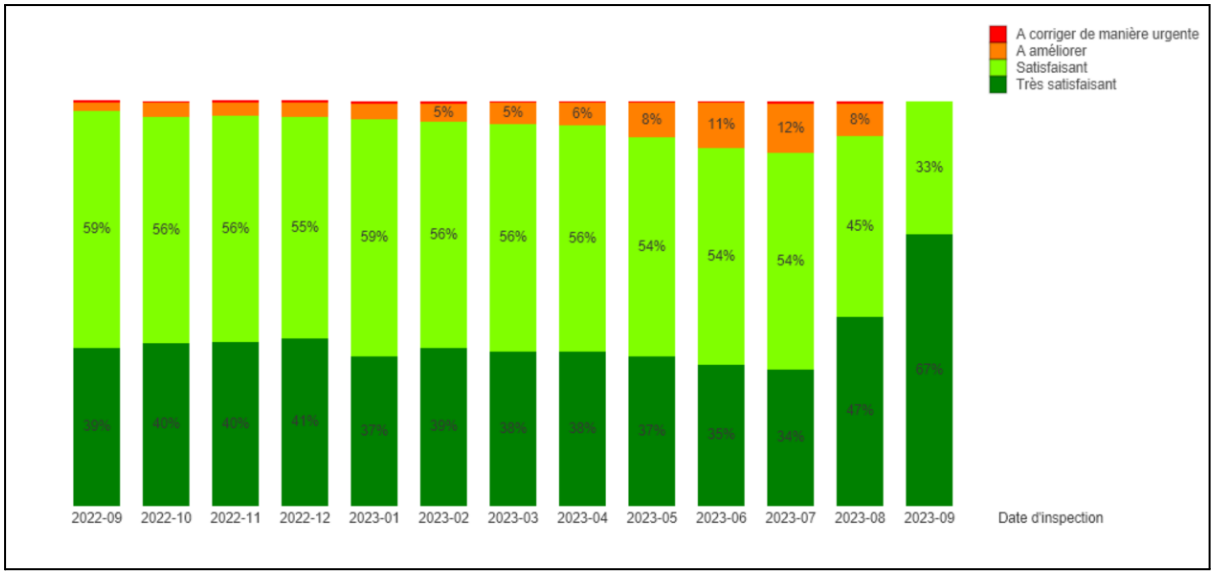

Figure 3 : Évolution des notes d'évaluation sanitaire au fil des mois

Nous remarquons sur le graphique ci-dessus qu'à l'approche de l'été, la satisfaction vis à vis du code sanitaire des établissements a tendance à baisser. Ce graphique pourrait ainsi soutenir notre hypothèse concernant l'impact de la saisonnalité sur l'évaluation sanitaire des établissements.

5. Modèles

5.1. Premier modèle

Le premier modèle que nous avons réalisé est un modèle de classification multiclass, celui du Random Forest, afin de prédire l'évaluation sanitaire qui se décline en 4 valeurs différentes. Pour ce faire, nous nous sommes basés sur la région, la temporalité et l'activité de l'établissement.

Nous avons tout d'abord conduit notre analyse en gardant toutes les valeurs de la colonne 'APP_Libelle_activite_etablissement', en incluant celles qui n'apparaissent qu'une seule fois, et nous avons obtenu un score de prédiction de 57%. Nous avons ensuite gardé que les valeurs qui apparaissent au moins une centaine de fois et nous avons obtenu une accuracy de 56%. Garder les valeurs les plus redondantes n'a donc pas d'incidence sur le résultat des prédictions.

Nous avons également testé ce même modèle en utilisant cette fois-ci la colonne "ods_type_activite" qui comporte bien moins de valeurs de 'APP_Libelle_activite_etablissement', néanmoins nous pouvons estimer que des informations sont perdues car beaucoup de valeurs de 'APP_Libelle_activite_etablissement' ont comme équivalent 'Autres' sous la colonne 'ods_type_activite'.

5.2. Second modèle

Pour notre deuxième modèle, nous souhaitons réaliser les mêmes prédictions, néanmoins nous chercherons à regrouper les types d'établissement car nous disposons d'un grand nombre de valeurs différentes sous la colonne 'APP_Libelle_activite_etablissement'.

Ainsi, nous utiliserons plutôt la colonne 'Filtre' et nous établirons un ordre de priorité dans les catégories que nous aurons définies, par exemple, les boucheries ont une plus grande priorité que les restaurants en ce qui concerne l'aspect sanitaire, donc si un établissement est un restaurant boucherie, la boucherie ayant une plus grande priorité, sera classé dans boucherie et non restaurant. Après avoir regroupé nos valeurs et entraîné notre modèle, nous obtenons cette fois-ci une accuracy de 59%.