Projet Science des Données 4
2026

CROENNE Victor 22308946,
FAIZANDIER Ambre 22301675,
MIRANDA Anthony 22301243,
VALENTIN Nina 22301255

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique

Université Paul Valéry, Montpellier

Mail informations
(semaine 1)

Bonjour,

Notre groupe pour le cours “Science des données 4” est composé de :

- CROENNE Victor 22308946,
- FAIZANDIER Ambre 22301675,
- MIRANDA Anthony 22301243 et
- VALENTIN Nina 22301255

Nous avons choisi ces deux bases de données :

- https://www.kaggle.com/datasets/shashwatwork/municipal-waste-management-cost-prediction
- https://www.kaggle.com/datasets/hassnainzaidi/garbage-classification

Notre problématique est la suivante : Comment l’intelligence artificielle peut-elle soutenir une gestion intelligente des déchets en combinant la prévision des flux de déchets communaux et l’assistance au tri pour les citoyens à partir d’images ?

En espérant que notre projet vous plaise autant qu’à nous !
Bonne fin de journée !
Cordialement,
L’ensemble du groupe

Compte Rendu 1

(semaine 2)

Pour cette deuxième semaine, nous nous sommes donné pour objectif de bien fixer la structure de notre projet en liant nos deux bases de données. Pour rappel, l'une d'elle est une base de données d'images de déchets (cartons, verre, métal, papier, plastique et poubelle) et l'autre est une base de données sur des déchets municipaux (région, nom de la municipalité, pourcentages de déchets par catégories, ….).

L'objectif de notre projet est de prévoir le taux de production de déchets dans chaque classe (déchet organique , papier , verre , bois , métal , plastique, Électrique/Électronique, textile , autres) pour une ville donnée avec ces caractéristiques.

Dans un premier temps, nous avons réfléchi à comment lier nos bases de données, puisqu'il y a des images d'un côté et des données quantitatives de l'autre.

Dans un second temps, nous avons testé des modèles.

Pour cela nous avons décidé d'effectuer une régression linéaire pour chaque classe de déchet afin de prédire son taux.

Dans le cadre d'un modèle très simplifié, nous avons commencé par effectuer une régression linéaire à l'aide d'un code python fonctionnant comme ceci: les données sont lues par le script et analysées. Le code demande alors (à titre d'exemple) une valeur concernant le PIB et la densité de population par km². Avec ces valeurs, une prédiction de production de déchets est calculée pour les trois types de déchets suivants: organique, métal et verre. Un coefficient de détermination est aussi calculé, et qui est dans notre cas assez haut pour chaque simulation montrant ainsi que le modèle ne reflète qu'une faible part de la variance et qu'il est nécessaire que ce dernier soit complexifié afin de prendre en compte véritablement ces prédictions.

Nous nous sommes également posés la question : Est- ce que les variables géographiques, telles que la région et la province, influencent de manière significative les variations de nos prédictions, ou si ces modèles pouvaient être généralisés à une plus grande échelle. Pour ce faire, nous avons utilisé un algorithme de Random Forest afin d'évaluer précisément le poids de ces facteurs par rapport aux autres indicateurs.

D'après nos résultats (annexes) nous avons vu que les variables sont peut être très corrélés aux prédictions que l'on veut effectuer. Nous allons encore réfléchir à comment est- ce que ça influence nos prédictions et comment nous allons nous adapter pour rendre cohérent notre projet.

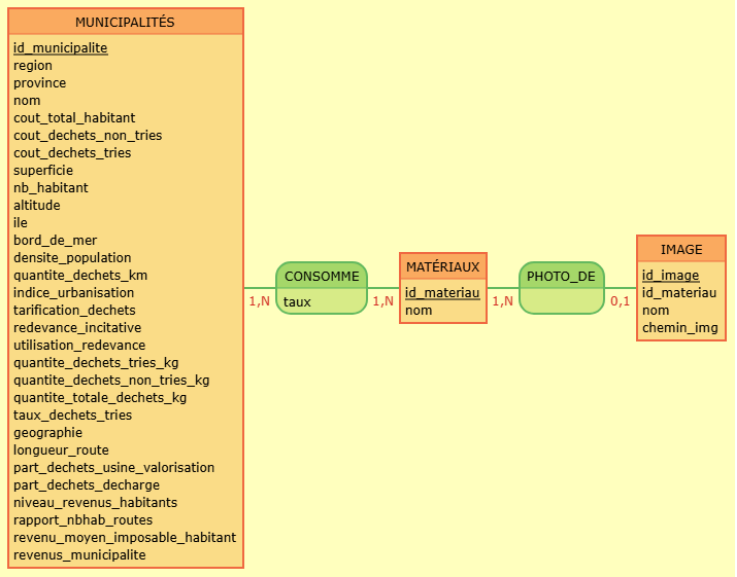
Compte Rendu 2

(semaine 3)

Pour cette troisième semaine, nous nous sommes donné pour objectif de bien structurer notre base de données, et de nous assimiler les données.

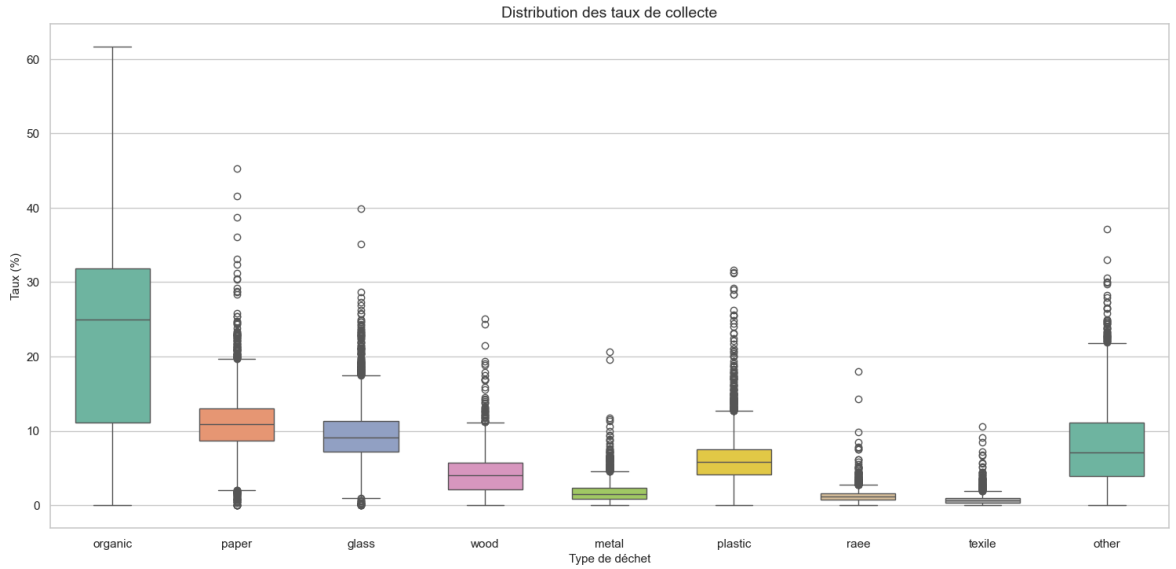
Nous avons, dans un premier temps, fait le MCD pour nous aider à visualiser et organiser nos données. Ensuite, nous avons fini de faire notre base de données (renommer variables, remplir tables, clés primaires/étrangères, …).

La difficulté principale a été de choisir une bonne façon de ranger nos données en liant les 2 bases, une avec des images et l'autre avec des données quantitatives. Par la suite, nous n'avons pas eu de difficultés particulières. Pour optimiser le stockage, nous avons choisi de conserver uniquement les chemins d'accès des images plutôt que leurs données binaires. Le format binaire n'étant pas primordial pour notre projet, cela nous a permis d'alléger notre base de données.

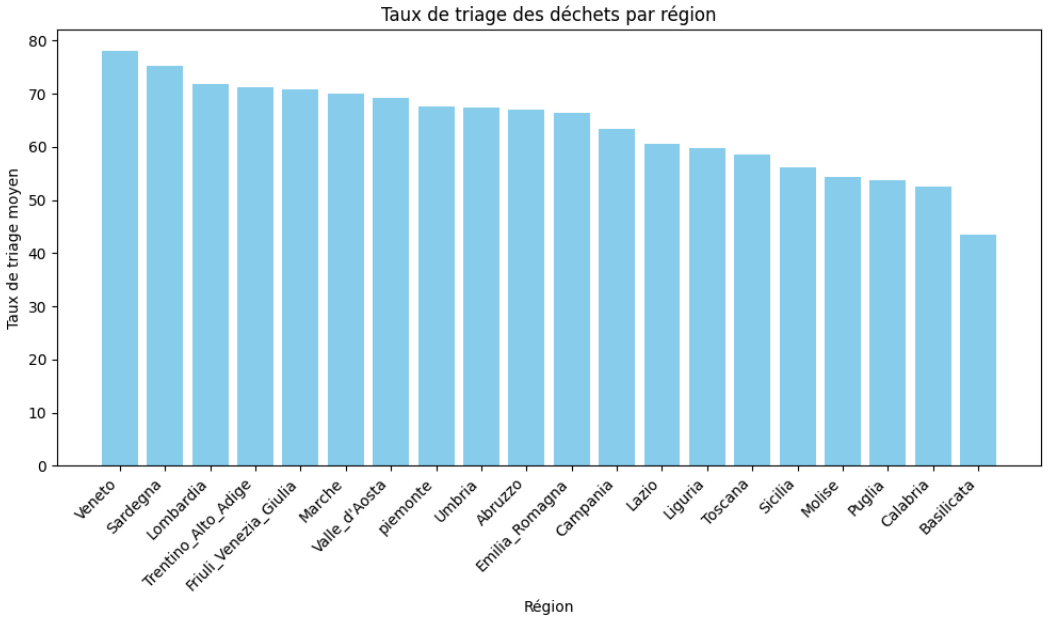


extrait du notre MCD

Nous avons aussi réalisé des premières visualisations afin de mieux comprendre nos données.

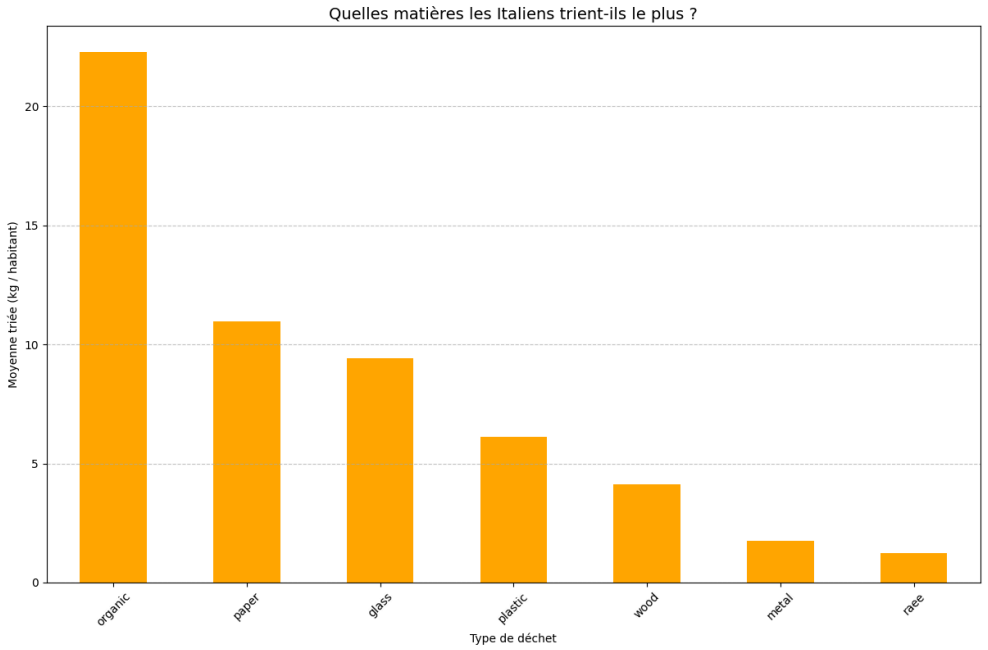


Analyse : Ce graphique représente le taux de déchets collecté de chaque aux communes en fonction des types de déchets. Les déchets organiques sont les plus collectés mais avec de fortes disparités. Le papier et le verre suivent avec une médiane autour des 10%. À l'inverse, le textile et l'électronique stagnent proche de 2%. Globalement, les déchets organiques sont les plus représentés et le reste des taux reste sous la barre des 10 %. On peut noter également que certaines communes obtiennent des résultats extrêmes bien aux dessus de la moyenne.



Analyse : Ce graphique représente le taux de triage des déchets en fonction de chaque région de notre base de données. L'axe des abscisses représente les régions italiennes et l'axe des ordonnées celui du taux de triage. Les barres sont classées par ordre décroissant, de la région la plus performante à celle la moins performante. Le Veneto est la région qui trie le plus avec environ 78% tandis que la Basilicata est dernière de ce classement avec environ 43% de déchets triés. On remarque une tendance comme quoi les régions du Nord

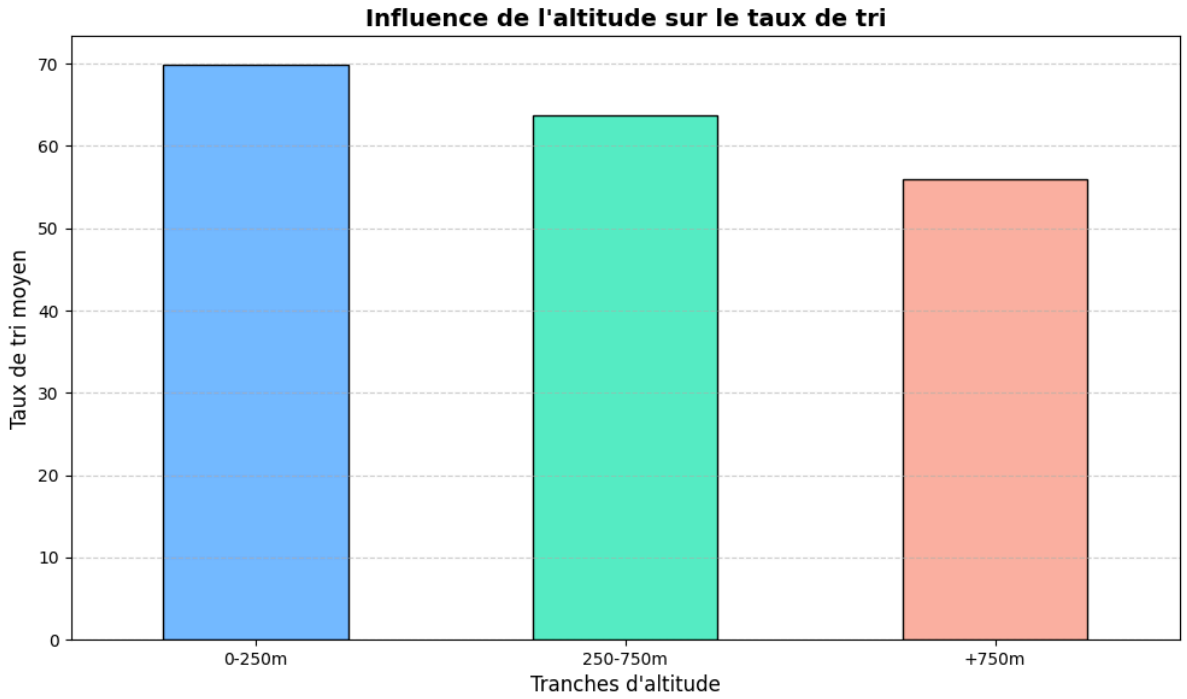
sont plus performantes au vu des résultats du classement. Le graphique montre une disparité assez importante entre le premier et le dernier du classement indiquant des différences importantes dans les politiques de gestion de déchets ou les infrastructures selon les régions.



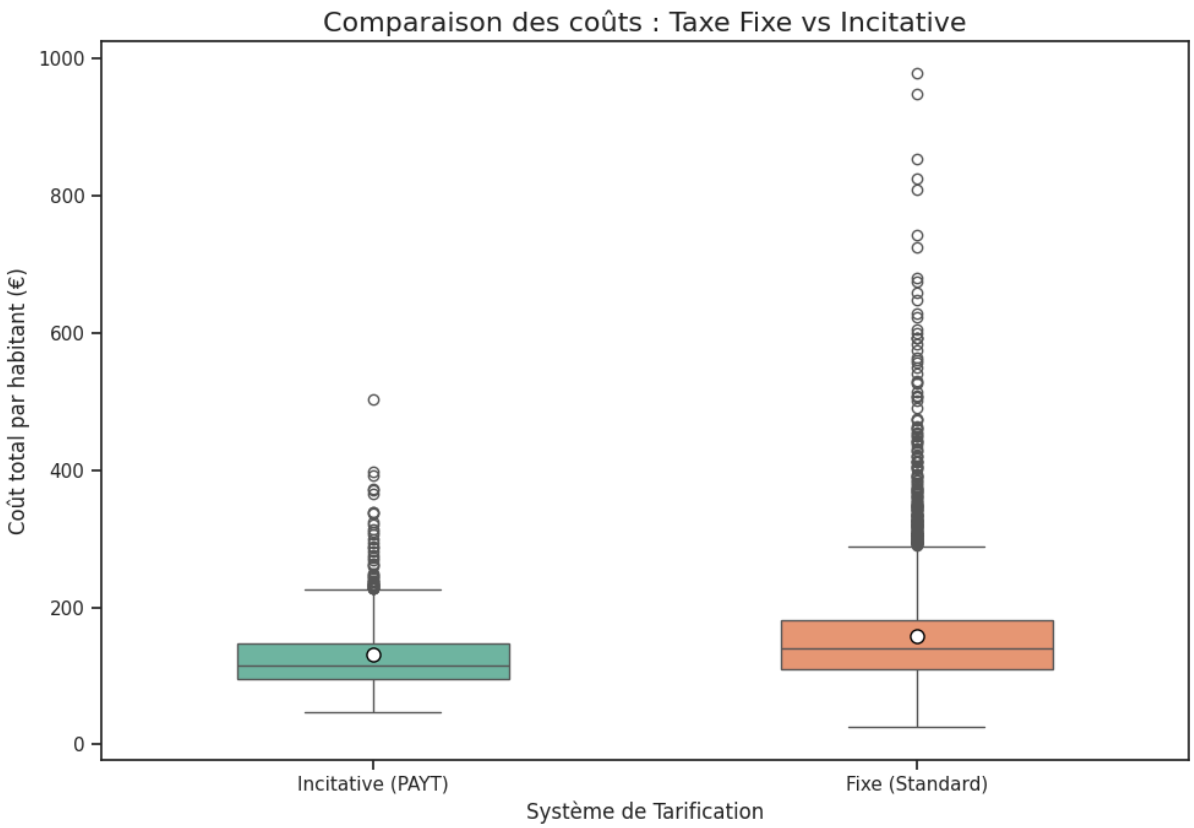
Analyse :

On compare la moyenne de chaque matériau pour voir ce que les gens mettent le plus dans les bacs de recyclage.

Ce qu'on remarque tout de suite avec ce graphique, c'est que le tri n'est pas du tout équilibré entre les matériaux : c'est l'organique qui gagne haut la main. Comme c'est la matière la plus lourde, c'est elle qui fait monter les scores de recyclage. Le papier et le verre suivent loin derrière, tandis que le plastique semble faible alors qu'il prend beaucoup de place : c'est simplement parce qu'il est très léger en poids. En conclusion, l'analyse montre que le tri n'est pas qu'une question de bonne volonté, c'est une question de poids.

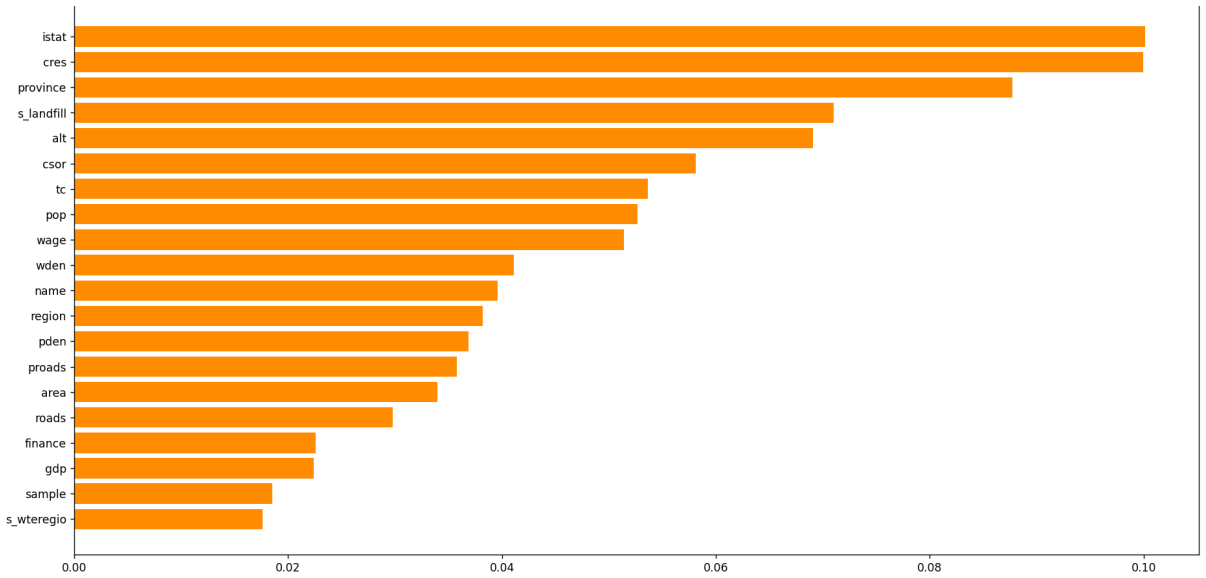


Analyse : Ce graphique en barre étudie une possible corrélation entre l'altitude des communes et leurs efficacités en termes de tri. Tout d'abord les communes ont été classées selon leurs altitudes en 3 catégories: - de 250 m, 250-750 m et + de 750 m. Une fois cela fait, on calcule pour chacune de ces catégories, la moyenne en % du taux de déchets triés. On peut supposer une corrélation donc entre ces deux facteurs : le taux de tri diminue à mesure que l'altitude augmente. De 0 à 250 m, le taux moyen atteint les 70% tandis qu'en montagne, le taux n'atteint "seulement" 56%. Cette tendance peut s'expliquer par plusieurs facteurs logistiques propres aux zones de haute altitude comme l'accessibilité (collecte plus complexe et coûteuse dans zones montagneuses, accès difficile pour les camion de ramassage), les infrastructures de tri (centres de tri et déchetteries situées plus loin des villages de hautes montagnes que des villes de plaine) où encore la saisonnalité (zones touristiques de montagnes subissent des pics de fréquentation saisonnière qui rend la gestion des déchets plus difficile).

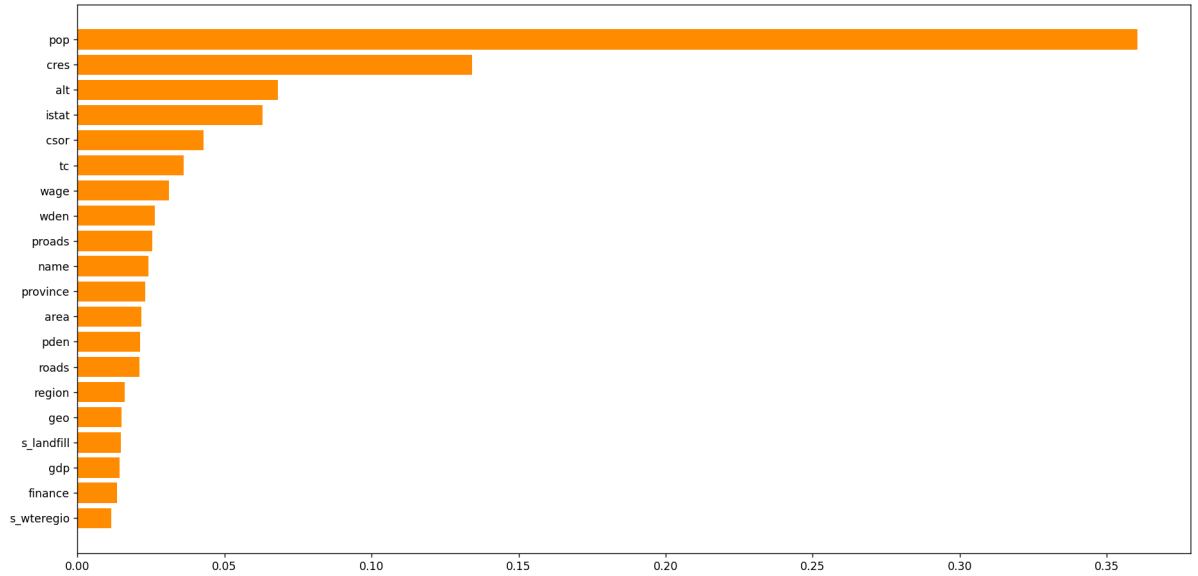


Analyse : la taxe au poids (PAYT) permet de mieux maîtriser les coûts que la taxe fixe. La "boîte" est plus basse, ce qui prouve que le système est efficace pour la majorité des villes. Par contre, les quelques points extrêmes nous montrent que la taxe dépend d'autres variables aussi.

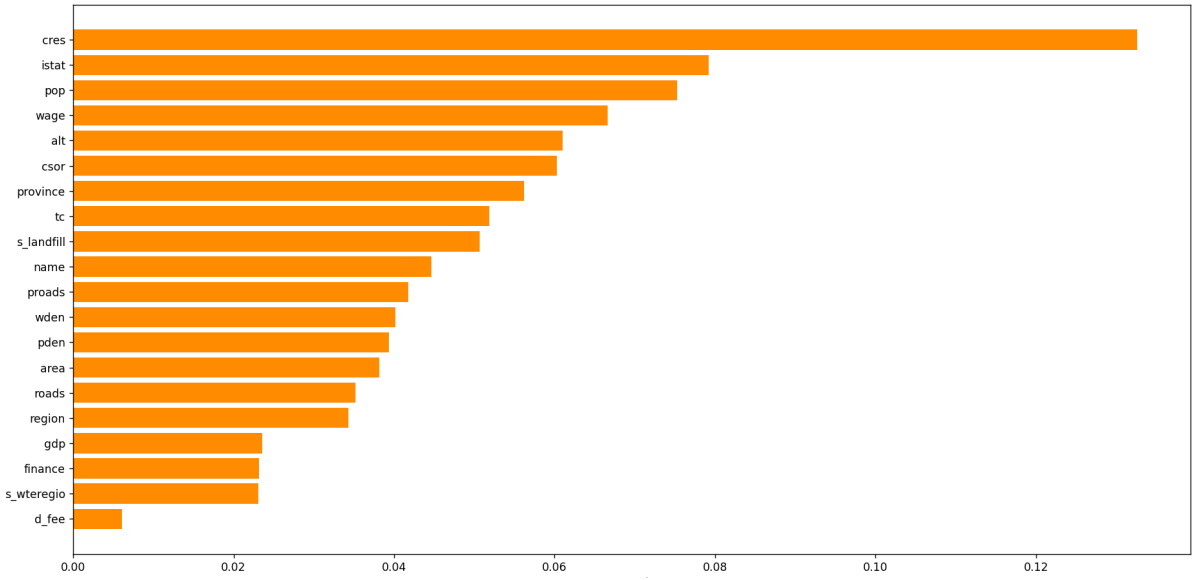
ANNEXE :



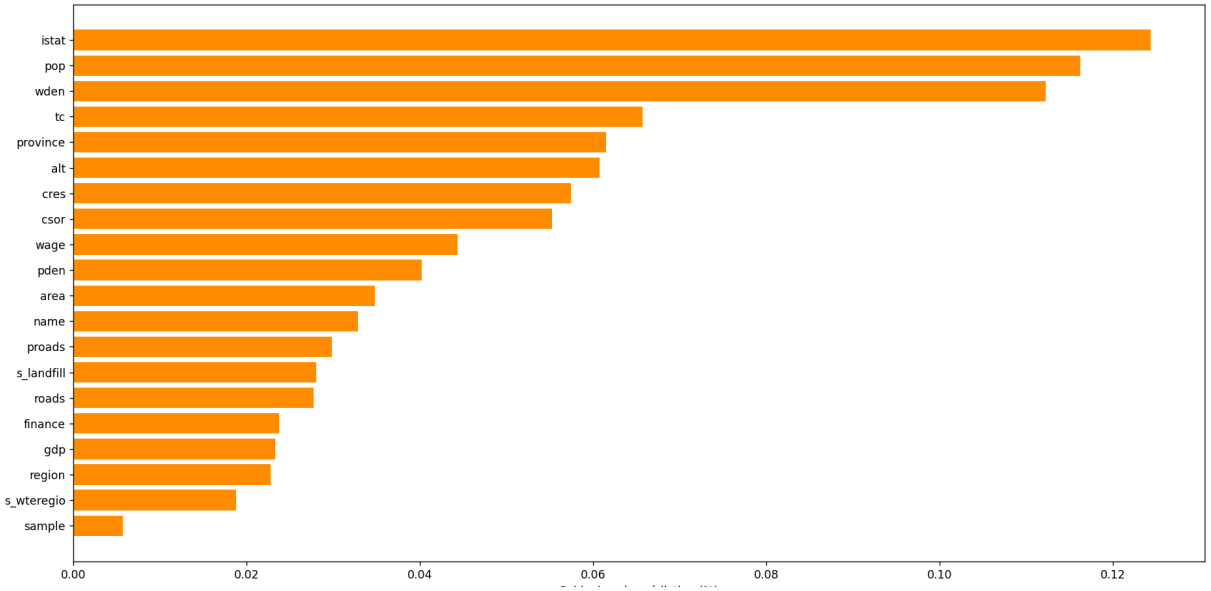
Poids des variables pour la prédiction du taux de plastique



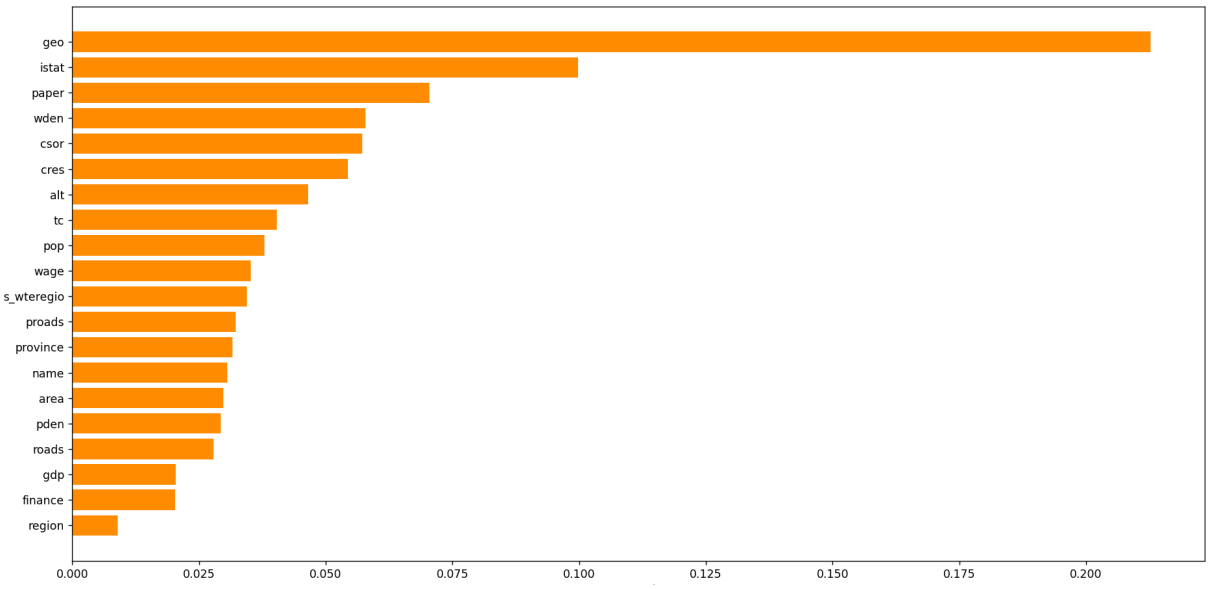
Poids des variables pour la prédiction du taux des déchets organiques



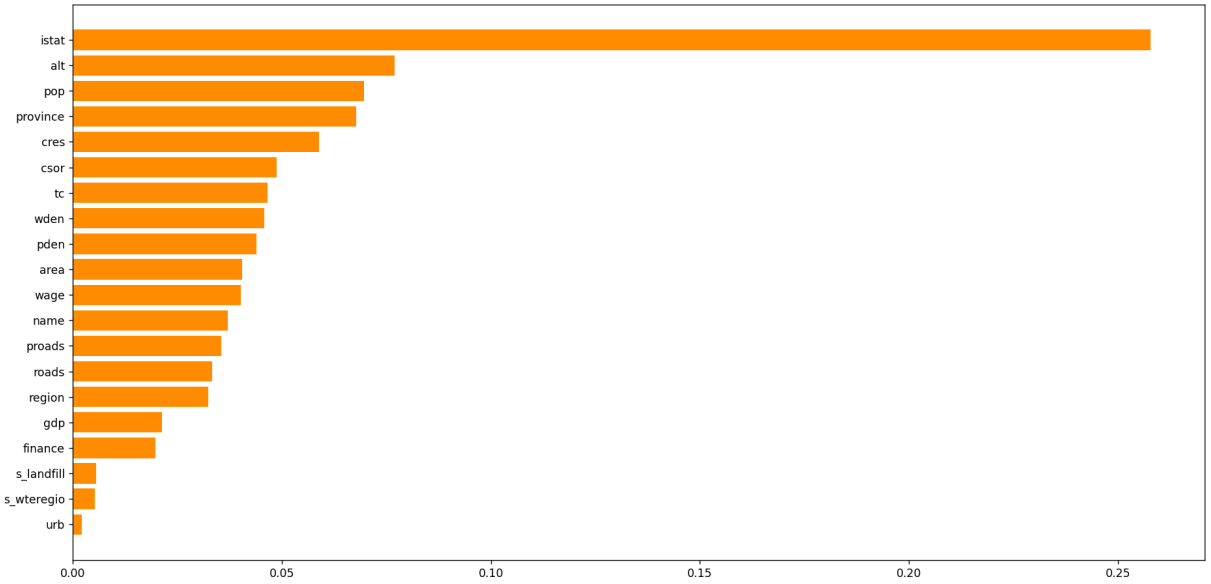
Poids des variables pour la prédiction du taux de papier



Poids des variables pour la prédiction du taux de verre



Poids des variables pour la prédiction du taux de bois



Poids des variables pour la prédiction du taux de bois