Rapport sur le pipeline de traitement et d'analyse des émissions

Objectif

Ce projet vise à construire un pipeline pour nettoyer, analyser et modéliser les données d'émissions fournies par des rapports clients. Nous avons effectué les étapes suivantes :

- Préparation des données.
- Analyse statistique et visualisation.
- Développement d'un modèle prédictif pour estimer les émissions futures.
- Conversion des valeurs dans une unité commune pour assurer la cohérence.

1. Préparation des données

a. Chargement et inspection

Nous avons chargé les données brutes à partir d'un fichier CSV contenant des informations variées. Une inspection initiale a permis de déceler des problèmes tels que des valeurs manquantes, des formats incohérents et des doublons.

b. Extraction des dates

Les informations de date étaient encodées dans la colonne Report Info. Nous avons extrait ces dates en utilisant une conversion à partir de dictionnaires, puis les avons transformées en un format temporel standard.

c. Nettoyage des données

- **Suppression des doublons** : Les entrées en double ont été supprimées.
- **Gestion des valeurs manquantes** : Les colonnes critiques comme Data Value ont été complétées en utilisant la médiane.

d. Unification des unités

Un dictionnaire de conversion a été mis en place pour transformer toutes les valeurs dans une unité commune (kg).

2. Analyse des données

a. Calcul des émissions annuelles

Nous avons calculé les émissions totales pour chaque année à Bruxelles en agrégeant les valeurs issues des rapports. Les données ont été ingérées dans une base de données SQLite, et des requêtes SQL ont été utilisées pour réaliser cette analyse.

b. Visualisation des tendances

Un graphique linéaire a été généré pour montrer les émissions totales annuelles de 2018 à 2025, avec une prédiction pour 2026. Les anomalies ont été éliminées (par exemple, les valeurs infinies ou manquantes).

3. Modélisation prédictive

a. Développement du modèle

Nous avons utilisé une régression linéaire simple pour modéliser les émissions en fonction des années. Les paramètres de la régression (a et b dans l'équation) ont été calculés et sauvegardés pour la prédiction.

b. Évaluation du modèle

Les métriques suivantes ont été calculées pour évaluer les performances du modèle :

- **RMSE (Root Mean Squared Error)** : Indique la précision moyenne des prévisions.
- **MAE (Mean Absolute Error)** : Reflète l'erreur absolue moyenne.

c. Prédiction pour 2026

La prévision pour l'année 2026 a été obtenue en utilisant le modèle, et un nouveau graphique a été créé pour inclure cette estimation.

4. Résultats et exportation

a. Résultats clés

- Les tendances montrent une augmentation progressive des émissions annuelles à Bruxelles.

- Le modèle prévoit des émissions totales d'environ **100,222.31 tonnes CO_2** pour 2026.

b. Sauvegarde

Les résultats, y compris les prédictions et les tendances historiques, ont été sauvegardés dans un fichier CSV. Les idées d'applications IA pour Tapio ont été documentées.

5. Défis rencontrés

- **Données manquantes** : Résolution via des valeurs de substitution basées sur la médiane.
- **Unités multiples** : Unification via un dictionnaire de conversion.
- **Valeurs infinies** : Élimination ou correction pour garantir des résultats fiables.

6. Conclusions

Ce projet a permis de développer un pipeline complet pour traiter, analyser et modéliser des données d'émissions. Les résultats obtenus sont exploitables pour la prise de décision et la prévision. Les futures étapes pourraient inclure l'utilisation de modèles avancés (séries temporelles, régressions polynomiales) et la création d'un tableau de bord interactif.