Rapport : Prédiction des Émissions de CO_2 des Véhicules commerciaux en France

Membre : - **WAMBO Jean Ernest**
- **JUMAA Dana**
Formation : Licence 3 INFO – UFR MIM
Tâche : Régression supervisée

1 Introduction et compréhension du dataset

Ce projet vise à modéliser et prédire les émissions de dioxyde de carbone (CO_2) des véhicules commercialisés en France. L'objectif est de comprendre comment les caractéristiques techniques (masse, puissance, transmission) influencent l'empreinte environnementale dès la conception du véhicule.

- **Origine :** Données issues de (`data.gouv.fr`).
- **Dimensions :** 55 044 observations et 29 variables initiales (quantitatives et qualitatives).
- **Objectif :** Développer un modèle de régression capable d'estimer précisément le taux de CO_2 (g/km).

2 Nettoyage des données (Data Cleaning)

Le jeu de données brut présentait de nombreuses incohérences et valeurs manquantes nécessitant un traitement adéquate :

- **Suppression des colonnes non pertinentes :** Retrait des colonnes vides (*Unnamed*), des identifiants administratifs (*cnit*, *dscom*, *tvv*) et des métadonnées (*date_maj*) afin d'éviter le surapprentissage sur des identifiants de modèles.
- **Traitement des fuites de données (Data Leakage) :** Suppression des variables de consommation (*conso_mixte*, etc.) et des polluants secondaires (*nox*, *hc*, *ptcl*), car ces données ne sont connues qu'après des tests en laboratoire.
- **Correction des types :** Conversion des variables numériques lues comme du texte (*object*) en nombres réels (*float*) via le remplacement des virgules par des points.
- **Imputation :** Comblement des valeurs manquantes par la médiane, le mode ou encore moyenne pour les variables numériques afin de préserver la taille du dataset.

3 Analyse exploratoire (EDA) et Feature Engineering

L'analyse exploratoire a permis d'identifier des tendances clés avant la phase de modélisation :

- **Distribution du CO_2 :** La majorité des véhicules émettent entre 150 et 250 g/km, avec une moyenne observée de 201,7 g/km.

— **Relations physiques :** Une forte relation linéaire est observée entre la masse du véhicule et son taux d'émission, ainsi qu'avec sa puissance maximale.
— **Analyse par carburant :** Les boxplots montrent une forte hétérogénéité des motorisations thermiques (outliers au-delà de 500 g/km pour l'essence), tandis que les motorisations alternatives sont plus homogènes.
— **Contribution de Marque :** l'affichage du tableau de contribution des marques permet de remarquer qu'il y a beaucoup plus de Mercedes et de Volkswagen dans notre data set.
— **Feature Engineering :**
  — Création d'une variable *masse_ moyenne* à partir des masses minimales et maximales.
  — Séparation de la transmission en deux variables : *type_ boite* et *nb_ rapports*.
  — Extraction de la norme environnementale numérique depuis la colonne *champ_ v9* (norme Euro).

4 Préparation des données et pipeline

Afin de rendre les données compatibles avec les algorithmes de machine learning :
— **Encodage :** Application du *One-Hot Encoding* sur les variables catégorielles (*lib_ mrq*, *cod_ cbr*, *carrosserie*, etc.) avec l'option `drop_first=True` afin d'éviter la multicolinéarité.
— **Découpage :** Séparation du dataset en un ensemble d'entraînement (80 %, soit 44 035 lignes) et un ensemble de test (20 %, soit 11 009 lignes).

5 Modélisation et évaluation

Six algorithmes ont été comparés afin d'identifier la meilleure approche prédictive. Les modèles ont été évalués sur le jeu de test (20 %) pour garantir leur capacité de généralisation.

Modèle	**MAE (g/km)**	**RMSE (g/km)**	R^2
Régression linéaire	9,85	13,35	0,8486
Ridge Regression	9,85	13,36	0,8485
KNN Regressor	6,53	10,74	0,9022
Gradient Boosting	6,21	7,96	0,9462
XGBoost	6,32	8,15	0,9436
Random Forest	**5,33**	**7,03**	**0,9580**

Analyse des performances : Le modèle *Random Forest* se distingue nettement avec un R^2 de 0,9580. Cette supériorité s'explique par sa capacité à capturer les interactions non linéaires (ex : l'impact conjoint de la masse et du type de carburant) que les modèles linéaires comme *Ridge* ne parviennent pas à modéliser aussi précisément. L'erreur moyenne (MAE) de 5,33 g/km est extrêmement faible au regard de la moyenne du dataset (201,7 g/km), validant la fiabilité du modèle.

6 Interprétation et Analyse des erreurs

— **Feature importance :** L'analyse du *Random Forest* révèle que la puissance maximale et la masse sont les prédicteurs dominants. Cela confirme que l'architecture physique du véhicule prime sur les facteurs commerciaux dans la génération de CO_2.

— **Analyse des Outliers :** Une étude de détection d'outliers par la méthode de l'Écart Interquartile (IQR) a été menée. Bien que ces points (véhicules de luxe ou sportifs) s'écartent statistiquement de la norme, leur suppression a entraîné une dégradation des performances (R^2 en baisse et MAE en hausse). Nous avons donc conclu que ces valeurs ne sont pas des erreurs de saisie mais des **réalités physiques extrêmes**. Les conserver permet au modèle de rester robuste sur l'intégralité du parc automobile français.

— **Analyse des erreurs :** Le graphique *Réel vs Prédit* montre une excellente linéarité. On note toutefois une légère sous-estimation pour les véhicules dépassant 450 g/km de CO_2, due au plus faible volume d'échantillons dans cette catégorie haut de gamme.

7 Conclusion

Ce projet démontre que les caractéristiques physiques d'un véhicule permettent de prédire son impact environnemental avec une grande précision. Un nettoyage rigoureux, notamment la suppression des données incohérentes et l'arbitrage en faveur du maintien des outliers réels, a été déterminant. Le modèle final constitue un outil d'aide à la décision fiable pour estimer l'impact carbone dès la phase de conception technique.