Rapport CO_2

Introduction

En 2019, 31% des émissions de GES (gaz à effet de serre) sont liées au secteur des transports. Le secteur routier contribue à la quasi-totalité des émissions du secteur des transports (94%)[1]. Ces émissions de gaz incombent à hauteur de 54 % aux véhicules particuliers, de 24 % aux poids lourds et de 20 % aux véhicules utilitaires légers.

Les mesures d'émission de CO_2 sont réalisées par le constructeur et sont enregistrées sur le certificat de conformité de chaque véhicule[2]. La valeur d'émission de CO_2 homologuée va déterminer le montant des bonus/malus écologiques (écotaxe) appliqués aux constructeurs au niveau européen et sur le prix de vente du véhicule au niveau national.

L'objectif de cette étude est de réaliser une analyse des caractéristiques techniques de tous les véhicules neufs immatriculés en France en 2019 recueillies par l'Agence européenne de l'environnement (AEE). En identifiant les caractéristiques techniques responsables du taux d'émission de CO_2, il serait possible d'établir un modèle prédictif de la pollution engendrée par de nouveaux types de véhicules (nouvelles séries de voitures par exemple).

La restitution du projet s'articule selon un axe chronologique. Elle débute avec la description des données sources (provenance, qualité) et balaye rapidement la réglementation encadrant la production des mesures du taux d'émission de CO_2. S'ensuit la description du *process* de réduction du dataset aux "véhicules acquis au moins une fois sur le territoire en 2019", avec en particulier la justification du choix de l'identifiant des véhicules. Les graphiques les plus informatifs extrait des analyses exploratoires sont ensuite commentés.
De cette phase de compréhension des données découle la décision de modéliser le taux d'émission de CO2 selon deux approches : régression et classification. La phase de préparation des données (*preprocessing*) est détaillée ainsi que les résultats des différents modèles déployés (performances et limites) pour chacune des approches. Pour finir, l'interprétation des modèles permet de répondre à la problématique soulevée, à savoir quelles caractéristiques des véhicules sont liées à leur taux d'émission de CO_2. Une discussion sur la portée des résultats vient clôturer le document.

Présentation des données

Données source

Le règlement (UE) n° 2019/631 oblige les pays à enregistrer des informations pour chaque voiture particulière et camionnette neuve immatriculée sur son territoire.

Chaque année depuis 2010, chaque État membre soumet à la Commission toutes les informations relatives à ses nouvelles immatriculations. En particulier, les informations suivantes sont requises pour chaque voiture particulière neuve immatriculée[3] : nom du constructeur,

---

[1] Source : CITEPA, rapport Secten 2020

[2] Le Certificat de conformité, CoC en abrégé fait référence à un numéro de châssis spécifique et certifie que ce véhicule correspond à un type approuvé dans l'UE.

[3] Les guidelines précisant le contenu du rapport à transmettre par les Etats membre sont disponibles en ligne : Guidelines on the monitoring and reporting of CO2 emissions from light-duty vehicles

numéro de réception par type, type, variante, version, marque et nom commercial, émissions spécifiques de CO_2 (protocoles NEDC et WLTP), masses du véhicule, roue base, largeur de voie, cylindrée et puissance du moteur, type et mode de carburant, éco-innovations et consommation d'électricité. Ces données sont nécessaires à la Commission pour connaître les émissions moyennes de CO_2 des voitures particulières neuves et pour fixer les objectifs d'émissions spécifiques qui doivent être atteints par les constructeurs automobiles.

L'AEE effectue pour la Commission le traitement des données. Elle réalise plusieurs contrôles sur les fichiers transmis par les États membres afin d'évaluer l'exactitude et la qualité de l'ensemble de données. Sur la base des vérifications et des réactions des États membres, l'AEE finalise et publie une base de données préliminaire. Parallèlement, des courriers de notification sont adressés aux industriels les informant de leurs performances CO_2 prévisionnelles. Les fabricants peuvent, dans les trois mois suivant la notification du calcul provisoire, notifier à la Commission toute erreur dans les données. Passé ces trois mois, les données et les objectifs finaux seront publiés sur le site Web de l'AEE.

Compte tenu de l'important volume de données générées au niveau européen, le projet s'est concentré sur l'analyse des données françaises de 2019. L'année 2020 a été écartée compte tenu de l'impact probable de l'épidémie de Covid sur la vente des véhicules neufs.

Ainsi, les informations complètes du **parc des véhicules particuliers et utilitaires légers neufs immatriculés en France pour l'année 2019** ont été téléchargées sur le site de l'AEE[4]. Ce jeu de données est composé de **2 305 720 immatriculations**.

Les items "date d'enregistrement", "consommation de fuel" et "autonomie des véhicules électriques" non renseignés en 2019 ont été écartés dès la phase d'importation des données. Les items utilisés pour filtrer les informations exportées à savoir "l'année d'immatriculation", le "pays membre" et le "statut des données" (provisoire versus définitif) ont également été supprimés. Au final, le jeu de données source contient 31 descripteurs des véhicules décrits dans le tableau 1.

Tableau 1 : Informations relatives aux immatriculations de véhicules neufs transmises par les Etats membres à l'AEE

Nom	Type	Descriptif
- Id	numérique	Identifiant d'enregistrement du véhicule (uniquement dans la base parc)
- r	numérique	Nombre d'enregistrement toujours égal à 1 (une ligne = une immatriculation)
Identification du véhicule		
- Cn	catégoriel	Nom commercial
- Ct	catégoriel	Catégorie du véhicule réceptionné
- Cr	catégoriel	Catégorie du véhicule immatriculé
- Tan	catégoriel	Numéro de réception (ou homologation) du véhicule
- T	catégoriel	Identifiant Type-Variante-Version : numéro de série spécifique de chaque véhicule
- Va	catégoriel	
- Ve	catégoriel	
- VFN	catégoriel	Identifiant de famille d'interpolation du véhicule
Constructeur		
- Mk	catégoriel	Marque (raison social du constructeur)

[4] https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-22

Nom	Type	Descriptif
- Mh	catégoriel	Nom du constructeur selon : - la dénomination standard EU - la déclaration du constructeur - la dénomination du registre national
- Man	catégoriel	
- MMS	catégoriel	
- Mp	catégoriel	Pool de constructeurs[5]
Caractéristiques techniques du véhicule		
- m	numérique	Masse en ordre de marche[6] (kg)
- Mt	numérique	Masse d'essai WLTP, véhicule complet et véhicule complété (kg)
- Ft	catégoriel	Type de carburant
- Fm	catégoriel	Mode de carburation
- ep	numérique	Puissance nette maximale (KW)
- ec	numérique	Cylindrée (cm^3)
- W	numérique	Empreinte au sol : empattement, largeur de voie de l'essieu directeur et largeur de voie de l'autre essieu (mm)
- At1	numérique	
- At2	numérique	
- z	numérique	Consommation d'énergie électrique (Wh/km)
Emission de CO_2 (g/km)		
- Enedc	numérique	Emissions spécifiques de CO_2 norme NEDC (g/km)
- Ewltp	numérique	Emissions spécifiques de CO_2 norme WLTP (g/km)
Eco-innovation		
- IT	catégoriel	Code(s) d'éco-innovation
- Ernedc	numérique	Réduction des émissions de CO_2 résultant d'éco-innovation (norme NEDC et WLTP en g/km)
- Erwltp	numérique	
Corrélation avec la simulation CO_2MPAS[7]		
- De	numérique	Facteur de déviation
- Vf	numérique	Facteur de vérification

Les émission de CO_2

L'information concernant l'émission de CO_2 des véhicules disponible est exprimée à partir de deux normes : l'ancien protocole NEDC ("New European Driving Cycle") et le nouveau protocole WLTP ("Worldwide Harmonized Light-Duty Vehicles Test Procedure").

A la différence de l'ancien protocole NEDC, la norme WLTP cherche à être au plus près des conditions réelles d'utilisation des véhicules. Ainsi, pour déterminer leur taux d'émission de CO_2, le protocole impose que le constructeur soumette les véhicules à des cycles de roulage variés, incluant différentes vitesses et des temps d'arrêts. Jusqu'ici, le taux d'émission était calculé à partir des émissions brutes du moteur, sans tenir compte de l'équipement du véhicule. La procédure WLTP étant plus réaliste que le cycle NEDC, les taux d'émissions de CO_2 estimés sont en moyenne plus élevés de 10% à 25% qu'avec l'ancienne procédure[8].

[5] Accord entre constructeurs pour mettre en commun les émissions de leur modèle dans le but de rester sous les quotas d'émission imposés par l'EU (les quotas s'imposant sur les émissions moyennes des modèles vendus, les véhicules à fortes émission ont tout intérêt à se positionner dans un pool avec des véhicules faible émetteur)

[6] La masse en ordre de marche d'un véhicule est sa masse incluant les consommables à un niveau prédéfini (exemple : 90% du carburant), les fluides fonctionnels comme l'huile et le liquide de refroidissement, les outillages et la roue de secours

[7] Pour 10% des véhicules enregistrés, un contrôle des déclarations est réalisé par l'UE. Les deux variables De et Vf rapportent les résultats du test.

[8] Informations sur les procédures d'essai reprises du site Wikipédia :
Procédure d'essai mondiale harmonisée pour les véhicules légers — Wikipédia

En 2019, peu de constructeurs ayant pu s'aligner sur le seuil d'émission fixé par la Commission en utilisant la nouvelle norme, il a été décidé que, pour cette année charnière, une valeur NEDC dite corrélée servirait de référence. Il s'agit d'un entre-deux, la valeur CO2 NEDC corrélée étant obtenue en transformant la mesure WLTP en utilisant le simulateur CO_2MPAS[9].

Les analyses produites porteront sur les émissions mesurées selon les deux protocoles (NEDC et WLTP), mais en toute logique les résultats obtenus seront très proche puisque les variables disponibles dans le dataset sont utilisées par le simulateur CO_2MPAS[10].

Pour finir, on retiendra qu'en 2019 la norme NEDC mesurée restait applicable pour les voitures de fin de série autorisées à la vente et certains utilitaires pour lesquels la mise en place du protocole WLTP avait été repoussée en 2020 et 2021.

Création du dataset "Véhicules"

Le *dataset* dénommé "Véhicules" constitué des **16 975 véhicules neufs acquis par au moins une personne en France en 2019** a été créé à partir du jeu de données source en supprimant les doublons sur un identifiant véhicule.

Identifiant véhicule : un véhicule a été spécifié comme l'ensemble des entités référencées par le triptyque type-variant-version unique (TVV) et appartenant à une même famille d'interpolation de véhicule. Cette spécification a été retenue sur la base des textes réglementaires définissant les numéros d'enregistrement disponibles[11] et compte tenu des objectifs fixés, sachant que :

- L'identifiant TVV est un numéro de série qui permet d'identifier un véhicule et de connaître, par exemple, ses finitions et spécificités.
- Une famille d'interpolation est définie à la demande d'un constructeur[12]. La famille d'interpolation étant déterminée sur des critères influençant les émissions de CO_2, il est apparu pertinent de conserver des véhicules disposant du même identifiant TVV mais qui appartiennent à deux familles d'interpolation différentes.

Ce choix d'identifiant composite "TVV - famille d'interpolation" entraîne l'élimination d'entités avec des valeurs différentes pour deux items :

- La masse essai WLTP qui peut varier en fonction des équipements. Dans ce cas, le choix a été de garder un seul véhicule par famille d'interpolation compte tenu que ces véhicules diffèrent sur des caractéristiques qui n'influencent pas les émissions de CO_2.
- Le numéro d'homologation : la réception par type ou homologation est un acte administratif qui atteste de la conformité technique d'un véhicule au regard de la réglementation. Le fabricant doit immédiatement informer l'autorité de réception de tout changement, autorité qui décidera si ces évolutions nécessitent une révision ou une

---

[9] Règlement d'exécution (UE) 2017/1152 de la Commission du 2 juin 2017 établissant une méthode de détermination des paramètres de corrélation nécessaires pour tenir compte de la modification de la procédure d'essai réglementaire en ce qui concerne les véhicules utilitaires légers et modifiant le règlement d'exécution (UE) no 293/2012 https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:02017R1152-20180724&qid=1536845081657&from=FR

[10] https://co2mpas.readthedocs.io/en/stable/glossary.html#term-EU-legislation

[11] En particulier pour la famille d'interpolation le Règlement (UE) 2017/1151 de la Commission du 1er juin 2017 complétant le règlement (CE) no 715/2007 du Parlement européen et du Conseil relatif à la réception des véhicules à moteur au regard des émissions des véhicules particuliers et utilitaires légers et aux informations sur la réparation et l'entretien des véhicules https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32017R1151&from=FR

[12] Pour une même famille d'interpolation, la procédure d'essai de mesures d'émissions est menée sur deux véhicules ; la méthode d'interpolation permet ensuite de calculer des estimations d'émission pour l'ensemble des véhicules de cette famille.

extension de la réception. En cas d'extension (modification mineure), le numéro d'homologation est tout de même modifié. En conservant cet identifiant, des entités extrêmement proches auraient donc été sélectionnées et ces véhicules présentant de nombreuses extensions auraient eu une influence plus grande sur les résultats.

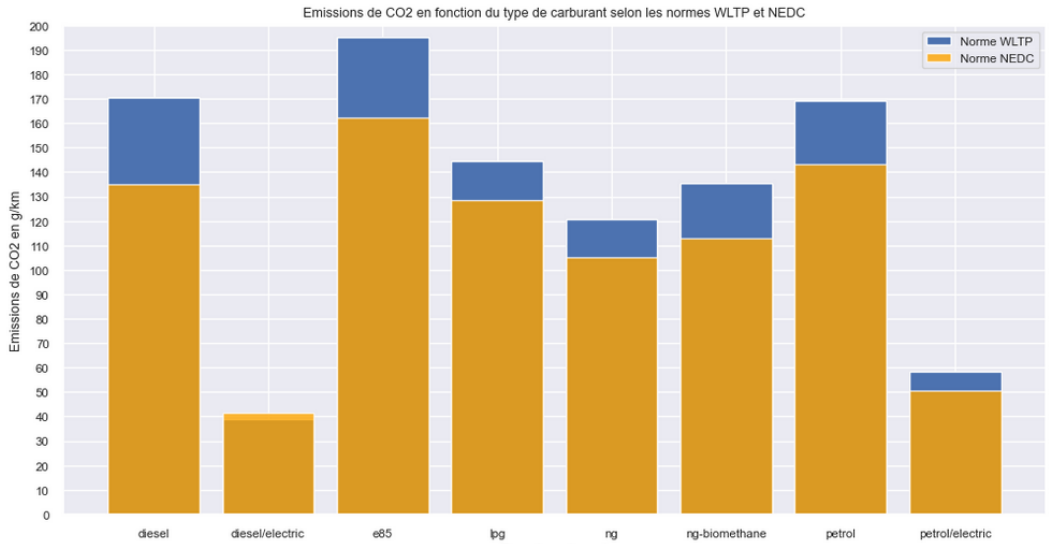
Exploration

Analyse des données et visualisation

Deux dataframes ont été créés spécifiquement pour les besoins de la data visualisation. Le dataframe nommé "parc_viz" regroupe les **véhicules neufs immatriculés en France pour l'année 2019** et a été créé à partir des données sources contenant les 31 descripteurs des véhicules décrits dans le tableau 1.

Le second dataframe dénommé "cars_viz" regroupe les **véhicules neufs acquis par au moins une personne en France en 2019** et a été créé à partir du dataset "Véhicules" mais en conservant les véhicules de tous types, c'est à dire les véhicules ne rejetant pas de CO2 (électrique, hydrogène) ainsi que les véhicules aux carburants rares.

Graphique 1 : Diagramme en barre des émissions de CO2 selon les normes NEDC et WLTP en fonction du type de carburant utilisé



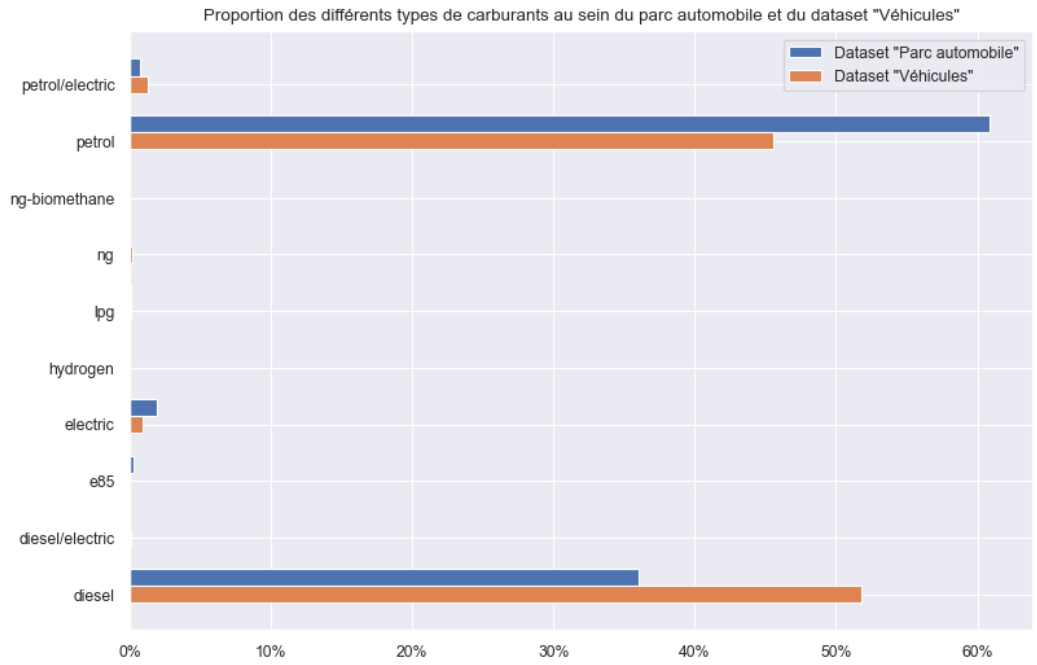
Cette première visualisation permet de confirmer le plus haut taux d'émission de CO2 estimé par la norme WLTP en comparaison avec le cycle NEDC corrélé, et ceci pour la plupart des types de carburants. Une exception est toutefois présente pour les véhicules de type diesel/électrique, dont l'émission de CO_2 selon la norme NEDC est très légèrement plus élevée.

Ce graphique met également en évidence la disparité du taux d'émission de CO_2 entre les différents types de carburants. Les véhicules roulant au carburant **E85**, également appelé **superéthanol** (un mélange constitué de biocarburant, d'éthanol et d'essence SP95) rejette beaucoup plus de CO_2 que des véhicules roulant au gaz naturel -ng- par exemple (~195 et ~120 g/km respectivement, selon norme WLTP).

Les véhicules diesel et essence (“petrol”), qui constituent la majorité des véhicules du dataset, émettent une quantité similaire de CO_2, aux alentours de 170 g/km pour la norme WLTP.

Enfin, les véhicules hybrides diesel/électrique sont ceux qui rejettent le moins de CO_2 (~40 g/km) suivis par les véhicules hybrides essence/électrique (~60 g/km).

Graphique 2 : Diagramme à barres horizontales des différents types de carburants au sein du parc automobile et du dataset “Véhicules”

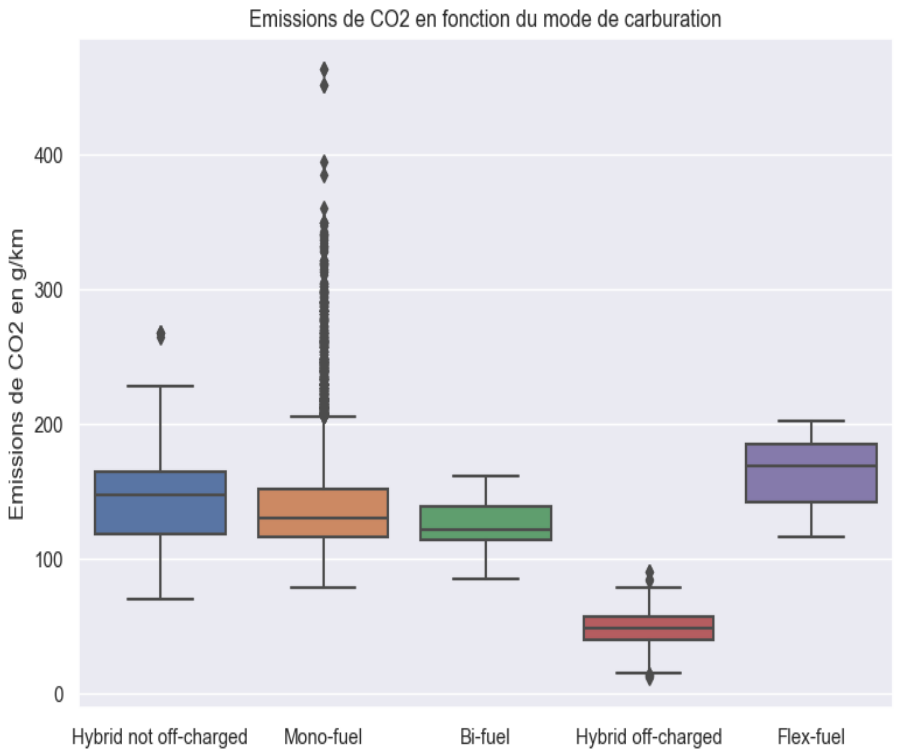


Ce diagramme en barre présente les proportions des différents types de carburants au sein du dataset “Parc automobile” (véhicules immatriculés en 2019 en France) et du dataset “Véhicules” ne contenant que les véhicules uniques. Les véhicules diesel et essence (“petrol”) sont largement représentés (plus de 96% pour les deux jeux de données).

Cette visualisation illustre également la rareté de certains types de carburant : les véhicules alimentés à l’hydrogène, au superéthanol (E85), au gaz naturel (ng), au mélange gaz naturel-biométhane (ng-biomethane), au gaz de pétrole liquéfié (lpg), ainsi que les véhicules diesel/électrique représentent une très faible proportion des deux jeux de données explorés (<0.4% tous types de carburant cumulés). Les véhicules électriques représentent ~1% à ~2% des jeux de données, et seront retirés dans la partie pre-processing (ne servent pas à la modélisation puisqu’ils ne rejettent pas de CO2).

Pour le dataset “Véhicules”, qui servira à établir les modèles de Machine Learning par la suite, la proportion de véhicules diesel est plus importante que la proportion de véhicules essence, contrairement au jeu de données du “Parc automobile”, qui lui, comprend une plus grande proportion de véhicules essence.

Graphique 3 : Boxplot des émissions de CO2 en fonction du mode de carburation



Cette visualisation en boîtes à moustache (boxplot) des émissions de CO2 en fonction du mode de carburation apporte plusieurs informations essentielles. Premièrement, la catégorie de véhicules fonctionnant avec un unique carburant (“mono-fuel”) présente de nombreuses valeurs au-delà des moustaches : ces valeurs ne sont pas aberrantes et sont associées à des véhicules très puissants (type “supercar”), représentés quasi-exclusivement par des moteurs essence dont la conception offre un meilleur rapport poids/performance.

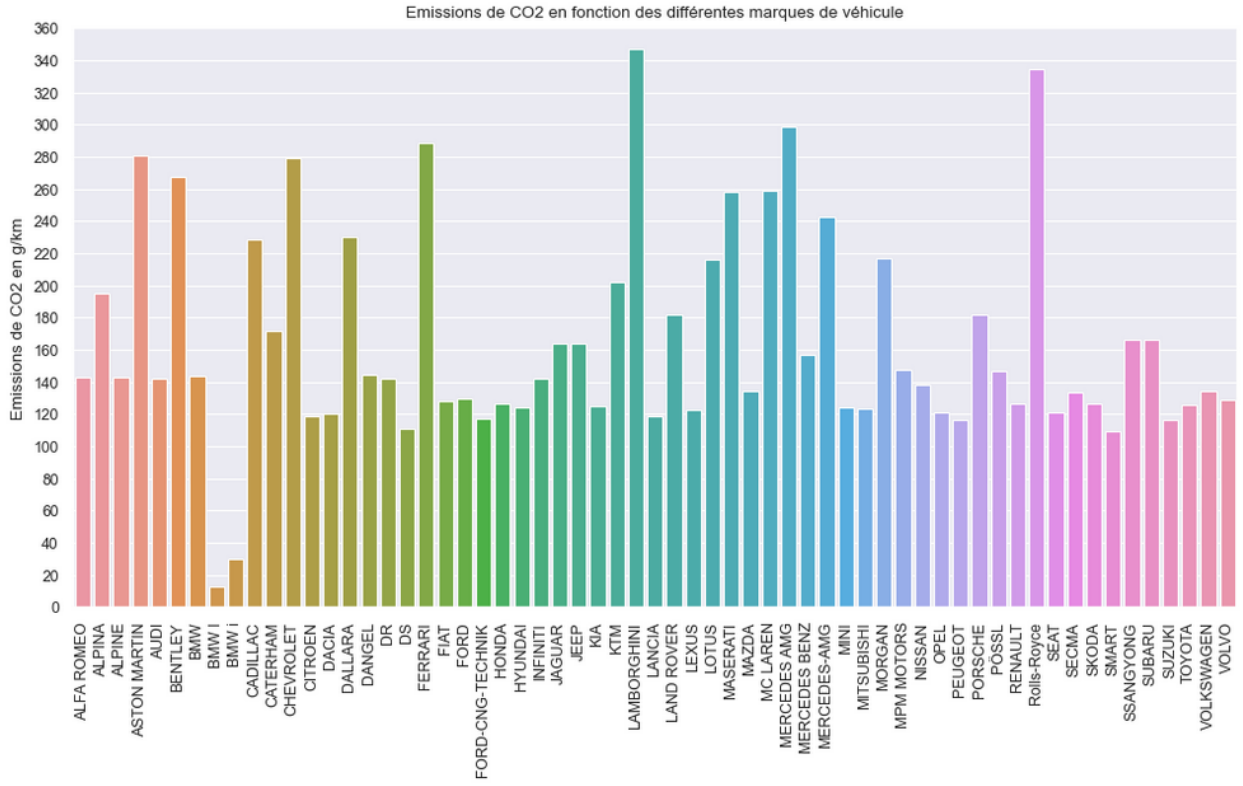
Deuxièmement, le corps des boxplots (médianes et quartiles) met en lumière une émission de CO2 plus élevée pour les véhicules hybrides non rechargeables (“hybrid not off-charged”), comparativement aux véhicules mono-fuel ou bi-fuel[13].

Enfin, les véhicules de type flex-fuel[14] ont la médiane la plus élevée (~180 g/km) et les véhicules hybrides rechargeables (“hybrid off-charged”) ont la médiane la plus faible (~50 g/km).

[13] véhicules disposant de deux réservoirs, pouvant fonctionner avec un carburant ou l’autre, non simultanément.

[14] véhicules fonctionnant avec plusieurs carburants, dans un même réservoir, très majoritairement essence et superéthanol.

Graphique 4 : Diagramme à barres des émissions de CO2 moyennes selon les marques de véhicule



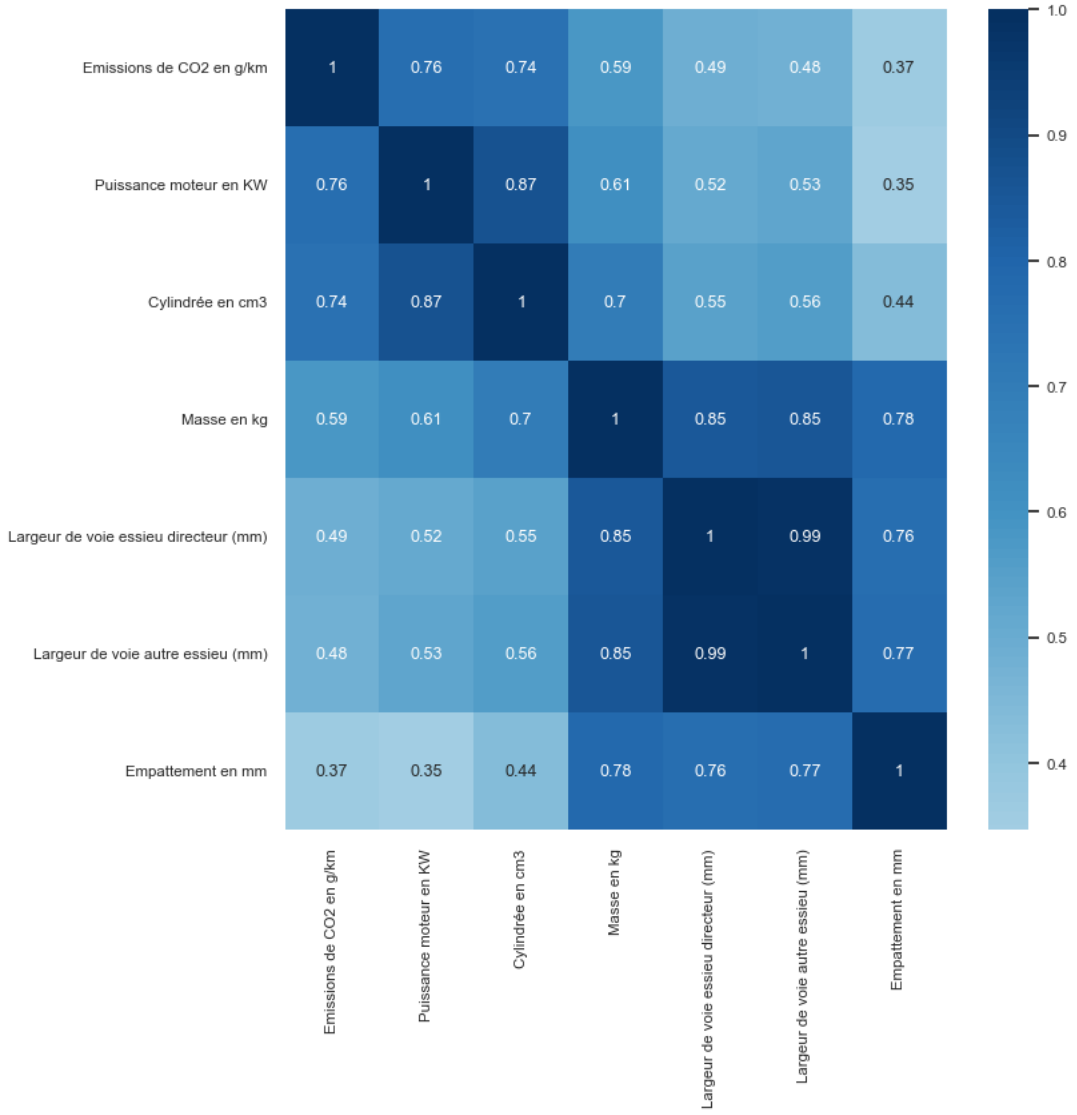
Ce diagramme en barre représente l'émission moyenne de CO2 en fonction des marques de véhicule du dataset. Les marques de véhicules électriques ont été retirées car ne rejetant pas de CO2 (ex : Tesla).

Alors qu'une grande partie des marques présente un taux d'émission moyen aux alentours de ~120 à ~160 g/km, certains pics sont bien plus importants : les marques Aston Martin, Bentley, Chevrolet, Ferrari, Lamborghini, Maserati, McLaren, Mercedes AMG ou encore Rolls-Royce présentent un taux d'émission moyen beaucoup plus important compris entre ~260 et ~350 g/km.

Ces valeurs extrêmes s'expliquent par la catégorie à laquelle tous ces véhicules appartiennent : ce sont des véhicules très haut de gamme dont la puissance moteur est une caractéristique vendeuse importante.

Enfin, la marque BMW i (ou BMW I), dispose du plus faible taux d'émission de CO2 moyen. En effet, BMW i est une filiale du groupe BMW consacrée aux véhicules hybrides et électriques.

Graphique 5 : Corrélation par paire entre les variables de masse, cylindrée, puissance moteur, largeur de voies d'essieux, d'empattement et d'émissions de CO2 (WLTP)



Cette carte de chaleur (ou heatmap) établit une corrélation par paire des caractéristiques techniques (variables continues) des véhicules du dataset, par la méthode de Pearson.
En s'intéressant aux corrélations avec notre variable cible (émissions de CO2), on s'aperçoit que la relation est modérée voire faible avec les variables d'empattement et de largeur de voies d'essieux (respectivement 0.37 et ~0.48-0.49).
En revanche, la dépendance est plus forte avec la variable de masse des véhicules (0.59), et les variables de puissance moteur et de cylindrée montrent la plus forte dépendance avec la variable cible (respectivement 0.76 et 0.74).

La matrice de corrélation met également en évidence une dépendance très importante (0.99) entre les deux variables de largeur de voies d'essieux. Cette relation s'explique notamment par le fait que 84,5% des valeurs de ces variables sont identiques. Nous y reviendrons par la suite dans la partie Pre-processing.

Pre-Processing

Suppression de variables

Les variables suivantes sont supprimées :

- Le numéro d'homologation et l'identifiant de la famille d'interpolation devenus inutiles ;
- Les variables en lien avec les contrôles de simulation CO_2MPAS disponible pour seulement un échantillon de 10% de véhicules ;
- La consommation d'énergie électrique qui ne concerne que les véhicules électriques non émetteurs direct de CO_2 et, de fait, hors du sujet de l'étude;
- la marque et les 3 variables correspondant au nom du constructeur, seul l'item pool de constructeur est conservé.

Suppression de véhicules

Les 3 404 véhicules (20%) suivants sont supprimés :

- 3 147 véhicules sans mesure d'émissions de CO_2 selon le protocole WLTP: ces véhicules sont essentiellement des véhicules de fin de série, la nouvelle réglementation ne leur étant pas imposée.
- 76 véhicules codés "out of scope" et "duplicated" pour le nom du constructeur selon la dénomination standard européenne (variable Mh). Ce codage est vraisemblablement réalisé par l'AEE pendant la phase de réception des données, les guidelines précisant que les caravanes, corbillards et ambulances sont considérés hors champs
- 153 véhicules électriques et 3 véhicules à hydrogène n'émettant pas de CO_2 et dont l'analyse est sans intérêt dans le cadre de la problématique développée.
- 41 véhicules alimentés avec un carburant d'utilisation marginale et dont l'impact sur les émissions de CO_2 ne pourra pas être exploré : gaz naturel (26), gaz de pétrole liquéfié (19), mélange gaz-naturel et biométhane (4), super-éthanol (2).

Le dataset "Véhicules" final contient **13 571 véhicules.**

Traitement des données manquantes

Les données sources sont produites selon un schéma de communication et de vérification exécuté par les administrations des États membres selon des procédures réglementaires. Les émetteurs disposent de *guidelines* et se doivent de transmettre des informations conformes et uniformes. Par ailleurs, la collecte des données ayant débuté en 2010 on peut considérer que le circuit de production est bien rodé. On observe donc sans surprise des taux de données manquantes très faibles (tableau 2).

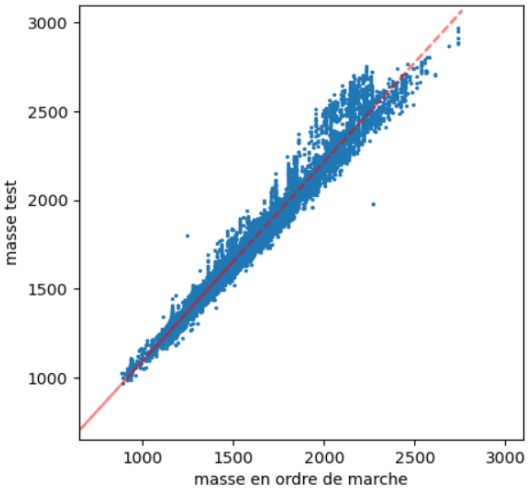
Tableau 2. Pourcentage de données manquantes

		données manquantes	
Item		n	%
Mt	masse test	698	5.12
Fm	mode de carburation	8	0.06
ep	puissance nette maximale	2	0.01

Le mode de carburation est manquant pour 8 véhicules BMW. L'exploration de ces véhicules montre que le carburant est lui aussi manquant car codé "unknown". Cependant, le nom commercial des véhicules indique la technologie de transmission des véhicules de la marque BMW[15]. Le carburant de ces véhicules 520d est donc recodé en "diesel" et leur mode de carburation en "M" pour monocarburant.

La puissance nette maximale manquante pour deux véhicules est recherchée sur internet. La BMW i3 dans sa version de base ou spéciale dispose d'une puissance identique de 125 kW.

La masse test utilisée pour le calcul de l'émission selon le protocole WLTP est manquante pour 698 véhicules. La masse test étant fortement corrélée à la masse en ordre de marche (coefficient de corrélation de Pearson égal à 0,98), les données manquantes sont implémentées par régression linéaire. Ainsi, les valeurs manquantes pour la masse test sont remplacées par les valeurs prédites à partir d'un modèle de régression simple de la masse test (Mt) par la masse en ordre de marche (m) : Mt = a+bm. Une indicatrice est créée afin de pouvoir filtrer le jeu de données sur les données implémentées ce qui permettra de vérifier que ces valeurs estimées ont peu d'impact sur la modélisation.



Recodage de variables

Seul un nombre limité de véhicules disposent d'éléments d'**éco-innovation** et plusieurs de ces innovations ne sont enregistrées que très rarement. L'information relative à ces dispositifs est donc recodée par une indicatrice indiquant leur présence ou non sur le véhicule.

Le type de carburant et le mode de carburation ont des modalités de codage qui induisent des croisements de modalités non réalisables. Par exemple, les véhicules à essence (mono-carburation) ne peuvent pas être des véhicules hybrides. L'impossibilité formelle d'observer des individus dotés de certaines caractéristiques peut provoquer des difficultés d'estimation des coefficients dans les modèles linéaires. Pour se prémunir de ces difficultés, le mode de carburation est simplifié et recodé en deux modalités, essence ou diesel. Le type de carburation reste codé mono-carburation, hybride rechargeable (OVC-HEV) et hybride non rechargeable (NOVC-HEV).

Les mesures de **largeur de voie** de l'essieu directeur et largeur de voie de l'autre essieu sont identiques pour 85% des véhicules. Ceci est étonnant, le site Citroën indiquant que "les autos qui tractent par l'avant auront, surtout pour les voitures les plus récentes, des voies avant plus larges qu'à l'arrière. Sur les 4X4 on aura souvent une équivalence alors que sur les propulsions on aura des voies arrière généralement plus larges qu'à l'avant". La largeur de voie de l'essieu directeur devrait donc être supérieure dans la grande majorité des cas et il est donc probable que cette caractéristique ait été mal renseignée. Quoi qu'il en soit, il est inutile de retenir les deux mesures et on conservera dans les modèles de régression la largeur de voie de l'essieu directeur. Une

[15] Caractéristiques des modèles BMW

variable binaire est créée pour indiquer une largeur de voie de l'essieu directeur plus petite, ce qui correspond à 14% des véhicules.

Concernant le constructeur, les regroupements présentés dans le rapport européen "Analyses of emission from new cars in 2020[16]" ont été repris : VW (4 002 véhicules), BMW (2 013), PSA-Opel (1 715), Daimler (1 581), RNM (996), FCA-Tesla (789), Ford-Werke GMBH (749), Toyota-Mazda (604), JLR (437), Kia (167), Hyundai (97), Suzuki (54). Les constructeurs non cités sont regroupés dans la modalité "autre" (418 véhicules).

Modélisation

La modélisation du taux d'émission de CO_2 en fonction des caractéristiques du véhicule est appréhendée selon deux approches : la régression et la classification.

Chacune des approches privilégie dans un premier temps la production de modèles classiques permettant de mieux comprendre les liens entre les items. Au moins un modèle de machine learning de capacité prédictive supérieure est ensuite implémenté. Le modèle prédictif le plus performant est exploité au mieux en calibrant finement ses hyperparamètres.

Régression

Par soucis de simplification, les résultats détaillés de construction des modèles ne seront présentés que pour la norme WLTP, seules les mesures synthétiques de prédiction seront présentées pour les deux normes.

Une analyse descriptive de la distribution des deux mesures du taux d'émission de CO_2 et de leur lien avec les prédicteurs est réalisée avant la modélisation. Cette phase exploratoire permettra de valider l'intérêt de mettre en œuvre un modèle de régression linéaire.

Pour la phase de modélisation, le *dataset* est partagé en données d'apprentissage (0,75%) et données de test (0,25%). Les **variables quantitatives sont centrées et réduites** à partir des estimations de la moyenne et de la variance réalisées sur l'échantillon d'entraînement. Les **variables qualitatives sont dichotomisées** en choisissant les modalités de référence suivantes : pas d'éco-innovation, largeur inférieur de l'essieu non directeur, carburateur à essence, monocarburation, deux roues motrices, pool de constructeur BMW.

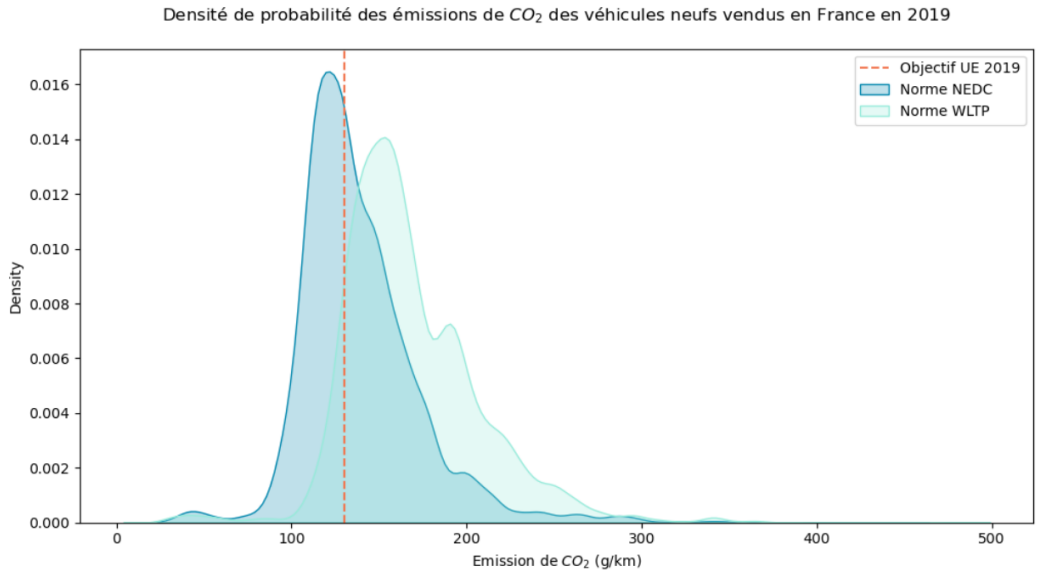
Variables d'intérêt

La distribution des émissions de CO_2 des véhicules montre des valeurs très regroupées présentant une dissymétrie de distribution avec un étalement plus marqué pour les valeurs élevées. La distribution est continue mais ne suit pas une loi normale. Sans que l'on puisse à proprement parler de distribution bimodale, on observe un pic, particulièrement prononcé pour la norme WLTP, vers 200g/km.

Comme attendu, les valeurs pour la norme WLTP sont supérieures. Elles sont aussi plus étendues, ce qui correspond à l'individualisation de la masse de chaque véhicule selon les

[16] https://www.transportenvironment.org/wp-content/uploads/2021/11/2021_11_car_co2_report_technical_annex.pdf

options choisies à l'achat. On note que l'objectif d'émission de CO_2 fixé par l'Union européenne adossé à la nouvelle norme WLTP est un nouveau défi pour les constructeurs.

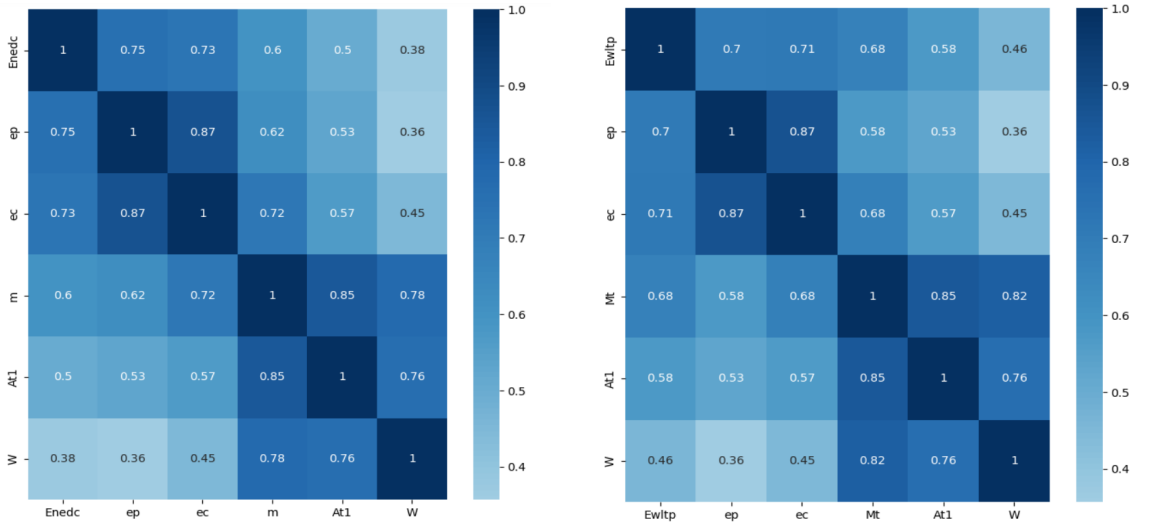


Liaisons entre variables

A chaque norme est associée une mesure de masse réglementaire : masse en ordre de marche pour la norme NEDC et masse test pour la norme WLTP. Pour modéliser chaque norme, on utilisera la mesure de masse correspondant.

Les heat maps montrent que le taux d'émission de CO_2 est corrélé aux 5 caractéristiques quantitatives : fortement pour la puissance, la cylindrée et le poids, un peu moins pour l'empreinte au sol. De plus, les variables de puissance et de cylindrée d'une part et les variables de poids et d'empreinte au sol d'autre part, sont très corrélées entre elles. Les modèles linéaires étant sensibles à la multicolinéarité, cette observation constitue un point de vigilance.

Corrélations linéaires (coefficient de Pearson) entre caractéristiques quantitatives des véhicules



Toutes les variables qualitatives sont liées au taux d'émission de CO_2 (tests de student ou Anova significatifs avec p<0,001) mais parfois pour des écarts entre les valeurs moyennes assez

faibles. Le tableau 2 indique les émissions moyennes pour chaque catégorie sachant que la valeur moyenne sur le *dataset* est de 139 g/km pour la norme NEDC et 168,4 g/km pour la norme WLTP.

Tableau 2 : Moyenne des émissions de CO2 selon les caractéristiques qualitatives des véhicules

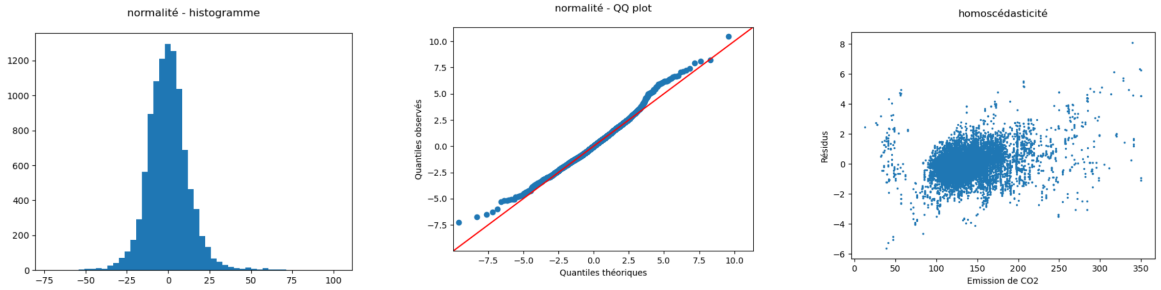
Variable	**Modalités**	**Emission moyenne de CO_2 (g/km)**	
		NEDC	**WLTP**
Largeur supérieure de l'essieu non directeur	non	137,8	166,8
	oui	146,2	178,8
Eco-innovation	non	139,6	168,8
	oui	137,6	167,5
Carburant	diesel	136,9	169,9
	essence	141,3	166,7
Carburation	mono	139,7	169,1
	NOVC-HEV	145,3	178,1
	OVC-HEV	50,6	55,8
Motricité	2 roues motrices	135,0	163,0
	4x4	176,8	220,1
Pool de constructeurs	BMW	142,4	173,1
	Daimler	160,3	186,6
	FCA-Tesla	137,6	162,3
	Ford-Werke GMBH	128,9	155,7
	Hyundai	128,2	148,5
	JLR	177,1	223,6
	KIA	124,9	143,9
	PSA-Opel	118,8	152,7
	RNM	131,4	155,3
	Suzuki	121,3	143,0
	Toyota	128,1	155,3
	VW	137,3	168,0
	Autres	165,0	187,8

Modèles de régression linéaire simple et régularisé

Un modèle de base a été produit à partir d'une **régression linéaire** multiple sans sélection de variables et sans paramétrage particulier (pour rappel, seuls les résultats détaillés pour la norme WTLTP sont produits).

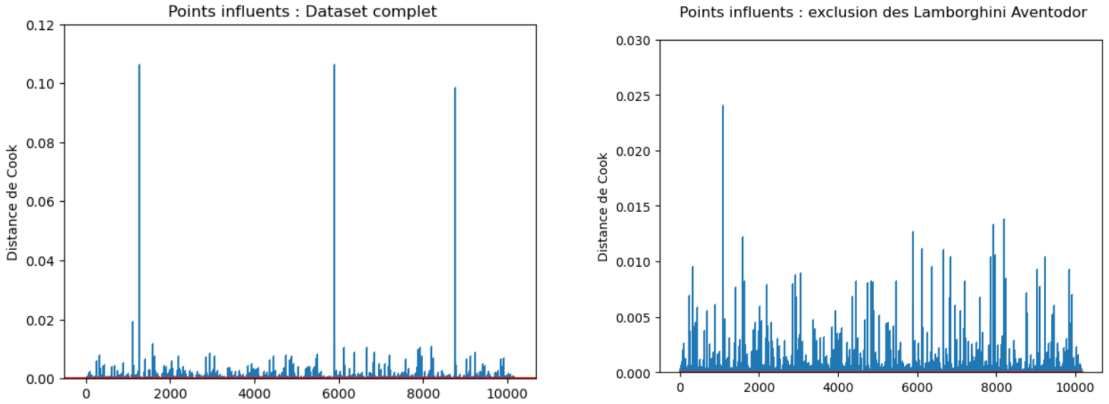
Les hypothèses de normalité et d'homoscédasticité des résidus ont été testées sur ce modèle de base. Si l'hypothèse de normalité peut être retenue, la variabilité des résidus est clairement dépendante des valeurs d'émission de CO_2 avec une variabilité plus élevée pour les valeurs hautes et basses. Le test de Breush-Pagan confirme que l'hypothèse d'homoscédasticité ne peut être retenue. L'allure du nuage de points indique que le modèle prédit moins bien les taux d'émission des véhicules avec des faibles ou des fortes valeurs.

Modèle de régression linéaire : vérification des hypothèses de distribution des résidus



La représentation des distances de Cook permet d'identifier trois véhicules particulièrement atypiques : il s'agit de trois différentes versions de la Lamborghini Avantador, un véhicule aux caractéristiques techniques effectivement hors norme. La régression linéaires étant très sensible aux valeurs extrêmes, ces véhicules ont été exclus du *dataset* d'entraînement. Néanmoins, on note que l'influence de certains véhicules dans l'estimation des coefficients de régression reste très marquées (figure de droite).

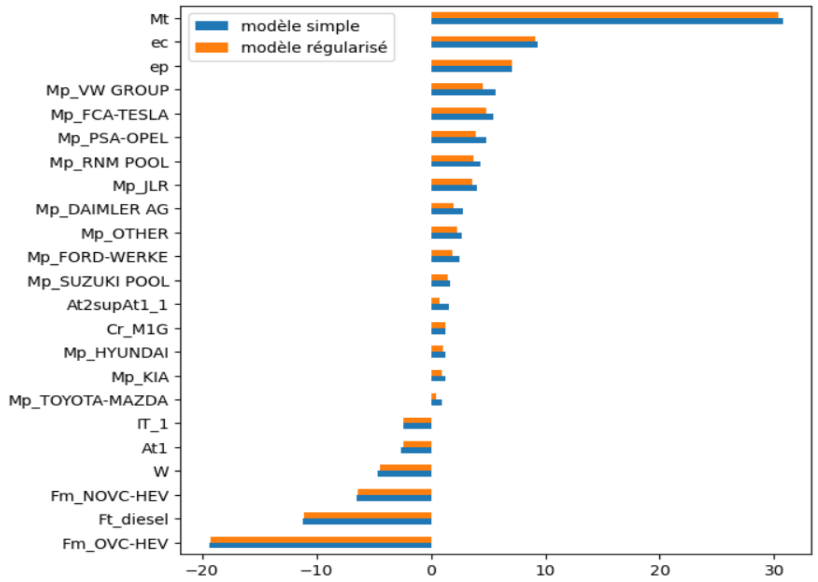
Analyse des points influents : distance de Cook



Les performances globale du modèle sont élevées : **coefficient de détermination de 0,89** sur les données d'entraînement et de test. Le modèle ne semble pas souffrir de sur-ajustement avec des écarts quadratiques moyens (RMSE) de 13,1 pour les données d'entraînement et 12,76 pour les données tests.

Compte tenu des corrélations fortes observées entre certaines caractéristiques techniques des véhicules, une régularisation est appliquée. Un modèle **Elastic Net** est implémenté avec une optimisation des paramètres de régularisation à partir d'une méthode *grid search*. Les valeurs de pénalités retenues modifient peu le modèle de base (pénalité L1 de 0,03 et L2 de 0,0003). La représentation graphique des coefficients obtenus pour le modèle simple et le modèle régularisé montre à la fois qu'aucun paramètre n'est exclu du modèle, et que les valeurs des coefficients restent très proches. Néanmoins, le modèle pénalisé permet bien de diminuer la différence des RMSE entre données d'entraînement (12,70) et données test (12,77).

Influence de la régularisation sur les coefficients de régression linéaire



Modèle XGBoost

XGBoost est un modèle d'ensemble très utilisé pour ces qualités prédictives. Ses principaux défauts sont le sur-apprentissage et un nombre très important d'hyperparamètres à optimiser.

Une optimisation simultanée des hyperparamètres à partir de la méthode *grid search* qui teste toutes les combinaisons de paramètres possibles demanderait un temps d'exécution faramineux. Une méthode séquentielle a donc été adoptée : les paramètres sont optimisés de façon isolée ou par paire, les modèles suivants intégrant les paramètres sélectionnés dans les étapes précédentes. Cette méthodologie ne permet pas d'obtenir les paramètres optimaux mais une sélection satisfaisante pour un temps de calcul acceptable.

Séquence d'optimisation des hyperparamètres du modèle XGBoost (métrique RMSE, 10 blocs) :

- Un taux d'apprentissage ou facteur de rétrécissement (learning_rate) élevé est fixé de façon à obtenir un nombre raisonnable d'arbres (n_estimator).
- Les paramètres de structure des arbres sont optimisés par validation croisée : profondeur maximale (max_depth) et nombre minimal d'individus requis dans chaque nœud (min_child_weight).
- Optimisation de l'hyperparamètre de régularisation de la profondeur des arbres (gamma).
- Les hyperparamètres d'échantillonnages sont calibrés à leur tour : ratio d'échantillonnage des variables lors de la construction de chaque arbre (colsample_bytree) et ratio de sous-échantillonnage de l'instance d'entraînement (subsample). Pour ce paramètre, une valeur de 0,5 signifie que XGBoost recueille de façon aléatoire la moitié des instances pour développer les arbres. Ces deux paramètres réduisent le surajustement.
- Les conditions de régularisation L1 (reg_alpha) et L2 (reg_lambda) sur les pondérations sont finalement optimisées de nouveau dans un souci d'éviter le surajustement.
- Lorsque tous les paramètres sont fixés, le taux d'apprentissage est réduit tout en augmentant le nombre d'arbres pour renforcer la précision du modèle.

Tableau 3 : Calibrage des hyperparamètres du modèle XGBoost par sélection séquentielle

	Valeurs testées	Valeurs retenues	Meilleur score	Durée d'exécution
Etape 0				
- learning_rate	0,9 - 0,5 - 0,2 - 0,1	0,2		
- n_estimator		100		
Etape 1				
- max_depth	2 à 16	15	-5,924	5mn55s
- min_child_weight	1 à 18	3		
Etape 2				
- gamma	0,5 - 5 - 10 - 15 - 20 - 25	20	-6,037	1mn41s
Etape 3				
- colsample_bytree	0,6 à 1	1	-6,010	3mn41s
- subsample	0,6 à 1	0,8		
Etape 4				
- reg_alpha	0 à 0,5	0,1	-6,063	7mn37s
- reg_lambda	1 à 2,5	2		

Pour le **modèle initial**, le taux d'apprentissage a été fixé à 0,2 et autres paramètres laissés aux valeurs par défaut. Le nombre d'arbres optimal obtenu par validation croisée est de 100. Les mesures de performances atteignent :

- entrainement : R^2 = 0,978 et RMSE = 5,937
- données test : R^2 = 0,971 et RMSE = 6,58

Pour le **modèle final**, des taux d'apprentissage de 0,01 et 0,005 ont été testés en augmentant fortement le nombre d'arbres, les valeurs des autres paramètres correspondant aux valeurs optimisées (tableau 3). Le taux de 0,005 donne le meilleur score. Les mesures de performances sont améliorées au détriment d'un sur-apprentissage :

- entrainement : R^2 = 0,994 et RMSE = 3,07
- données test : R^2 = 0,980 et RMSE = 5,50

Synthèse régression

Modèle de régression	Données d'entraînement			Données test		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Norme WLTP						
- Modèle linéaire simple	0,890	13,01	9,43	0,892	12,76	9,27
- Modèle linéaire régularisé	0,895	12,77	9,27	0,892	12,70	9,38
- Modèle XGBoost	0,994	3,07	2,24	0,980	5,50	3,79
Norme NEDC						
- Modèle linéaire simple	0,864	13,02	9,32	0,862	12,76	9,16
- Modèle linéaire régularisé	0,867	12,71	9,25	0,861	12,77	9,13
- Modèle XGBoost	0,996	2,28	1,41	0,986	4,12	2,42

Les résultats obtenus pour les deux normes sont similaires conformément à ce que la lecture des textes réglementaires laissait présager pour la grande majorité des véhicules, la norme

NEDC en 2019 est une prédiction réalisée à partir des mesures réalisées pour la norme WLTP (lire le chapitre “les émissions de CO_2“ pour plus de détail).

La seule différence notable est un meilleur R^2 pour le modèle linéaire avec la norme WLTP (+3%) très certainement en lien avec une mesure plus fine de la masse et de l’émission de CO_2, les deux paramètres étant très fortement corrélés.

Le taux d’émission de CO_2 est très bien prédit à partir des données disponibles:

- Le modèle de régression linéaire donne d’emblée de bons résultats sans nécessité de régularisation mais l’hypothèse d'homoscédasticité des résidus est prise en défaut.
- Le modèle XGBoost prédit beaucoup mieux la valeur d’émission de CO_2 mais au prix d’un léger surajustement.

Pour aller plus loin :

- le modèle de régression linéaire pourrait être exploré plus avant :
    - en testant certaines interactions entre facteurs prédictifs
    - en transformant en quantile la variable de cylindrée qui n’est pas continue (80% des véhicules prennent seulement 20 valeurs)
    - en transformant la masse pour explorer une relation non linéaire (utilisation de polynômes fractionnaires)
- le modèle XGBoost pourrait être amélioré :
    - en utilisant une métrique moins sensible aux valeurs extrêmes pour le calibrage des hyperparamètres (RMSLE)
    - en sélectionnant seulement les facteurs prédictifs les plus importants pour éviter le surajustement
    - en transformant en facteur via une ACP les paramètres quantitatifs très corrélés de nouveau pour diminuer le surajustement
    - l’optimisation séquentielle sélectionne une profondeur d’arbre élevée régulée par une forte valeur de gamma sans pour autant éviter un léger surajustement. Un paramétrage mieux ajusté aux données pourrait être recherché en examinant l’évolution des fonctions de pertes pour les données d'entraînement et de test[17].

Les deux modèles sont concordants sur l’importance des facteurs de prédiction. Ce point sera exploré plus avant dans le chapitre interprétabilité.

[17] https://medium.com/data-design/xgboost-hi-im-gamma-what-can-i-do-for-you-and-the-tuning-of-regularization-a42ea17e6ab6

Classification

Variables d'intérêt

Les variables d'intérêt d'émission de CO_2 ont été découpées en quartiles afin de produire des modèles de classification. De cette façon, il n'y a pas de déséquilibre de classes, bien qu'en raison de la dissymétrie de distribution des valeurs, les classes intermédiaires (2ème et 3ème quartiles) s'étendent sur une plage étroite de valeurs, ce qui rendra leur détection par le modèle probablement plus difficile.

Pour la norme NEDC, les classes représentent les intervalles suivants :

- De 12,9 à 117 g/km, classe 1.
- De 117 à 132 g/km, classe 2.
- De 132 à 154 g/km, classe 3.
- De 154 à 464 g/km, classe 4.

Pour la norme WLTP :

- De 3,9 à 143 g/km, classe 1.
- De 143 à 160 g/km, classe 2.
- De 160 à 190 g/km, classe 3.
- De 190 à 499 g/km, classe 4.

Comme pour la régression, les résultats détaillés de construction des modèles seront présentés avec la norme WLTP. Les résultats pour les deux normes seront présentés en synthèse.

Modèle de régression logistique

Un modèle de base a été produit à partir d'une **régression logistique multinomiale.** Il s'agit d'un modèle linéaire dont chaque variable explicative est associée à des coefficients qui permettent de comprendre l'impact de chaque variable sur le choix (entre 0 et 1). Ce modèle étant très simple, il y a peu de risques de sur-apprentissage et les résultats ont tendance à avoir un bon pouvoir de généralisation.

Le modèle initié avec ses paramètres par défaut présente un problème de convergence. De ce fait, le nombre d'itérations max du modèle ("max_iter") a été augmenté de 100 à 500. Les autres hyperparamètres ont été attribués par défaut.

Le score du modèle obtenu sur le jeu de test est de **0.77**, il est très légèrement supérieur au score du jeu d'entraînement (**0.76**). Le modèle ne semble pas souffrir de sur-apprentissage.

L'affichage de la matrice de confusion ci-contre illustre très bien la difficulté du modèle à prédire les classes intermédiaires, et plus particulièrement la deuxième classe, qui pour rappel, s'étend sur une plage très restreinte de valeurs : 143 à 160 g/km (taux d'émission de CO2).

Predictions	1	2	3	4
EwltpQuart				
1	527	103	15	0
2	124	445	116	0
3	12	136	497	55
4	3	2	60	620

Modèle des K plus proches voisins

Le modèle des K plus proches voisins est un modèle qualifié comme paresseux (Lazy Learning) car il n'apprend rien pendant la phase d'entraînement. Pour prédire la classe d'une nouvelle donnée d'entrée, il va chercher ses K voisins les plus proches et choisira la classe des voisins majoritaires. C'est un algorithme simple et déployable rapidement, qui néanmoins devient de mauvaise qualité lorsque le nombre de variables explicatives devient important.

Le modèle est dans un premier temps initié avec ses paramètres par défaut. Le score obtenu sur le jeu d'entraînement est de **0.90** et de **0.84** sur le jeu de test.
Le modèle peut potentiellement être amélioré en testant des métriques de distance différentes : Minkowski, Manhattan et Chebyshev. Le nombre de voisins sera également déterminé en fonction du score obtenu sur ces 3 métriques.

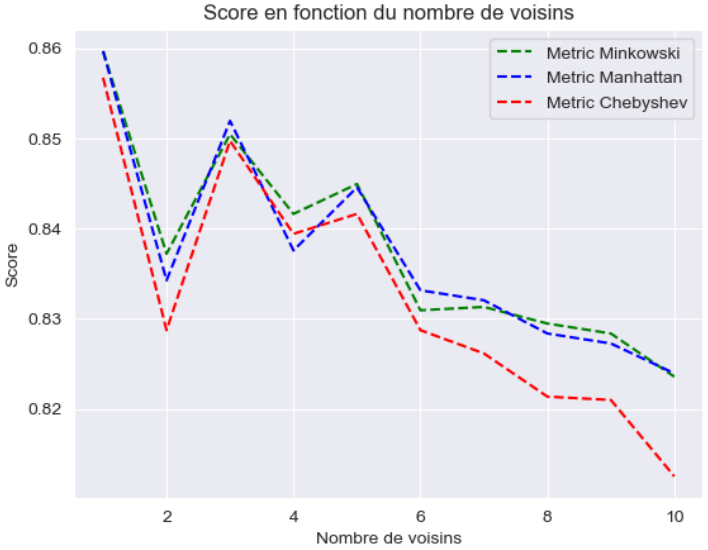
Le graphique ci-contre représente le score obtenu pour chacune des métriques citées précédemment en fonction du nombre de voisins, sur le jeu de test.

Le meilleur score est obtenu en choisissant la métrique Manhattan et en définissant un nombre de voisins égal à 1.

Cependant, choisir un nombre de voisins égal à 1 introduirait un biais important, car le modèle s'ajustera uniquement au voisin le plus proche. Cela signifie que le modèle sera très proche et dépendant des données d'entraînement.
Ainsi, en fournissant de nouvelles données de test, les prédictions risquent d'être très différentes à chaque itération.

Afin d'établir un modèle plus pertinent et potentiellement réutilisable, le nombre de voisins choisi sera de 3, avec la métrique Manhattan.

De cette façon, le modèle obtient un score de **0.92** sur le jeu d'entraînement et de **0.85** sur le jeu de test.



Représentation du score pour chaque métrique en fonction du nombre de voisins

Modèle des arbres de décision

L'algorithme de l'arbre de décision est un modèle simple et très interprétable utilisé pour représenter visuellement et explicitement les décisions et la prise de décision pour des problèmes de classification ainsi que pour des problèmes de régression. Il représente aussi l'élément de base de plusieurs modèles comme le Random Forest ou XGBoost, utilisé précédemment pour la régression et qui fera également suite à ce modèle pour la classification. Cependant, le risque de sur-apprentissage (créer un arbre avec une très grande profondeur) est très élevé pour les modèles non-paramétriques, dont il fait partie.

Le modèle est initié avec ses paramètres par défaut. Le score obtenu sur le jeu d'entraînement est de **0.99** et met tout de suite en évidence un problème de sur-apprentissage. Le score obtenu sur le jeu de test est de **0.83**.

Dans un souci d'optimisation et afin de réduire le sur-apprentissage du modèle, les hyperparamètres optimaux seront sélectionnés par validation croisée dans une grille de recherche : méthode GridSearchCV.

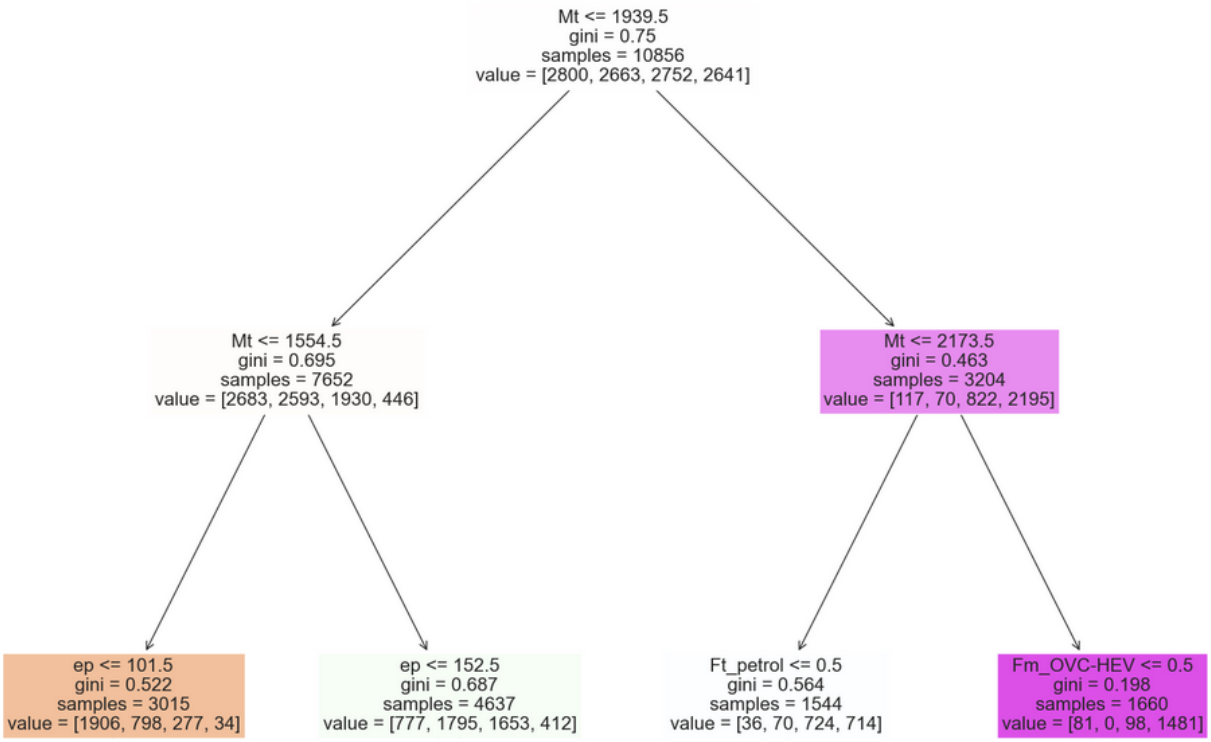
En limitant la profondeur de l'arbre de décision à 15 branches (paramètre optimal déterminé par GridSearchCV), le score obtenu pour le jeu d'entraînement est de **0.95** et de **0.85** pour le jeu de test. Le sur-apprentissage a été réduit et le score du jeu de test amélioré.

Pour connaître les variables qui ont le plus contribué à déterminer le classement des véhicules, l'attribut "feature_importances_" retourne l'importance normalisée de chaque variable dans la construction de l'arbre.
Cette importance est définie comme le décroissement total entre un nœud et les deux suivants du critère d'impureté utilisé pour diviser le nœud (ici "gini"). Plus l'écart entre l'impureté calculée pour un nœud et ses nœuds 'fils' est élevé, plus la variable utilisée pour diviser le nœud est importante.

	Importance
Mt	0.496196
ep	0.127802
W	0.113435
At1	0.087773
ec	0.054521
Fm_OVC-HEV	0.026483
Ft_petrol	0.023909
Ft_diesel	0.023279

La variable 'Mt', qui correspond à la masse d'un véhicule dans les conditions du test de masse de la norme WLTP, se détache fortement des autres variables explicatives, avec une importance proche de 50%.
La variable 'ep', qui correspond à la puissance moteur d'un véhicule, présente une importance d'environ 13%.
Enfin, les variables associées à l'empattement, à la largeur de voie de l'essieu directeur et à la cylindrée ont une importance respective de 11%, 9% et 5% environ.

La visualisation des premières branches de l'arbre confirment l'importance de la variable 'Mt' et 'ep' dans le calcul pour la division des noeuds, et met également en évidence deux autres variables importantes dans les premières étapes de l'algorithme : les variables catégorielles 'Ft_petrol' et 'Fm_OVC-HEV', qui identifient respectivement les véhicules essence et les véhicules hybrides rechargeables.



Représentation des premières branches de l'arbre de décisions

Modèle XGBoost

Comme abordé précédemment dans la partie régression, XGBoost est un modèle très performant mais contient un nombre important d'hyperparamètres à optimiser. L'optimisation via GridSearchCV évalue chaque combinaison d'hyperparamètres pour le modèle. Elle prend donc énormément de temps lorsqu'il y a beaucoup de combinaisons à tester.
Une autre méthode de réglage des hyperparamètres sera ici explorée, l'optimisation bayésienne, à l'aide de la librairie Hyperopt. L'optimisation bayésienne utilise les résultats de l'étape précédente pour décider de la combinaison d'hyperparamètres à évaluer ensuite.
En d'autres termes, plutôt que d'oublier toute l'information récupérée comme le ferait une méthode de recherche par grille (GridSearchCV), celle-ci est conservée et oriente l'algorithme au fil des itérations.

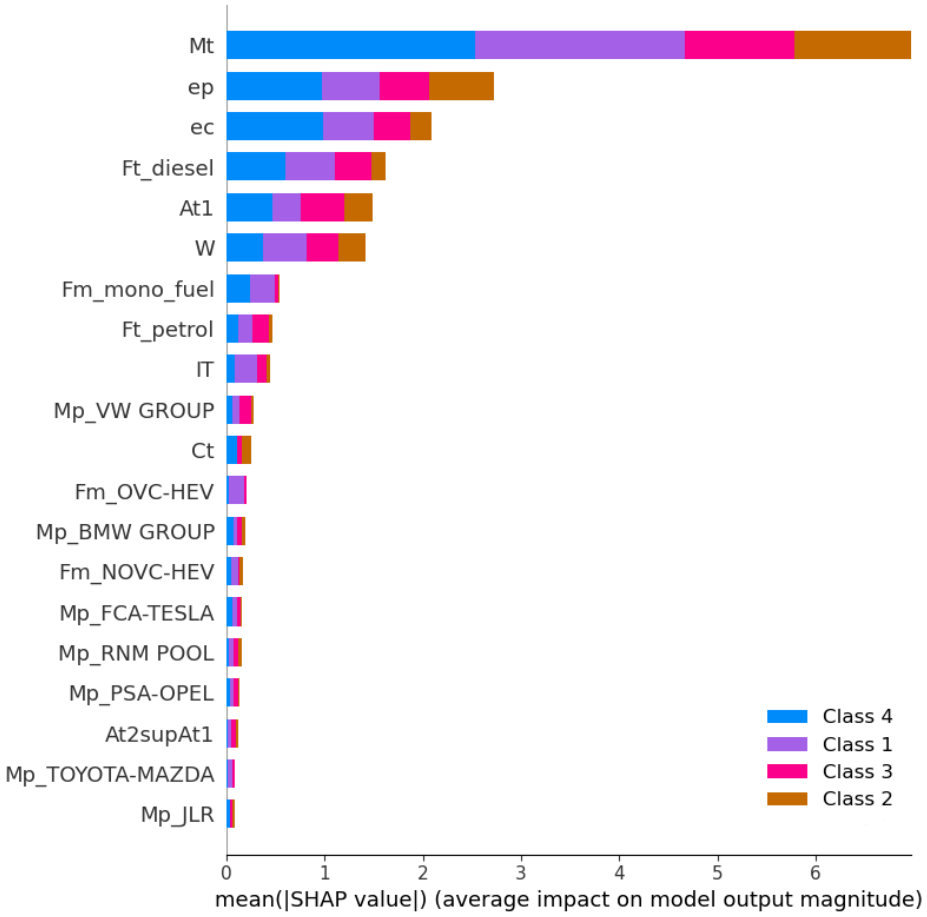
Dans un premier temps, un modèle XGBoost simple est instancié sans recherche de paramètres optimaux. Le score obtenu sur le jeu d'entraînement est de **0.95** et de **0.87** sur le jeu de test (norme WLTP, pour rappel). Celui-ci est supérieur à tous les autres modèles testés, sans même avoir procédé à l'optimisation des hyperparamètres. Les scores obtenus sont encore meilleurs pour la norme NEDC : **0.94** sur le jeu d'entraînement et **0.90** sur le jeu de test.

Les paramètres à faire varier sont les mêmes que ceux définis précédemment dans la partie régression. L'étendue des valeurs sur laquelle chaque paramètre est testé ainsi que la valeur retenue après optimisation sont résumées dans le tableau ci-dessous.

		Valeurs testées	Valeurs retenues
-	learning_rate	0,1 à 1	0,4
-	max_depth	2 à 16	9
-	min_child_weight	1 à 10	1
-	gamma	0 à 0,3	0,12
-	colsample_bytree	0,1 à 1	0,9
-	subsample	0,1 à 1	0,8
-	reg_alpha	0 à 1	0,25
-	reg_lambda	0 à 2,5	0,6

Tableau 1 : Calibrage des hyperparamètres du modèle XGBoost par optimisation bayésienne

L'optimisation du modèle a permis d'améliorer sensiblement les performances de prédiction.
Pour la norme WLTP, le score obtenu sur le jeu de test est passé de **0.87** à **0.88**.
Pour la norme NEDC, le score obtenu sur le jeu de test est passé de **0.90** à **0.91**.



Importance des variables selon les valeurs de Shap

Synthèse classification

Modèle de classification : Norme NEDC	Données d'entraînement F1-Score					Données test F1-Score				
	Classe 1	Classe 2	Classe 3	Classe 4	Average	Classe 1	Classe 2	Classe 3	Classe 4	Average
Modèle régression logistique	0.79	0.61	0.67	0.89	**0.74**	0.80	0.60	0.66	0.90	**0.74**
Modèle KNN	0.94	0.88	0.90	0.96	**0.92**	0.91	0.81	0.86	0.93	**0.88**
Modèle Decision Tree	0.95	0.93	0.94	0.97	**0.95**	0.92	0.84	0.90	0.95	**0.90**
Modèle XGBoost	0.95	0.93	0.94	0.97	**0.95**	0.92	0.85	0.90	0.95	**0.91**

Modèle de classification : Norme WLTP	Données d'entraînement F1-Score					Données test F1-Score				
	Classe 1	Classe 2	Classe 3	Classe 4	Average	Classe 1	Classe 2	Classe 3	Classe 4	Average
Modèle régression logistique	0.81	0.63	0.70	0.90	**0.76**	0.80	0.65	0.72	0.91	**0.77**
Modèle KNN	0.94	0.87	0.91	0.97	**0.92**	0.87	0.77	0.83	0.94	**0.85**
Modèle Decision Tree	0.94	0.88	0.93	0.99	**0.93**	0.88	0.77	0.80	0.93	**0.84**
Modèle XGBoost	0.96	0.92	0.94	0.99	**0.95**	0.90	0.80	0.86	0.95	**0.88**

Les résultats obtenus sont meilleurs pour la norme NEDC, à l'exception du modèle de régression logistique, qui obtient de meilleurs scores selon la norme WLTP.
La meilleure performance est atteinte avec le modèle XGBoost (norme NEDC) après optimisation bayésienne (91% d'accuracy sur le jeu de test).

Malgré de très bons résultats, l'approche de classification paraît moins pertinente pour répondre à une problématique d'estimation du taux d'émission de CO_2.
Néanmoins, il est tout à fait possible qu'un système de classes pour les véhicules plus ou moins polluants soit mis en place à l'avenir, à l'instar des classes énergétiques que l'on peut trouver sur la plupart des produits électroniques aujourd'hui.
On pourrait notamment penser au dispositif Crit'air, mis en place en 2015, qui est censé catégoriser les véhicules selon la pollution qu'ils engendrent, mais les catégories Crit'air ne prennent pas en compte l'émission de gaz à effets de serre ou de particules fines d'un véhicule mais seulement le type de carburant et l'année de construction du véhicule.

Synthèse modélisation

Le modèle XGBoost est sans surprise le meilleur modèle de prédiction aussi bien pour l'approche régression que classification.

Pour chaque approche mises en œuvre - régression et classification - une stratégie différente d'optimisation automatique des hyperparamètres du modèle XGBoost a été développée : sélection séquentielle pour la régression et bayésienne (librairie Hyperopt) pour la classification. Profitant de la mise au point des deux méthodes sur nos données, une comparaison a été réalisée sur le modèle de régression.

Les conclusions sont les suivantes :

- La méthode bayésienne est deux fois plus rapide : la durée d'exécution est de 9 minutes et 20 secondes alors que la méthode séquentielle nécessite 18 minutes et 33 secondes.
- Les valeurs des hyperparamètres sélectionnés sont proches excepté pour les paramètres de régularisation : le paramètre gamma est quatre fois plus faible (5 vs 20), le paramètre alpha 10 fois plus fort (1 vs 0,1).
- Les résultats des mesures de performance des modèles sélectionnés selon les deux méthodes sont identiques.

La sélection bayésienne de paramétrage du modèle XGBoost ne permet pas d'améliorer les performances du modèle mais elle est cependant plus rapide à mettre en œuvre. Pour aller plus loin, il serait intéressant d'examiner les véhicules impactés par les changements de paramètres.

Interprétation

Les deux approches (régression et classification) et les deux normes (NEDC et WLTP) donnent des résultats similaires. Pour ne pas alourdir le propos, nous avons choisi de présenter les résultats de la régression, puisque cette approche correspond mieux à la nature du taux d'émission de CO_2 (variable quantitative), tout en privilégiant la norme WLTP qui est dorénavant la mesure de référence.

Les résultats obtenus correspondent aux connaissances établies : la masse du véhicule détermine majoritairement le taux d'émission de CO_2 des véhicules. Selon la régression linéaire, la masse expliquerait un peu plus de 30% des émissions.

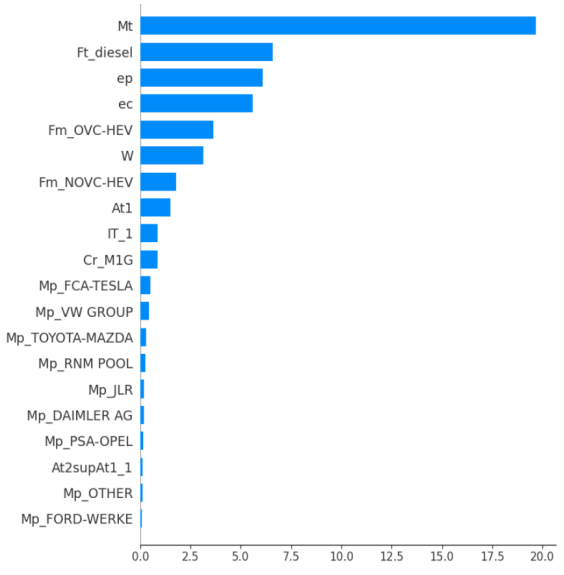
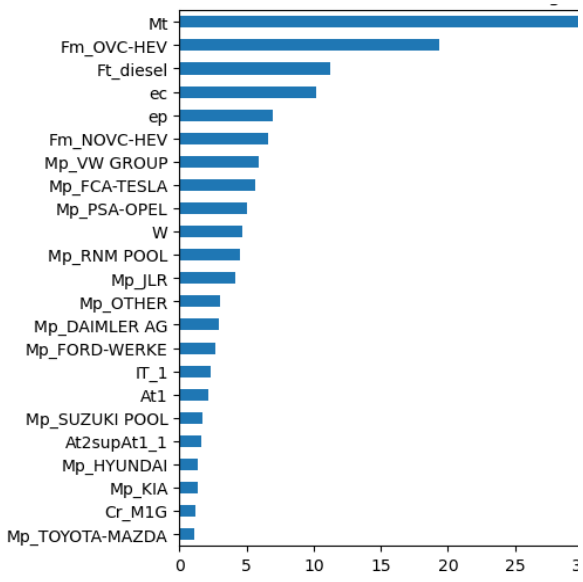
Les deux modèles produisent des informations très proches. Toutefois, nous relevons deux points :

- Concernant les caractéristiques techniques : les deux modèles donnent une importance différente au mode de carburation. En effet, pour la régression linéaire, la modalité "véhicules hybrides rechargeables" arrive immédiatement derrière le critère de masse et avant le triptyque cylindré/puissance/carburant ; le modèle XGBoost inverse ce positionnement.
  Il s'agit probablement d'une interaction non traitée dans le modèle linéaire, vraisemblablement entre le mode de carburation (véhicule hybride ou non) et le type de carburant (essence ou diesel). Ce point est à examiner.

- L'importance accordée au constructeur est différente et ne produit pas le même classement. Le modèle XGBoost semble le plus cohérent. Il nous indique que les caractéristiques techniques des véhicules relevées par la Commission sont prépondérantes (à une exception près, cette exception correspondant à la différence de mesure de voie, un item dont la qualité de recueil a été mise en doute lors de la phase exploratoire). Une fois les caractéristiques techniques prises en compte, le constructeur semble jouer un rôle résiduel. On pense ici à l'impact de caractéristiques telles que l'aérodynamisme, les réglages des appareils mécaniques ou électroniques qui relèvent d'un savoir-faire propre à certains industriels.

  Ce qui est particulièrement cocasse dans les résultats obtenus est que les deux constructeurs qui arrivent en tête sont les deux groupes industriels entachés de "scandale" sur les taux d'émission de CO_2. Selon le *Finantial Times* au printemps 2019, FCA a accepté de payer à Tesla "des centaines de millions de dollars" pour que les émissions de CO_2 - évidemment nulles - des véhicules de la marque Tesla soient prises en compte dans son périmètre et donc limite le montant des pénalités exigées par la Commission européenne. De son côté, le groupe Volkswagen subissait en 2015 le "*dieselgate*", scandale industriel lié à l'utilisation de différentes techniques visant à réduire frauduleusement les émissions polluantes de certains de ses moteurs lors des essais d'homologation.

Influence des facteurs dans la construction du modèle

XGBoost **moyenne des valeurs absolues des Shapley**	**Régression linéaire :** **valeur absolue des coefficients de régression**
	

Conclusion

Cet exercice nous a permis de mettre en œuvre de nombreuses techniques et malgré des performances exceptionnelles, quelques pistes d'améliorations peuvent être proposées, notamment une sélection de variable ou une réduction de dimension pour les caractéristiques fortement corrélées entre elles.

Les résultats obtenus ont peu de développement possible, l'AEE ayant fait développer un simulateur de pointe disponible sous forme d'API. Néanmoins, pour mettre à l'épreuve le meilleur modèle obtenu, il serait intéressant de tester les performances du modèle sur l'ensemble du parc européen, et notamment sur les véhicules non vendus en France.

Enfin, Il nous paraît important d'aborder plusieurs points éthiques et écologiques :

Nous avons pu voir dans la partie exploratoire que les véhicules hybrides rechargeables (VHR) rejettent en moyenne beaucoup moins de CO_2 que les autres types de véhicule. Cependant, il convient de relativiser ce résultat : ces valeurs d'émission de CO_2 sont établies selon un usage optimal du véhicule, c'est-à-dire en considérant que la batterie est très régulièrement rechargée. Hors, l'actualisation de l'ICCT en juin 2022 sur les usages des VHR montre que le mode électrique n'est utilisé que pour la moitié des distances parcourus pour les particuliers, et seulement 11 à 15 % pour les véhicules d'entreprises, la faute au manque d'infrastructure de recharge ou au non branchement du véhicule lorsque c'est nécessaire. Dans les deux cas, on est bien loin des hypothèses de 70 à 85 % des distances considérées dans les procédures d'homologations de ces véhicules.

De plus, la présence de 2 systèmes de motorisations rendent les véhicules hybrides rechargeables très lourds. Au niveau européen en 2020, ils pesaient plus de 1 900 kg en moyenne (masse en état de marche), soit un surpoids de plus de +600 kg par rapport à la moyenne des modèles essence, quasiment +300 kg par rapport au diesel et même plus de +200 kg de plus que les modèles électriques. Or, il est bien connu que la masse d'un véhicule est le facteur le plus corrélé à l'émission de CO_2.

D'un point de vue purement écologique, il est également important de préciser que le rejet de CO_2 n'est pas le seul facteur polluant associé à l'utilisation d'un véhicule (les véhicules diesels moins émetteurs en CO2 le sont davantage en particules fines).
Par ailleurs, la conception, l'extraction, le transport et la transformation des matières premières ainsi que la fabrication du produit sont autant d'étapes qui utilisent une énergie "invisible" dénommée **énergie grise**.

De plus, la production de batteries de voitures électriques est énergivore. L'extraction et le raffinage du lithium et du cobalt dont elle est composée nécessitent l'utilisation intensive de produits chimiques ayant un impact environnemental.
Néanmoins, une fois cette phase de fabrication passée, le véhicule électrique est bien moins polluant sur toute la durée de son cycle de vie. Selon une **étude T&E, le VE émet 77% moins de CO2** sur toute sa durée de vie que son équivalent thermique.