

Sommaire

Introduction	page 3
1– Cadre de l'étude	page 5
1.1 Jeu de données sélectionné	
1.2 Limitations	
2- Pre-processing et feature engineering	page 11
2.1 Traitement des doublons	
2.2 Gestion des données manquantes	
2.2.1 Nettoyage des valeurs manquantes de la variable 'cylindrée'	
2.2.2 Nettoyage des valeurs manquantes de la variable 'autonomie électrique'	
3- Visualisations et statistiques	page 15
3.1 Nos objectifs dans le cadre de la phase d'exploration	
3.2 Distribution du CO2	
3.3 Etude des variables explicatives	
4- Modélisation	page 23
4.1 Visualisations de la distribution des variables dans le dataset nettoyé	
4.2 Modèles de classification	
4.3 Modèles de régression	
4.4 Choix de modèle le plus adapté au contexte métier	
5. Production Finale	page 45
6. Les limites et suite du projet	page 46
6.1 Les limites	
6.2 Suite du projet	
7. Les difficultés du projet	page 48
8. Conclusion	page 49
9. Annexes	page 50

Introduction

Le changement climatique est une réalité incontestable qui affecte notre quotidien.

Le réchauffement de la France est déjà de +1,7°C par rapport aux années 1850. Si on ne change rien, la trajectoire actuelle amène le monde vers une élévation des températures à +3°C en 2100 ce qui signifie, par exemple, +4°C pour le territoire français.

Quel que soit la géographie, les conséquences pour la vie humaine sont considérables. Les causes du réchauffement climatiques sont globalement communes avec celles de l'augmentation de la pollution.

En effet, les émissions de GES, telles que le dioxyde de carbone (CO2), le méthane (CH4) et le protoxyde d'azote (N2O), sont des facteurs majeurs de la pollution atmosphérique et du changement climatique.

Les émissions mondiales de CO2, principalement dues à la combustion de combustibles fossiles, ont augmenté de manière significative au cours des dernières décennies, contribuant ainsi au réchauffement climatique.

La pollution de l'air, principalement causée par les émissions des véhicules, des centrales électriques, de l'industrie et d'autres sources, est un problème majeur dans de nombreuses régions du monde. Les particules fines, les oxydes d'azote, le dioxyde de soufre et d'autres polluants atmosphériques peuvent avoir des effets néfastes sur la santé humaine et l'environnement.

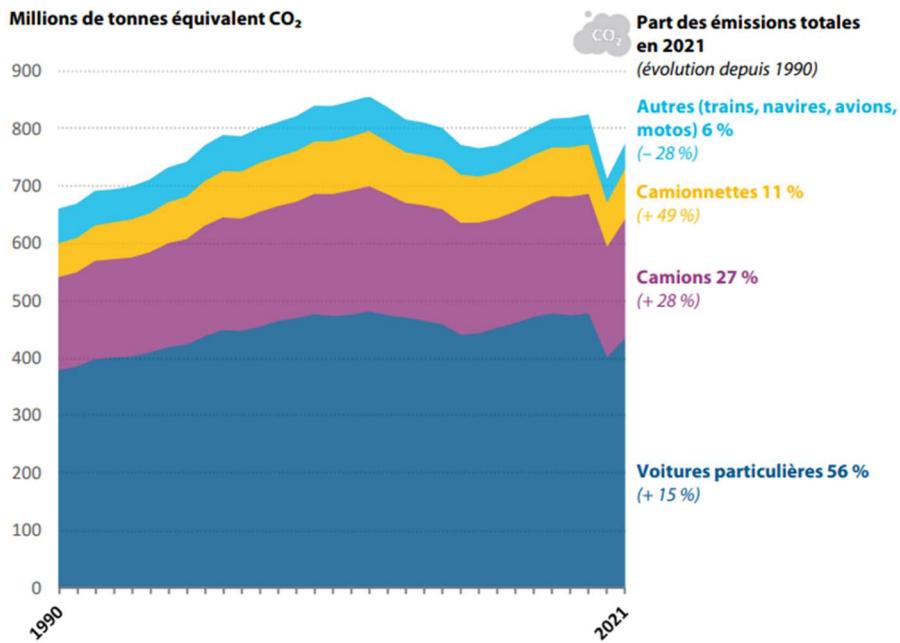
Le consensus qui est en train de s'établir auprès de toutes les parties prenantes est qu'il est nécessaire de s'adapter et d'atténuer significativement les émissions de CO2 et de polluants.

« S'adapter » signifie réduire notre vulnérabilité face aux impacts du changement climatique (protection des personnes, préparation des territoires, préservation des milieux naturels et du patrimoine culturel...).

« Atténuer » signifie diminuer nos impacts dont, entre autre, réduire les polluants et les émissions de gaz à effet de serre d'origine anthropique.

L'effet de serre en lui-même n'est pas problématique, puisque c'est grâce à ce mécanisme physique que les conditions de vie humaine sont possibles sur Terre.

C'est l'excès croissant de CO2 généré par les activités humaines depuis 150 ans qui est alertant. En effet, il induit un forçage radiatif qui concentre des gaz à fort pouvoir réchauffant. Le réchauffement climatique induit des conséquences trop impactantes pour le vivant. Il s'agit donc d'un enjeu essentiel à traiter.

Ventilation des émissions de CO₂ issues des transports (1990- 2021)

Source: Cour des comptes européenne, sur la base de données de l'AEE (Greenhouse gases – data viewer, consulté le 22 juin 2023).

Force est de constater que le secteur des transports a de lourds impacts environnementaux.

Et pour cause, en prenant l'exemple de la France, il fait partie des principaux émetteurs de gaz à effet de serre (GES) dont le dioxyde de carbone (CO₂).

Selon les chiffres de l'Agence de l'Environnement et de la Maîtrise de l'Énergie (Ademe) datant d'avril 2018, les transports sont responsables de 39 % des émissions totales de gaz à effet de serre dans le pays.

Dans le détail, plus de la moitié de ces émissions sont produites par les voitures, 20 % des GES sont émis par les poids lourds, et 17 % par les petits véhicules utilitaires. La part restante concerne les deux-roues, les avions et les transports ferroviaires, maritimes et fluviaux. La quasi-totalité des émissions de gaz à effet de serre (93 %) est liée au transport routier.

Un chiffre qui n'a rien de surprenant puisque près de neuf trajets sur dix, en France, sont réalisés en voiture (particuliers et professionnels confondus). De même, 90 % du transport de marchandises du pays se réalise via les différents axes routiers.

Parmi les actions pour mettre en route rapidement la trajectoire d'atténuation, il est nécessaire d'identifier les véhicules qui émettent le plus de CO₂ pour comprendre les caractéristiques techniques qui jouent un rôle dans la pollution

Prédire cette pollution permet de prévenir dans le cas d'apparition de nouveaux types de véhicules (nouvelles séries de voitures par exemple).

1– Cadre de l'étude

Pour travailler notre sujet, nous disposons de jeux de données initiaux issus de deux fichiers :

- cl_JUIN_2013-complet3.csv : ce fichier contient les caractéristiques techniques des véhicules commercialisés en France en 2013 provenant du site officiel data.gouv.fr.
- data.csv : ce fichier contient les caractéristiques techniques des véhicules commercialisés en 2019 en Europe

Les deux fichiers contiennent des données intéressantes et utiles pour nourrir le sujet.

Toutefois, nous avons noté des limites sur les deux fichiers qui nous ont amené à rechercher dans un premier temps un nouveau set de data et à faire le choix d'un set particulier.

Concernant le fichier 'cl_JUIN_2013-complet3.csv', les données datant de 2013 nous semblent trop éloignés de la réalité actuelle.

En effet, lors de nos premières analyses exploratoires des deux fichiers, nous avons notamment constaté avec l'analyse de la distribution du CO2 que la médiane était bien plus élevée en 2013 (203 g/km) comparée à celle de 2019 (119 g/km). Cela signifie qu'il y a un gap technique conséquent.

Un exemple significatif est la quasi-inexistence de données relatives aux véhicules hybrides car peu développés à cette période alors que sur le fichier de 2019, la présence de ces véhicules a un véritable impact sur la distribution d'émission de CO2.

En explorant le fichier 'data.csv' datant de 2019, nous avons identifié deux mesures de CO2 relevant d'un changement de norme qui s'est opéré en 2019.

En analysant les deux normes (NEDC avant 2019 et WLTP à partir de 2019), nous avons compris qu'avec la norme WLTP les tests étaient plus poussés et plus proches de la réalité d'un usage quotidien d'un véhicule.

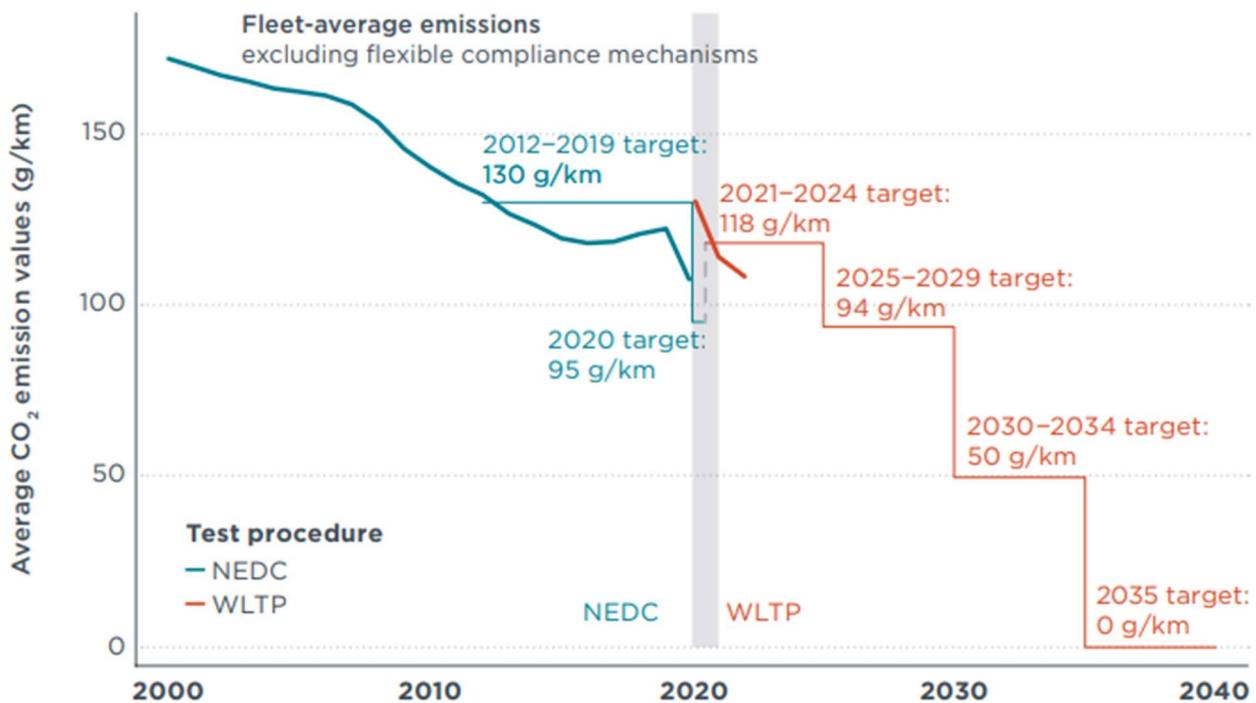
L'infographie ci-dessous compare les modalités des deux normes NEDC et WLTP :

NEDC	WLTP
Cycle d'essai unique	Cycle dynamique, davantage représentatif de la conduite réelle
20 minutes	30 minutes
11 kilomètres	23,25 kilomètres
2 phases: 66 % en milieu urbain; 34 % en dehors	4 phases plus dynamiques: 52 % en milieu urbain; 48 % en dehors
34 km/h	46,5 km/h
120 km/h	131 km/h
Pas de prise en considération de l'impact sur le CO ₂ et la consommation	Prise en considération des équipements optionnels (qui peuvent différer d'une voiture à l'autre)

Source: Cour des comptes européenne, sur la base des informations de l'ACEA.

Projet nov23-cda-CO2

Objectifs de l'Union Européenne et émissions moyennes de CO2 des véhicules neufs mesurés en laboratoire



Source: Cour des comptes européenne, sur la base de données de l'AEE, du JRC et des règlements européens sur les émissions de CO₂ des voitures particulières.

Nous avons alors décidé de rechercher des données plus récentes embarquant une mesure plus large et plus précise avec cette nouvelle norme.

Nos recherches nous ont amené à consulter le site de l'ADEME. Nous avons identifié un fichier 'ADEME-CarLabelling.csv'.

Après pre-processing des données, il s'est avéré que le dataset était trop réduit (759 données) pour faire une modélisation correcte.

Nous avons ensuite sollicité l'ADEME par mail pour obtenir un fichier enrichi.

L'ADEME a répondu qu'elle n'était pas responsable des données publiées sur la plateforme et nous conseillait de contacter directement le producteur des données en se proposant d'appuyer notre demande.

En parallèle, nous avons continué nos investigations sur les sites Européens avec des données plus larges.

Notre veille de données a été payante puisqu'en mars 2024, le site www.eea.europa.eu a publié un dataset enrichi qui répondait à tous nos points cette fois-ci.

Jeu de données sélectionné

Notre nouveau jeu de données nommé 'data_2024.csv' comporte 9 479 544 données, 38 variables.

Le tableau ci-dessous présente la description succincte des variables :

Nom	Désignation	Description	Type	Taux de NA
ID	Identification Number	numéro ID	int64	0,00%
Country	Country	Code Pays	object	0,00%
VFN	Vehicle Family Number	Codification du véhicule (Vehicle interpolation family identifier), numéro sur la carte grise	object	8,70%
Tan	EU standard denomination	Nom du Constructeur -- Standard Européen	object	6,50%
Mh	Manufacturer brand	Marque	object	0,00%
Man	Manufacturer denomination	Nom du Constructeur -- Déclaration du constructeur	object	0,00%
MMS	National Registry denomination	Nom du constructeur - registre national	float64	100,00%
Tan	Type approval number	Le numéro d'homologation - certification des produits à une norme	object	0,20%
T	Type	1er champ du certificat européen de conformité.	object	0,00%
Va	Variant	2e champ du certificat européen de conformité.	object	0,20%
Ve	Version	3e champ du certificat européen de conformité.	object	0,50%
Mk	Make	la marque - Constructeur	object	0,00%
Cn	Commercial Name	Désignation Commerciale	object	1,50%
Ct	Category of the vehicle type approved	classification de véhicules en fonction de sa capacité d'accueil du nombre de passagers	object	0,20%
Cr	Category of the vehicle registered	genre national du véhicule	object	0,00%
r	Registrations	valeur unique par défaut	int64	0,00%
m (kg)	Mass in running order	Poids en ordre de marche. Poids à vide national (PV)	float64	0,00%
Mt	WLTP test mass	Poids de référence utilisé lors des tests conformément à la procédure d'essai WLTP	float64	11,20%
Enedc (g/km)	Specific NEDC CO2 emissions	Emission CO2 selon norme NEDC (Avant 2019)	float64	83,90%
Ewltp (g/km)	Specific WLTP CO2 emissions	Emission CO2 selon norme WLTP (A partir de 2019)	float64	0,20%
W (mm)	Wheelbase	Wheelbase = l'empattement qui est une des dimensions principales d'un véhicule.	float64	0,20%
At1 (mm)	track width of steering axle	Distance entre les roues sur le même essieu	float64	2,20%
At2 (mm)	width of other axles	Autre écartement entre les roues	float64	2,40%
Ft	Fuel type	Type de carburant	object	0,00%
Fm	Fuel mode	Mode carburant :	object	0,00%
ec (cm3)	Engine capacity	Cylindrée en cm3	float64	13,50%
ep (KW)	Engine power	Puissance (Kw)	float64	0,30%
z (Wh/km)	Electric energy consumption	Conso electricité (Wh/km)	float64	78,00%
IT	Code of the eco-innovation	Code d'éco innovation	object	37,80%
Ernedc (g/km)	NEDC CO2 emission savings due to eco-innovation(s)	Réduction des émissions de CO2 grâce des innovations avec la norme NEDC	float64	100,00%
Erwltp (g/km)	WLTP CO2 emission savings due to eco-innovation(s)	Réduction des émissions de CO2 grâce des innovations avec la norme WLTP	float64	46,50%
De	Deviation factor	Mesure d'écart entre déclaration du constructeur et mesure physique des émissions de CO2	float64	100,00%
Vf	verification factor	Facteur de vérification	float64	100,00%
Status	status	pas de référence dans le document de l'Europe	object	0,00%
year	year	année	int64	0,00%
Date of registration	Date of registration	Date d'enregistrement 365 dates toutes en 2022	object	1,70%
Fuel consumption	Fuel consumption	consommation de carburant	float64	23,60%
Electric range (km)	Electric range (km)	Autonomie de la batterie électrique en km (entre 13 km et 739 km)	float64	83,00%

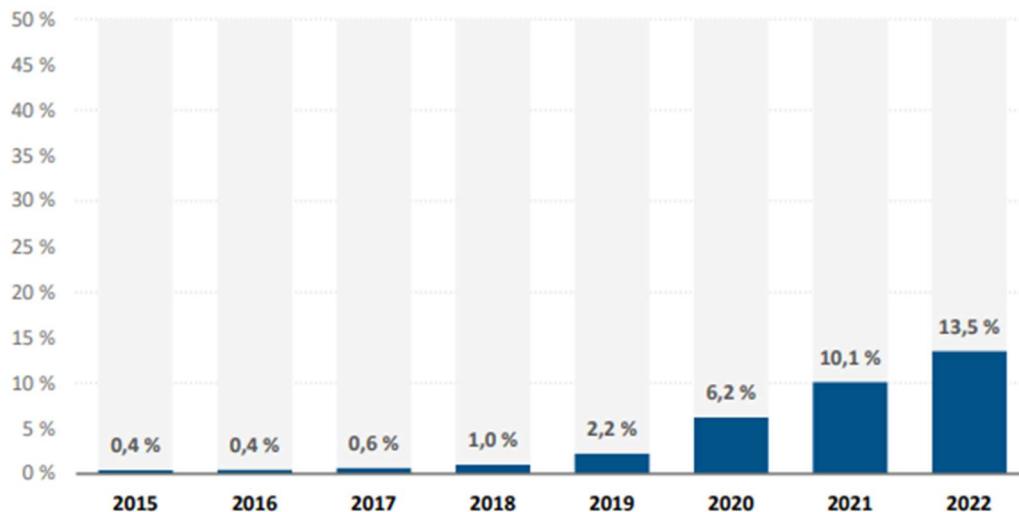
Ce jeu de données reflète mieux l'électrification et l'hybridation du marché qui se sont opérées entre 2013 et 2022.

En effet, c'est une tendance structurelle que l'on doit prendre en compte.

Projet nov23-cda-CO2

Pour illustrer l'ampleur de la tendance, les deux représentations graphiques ci-dessous explicitent l'évolution des ventes de véhicules électriques en Europe et électrifiés en France.

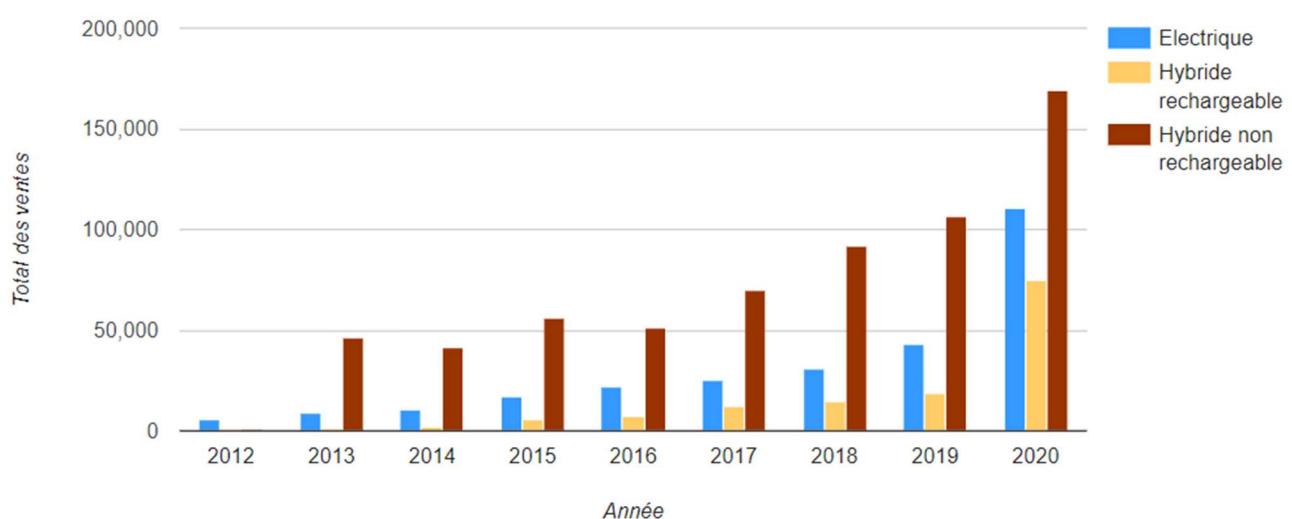
Part des véhicules électriques parmi les immatriculations de véhicules neufs en EUROPE



Remarque: EU-27 plus l'Islande, la Norvège et le Royaume-Uni (inclus jusqu'en 2020).

Source: Cour des comptes européenne, sur la base de données sur les voitures particulières nouvellement immatriculées communiquées par l'AEE.

Part des véhicules électrifiés parmi les immatriculations de véhicules neufs en FRANCE



Source ADEME

L'électrification se décline en **plusieurs typologies de moteur** :

- **Le moteur 100% électrique** - utilisation ne générant aucune émission de CO₂.
Dans nos données, nous la retrouvons dans la variable Fuel Mode au niveau de la modalité "E" (battery electric vehicles : BEV)

- **Le moteur hybride** se répartissant principalement en deux sous-catégories :

- ✓ **Hybride rechargeable** sur borne électrique.

C'est la modalité "P" de la variable Fuel Mode (Off vehicle charging hybrid electric vehicles : OVC-HEV)

- ✓ **Hybride non rechargeable** sur borne électrique.

Ce type de moteur fonctionne en combinant un moteur électrique et un moteur thermique simultanément.

Le moteur électrique ne fait fonctionner le véhicule en autonomie 100% électrique que dans les phases de démarrage, le reste du fonctionnement est une association simultanée entre les deux moteurs (électrique et thermique).

Nous retrouvons les données concernant ce type de véhicule dans la modalité "H" (Not-Off vehicle charging hybrid electric vehicles : NOVC-HEV) de la variable Fuel Mode.

A noter qu'en 2020 en France, la progression des véhicules :

- Électriques est de 159%
- Hybrides rechargeables est de 301%
- Hybrides non rechargeables est de 58,1%

Les trois facteurs majeurs expliquant cette forte hausse sont :

- Les incitations fiscales à l'achat de ce type de véhicule (bonus, prime à la conversion)
- La maturité technologique de ces motorisations et le développement de l'offre des constructeurs
- la règlementation européenne qui limite les émissions de CO2

Notre sujet porte sur l'identification des caractéristiques techniques des véhicules émettant le plus de CO2.

Le dataset comporte de nombreuses variables dont certaines n'ont pas d'utilité au regard du sujet à traiter (exemple : marque, numéro de VIN...). Nous ne retiendrons que les variables portant sur les caractéristiques techniques influençant l'émission de CO2.

Parmi toutes ces caractéristiques, nous faisons face à deux situations qui nous amènent à exclure des variables :

Soit les données sont inexistantes et/ou inexploitables (exemple : Ernedc (g/km), Enedc (g/km))

Soit la variable n'est pas une cause racine de l'émission de CO2.

En effet, si nous prenons le cas de la variable «Fuel Consumption » qui est fortement corrélée à l'émission de CO2, il existe un rapport direct entre la consommation d'une voiture et ses émissions de dioxyde de carbone (CO2) : plus on consomme et plus on rejette. Ce rapport est quasi-fixe.

Cette variable pourrait être considérée comme une variable à intégrer dans notre modèle. Or l'émission de CO2 est liée à la combustion de carburant, donc à sa consommation.

Les inducteurs de cette consommation sont les caractéristiques intrinsèques du véhicule ce qui ne fait de la consommation de carburant qu'un processus intermédiaire ne pouvant être considéré comme un facteur primaire à considérer.

Le carburant peut être simplifié sous la forme d'une molécule CxHy (où x et y varient en fonction du type de carburant et de sa nature).

Si on simplifie un maximum (on considère qu'on a une combustion complète), on a :

$$\text{CxHy} + (\text{x}+\text{y}/4).\text{O}_2 \rightarrow \text{x.CO}_2 + (\text{y}/2).\text{H}_2\text{O}$$

On voit bien que les émissions de CO2 sont proportionnelles à la consommation de carburant sans être pour autant une caractéristique technique de véhicule.

Par conséquent, nous avons conservé les variables suivantes dans un dataset de départ que l'on nomme data_target :

Nom	Désignation	Description	Type	Nb NA	Taux de NA
m (kg)	Mass in running order	Poids en ordre de marche. Poids à vide national (PV)	float64	131	0,00%
Ewltp (g/km)	Specific WLTP CO2 emissions	Emission CO2 selon norme WLTP (A partir de 2019)	float64	14 718	0,20%
W (mm)	Wheelbase	Wheelbase = l'empattement qui est une des dimensions principales d'un véhicule.	float64	21 802	0,20%
Ft	Fuel type	Type de carburant	object	0	0,00%
Fm	Fuel mode	Mode carburant :	object	29	0,00%
ec (cm3)	Engine capacity	Cylindrée en cm3	float64	1 277 808	13,50%
ep (KW)	Engine power	Puissance (Kw)	float64	23 993	0,30%
Electric range (km)	Electric range (km)	Autonomie de la batterie électrique en km (entre 13 km et 739 km)	float64	7 866 906	83,00%

Limitations

- Impact de la carrosserie : dans nos variables retenues, nous avons seulement l'empattement de la carrosserie (Wheel base) et non pas la carrosserie, plus significante, qui était présente dans le fichier « data-2013 » mais avec une couverture locale France et en date de 2013.
- Impact de la boite de vitesse : dans nos variables retenues, cette variable est absente. Elle était présente dans le fichier data-2013
- Le détail des polluants et particules fines n'est pas intégré dans nos variables. Ces données étaient présentes dans le fichier data_2013 sur un périmètre France. Après une exploration de ces données, nous avons constaté une bonne corrélation entre le volume de CO2 émis et celui des polluants. Ce qui signifie que suivre l'émission de CO2 est un marqueur clef global pour l'ensemble de la pollution.
- Impact des éco-innovations : les données à ce sujet correspondaient à un agglomérat de plusieurs innovations avec un résultat de gain en face. Cette variable agrégat ne permet pas d'isoler chaque éco-innovation et d'en mesurer l'impact. Malgré différents tests, cette variable s'est révélée inexploitable

2 – Pre-processing et feature engineering

Notre dataset 'data_target' que nous avons sélectionné comporte 9 479 544 lignes, 8 variables dont 6 numériques (float64) et 2 catégorielles (object).

Pour faciliter la lecture, nous renommons les variables à l'aide d'un dictionnaire qui donne les correspondances suivantes :

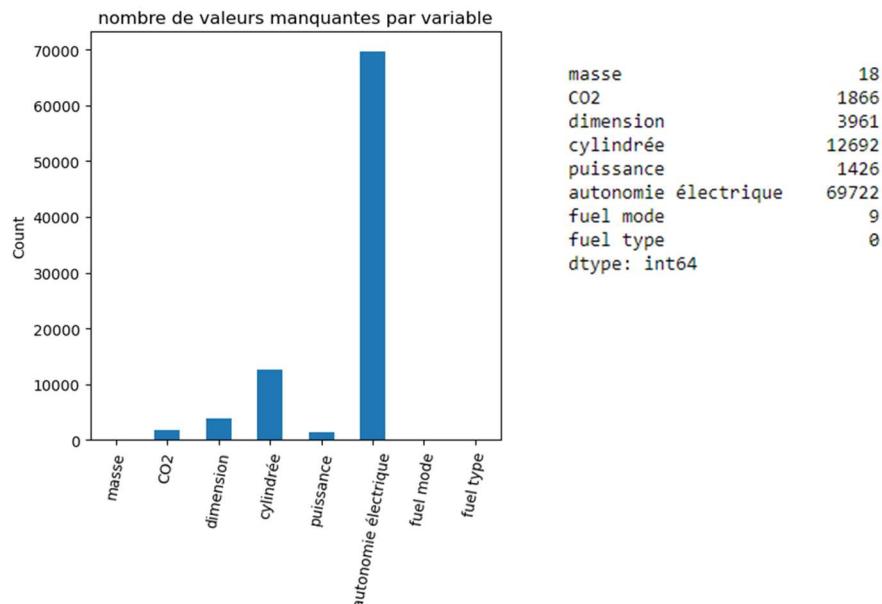
- 'm (kg)': 'masse',
- 'Ewltp (g/km)': 'CO2'
- 'W (mm)': 'dimension',
- 'ec (cm3)': 'cylindrée',
- 'ep (KW)': 'puissance',
- 'Electric range (km)': 'autonomie électrique',
- 'Fm' : 'fuel mode',
- 'Ft': 'fuel type'

2.1- Traitement des doublons

Nous identifions et gérons les doublons du dataset data_target.

Il comporte 9 392 502 doublons que nous supprimons, ce qui amène le dataset à 87 042 lignes à ce stade.

2.2- Gestion des données manquantes

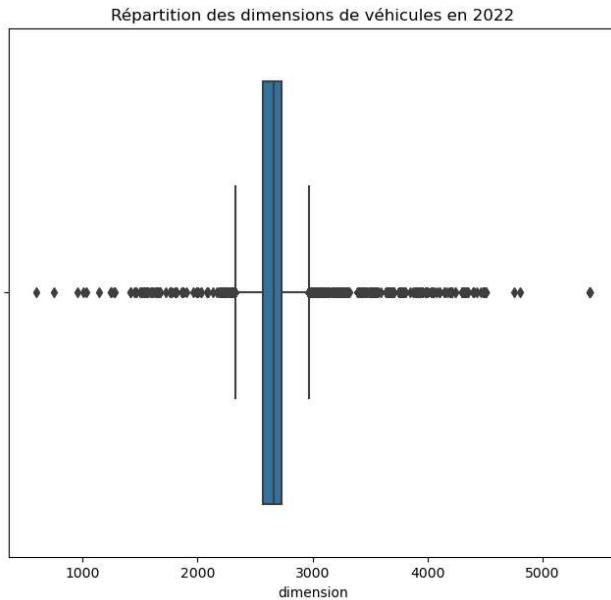


Nous avons constaté que certaines données manquantes étaient faibles en volume et pouvaient être tout simplement supprimées. Il s'agit des variables suivantes :

- Masse
- CO2
- Dimension
- Puissance
- Fuel Mode

Le dataset passe alors à 81 408 lignes.

Avant de décider d'enlever les données manquantes de la variable ‘Dimension’, nous nous sommes posés la question de les remplacer par la médiane de la variable. En analysant la répartition des valeurs, on constate une grande dispersion avec une quantité importante d’outliers.



Au lieu de prendre le risque de mettre une valeur médiane qui pourrait être en écart avec la réalité du métier, nous avons préféré exclure ces valeurs manquantes, le set de données restant étant largement suffisant pour réaliser une modélisation.

Au regard du nombre et de la diversité des motorisations de voitures impactant les deux variables ‘cylindrée’ et ‘autonomie électrique’, cela nécessite un approfondissement pour bâtir une stratégie de nettoyage plus fine.

2.2.1 – Nettoyage des valeurs manquantes de la variable ‘cylindrée’

Le nombre de données manquantes pour la variable ‘cylindrée’ est de 10 885.

On cherche à analyser la composition de ces données manquantes au regard de la variable ‘fuel mode’. En effet, l'idée est de voir si les données manquantes de cette variable viennent uniquement du fait que ce sont des véhicules 100% électriques : ces véhicules n'ayant pas de cylindrée. Nous décidons de les remplacer par la valeur 0.

Le nombre de données manquantes sur la variable cylindrée et de fuel mode électrique est de 10 868. Cela signifie qu'il faut identifier les autres catégories de ‘fuel mode’ qui viennent s'additionner aux électriques pour atteindre nos 10 885.

En filtrant sur le fuel mode ‘M’ (véhicules mono carburant) on constate qu'il y a 16 lignes concernées et 1 ligne sur le fuel mode ‘P’ (véhicules hybrides rechargeables). Nous décidons de supprimer ces 17 lignes au regard de leur très faible poids.

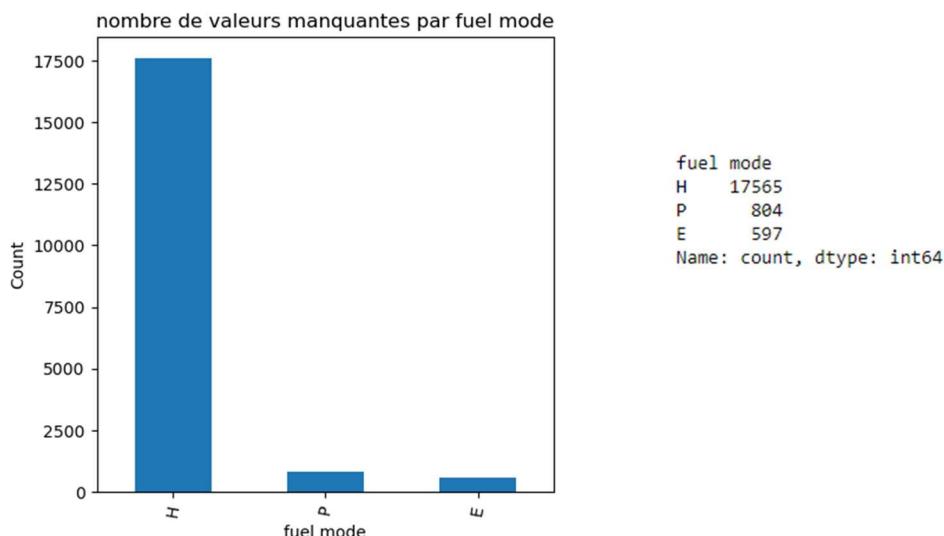
2.2.2 – Nettoyage des valeurs manquantes de la variable ‘autonomie électrique’

Au sein de la variable ‘autonomie électrique’, l’objectif est :

- Identifier les véhicules avec une batterie électrique qui sont d’un fuel mode ‘E’ (100% électrique) ou ‘P’ (hybride rechargeable) ou ‘H’ (hybride non rechargeable) qui apparaissent dans les données manquantes,
- Identifier les véhicules 100% thermiques (qui n’ont pas de batterie électrique) qui apparaissent dans les données manquantes.

La stratégie sera différente selon ces deux cas de figure.

Concernant les véhicules ayant une batterie électrique, le volume de données manquantes par fuel mode est le suivant :



Trois cas de figures pour la gestion des données manquantes de cette catégorie :

- **Fuel mode ‘P’** : nous décidons de remplacer les valeurs manquantes par la médiane. La valeur médiane de l’autonomie électrique des véhicules hybrides rechargeable sur une prise est de 59 km.
- **Fuel mode ‘H’** : le dataset ne présentant aucune valeur, nous avons sourcedes valeurs sur internet. L’autonomie en 100% électrique de ce type de véhicule est très faible. En effet, la batterie est utilisée en complément du moteur thermique et l’usage en 100% électrique se réduit le plus souvent au démarrage du véhicule.

Le consensus des constructeurs annonce une moyenne autour de 2 km, valeur que nous retenons pour remplacer les valeurs manquantes de cette catégorie.

- **Fuel mode ‘E’** : ce sont les véhicules 100% électrique qui n’émettent pas de CO2 dans leur usage. Au regard de notre sujet, nous décidons de supprimer les modalités qui concernent les véhicules 100% électrique.

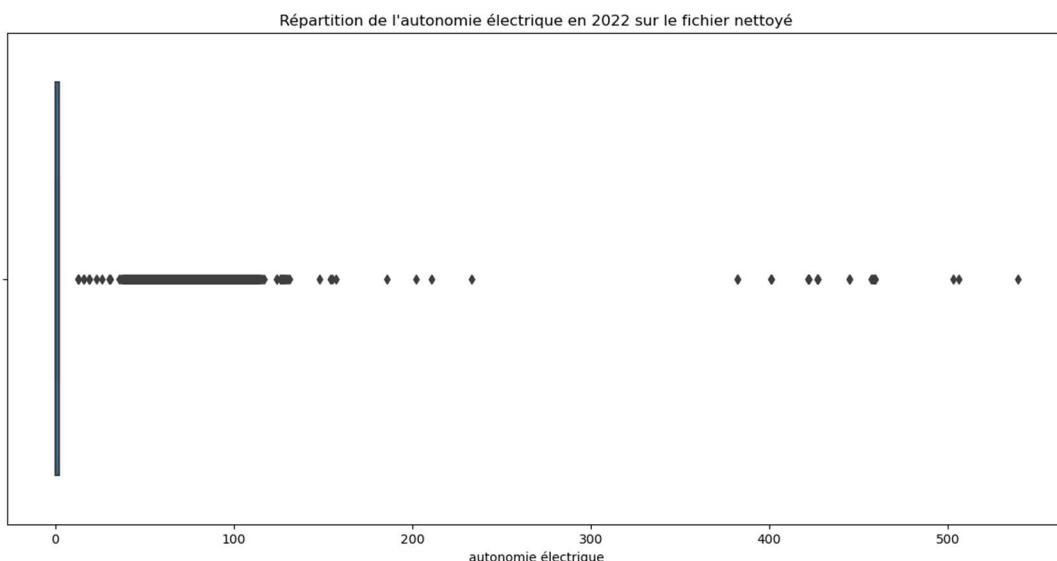
Projet nov23-cda-CO2

Concernant les véhicules 100% thermique, n'ayant pas de batterie électrique, nous remplaçons les valeurs manquantes de la variable «autonomie électrique » par 0.

Nous avons vérifié par précaution que toutes les données étaient renseignées au bon format.

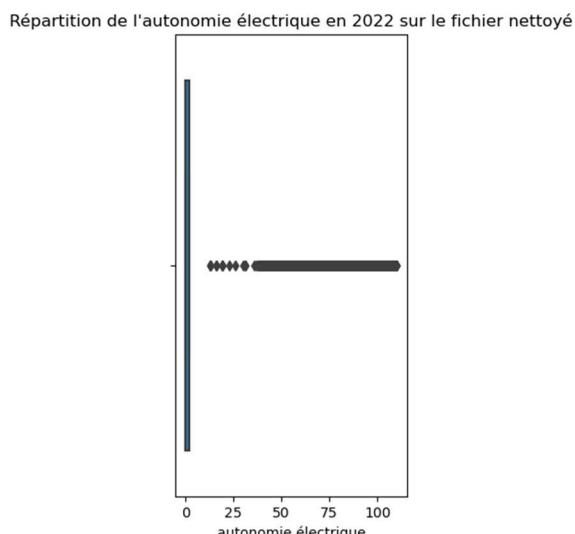
Nous avons constaté des outliers pour lesquels nous allons mener une étude. A ce stade du nettoyage, notre jeu de données comprend 70 523 lignes.

En analysant la distribution de l'autonomie électrique sur notre jeu de données, nous remarquons la présence de valeurs aberrantes. En effet, notre jeu de données ne contient plus de véhicules 100% électrique et nous constatons des valeurs d'autonomie électrique supérieures à 500 km. Or l'autonomie maximum des véhicules hybrides rechargeables est de 110 km. Le graphique ci-dessous illustre la présence des valeurs aberrantes.



Nous identifions 105 véhicules dont 3 en modalité 'M' (monocarburant) et 102 en modalité 'P' (hybride rechargeable). Les véhicules monocarburant n'ayant pas d'autonomie électrique, nous supprimons ces 3 lignes. Concernant les 102 hybrides rechargeables, nous remplaçons les valeurs aberrantes par la médiane des véhicules hybrides rechargeables dont l'autonomie électrique est inférieure à 110 km. Ces changements génèrent 154 doublons que nous supprimons.

Après ces opérations, notre jeu de données est de 70 366 lignes. La répartition de l'autonomie électrique est la suivante :



3.Visualisations et statistiques

3.1- Nos objectifs dans le cadre de la phase d'exploration

Notre sujet portant sur l'identification des véhicules qui émettent le plus de CO2, notre variable cible est clairement identifiée : le taux d'émission de CO2 en gramme par kilomètre (g/km).

Nous cherchons à identifier visuellement les variables qui influencent le niveau d'émission de CO2. Nous nous concentrerons sur la sélection des caractéristiques techniques des véhicules qui ont un impact significatif sur l'émission de CO2.

Le but est de pouvoir prédire le niveau de pollution qui sera émis par les nouveaux véhicules produits.

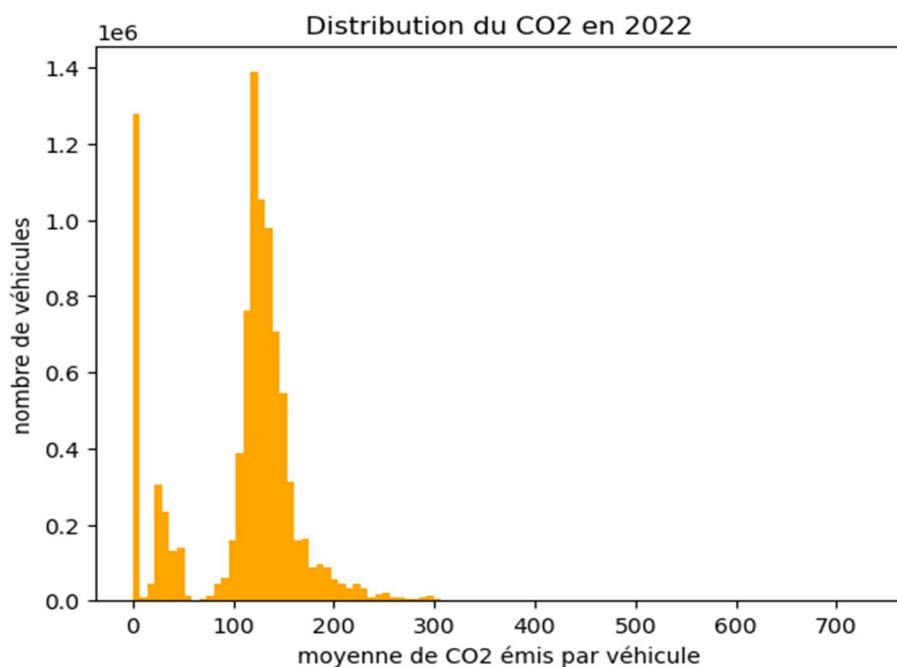
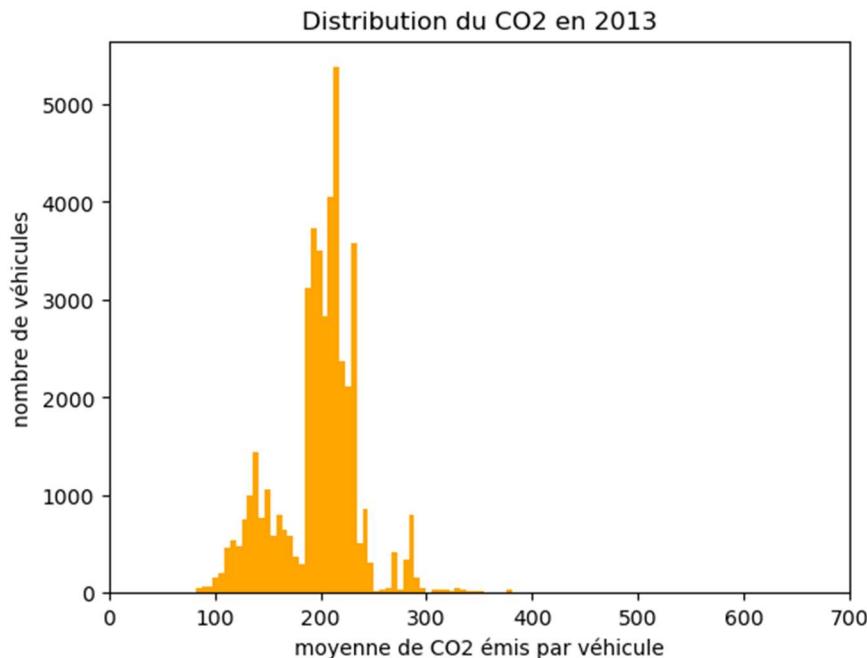
Cela permettra aussi d'un point de vue plus global de pouvoir mettre en perspective l'évolution de la pollution globale comparée aux objectifs de réduction fixés ayant pour but ultime la neutralité carbone.

3.2- Distribution du CO2

En comparant la distribution du CO2 sur le fichier 2013 et celui que nous avons choisi pour notre étude qui date de 2022 nous avons constaté que le marché de l'automobile avait fortement évolué entre 2013 et 2022,

Les deux distributions ci-après montrent :

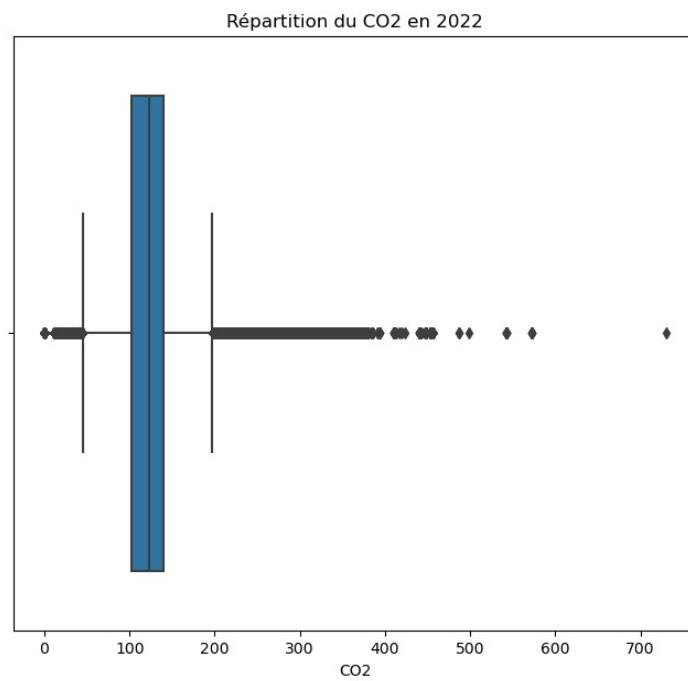
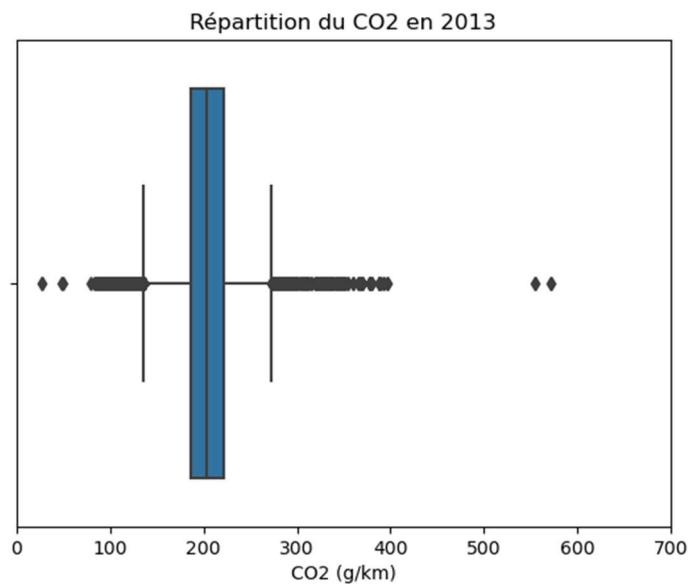
- La réduction significative du niveau d'émission de CO2 entre 2013 et 2022 (ce qui se voit par la translation de la distribution vers la gauche)
- L'existence de grandes familles de véhicules qui ont un impact sur la distribution du CO2
- L'apparition d'une nouvelle famille de véhicules visible sur les données de 2022 amenant une émission de CO2 à zéro, ce que l'on peut identifier facilement par les véhicules électriques.



Malgré l'écart de nombre de véhicules entre nos données de 2013 et 2022, cette visualisation permet d'avoir une vue macroscopique de l'évolution du marché.

En dépit d'une forte amélioration de la réduction des émissions de CO2 entre 2013 et 2022, le parc des véhicules neufs semble être toujours assez fortement dispersé.

Nous voyons, en effet, dans les boîtes à moustache ci-dessous une forte amélioration de la médiane mais les valeurs au-delà du 3^{ème} interquartile sont au moins aussi nombreuses.

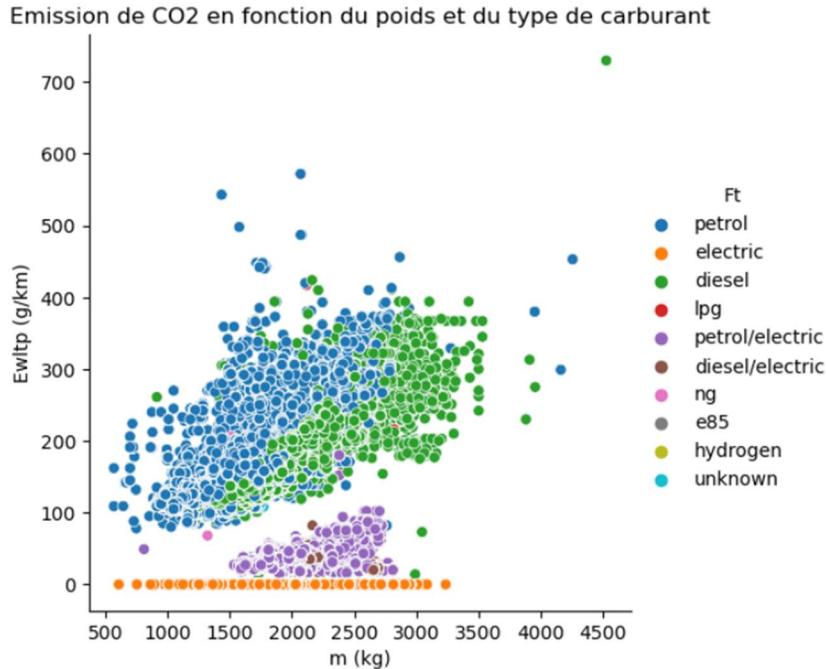


Au travers de ces graphiques, on constate clairement que les constructeurs ont fourni un effort au niveau de la conception des véhicules pour réduire les émissions de CO2.

Nous allons à présent chercher les leviers techniques qui permettent d'influencer les émissions et dans quelle proportion.

3.3- Etude des variables explicatives

L'augmentation du poids des véhicules a un impact direct sur le niveau des émissions de CO2.



On identifie trois groupes de véhicules dans la combinaison poids et mode de carburant des véhicules :

- Les véhicules électriques dont l'augmentation du poids n'a aucun impact sur les émissions de CO2
- Les véhicules hybrides rechargeables forment un groupe bien distinct dont l'émission de CO2 est bien en dessous de celle des véhicules mono-carburant au regard de l'augmentation du poids
- Les véhicules mono-carburant (essence ou diesel) sont les plus émissifs et leur niveau d'émission est sensible au poids.

A l'intérieur de cette famille de véhicules, on observe une catégorisation en trois parties :

- A moins de 1500 kg, les véhicules sont majoritairement en motorisation essence.
- Entre 1500 et 2700 kg, il y a un mixte de motorisation essence et diesel.

A poids équivalent, on remarque une émission de CO2 très différente.

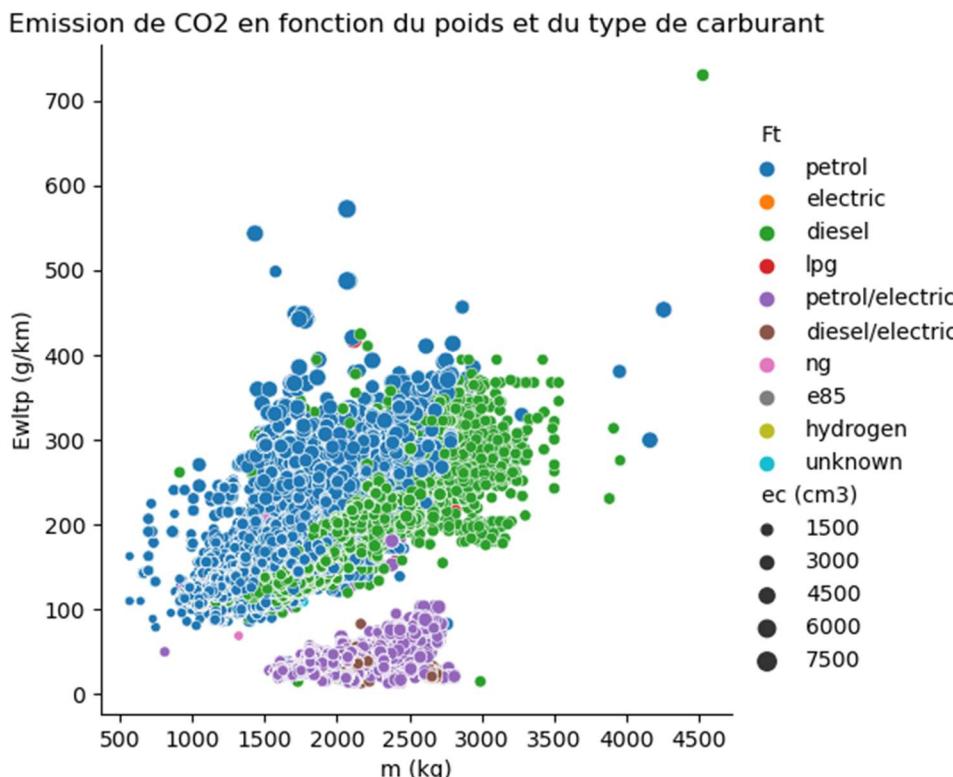
Une majorité de véhicules essence semblent émettre plus que ceux à motorisation diesel. Or le diesel est plus émissif, ce qui suggère que d'autres configurations techniques pourraient influer fortement sur le niveau d'émission.

Cela se vérifie en effet dans le graphique ci-dessous qui retrace l'émission de CO2 en fonction du poids, du type de carburant et de la cylindrée : à poids égal on constate une augmentation directe d'émission de CO2 corrélée à l'augmentation de la cylindrée.

- A plus de 2700 kg, on constate que les véhicules sont principalement équipés d'un moteur diesel.

Cela s'explique par le besoin de couple c'est-à-dire une capacité à générer une force de traction. C'est ce qu'un véhicule diesel procure, pour tracter le poids d'un véhicule plus lourd.

A l'opposée, un véhicule essence se caractérise par sa puissance (capacité à déplacer la voiture rapidement et atteindre une vitesse élevée en un temps très court) et non pas par son couple. Cela permet d'avoir une reprise plus importante à poids plus restreints, notamment sur les véhicules citadins généralement plus petits et plus légers.

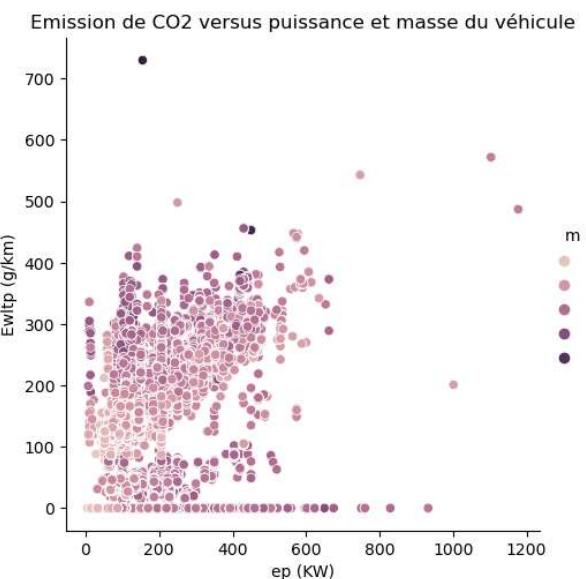
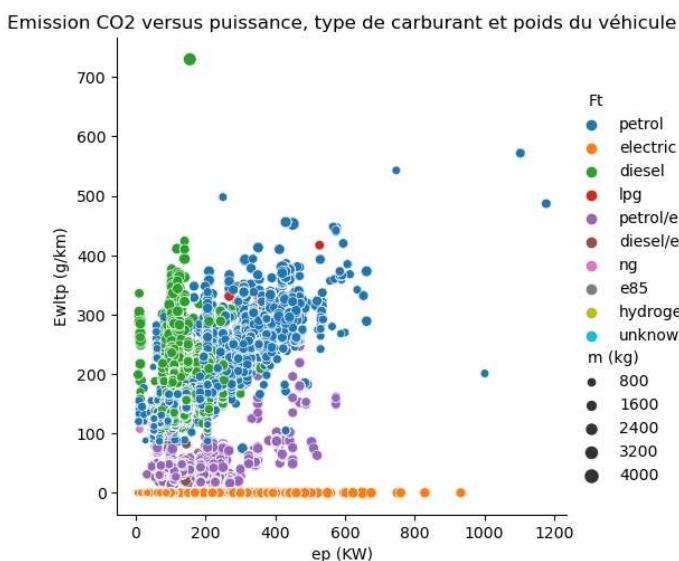


En étudiant l'impact de la puissance, le poids reste une influence plus prépondérante, notamment sur la motorisation diesel. A la différence des véhicules à motorisation essence où l'impact du poids et de la puissance sont plus proportionnels.

Sans surprise, la puissance n'a pas d'impact sur les émissions de CO2 des véhicules électriques.

Quant aux hybrides, on reste dans le même constat que précédemment : un impact maîtrisé se situant entre les véhicules électriques non émissifs et les véhicules mono-carburant.

Projet nov23-cda-CO2



Le type et le mode de carburant sont des facteurs qui influencent l'émission de CO2.

Le type de carburant (variable 'fuel type') correspond au carburant et/ou d'énergie mis dans le véhicule pour le faire fonctionner.

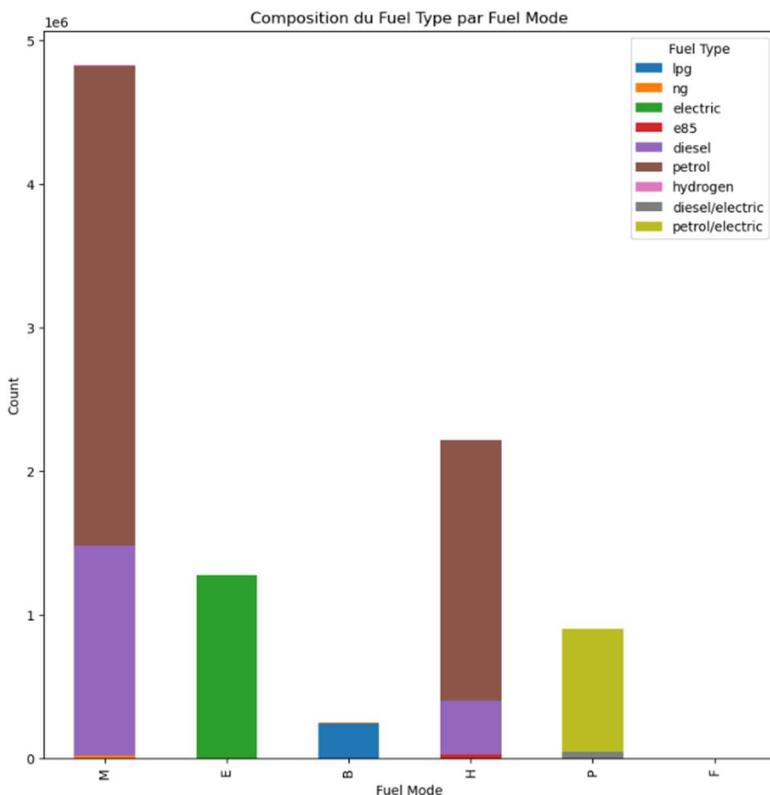
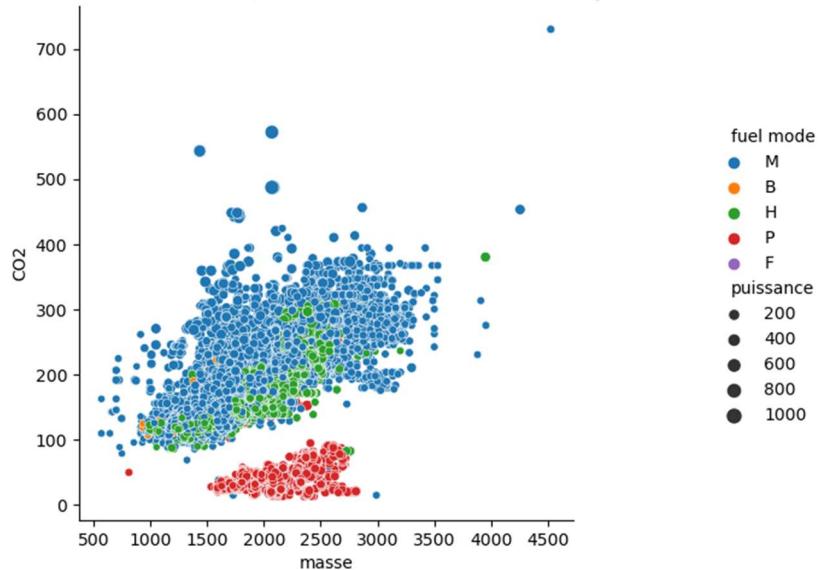
Le mode de carburation (variable 'fuel mode') correspond à différentes catégories :

- la catégorie 'M' pour Mono carburant (véhicules 100% essence, diesel, hydrogène ou gpl)
- la catégorie 'E' pour 100% électrique
- la catégorie 'B' pour bi-carburation pour les véhicules ayant deux réservoirs pour accueillir deux types de carburant différent
- la catégorie 'F' pour la carburation 'flexible' pour les véhicules qui associent au sein d'un même réservoir un mélange de plusieurs types de carburant (exemple : essence combinée avec éthanol...)
- la catégorie 'H' pour les véhicules hybrides non rechargeables sur prise utilisant une motorisation essence ou diesel.
- la catégorie 'P' pour les véhicules hybrides rechargeables utilisant une motorisation essence ou diesel.

La distribution du mode de carburant montre la présence toujours soutenue des véhicules mono-carburant. Toutefois, cette hégémonie est remise en cause par la montée rapide et conséquente des véhicules avec de nouvelles motorisations (hybrides et électriques notamment).

Projet nov23-cda-CO2

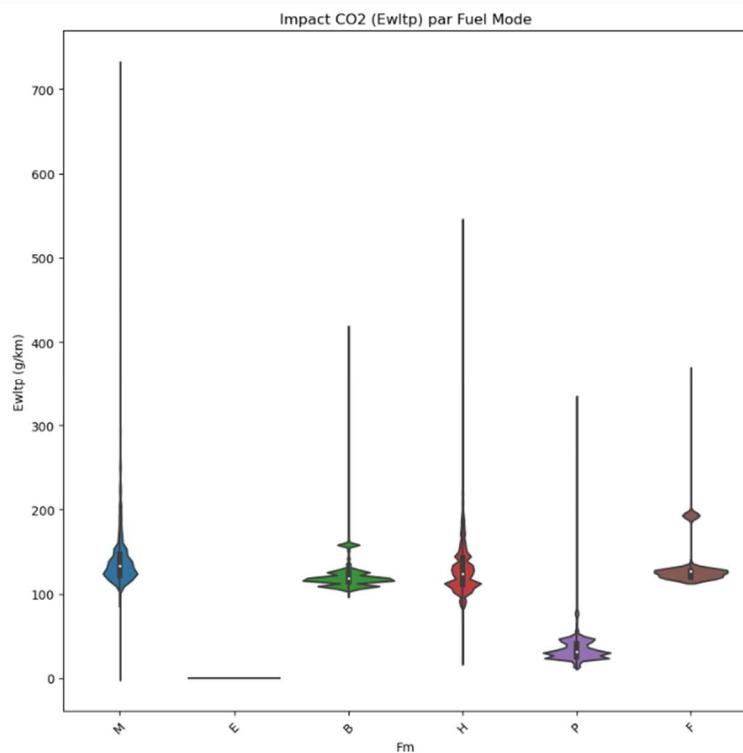
Emission de CO2 en fonction de la masse, du mode de carburation et de la puissance du véhicule



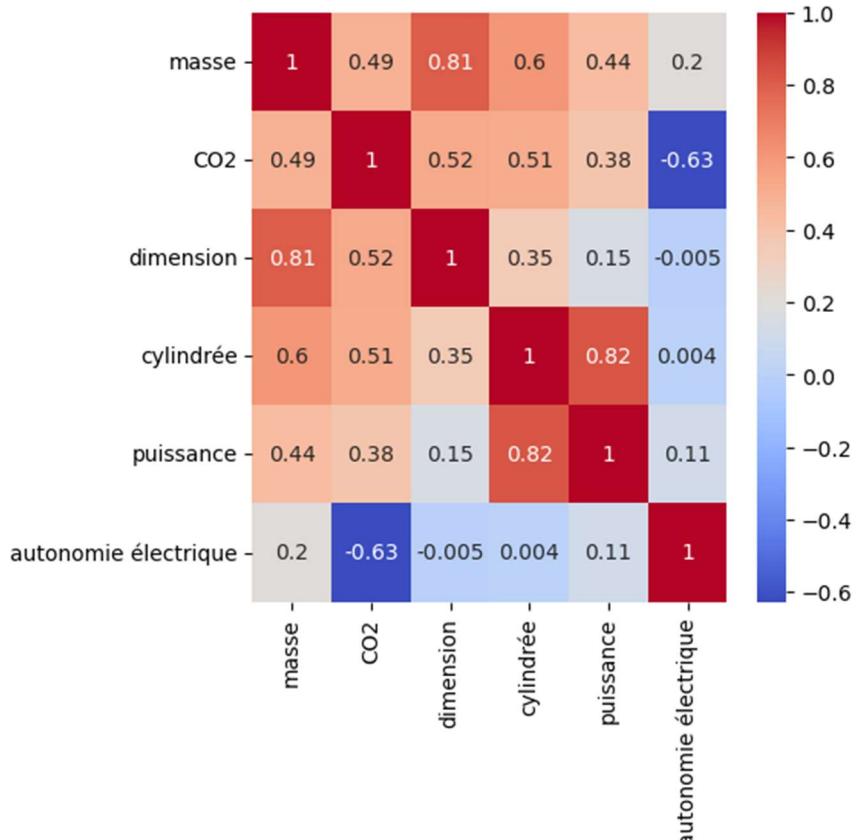
Les véhicules hybrides rechargeables et les véhicules 100% électriques sont les deux catégories qui sont classées sur des taux d'émission de CO2 nettement plus faibles que les autres modes.

Hormis les véhicules 100% électriques, tous les véhicules ont une forte dispersion dans leur niveau d'émission de CO2, avec des outliers très émissifs.

Projet nov23-cda-CO2



Concernant les variables numériques ‘masse’, ‘dimension’, ‘cylindrée’, ‘puissance’ et ‘autonomie électrique’, nous avons identifié visuellement une bonne corrélation avec l’émission de CO2. Celle-ci est confirmée via le heatmap.

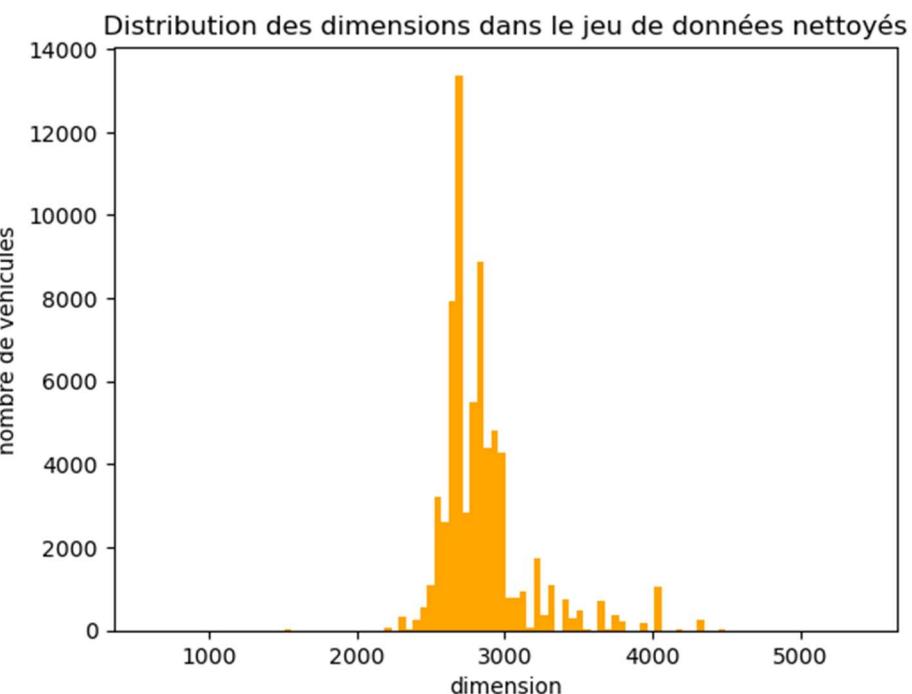
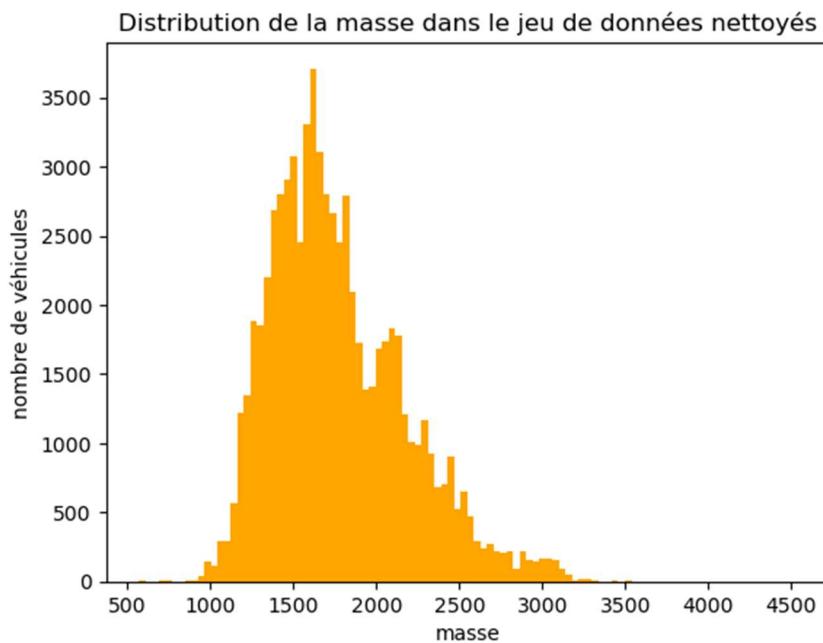


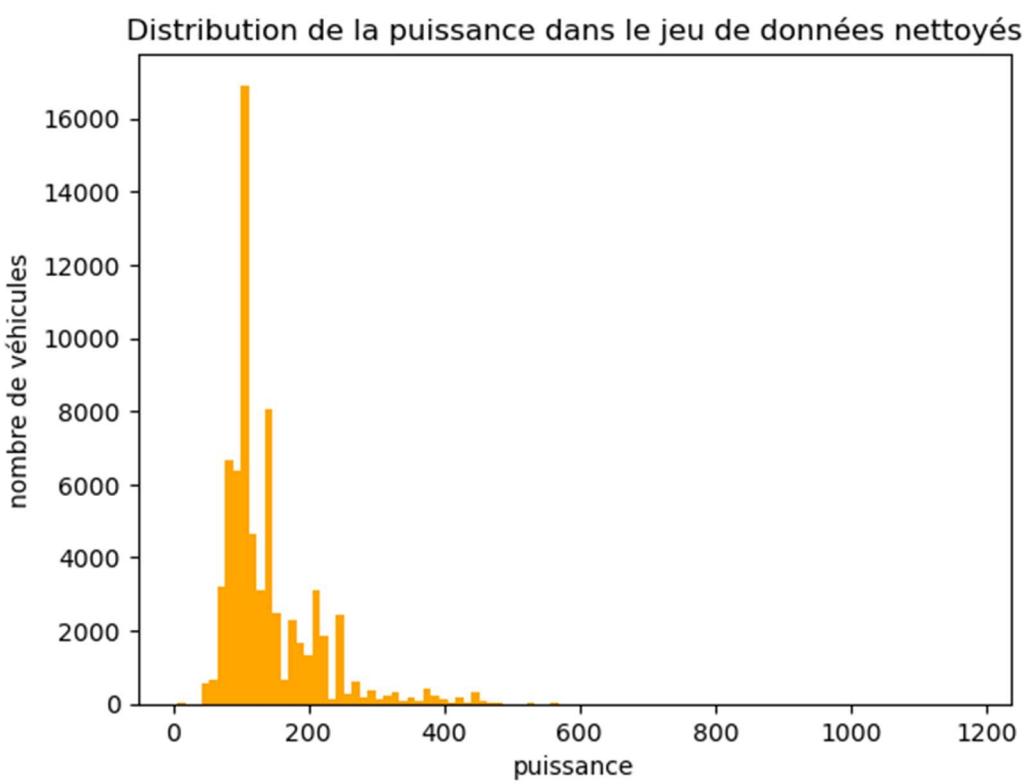
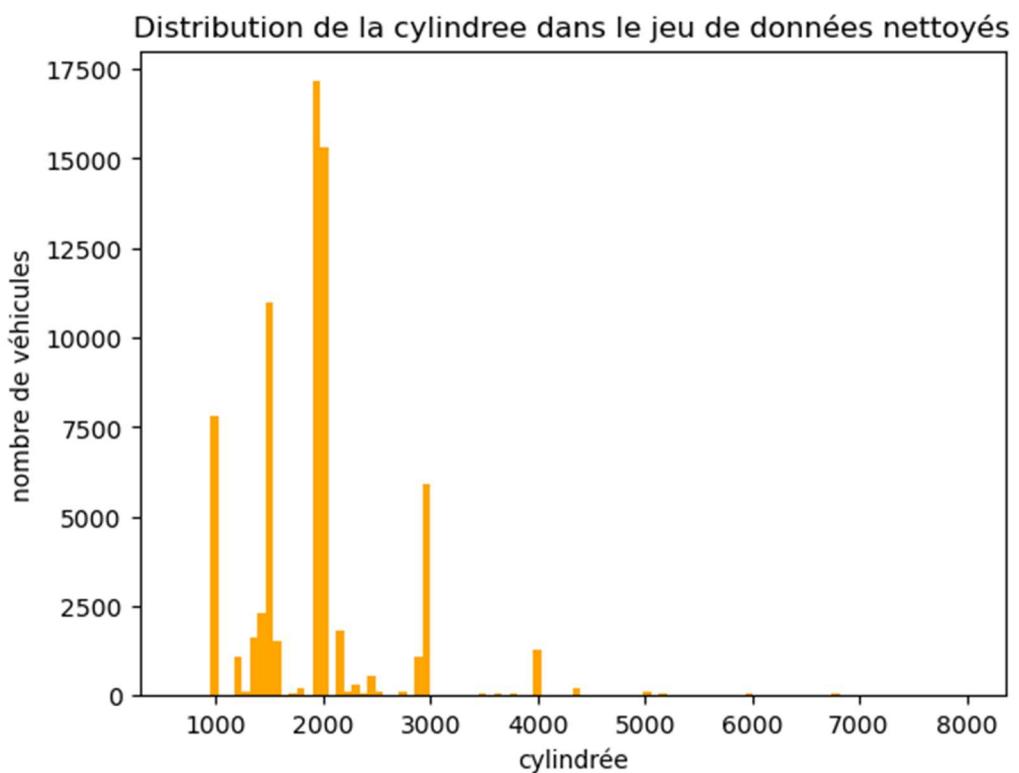
Heatmap sur les variables retenues sur le jeu de données nettoyé

4.Modélisations

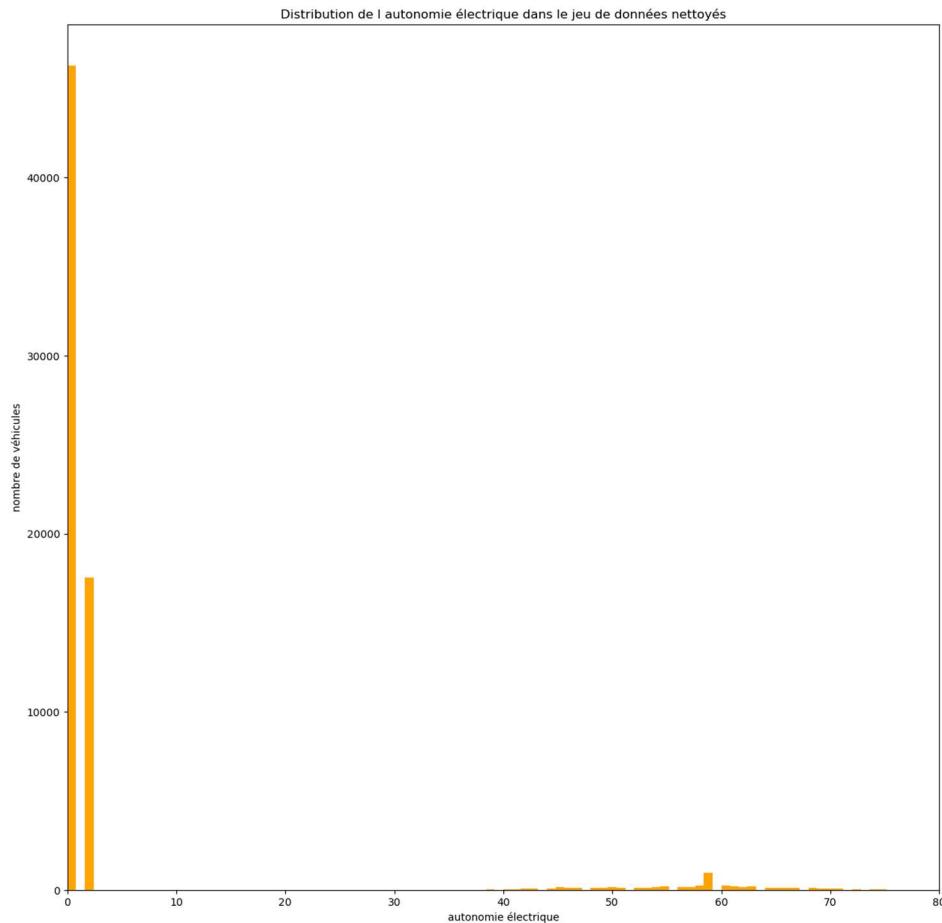
4.1 – Visualisations de la distribution des variables dans le dataset nettoyé

Ci-dessous les distributions de nos quelques variables explicatives les plus impactantes :





Distribution de l'autonomie électrique sur le jeu de données nettoyées



4.2 – Modèles de Classification

Jusqu'à ce jour, sur le marché de l'automobile, les constructeurs annoncent les émissions de véhicules sous forme de classes (exemple : la classe A concerne les véhicules émettant jusqu'à 100 g/km de CO2).

Pour être en lien avec le marché actuel, nous démarrons avec une approche basée sur des modèles de classification. Il s'agit alors de prédire la classe d'émission de CO2 d'un véhicule selon ses caractéristiques physiques (masse, dimension, cylindrée, puissance, autonomie électrique, fuel mode (mode carburant) et le fuel type (type carburant)).

Plusieurs familles de modèles ont été testées pour la classification, à savoir la régression logistique, l'arbre de décision, le RandomForest Classifier, le SVM(Support Vecteur Machine) et le BalancedRandomForestClassifier.

Le dataset (data_target13.csv) retenu pour l'étude après nettoyage pour la modélisation est constitué de 70366 entrées et de 8 variables dont 7 explicatives et 1 variable cible (CO2) à prédire.

L'aperçu des 5 premières lignes de données du dataset donne ce qui suit :

	masse	CO2	dimension	cylindrée	puissance	autonomie électrique	fuel mode	fuel type
0	993.0	B	2492.0	999.0	48.0	0.0	M	petrol
1	1015.0	B	2492.0	999.0	48.0	0.0	M	petrol
2	1015.0	B	2492.0	999.0	48.0	0.0	M	petrol
3	993.0	B	2492.0	999.0	48.0	0.0	M	petrol
4	984.0	B	2492.0	999.0	48.0	0.0	M	petrol

A – Modélisation initiale

A.1 Démarche

Après avoir scindé le dataset data_target13.csv en deux dataframe :

- Feats (X), regroupant toutes les colonnes explicatives : masse, dimension, cylindrée, puissance, autonomie électrique, fuel mode et fuel type
- Target (Y) contenant uniquement colonne CO2, la cible à prédire.

Les deux sous dataframe ont été chacun partagés ensuite en 1 set d'entraînement et 1 set de test selon les proportions suivantes :

Jeux d'entraînement (X_train, y_train) : 56293 lignes soit 80% du dataset

Jeux de test (X_test, y_test) : 14073 lignes soit 20% du dataset

Il n'y avait pas de NAN à gérer étant donné que le dataset final data_target13 n'en contenait plus.

Afin de procéder à l'encodage et standardisation des données d'entraînement et de test(X_train et X_test), les colonnes catégorielles ont été séparées de celles numériques. Ainsi, quatre sous dataframe (num_train, num_test, cat_train, cat_test) ont été mis en place.

A l'aide des méthodes « transform » et « fit. transform » de OneHotEncoder de la librairie sklearn.preprocessing, les données catégorielles fuel mode et fuel type ont été encodés selon leurs différentes modalités sur les dataframe cat_train, cat_test. Ainsi autant de colonnes que de modalités constituant les variables fuel mode et fuel type ont été générés avec des valeurs 0 ou 1 à l'intérieur de chaque colonne selon que la donnée appartienne à la catégorie concernée (valeur 1) ou non (valeur 0)

A l'aide des méthodes « fit. transform » et « transform » de StandardScaler de la librairie sklearn.preprocessing, les données numériques (masse, dimension, cylindrée, puissance, autonomie électrique) ont été standardisés sur les dataframe num_train, num_test.

A l'aide des méthodes « fit. transform » et « transform » de LabelEncoder de la librairie sklearn.preprocessing, les données de la variable cible CO2 ont été labellisés sur les données y_train et y_test.

Suite aux différents encodages et standardisations des données catégorielles et numériques, les données finales d'entraînement (X_train) et de test(X_test) à soumettre aux modèles ont été reconstituées dans de nouveaux dataset X_train_new et X_test_new par concaténation des données encodées et standardisées d'un côté de num_train et cat_train et de num_test et cat_test de l'autre.

Les dimensions des jeux d'entraînement et de test finaux sont comme suit :

Train Set : X_train_new (56292, 17) , avec 56293 lignes et 17 colonnes
y_train (56292, 1) , avec 56293 lignes et 1 colonne

Test Set : X_test_new (14074, 17) , avec 14073 lignes et 17 colonnes
y_test (14074, 17), avec 14073 lignes et 17 colonnes

Les modèles reglog1(régression logistique), dectree1(arbre de décision) et rfcl1(RandomForest) ont été entraînés sur les sets de train (X_train_new et y_train)

Les prédictions de chacun des modèles ont été réalisées sur le jeu de test X_test_new.

Les scores Accuracy sur les jeux d'entraînement et de test ont été évalués pour chaque modèle ainsi que les matrices de confusion et les rapports de classification.

Une représentation graphique des variables les plus importantes prises en compte dans la classification a été établie.

A.2 Résultats et Analyses

Accuracy Scores :

```
Pour le modèle LogisticRegression Acuracy Score sur ensemble train 0.6960314076600582
Pour le modèle LogisticRegression Acuracy Score sur ensemble test 0.696603666335086
Pour le modèle DecisionTreeClassifier Acuracy Score sur ensemble train 0.9175193633198323
Pour le modèle DecisionTreeClassifier Acuracy Score sur ensemble test 0.8534887025721188
Pour le modèle RandomForestClassifier Acuracy Score sur ensemble train 0.917483834292617
Pour le modèle RandomForestClassifier Acuracy Score sur ensemble test 0.8573255648713941
```

Du point de vue des scores accuracy qui indiquent pour un modèle le taux de prédictions correctes des classes sur le total des prédictions, le modèle RandomForest rfcl1 obtient le meilleur score de 91,74% de prédictions correctes sur l'ensemble train et de 85,73 % sur l'ensemble test.

Le modèle dectree1 (Arbre de classification) obtient un score test moindre (85,34%) comparé à celui du RandomForest et un score train de 91,75%.

Cependant on note un surapprentissage de l'ordre de 6% pour le modèle rfcl1 contre 6,41% pour le modèle dectree1.

Malgré une absence de surapprentissage sur le modèle de régression logistique reglog1, ce dernier s'avère moins précis dans ses prédictions à la fois sur les ensembles train et test (score train : 69,60% ; score test : 69,66%) en faisant une comparaison par rapport aux deux autres modèles.

En calculant les matrices de confusion et les rapports de classification des 3 modèles, les résultats suivants sont obtenus :

- Pour la régression logistique reglog1

Prédiction	A	B	C	D	E	F	G
Realité							
A	1288	27	6	0	1	3	16
B	5	302	714	13	1	2	0
C	10	117	2666	479	22	9	0
D	6	5	546	1753	496	18	0
E	4	3	26	537	2112	283	15
F	1	6	7	13	551	958	121
G	0	1	9	0	13	184	725

	precision	recall	f1-score	support
A	0.98	0.96	0.97	1341
B	0.66	0.29	0.40	1037
C	0.67	0.81	0.73	3303
D	0.63	0.62	0.62	2824
E	0.66	0.71	0.68	2980
F	0.66	0.58	0.62	1657
G	0.83	0.78	0.80	932

accuracy			0.70	14074
macro avg	0.73	0.68	0.69	14074
weighted avg	0.70	0.70	0.69	14074

La classe B est mal classifiée par le modèle reglog1 en classant la majorité des véhicules catégorisés dans la réalité en classe B en classe C (714 véhicules).

Avec un taux de précision de 66%, le taux de rappel de 29 % est très faible par rapport à celui des autres classes.

La classe A est cependant la mieux classée avec 98% de bonnes prédictions et le meilleur taux de rappel de 96%. Sur 1341 véhicules, 1288 ont été correctement prédits.

- Pour l'arbre de décision dectree1

Prédiction	A	B	C	D	E	F	G
Realité							
A	1325	8	3	2	2	0	1
B	11	757	255	8	5	0	1
C	4	223	2803	260	9	2	2
D	7	3	296	2295	221	0	2
E	4	4	14	260	2548	135	15
F	3	0	2	10	151	1433	58
G	0	0	2	0	21	58	851

	precision	recall	f1-score	support
A	0.98	0.99	0.98	1341
B	0.76	0.73	0.75	1037
C	0.83	0.85	0.84	3303
D	0.81	0.81	0.81	2824
E	0.86	0.86	0.86	2980
F	0.88	0.86	0.87	1657
G	0.92	0.91	0.91	932
accuracy			0.85	14074
macro avg	0.86	0.86	0.86	14074
weighted avg	0.85	0.85	0.85	14074

L'arbre de décision dectree1 prédit mieux la classe B comparé au modèle précédent de régression logistique reglog1 avec 757 bonnes prédictions sur 1037 au total soit une précision 76% et un taux de rappel de 73% contre 29% constaté auparavant.

La classe B reste toujours la classe la moins bien prédite et la classe A la mieux classifiée (1325 véhicules correctement prédits en classe A sur 1341 véhicules de classe A réels).

Les classes C à G sont correctement prédites à minima à 80% avec en tête la classe G catégorisée correctement avec une précision de 92% et un coefficient de rappel de 91%.

- Pour le RandomForest rfcl1

Prédiction	A	B	C	D	E	F	G
Realité							
A	1320	12	3	0	3	1	2
B	9	742	276	8	2	0	0
C	3	188	2823	279	10	0	0
D	7	1	274	2309	232	0	1
E	1	2	12	235	2557	159	14
F	2	0	3	4	137	1448	63
G	0	0	0	0	15	50	867
	precision	recall	f1-score	support			
A	0.98	0.98	0.98	1341			
B	0.79	0.72	0.75	1037			
C	0.83	0.85	0.84	3303			
D	0.81	0.82	0.82	2824			
E	0.87	0.86	0.86	2980			
F	0.87	0.87	0.87	1657			
G	0.92	0.93	0.92	932			
accuracy			0.86	14074			
macro avg	0.87	0.86	0.86	14074			
weighted avg	0.86	0.86	0.86	14074			

Par le modèle RandomForest rfcl1, on note une amélioration sur les taux de rappel pour les classes D, F et G.

Ainsi, comparé à la réalité, les nombres d'individus correctement prédis par ce modèle est meilleur par rapport à ceux obtenus précédemment.

Le coefficient de rappel de la classes B s'est dégradé au profit du taux de précision(79%) qui reste meilleur par rapport à celui des deux autres modèles. Les taux de rappel sont meilleurs pour les classes D, F et G.

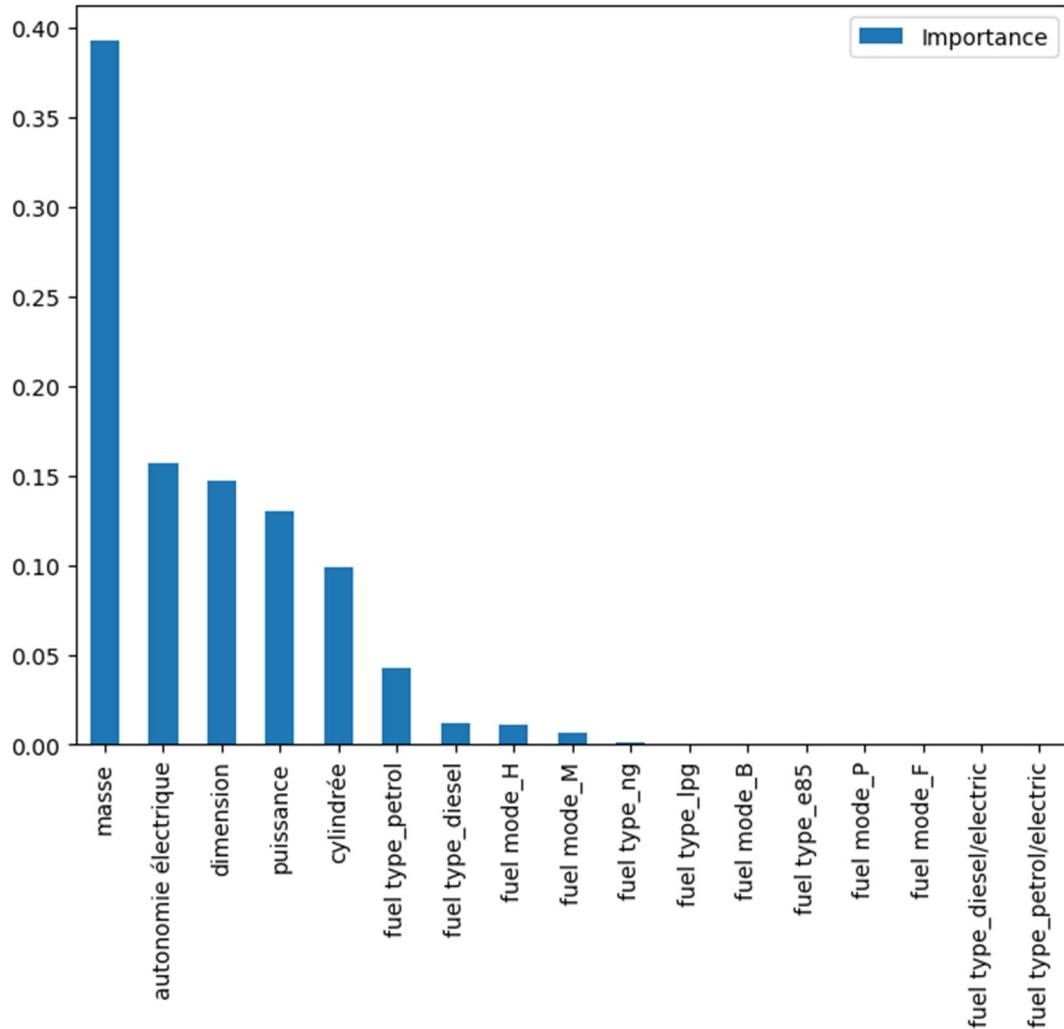
De la classe A à la classe G, le modèle RandomForest rfcl1 se trompe moins en réalisant 86% de bonnes prédictions en moyenne. Ainsi sur un jeu de test de 14074 véhicules :

- Pour la classe A : 1320 de prédictions correctes sur un total de 1341 véhicules A
- Pour la classe B : 742 de prédictions correctes sur un total de 1037 véhicules B
- Pour la classe C : 2823 de prédictions correctes sur un total de 3303 véhicules C
- Pour la classe D : 2309 de prédictions correctes sur un total de 2824 véhicules D
- Pour la classe E : 2557 de prédictions correctes sur un total de 2980 véhicules E
- Pour la classe F : 1448 de prédictions correctes sur un total de 1657 véhicules F
- Pour la classe G : 867 de prédictions correctes sur un total de 932 véhicules G

Le modèle RandomForest rfcl1 paraît plus intéressant parmi les trois modèles étudiés de part son score accuracy de 86%, son taux de surapprentissage de 6% et de ses coefficients de précision et de rappel meilleurs.

A.3 Les variables importantes de la classification

Le graphique suivant récapitule les variables les plus importantes pris en compte dans la classification



La masse du véhicule représente la variable la plus importante prise en compte dans la classification en matière d'émission de CO2(environ 40%).

Viennent ensuite l'autonomie électrique, la dimension, la puissance et la cylindrée (entre 9 et 16%). Les types de carburant pétrole , diesel ainsi que les véhicules de mode M(Mono-carburant) et H(Hybride non rechargeable sur borne électrique) jouent un rôle dans cette classification, cependant leur importance reste relativement faible(inférieure à 5 %).

Dans la section Optimisations suivante, nous allons essayer d'entrainer d'autres types de modèles sur le jeu de train traité avec d'autres méthodes.

B – Optimisations

B.1 Démarche

- Les modèles reglog1, dectree1 et rfcl1 suivants ont été déjà étudiés ci-dessus.

reglog1 : Modèle régression logistique entraîné avec toutes les variables explicatives retenues,

dectree1 : Modèle arbre de décision entraîné avec toutes les variables explicatives retenues ,

rfcl1 : Modèle random forest entraîné avec toutes les variables explicatives retenues,

- Les modèles ci-dessous énumérés constituent ceux étudiés dans le cadre des tentatives d'optimisation.

reglog2 : Modèle régression logistique entraîné avec les 5 premières Variables importantes,

reglog3 : Modèle régression logistique entraîné avec les 6 premières Variables importantes,

reglog4 : Modèle régression logistique entraîné avec les 7 premières Variables importantes,

dectree2 : Modèle arbre de décision entraîné avec les 5 premières Variables importantes,

dectree3 : Modèle arbre de décision entraîné avec les 6 premières Variables importantes,

dectree4 : Modèle arbre de décision entraîné avec les 7 premières Variables importantes,

rfcl2 : Modèle random forest entraîné avec les 5 premières Variables importantes,

rfcl3 : Modèle random forest entraîné avec les 6 premières Variables importantes,

rfcl4 : Modèle random forest entraîné avec les 7 premières Variables importantes,

rfcl5 : Modèle random forest entraîné avec le jeu de train retraité avec la méthode Normalized Oversampling,

svm_ro : Modèle SVM entraîné avec le jeu de train retraité avec la méthode Oversampling Randomsampler,

svm_sm: Modèle SVM entraîné avec le jeu de train retraité avec la méthode Oversampling SMOTE,

svm_ru : Modèle SVM entraîné avec le jeu de train retraité avec la méthode Undersampling Randomsampler,

svm_cc : Modèle SVM entraîné avec le jeu de train retraité avec la méthode Undersampling ClusterCentroids,

svm_probas : Modèle SVM entraîné avec le jeu de train retraité avec la méthode Undersampling Probabilités appartenance aux classes,

svm_poids : Modèle SVM entraîné avec le jeu de train retraité avec la méthode Poids des classes,

bclf: Modèle BalancedRandomForest entraîné avec toutes les variables explicatives retenues.

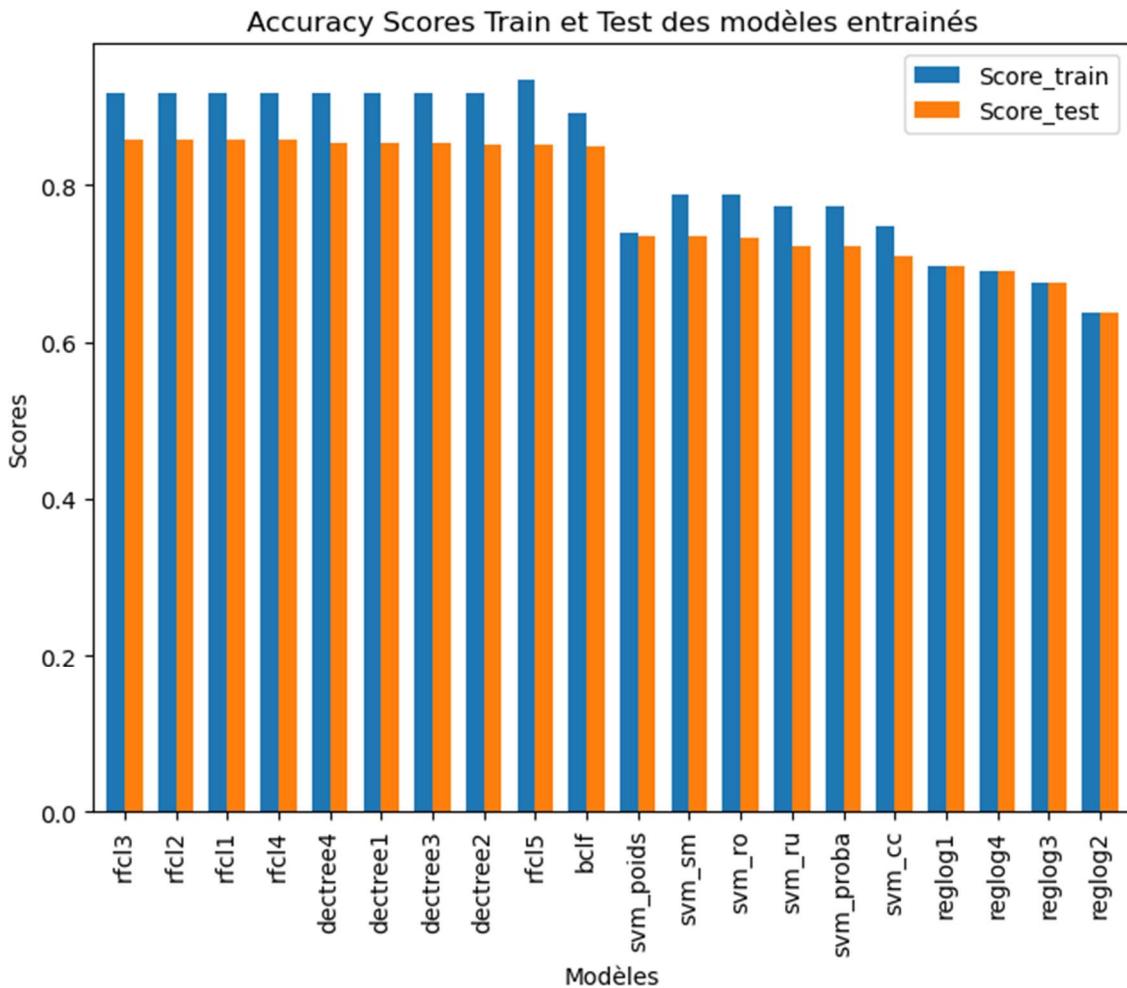
La démarche a été de retraiter le jeu d'entraînement avec différentes méthodes (variables les plus importantes, rééchantillonnages par oversampling, undersampling, smote, clustercentroids, probabilités d'appartenance aux classes, poids des classes) et d'entrainer en plus des modèles

régression logistique, arbre de décision et RandomForest, de nouveaux modèles (SVM et BalancedRandomForest) avec les nouvelles données. Les tests ont été réalisés sur le jeu de test initial. Les scores Accuracy train et test ont été calculés et les matrices de confusion et rapport de classification ont été établis.

B.2 Résultats et Analyses

Pour tous les modèles testés, les différents scores accuracy test obtenus sont les suivants :

	Modele	Score_train	Score_test	Details
0	reglog1	0.696031	0.696604	Modèle regression logistique avec toutes Variables
1	reglog2	0.636627	0.637203	Modèle regression logistique avec les 5 premières Variables
2	reglog3	0.676082	0.675288	Modèle regression logistique avec les 6 premières Variables
3	reglog4	0.690773	0.691133	Modèle regression logistique avec les 7 premières Variables
4	dectree1	0.917519	0.853489	Modèle arbre de décision avec toutes Variables
5	dectree2	0.916773	0.852210	Modèle regression logistique avec les 5 premières Variables
6	dectree3	0.917324	0.853276	Modèle regression logistique avec les 6 premières Variables
7	dectree4	0.917324	0.853631	Modèle regression logistique avec les 7 premières Variables
8	rfcl1	0.917484	0.857326	Modèle random forest avec toutes Variables
9	rfcl2	0.916755	0.858391	Modèle random forest avec les 5 premières Variables
10	rfcl3	0.917324	0.858462	Modèle random forest avec les 6 premières Variables
11	rfcl4	0.917306	0.857255	Modèle random forest avec les 7 premières Variables
12	rfcl5	0.933087	0.850931	Modèle random forest avec Normalized Oversampler
13	svm_ro	0.787147	0.733338	Modèle SVM Oversampling Randomsampler
14	svm_sm	0.787049	0.735399	Modèle SVM Oversampling SMOTE
15	svm_ru	0.772749	0.722893	Modèle SVM Uversampling Randomsampler
16	svm_cc	0.748358	0.708754	Modèle SVM Uversampling ClusterCentroids
17	svm_proba	0.772749	0.722893	Modèle SVM Uversampling Probabilités appartenances
18	svm_poids	0.738400	0.735612	Modèle SVM Poids classe
19	bclf	0.890873	0.848799	Modèle BalancedRandomForest



Parmi les modèles testés dans le cadre de l'optimisation, les meilleurs scores test sont obtenus par les modèles de type randomforest(scores entre 85 et 86%) notamment le modèle rfcl3 qui comparé au modèle rfcl1 étudié plus haut améliore non seulement le score test (85,84% vs 85,73%) mais réduit aussi le surapprentissage (5,89% vs 6%)

Les modèles de régressions logistiques affichent des scores train et test inférieurs à 70% mais n'ont pas de surapprentissage.

Les scores des modèles SVM se situent entre 70% et 73% avec une particularité pour le modèle svm_poids entraîné selon le poids des différentes classes qui donne en revanche un score train et score test quasi identiques de 73% avec un surapprentissage quasi nul.

Les modèles d'arbre de décision affichent des scores test au-dessus de 85% avec du surapprentissage supérieur à 6%.

Le modèle BalancedRandomForest bclf, avec un score train de 89% et test de 84,87% présente un surapprentissage de 4,21% qui reste moindre par rapport à celui des modèles randomforest, arbre de décision et svm.

Focus sur le Modèle RandomForest rfcl3

Prédiction	A	B	C	D	E	F	G
Realité							
A	1318	12	3	2	3	1	2
B	6	744	279	6	2	0	0
C	3	192	2823	275	10	0	0
D	7	1	266	2313	236	0	1
E	1	2	9	239	2572	143	14
F	2	0	3	3	141	1445	63
G	0	0	0	0	18	47	867

	precision	recall	f1-score	support
A	0.99	0.98	0.98	1341
B	0.78	0.72	0.75	1037
C	0.83	0.85	0.84	3303
D	0.82	0.82	0.82	2824
E	0.86	0.86	0.86	2980
F	0.88	0.87	0.88	1657
G	0.92	0.93	0.92	932
accuracy			0.86	14074
macro avg	0.87	0.86	0.87	14074
weighted avg	0.86	0.86	0.86	14074

L'analyse de la matrice de confusion et du rapport de classification du modèle rfcl3 par rapport au modèle rfcl1 étudié plus haut permet de constater une amélioration des coefficients de rappel des classes B , D et E augmentant ainsi les nombre de bonnes prédictions de ces trois classes.

Les taux de précision et de rappel des autres classes restent néanmoins relativement stables.

Ainsi, par ce modèle rfcl3, nous obtenons :

- Pour la classe A : 1318 vs 1320 de prédictions correctes pour le modèle rfcl1
- Pour la classe B : 744 vs 742 de prédictions correctes pour le modèle rfcl1
- Pour la classe C : 2823 vs 2823 de prédictions correctes pour le modèle rfcl1
- Pour la classe D : 2313 vs 2309 de prédictions correctes pour le modèle rfcl1
- Pour la classe E : 2572 vs 2557 de prédictions correctes pour le modèle rfcl1
- Pour la classe F : 1445 vs 1448 de prédictions correctes pour le modèle rfcl1
- Pour la classe G : 867 vs 867 de prédictions correctes pour le modèle rfcl1

4.3 - Modèles de Régression

D'un point de vue technique et malgré différentes méthodes d'optimisation de classification, nous observons toujours le phénomène de surapprentissage dans le modèle de classification dû notamment au déséquilibre de chacune des classes de la variable.

D'un point de vue métier, le choix d'un modèle de classification permettait de simplifier et de faciliter la lecture par tranche d'émission de CO2 pour les véhicules sur le marché.

Or notre variable cible, l'émission de CO2, est une variable continue qui, au regard des enjeux structurants de réduction, a besoin d'être analysée plus finement et actualisée régulièrement.

Mettre un nouveau référentiel de classes qui serait susceptible d'être à nouveau modifié pourrait provoquer l'incompréhension du grand public. Mesurer l'émission sur des tranches de catégories limite aussi la précision de la mesure de l'impact.

C'est pour cela que nous avons souhaité élargir notre étude à des modèles de régression .

Les étapes de preprocessing et d'encodage des variables sont les mêmes que celles opérées pour les modèles de classification, mise à part l'étape de catégorisation de la variable cible.

Nous passons dans une logique de variable continue.

Par conséquent, les classes de CO2 de A à F selon la norme en vigueur, n'ont plus lieu d'être dans ce modèle. Nous laissons donc la variable cible en valeur numérique.

Nous entraînons nos données sous 3 modèles :

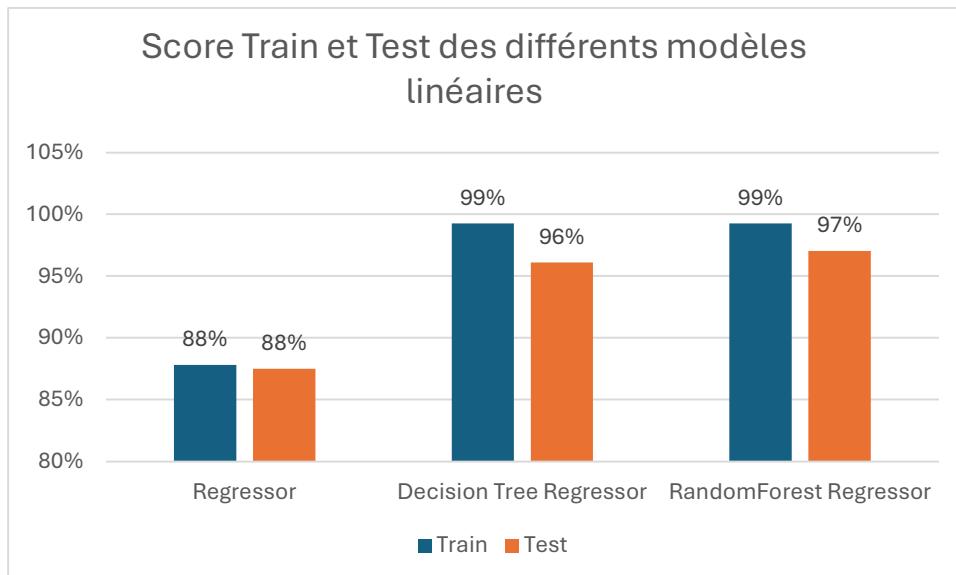
- Régression Linéaire,
- Decision Tree Regressor,
- Random Forest Regressor.

Le principe de séparation des jeux de données de train et de test est identique par rapport aux modèles de classification présentés plus haut.

Après l'entraînement, les 3 modèles donnent de très bons résultats avec des coefficients de détermination qui s'approchent de 1.

A noter que sur le modèle régression linéaire, nous avons une absence de phénomène de surapprentissage.

Sur les deux modèles, Decision Tree Regressor et RandomForest Regressor, nous observons un léger phénomène de surapprentissage. Touefois, l'écart est très faible (nous sommes sur 0,02/0,03 points par rapport à 0,06 points comparés aux modèles de classification).

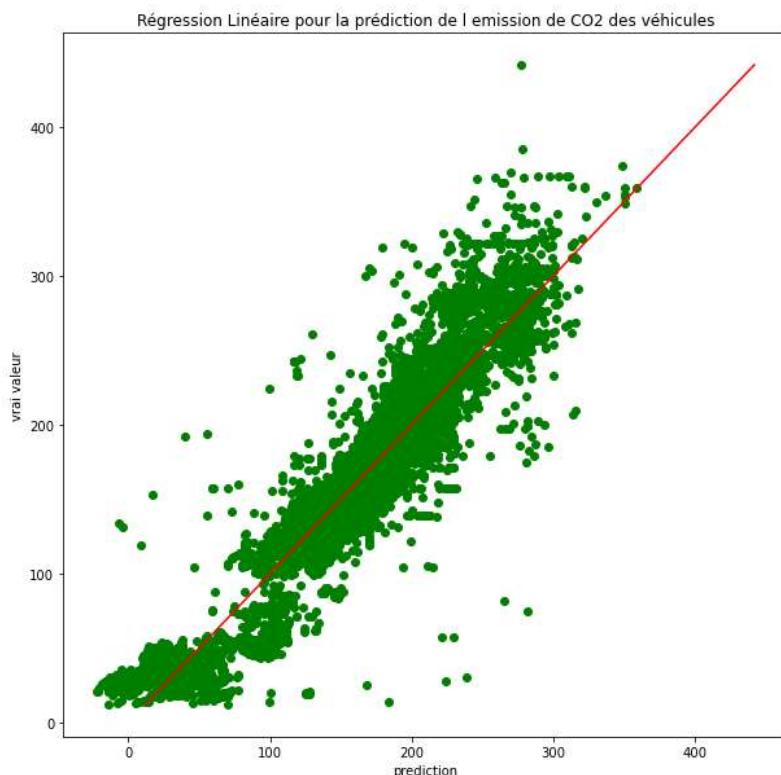


1. Régression linéaire :

Comme précisé plus haut, le modèle de Régression Linéaire donne un score de train de 87,8% et de test de 87,5%, ce qui est un bon score. Toutefois, ce modèle se place en dernière position au regard de sa performance parmi les 3 modèles. Cependant, nous notons une absence de surapprentissage sur ce modèle.

Le graphique ci-dessous montre que les données sont globalement alignées avec la droite de régression, ce qui confirme bien la corrélation entre les valeurs prédites et les valeurs réelles. Quelques écarts observés concernent les valeurs extrêmes des véhicules qui émettent plus de 400 g/km de CO₂.

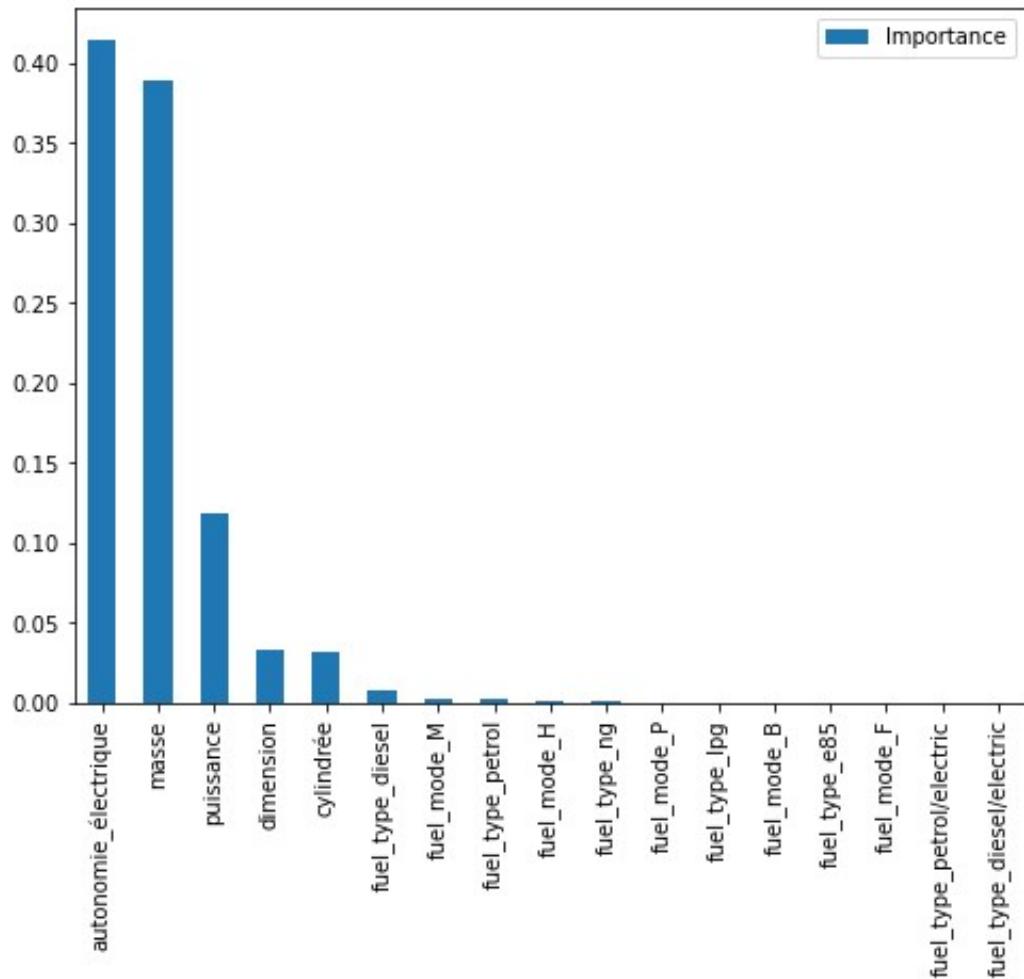
On observe quelques outliers qui sont certains véhicules de carburation essence et diesel en mode M (mono carburant). mode M (mono carburant).



2. Decision Tree Regressor:

Le modèle Decision Tree donne un score de train de 99,27% et de test de 96,09%. Ce modèle donne un très bon score. Cependant, nous observons un surapprentissage avec l'écart entre le score de train et le score de test de 0,3 points.

Les variables qui impactent le plus dans l'émission CO2 sont représentés dans le graphique de feature importance ci-dessous :



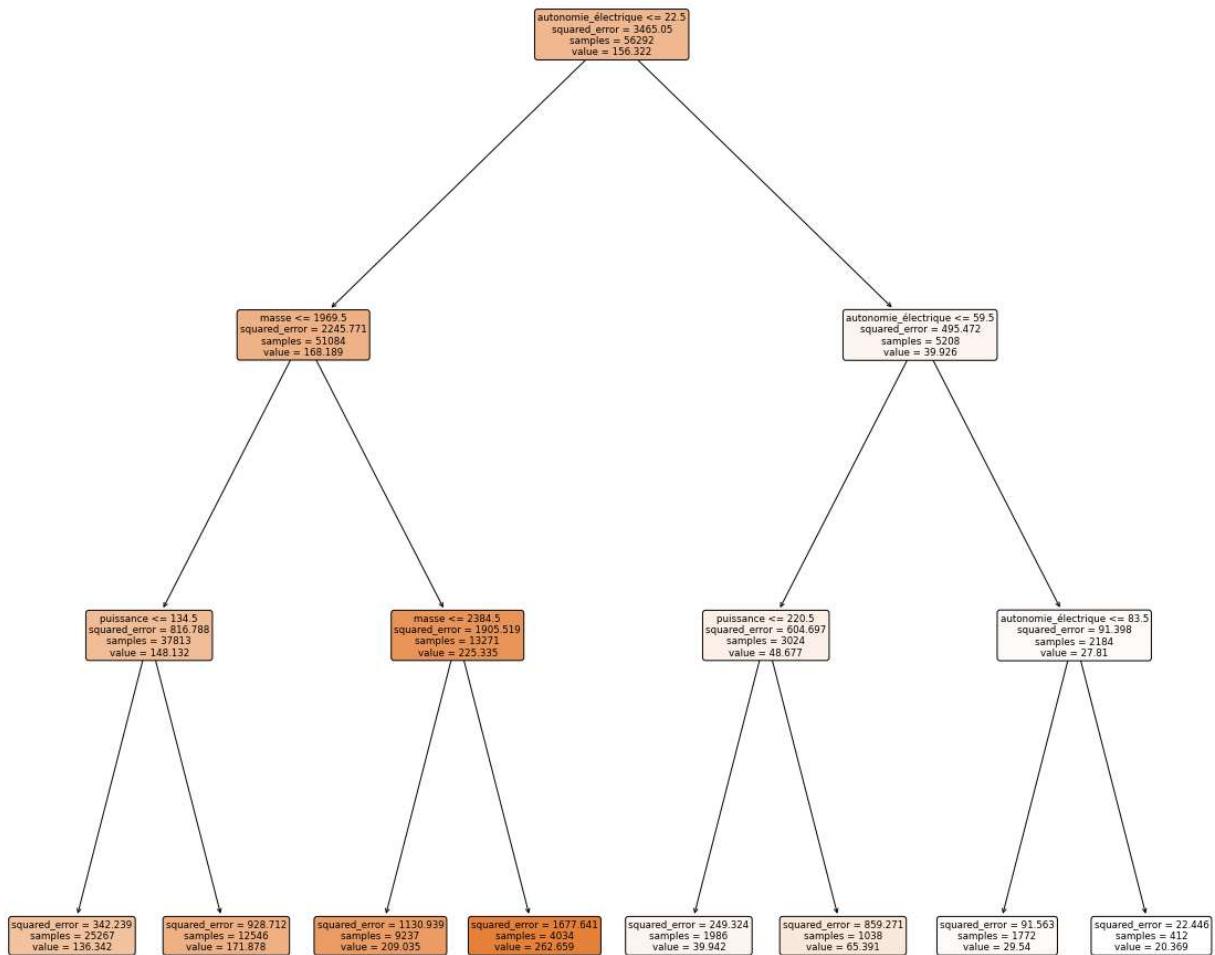
Soit en résumé les principaux :

- L'autonomie électrique
- La masse
- La puissance
- La cylindrée
- La dimension

Le modèle de DecisionTree Regressor nous permet de dessiner l'arbre.

Nous pouvons les interpréter de manière suivante :

- Un véhicule avec une autonomie électrique inférieure ou égale à 22,5 km aurait une émission à 156 g/km de CO₂.
- Une masse inférieure ou égale à 1969 kg aurait une émission de 168 g/km
- Une puissance inférieure ou égale à 134.5 kW aurait une émission de 148 g/km

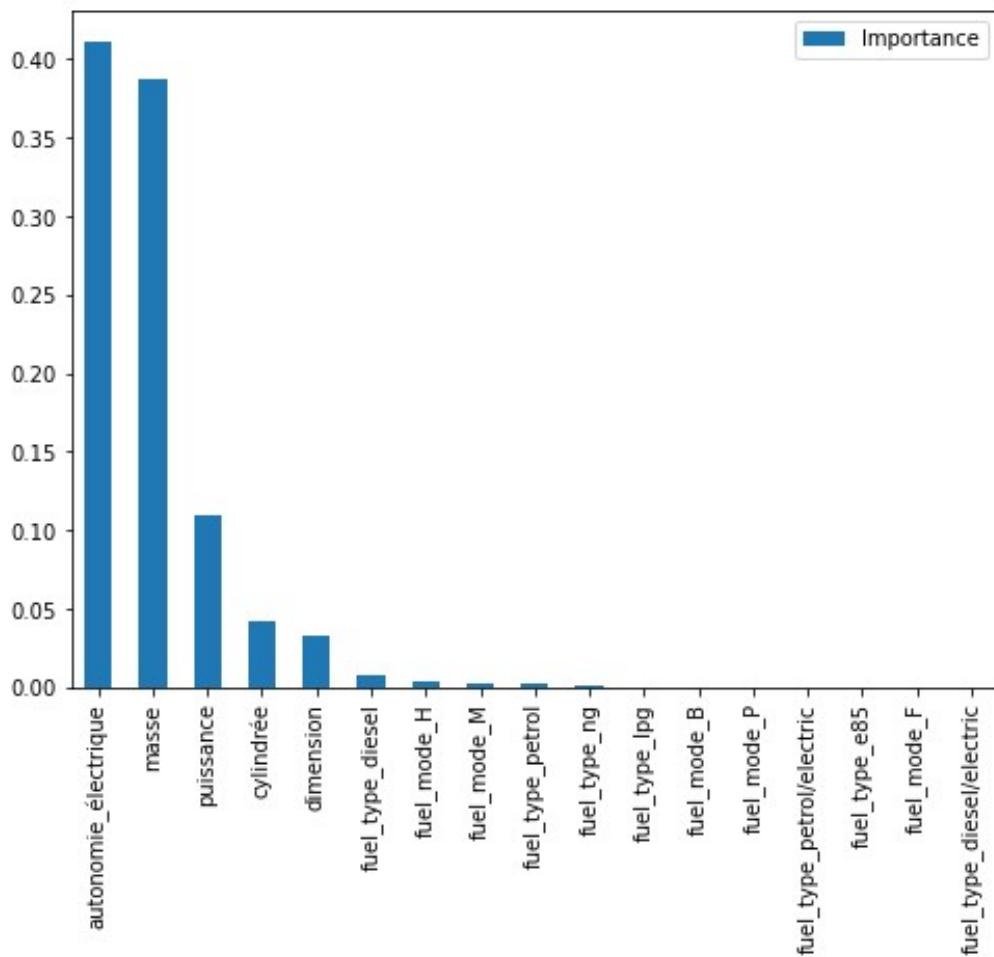


3. Random Forest Regressor :

Le modèle Random Forest Regressor donne un score de train de 99,04% et de test de 97,02%. Par rapport aux deux modèles précédents, ce modèle donne un excellent score et avec un surapprentissage moindre que le Decision Tree Regressor (écart entre le score de train et le score de test de 0,2 points).

Ce modèle semble un bon compromis avec un score proche de 100% et un surapprentissage de 0,2 points qui est très faible.

Les variables qui impactent le plus dans l'émission CO2 sont représentés dans le graphique de feature importance ci-dessous :



Le feature importance du modèle Random Forest Regressor est assez similaire à celui du modèle Decision Tree Regressor avec une petite différence.

En effet, la variable cylindrée est située avant la variable dimension alors que cette dernière est avant la cylindrée dans le modèle Decision Tree Regressor.

3. Évaluer la performance des modèles Decision Tree et Random Forest par l'étude de résidus :

Nous avons utilisé plus haut le score pour comparer les différents modèles entre eux.

Cependant, cette valeur n'est pas forcément la plus optimale et la plus pertinente pour évaluer des modèles non linéaires tels que le Decision Tree Regressor et Random Forest Regressor.

En effet, ces méthodes présentent une grande variance et cela pourrait biaiser le score.

Pour évaluer la performance de ces deux modèles, nous interprétons les métriques MSE (Mean Squared Error), MAE (Mean Absolute Error) et RMSE (Root Mean Squared Error).

Ci-dessous les résultats des résidus du modèle Decision Tree Regressor et Random Forest Regressor:

La mesure MSE est élevée, mais cela est à mettre en lien avec l'ordre de grandeur de la variable cible : l'émission CO2 est entre 128 à 393 g/km.

	MAE Train	MAE Test	MSE Train	MSE Test	RMSE Train	RMSE Test
Decision Tree	2.64	4.99	25.02	133.69	5	11.56
Random Forest	3.18	4.755	32.9	101.95	5.73	10.09

En regardant la métrique la plus interprétable (MAE), nous pouvons conclure qu'en moyenne, le modèle Decision Tree se trompe seulement de 4,9 unités sur la prédiction de l'émission de CO2 et le modèle Random Forest de 4,75 unités. Ce chiffre peut être considéré comme très bas.

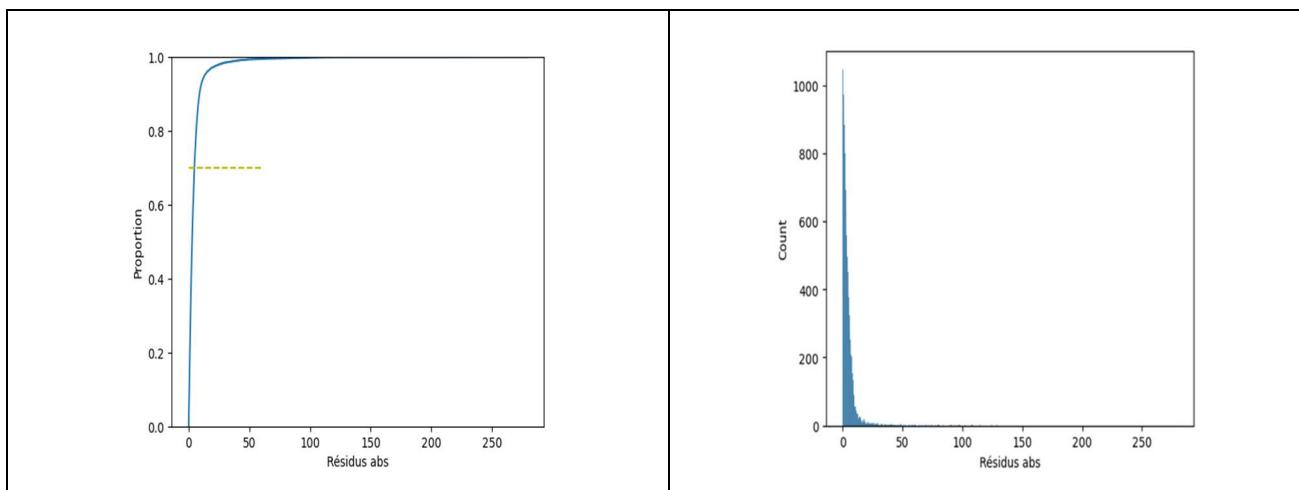
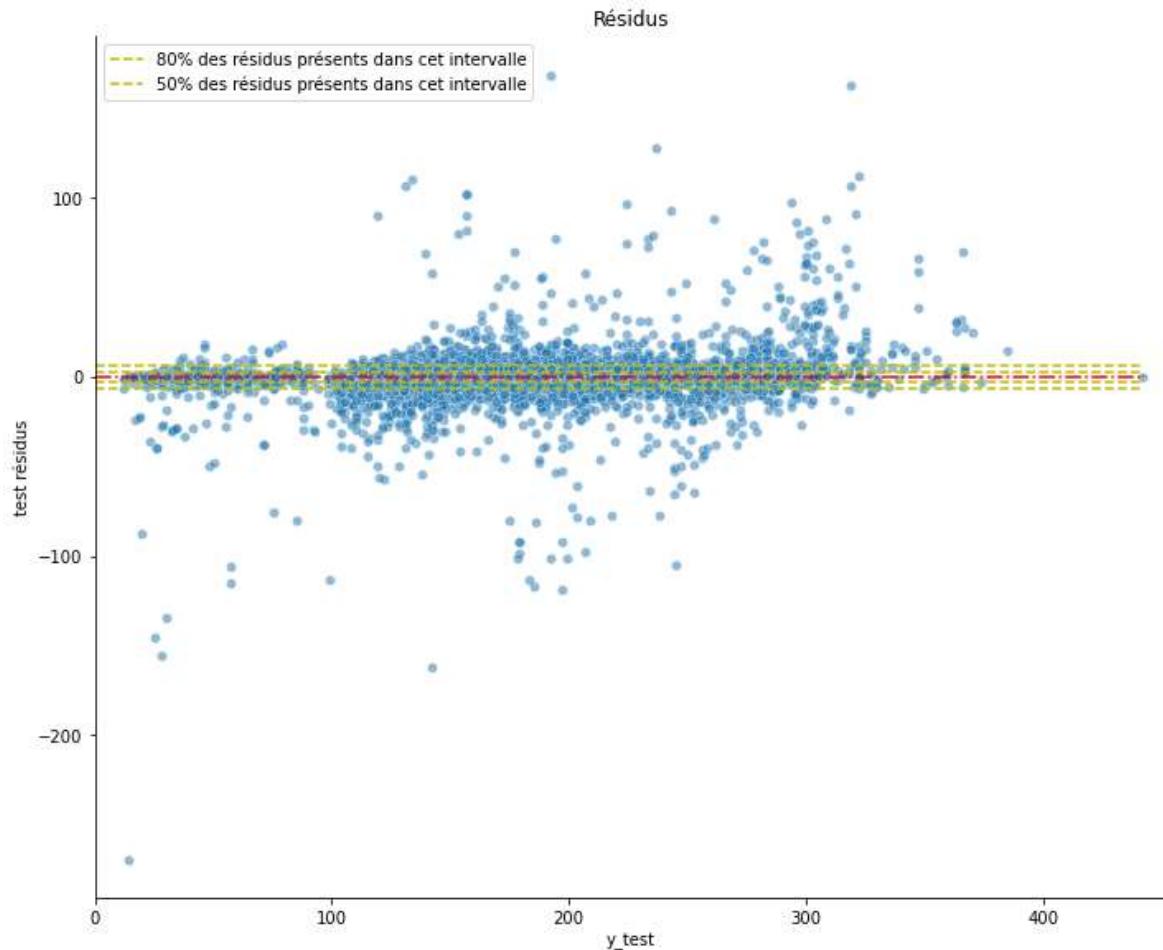
Nous pouvons également voir que le score de RMSE est également bas, ce qui nous permet de dire que notre modèle est assez performant.

La comparaison des scores donne les avantages et inconvénients suivants :

- Decision Tree :
 - Avantages : Meilleure performance sur les données d'entraînement.
 - Inconvénients : tendance au surapprentissage et moins performant sur les données de test.
- Random Forest :
 - Avantages : Meilleure généralisation sur les données de test, plus robuste et moins susceptible de surapprentissage.
 - Inconvénients : Légèrement moins performant sur les données d'entraînement.

On choisit d'approfondir l'étude des résidus du modèle Random Forest dont l'overfitting est plus faible.

Les graphiques ci-dessous nous montrent le rapport entre les valeurs réelles et les valeurs prédictes. Nous constatons que 80% des résidus ont des valeurs très faibles qui sont très proches de 0.



En analysant de plus près des résidus, nous obtenons les quantiles suivants qui confirme les valeurs faibles des résidus :

<pre>0.10 -6.182078 0.25 -2.908536 0.75 2.806312 0.90 6.126126 Name: residus, dtype: float64</pre>	<pre>count 14074.000000 mean 4.755909 std 8.907354 min 0.000000 25% 1.233417 50% 2.852375 75% 5.361999 max 269.262500 Name: Résidus abs, dtype: float64</pre>
--	--

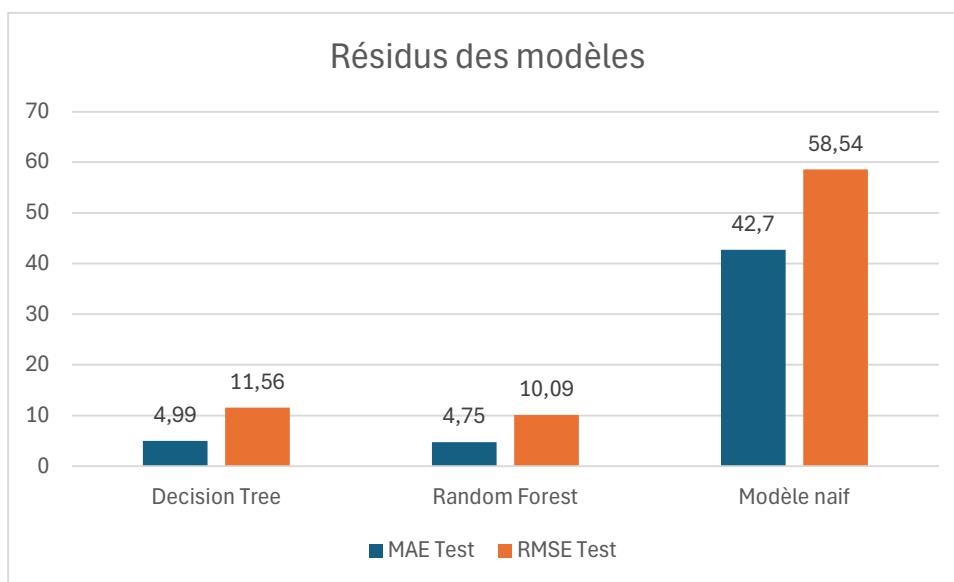
Pour conclure, nous essayons de les évaluer à travers une étude de « benchmark ».

En effet, nous allons comparer nos modèles non linéaires avec un modèle dit « naïf » qui prend tout simplement la moyenne de la variable d'entraînement y_{train} .

Nous obtenons les erreurs comme suit :

- MAE naïf de 42.7
- RMSE naïf de 58.54

Nous pouvons constater que les résidus du modèle « naïf » est très élevé par rapport aux deux modèles non linéaires que nous avons entraînés. Cela veut dire que nos modèles ne sont pas du tout « naïfs » et qu'ils ont une bonne performance en captant des relations complexes.



4.4. Choix de modèle le plus adapté au contexte métier

Rappel du contexte métier :

Nous sommes sur un marché qui évolue constamment et cela est d'autant plus nécessaire qu'en 2035, l'Europe a pris un engagement de ne plus vendre les véhicules qui émettent de CO2.

Ce sera donc un contexte d'amélioration continue pour lequel la précision des mesures sera clef. Il est donc essentiel de choisir un modèle qui soit suffisamment sensible à toutes les améliorations/optimisations apportées par les constructeurs et qui s'adapte à de nouvelles données.

Choix du modèle :

Le modèle de régression linéaire n'entraîne pas de surapprentissage car il est moins sensible aux bruits, aux changements et fluctuations des données de marché. Il est donc plus généralisable et stable sur les nouvelles données.

Toutefois, il entraîne un manque de précision qui est structurant dans les choix que les constructeurs automobiles doivent mener. La limite étant que si la relation des variables explicatives avec la variable cible n'est pas complètement linéaire, le modèle linéaire peut ne pas capturer aussi bien qu'un modèle plus complexe. Des tests concrets, nous ont démontré trop d'imprécisions.

Comme constaté plus haut, les deux modèles Decision Tree Regressor et Random Forest Regressor sont capables non seulement de capturer les tendances générales au travers des données d'apprentissage et les relations complexes.

Ils seront très précis mais moins robustes si les conditions de marché changent rapidement. Pour améliorer la robustesse, il est essentiel d'adopter des stratégies qui prennent en compte la dynamique changeante du marché, la variabilité des données et les futures tendances possibles.

Parmi les éléments clefs de la stratégie d'apdataion, nous pouvons citer :

- l'incorporation de nouvelles données régulièrement
- l'analyse régulière de l'importance des variables
- la surveillance en continue et le ré-entraînement régulier du modèle
- la détection régulière d'anomalies et outliers

En conclusion, nous choisissons **le modèle Random Forest Regressor** qui apporte une excellente précision et un surapprentissage le plus faible parmi les modèles les plus performants analysés.

Notre choix s'accompagne d'une **recommandation de mise en oeuvre d'une stratégie intégrant les éléments clefs cités ci-dessus** pour assurer la bonne adaptation au marché changeant.

5. Production Finale

Nous avons développé une application sur Streamlit basée sur le modèle Random Forest Regressor.

Cette application permet de prédire la valeur de l'émission de CO2 en entrant les données des caractéristiques telles que la masse, la puissance, la dimension, le cylindrée, la puissance, l'autonomie-électrique, le fuel mode et le fuel type.

Elle permettra aux constructeurs automobiles de simuler les caractéristiques sélectionnées et d'en prévoir l'émission de CO2 avant de décider une mise sur le marché.



Cette application peut être renforcée par l'ajout de nouvelles caractéristiques impactant l'émission de CO2 lors de l'usage des véhicules (type de boîte de vitesse...).

Nous allons aborder ces prolongements dans la section suivante qui traite des limites.

6. Les limites et suite du projet

6.1. Les Limites

Nous avons certes de belles performances basées sur un set de données conséquent.

Après différentes itérations, nous avons opté pour un modèle de régression, le modèle Random Forest Regressor, répondant largement à la nature des données. Pour assurer une bonne adaptation du modèle avec le contexte évolutif et à l'enjeu du marché, nous le complétons d'une stratégie d'adaptation (Cf point 4.4)

Toutefois, des limites à notre étude et au set de données sont à considérer.

➤ **Certaines caractéristiques techniques et technologiques des véhicules ne sont pas prises en compte :**

- Le **volume du véhicule** que l'on peut appréhender par type de segment (berline, citadine, van, coupé, SUV...).

Celui-ci a un impact sur la consommation de carburant en lien avec la résistance au vent du véhicule (C_x), par conséquent sur l'émission de CO2.

- Le **type de boite de vitesse** (boite automatique ou manuelle).

Dans le passé, les boites automatiques consommaient plus de carburant. Désormais de gros progrès permettent d'atténuer fortement ce différentiel.

- La présence d'un **turbo compresseur**

A puissance équivalente, il permet d'avoir un moteur avec une cylindrée plus faible, par conséquent une consommation de carburant réduite, et donc de CO2.

- **L'injection directe**

Les moteurs à essence bénéficient désormais de ce dispositif qui contribue aussi à baisser la consommation de carburant

- **Le régulateur de vitesse**

Il permet d'avoir une conduite plus homogène automatiquement, par conséquent moins d'accélérations qui génèrent de la consommation, notamment dans les contextes de trafics élevés (villes, embouteillages)

- Le **start and stop** qui permet de couper le moteur à l'arrêt (feu, stop, embouteillages),
- Le **filtre à particules et les catalyseurs** qui permet de traiter les émissions en sortie du moteur pour réduire les impacts au niveau de la pollution
- La **climatisation intelligente** qui permet d'avoir une gestion optimisée des températures.

La climatisation est un grand consommateur d'énergie qui augmente la consommation de carburant.

- **Les freins régénératifs** pour les véhicules hybrides

Ils permettent de récupérer l'énergie cinétique au freinage. Cela permet de réduire l'utilisation du moteur à combustion et réduit donc les émissions de CO2

➤ **Certaines caractéristiques relatives à l'usage :**

- Le type de conduite (éco-conduite/conduite sportive)
- Le taux de remplissage des véhicules (co-voiturage)
- Les conditions liées à la géographie (conduite en plaine, en montagne)
- Les conditions liées à la météo (vent, chaleur)
- Les conditions liées à la densité du trafic

Tous ces paramètres sont des variables qui ont un impact sur les émissions de CO2.

6.2. Suite du projet

La suite du projet consisterait à enrichir le dataset avec un ensemble de caractéristiques mentionnées précédemment et intégrer le volet dynamique du marché.

Nous pourrions envisager :

➤ A court terme :

- Enrichir le set de données en opérant un Webscraping sur les sites des constructeurs automobiles qui répertorient les gammes de véhicules par segment (Berline, Citadine...). Nous pourrions alors évaluer l'impact du volume sur les émissions de CO2.

➤ A moyen et long terme :

- Surveiller le contenu des sites constructeurs et des publications des organismes de référence (ADEME, Europea) pour identifier des publications futures contenant ces caractéristiques
- Monitorer et ré-entrainer périodiquement notre modèle au regard des évolutions du marché

7. Les difficultés du projet

La première difficulté rencontrée a été d'accepter de prendre un temps conséquent dans la phase d'exploration et de compréhension de la problématique.

En effet, nous avons consacré huit semaines à :

- Lever les problèmes de volumétrie de données pour accéder aux données proposées
- Trouver des sources de données de qualité (données actuelles et référencées par organismes reconnus (ADEME, Europea..))
- Comprendre et bien assimiler les données et leur signification métier
- Sélectionner les caractéristiques à conserver

Cette étape a permis d'approfondir le sujet et de nous faire monter en connaissance sur le métier et la problématique du sujet.

En complément, on peut noter que le Module Machine learning arrive dans la formation avec un léger décalage par rapport au démarrage du projet ce qui amène quelques difficultés à se projeter sur les étapes suivantes.

Toutefois, cela a été aussi une opportunité car l'apprentissage du cours a été jumelé avec l'application concrète du projet.

Enfin, d'un point de vue de l'organisation de l'équipe, nous sommes quatre dont une personne sur un fuseau horaire différent (6 heures de décalage), ce qui positionnait les horaires de réunion à partir de l'après-midi voire la soirée.

A souligner malgré cela l'investissement et la régularité de tous les membres à tous nos rendez-vous.

Notre sujet est un thème qui nous touche tous dans notre vie (pollution, santé, environnement) et notre avenir.

La sensibilité et les expertises différentes de chacun auraient pu être clivantes d'autant plus que sans hiérarchie par définition dans un tel projet, le consensus aurait pu être compliqué et freiné l'avancement.

Cela n'a pas été le cas. Les échanges ont été nourris de points de vue différents et constructifs.

Chacun a su et pu se positionner naturellement avec une belle complémentarité de compétences, aussi bien de savoir-faire que de savoir être, ce qui a amené de la richesse et de la nuance dans nos échanges.

8. Conclusion

Avec notre modèle, nous pouvons prédire un premier niveau de pollution dans le cas de la mise sur le marché de nouveaux types de véhicules. Celui-ci pourra être affiné avec l'intégration des caractéristiques complémentaires citées précédemment.

Au regard de l'enjeu structurant concernant le climat et les décisions prises en 2023 par l'Europe (en 2035 les nouveaux véhicules ne devront plus émettre de CO2), notre analyse de données a démontré que le véhicule 100% électrique s'avérait être la solution à planifier par les constructeurs.

En attendant cet horizon, les constructeurs doivent faire l'effort de transformer progressivement les véhicules thermiques pour évoluer vers des émissions qui tendent vers zéro.

Cette trajectoire de décroissance doit être significative dans les onze années à venir.

L'outil permet de donner des clefs pour développer des véhicules avec des caractéristiques qui répondent aux objectifs annuels de réduction d'émissions de CO2.

Notre sujet a porté sur les émissions liées à l'usage des véhicules.

Toutefois, il faudrait garder à l'esprit qu'en vue systémique et chaîne de valeur, la réduction des émissions de CO2 dans le domaine du transport, devra intégrer la partie amont et la partie aval des constructeurs automobiles.

En effet, la production des véhicules et le fret amont tout comme la partie logistique transport et l'économie circulaire des véhicules sont très impactant en termes d'émissions de CO2 et de pollution.

9. Annexes

- Fichier d'analyse et rapport d'exploration du dataset et des variables en amont

https://docs.google.com/spreadsheets/d/1AyNcngEq_MDM_k2awOdoB2AQQqbqhXqi/edit?gid=1361603071#gid=1361603071

- Synthèse codes projet réalisés dans GitHub

<https://github.com/mhtran158/Prediction-CO2>

- Fichier retenu

<https://www.eea.europa.eu/en/datahub/datahubitem-view/fa8b1229-3db6-495d-b18e-9c9b3267c02b?activeAccordion=>

2022 | Monitoring of CO2 emissions from passenger cars, 2022 - Final data | ascii (.csv, .txt, .sql)

Monitoring of CO2 emissions from passenger cars Regulation (EU) 2019/631

Prod-ID: DAT-116-en | Published 03 Aug 2023 | Last modified 18 Mar 2024