



UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE
FACULTÉ DES SCIENCES ÉCONOMIQUES, SOCIALES ET DE GESTION
FACULTÉ DES SCIENCES EXACTES ET NATURELLES

**Deuxième année de Master Statistique pour l'Évaluation et la
Prévision**

Gestion de projet digital

**CARBOMILES, une application de prédiction des émissions de GES
moyennes hebdomadaires par personne pour les trajets domicile-travail**

Réalisé par **Emilie LI, Brieuc DAXHELET, Axel DUBREUIL, Moussa DIALLO et
Imane SAHNOUNE**

Sous la direction de **Morgan Cousin, Arona Diene et Amor Keziou**

Année académique 2024-2025

Table des matières

Remerciements	1
Introduction.....	2
I. Présentation du produit et de l'interface utilisateur	3
1. CARBOMILES, c'est quoi ?	3
2. Interface d'accueil	4
3. Formulaire	4
4. Interface de sortie	5
II. Les fonctionnalités implémentées dans CARBOMILES	7
1. Les fonctionnalités générales de l'application	7
2. Les fonctionnalités du formulaire	8
3. Construction du pipeline VBA-Python.....	10
III. Préparation et analyse des données	11
1. Présentation de la base de données	11
2. Nettoyage de la base de données	13
3. Exploration des données	14
IV. Elaboration d'un modèle de prédiction	17
1. Sélection des variables	17
2. Méthodes de prédiction testées et sélection du modèle	18
3. Les limites du modèle sélectionné	19
Conclusion	20
Pour aller plus loin.....	20
Bibliographie	21
Annexes	22
Annexe 1 : Roadmap	22
Annexe 2 : Dictionnaire des variables	24
Annexe 3 : Tableau des statistiques descriptives de la variable cible	39
Annexe 4 : Analyse des corrélations entre les variables quantitatives explicatives.....	40
Annexe 5 : Analyse des corrélations entre les variables quantitatives explicatives et la variable cible.....	40
Annexe 6 : Analyse des corrélations entre les variables catégorielles explicatives.....	41
Annexe 7 : Comparaison des erreurs MSE (Mean Squared Error) et MAE (Mean Absolute Error) du modèle lasso et de l'arbre de régression en grammes de CO2e hebdomadaire .	41

Table des figures

Figure 1 : Page d'accueil de l'application CARBOMILES	4
Figure 2 : Formulaire de saisie à remplir par l'utilisateur	5
Figure 3 : Fenêtre de chargement.....	6
Figure 4 : Interface de sortie affichant les résultats de l'estimation de l'empreinte carbone de l'utilisateur.....	7
Figure 5 : Formulaire de saisie contenant des messages par défaut dans les champs	9
Figure 6 : Exemple de valeur non-valide dans le champ de la distance	10
Figure 7 : Emissions de CO2 hebdomadaires moyennes par personne et distance domicile-travail moyenne pour chaque catégorie socio-professionnelle	15
Figure 8 : Emissions de CO2 hebdomadaires moyennes par personne et distance domicile-travail moyenne par tranches d'âge	16
Figure 9 : Quantité moyenne de CO2 hebdomadaire par personne selon la distance domicile-travail.....	16

Remerciements

Nous tenons à exprimer notre sincère gratitude aux enseignants qui nous ont accompagnés et guidés tout au long de ce projet de développement d'une application. Chacun de ces enseignants a contribué de manière unique et complémentaire à l'aboutissement de ce projet, et nous leur en sommes profondément reconnaissants. Leurs conseils et leur expertise ont été essentiels pour la réussite de ce travail, et leurs encouragements nous ont permis d'avancer avec confiance et rigueur.

Nous adressons tout d'abord nos remerciements à Monsieur Morgan COUSIN pour sa contribution significative en matière de méthodes et d'outils de travail, en nous initiant aux meilleures pratiques de la gestion de projet agile. Ses retours constructifs et détaillés à l'issue de nos présentations nous ont permis d'améliorer notre travail à chaque sprint et d'approfondir notre compréhension des principes agiles. Sa vision et ses conseils avisés ont été un des repères précieux pour structurer et organiser notre projet de manière optimale.

Nous souhaitons également remercier Monsieur Amor KEZIOU, dont les compétences en statistiques ont été d'une aide précieuse pour la conception et la sélection d'un modèle de prédiction. Son accompagnement nous a permis de mieux appréhender l'aspect analytique de notre application et de faire des choix judicieux quant aux outils et aux méthodes statistiques à utiliser. Grâce à ses conseils, nous avons pu ajouter une dimension de prédiction solide à notre projet.

Enfin, nous exprimons toute notre reconnaissance à Monsieur Arona DIENE qui nous a guidé dans le développement de l'interface utilisateur et dans la liaison entre VBA et Python. Son expertise technique en VBA a été déterminante pour améliorer l'ergonomie et l'expérience utilisateur de notre application. Ses recommandations ont facilité la prise en main et la fluidité de notre interface, ainsi que le renforcement des fonctionnalités de notre application.

Introduction

Dans le cadre de la transition écologique, la réduction des émissions de CO₂ est un enjeu majeur pour limiter les effets du changement climatique et préserver l'environnement. La prise de conscience croissante à l'échelle mondiale, qu'il s'agisse des gouvernements, des entreprises ou des citoyens, souligne l'importance de réduire l'empreinte carbone. En particulier, les trajets domicile-travail représentent une part significative des émissions de gaz à effet de serre, en raison de la dépendance aux transports individuels et à l'utilisation de véhicules polluants.

Cependant, bien que ces trajets contribuent largement à l'empreinte carbone quotidienne, il reste difficile pour les individus de mesurer avec précision l'impact environnemental de leurs déplacements. Beaucoup ne disposent pas d'outils permettant d'évaluer les émissions de CO₂ de leurs trajets ou pour comparer leur situation à celle d'autres usagers. En outre, trouver des solutions concrètes pour réduire cet impact, telles que l'adoption de modes de transport plus écologiques ou l'optimisation des trajets, reste complexe.

Ce constat met en avant l'urgence de développer des solutions accessibles et compréhensibles, qui permettent à chacun d'évaluer son empreinte carbone et de modifier ses comportements de déplacement de manière responsable. Les outils numériques et les applications mobiles représentent une opportunité efficace pour aider les individus à quantifier et réduire leur impact écologique, contribuant ainsi à la transition vers une société plus durable. Ainsi, nous avons créé l'application CARBOMILES qui vise à prédire et mesurer les émissions de gaz à effet de serre (GES) moyennes hebdomadaires liées aux déplacements domicile-travail, en fournissant aux utilisateurs des informations sur leur impact carbone dont ils sont curieux de connaître ou tout simplement soucieux afin de mettre en place des actions plus respectueuses de l'environnement.

Afin de mener ce projet à bien, une responsabilité a été attribuée à chacun des membres de l'équipe et dont les rôles sont complémentaires, chacun apportant alors une contribution essentielle à sa réalisation. Les contributeurs du projet se décomposent de la manière suivante :

- **Brieuc DAXHELET (Product Owner)** : Il dirige l'équipe en définissant la vision du produit, priorisant les fonctionnalités, et veillant à ce que l'application réponde aux besoins des utilisateurs. Avec une expertise en gestion de projets, il est le moteur derrière notre application de prédiction de GES émis lors des déplacements domicile-travail.
- **Axel DUBREUIL (Data Analyst/Data Governance + Scrum Master)** : En tant que data analyst, il est responsable des analyses et de l'interprétation de nos données et résultats, ainsi que de leur visualisation. Parallèlement, son rôle en tant que Scrum

Master est d'assurer la collaboration au sein de l'équipe et veiller à ce que les objectifs du projet soient atteints dans les délais impartis.

- **Moussa DIALLO (Data Engineer)** : Son rôle essentiel est la collecte, la préparation, et la structuration des données qui alimentent l'application, assurant ainsi la disponibilité de données de qualité pour des prédictions précises.
- **Imane SAHNOUNE (Data Scientist)** : Elle se concentre sur l'élaboration de l'algorithme de prédiction au cœur de l'application, en développant des modèles analytiques et statistiques, contribuant ainsi à faire de notre application un outil fiable.
- **Émilie LI (Front End/User Interface)** : Son rôle est de concevoir et de développer l'aspect visuel et interactif de l'application. Elle crée une interface utilisateur intuitive et ergonomique, permettant aux utilisateurs d'accéder facilement aux fonctionnalités, comme la saisie de données et la visualisation des résultats.

Ce rapport sera composé d'une première partie consacrée à la présentation du produit, en tant que réponse au besoin client, et à la conception de l'interface utilisateur. Nous présenterons ensuite dans une seconde partie, les fonctionnalités implémentées dans l'application ; puis nous enchaînerons sur la préparation des données de notre jeu de données et son analyse dans une troisième partie. Enfin, nous terminerons par l'élaboration du modèle de prédiction de notre application CARBOMILES. Ce processus détaillé permettra de créer une application pouvant sensibiliser les personnes actives à réduire leur empreinte carbone et modifier leurs comportements de déplacement, en privilégiant notamment des solutions de mobilités douces.

I. Présentation du produit et de l'interface utilisateur

1. CARBOMILES, c'est quoi ?

CARBOMILES est une application digitale visant à donner une estimation des émissions de GES hebdomadaires moyennes, mesurées en grammes de CO2 équivalent, pour les trajets domicile-travail. De ce fait, cette application s'adresse plus particulièrement aux personnes en activité, qui souhaiteraient comprendre et mesurer l'empreinte carbone de leurs déplacements domicile-travail, que ce soit par curiosité ou par souci de l'environnement, afin de mettre éventuellement en place des actions plus durables. Pour répondre au mieux à ce besoin, notre application apporte une estimation personnalisée de l'impact carbone pour chaque utilisateur, tout en lui offrant la possibilité de se situer en moyenne par rapport à d'autres personnes ayant des profils ou des habitudes similaires.

Le développement de notre application s'est réalisé sous la méthode Agile : des sprints reviews ont été réalisés chaque quinzaine afin de montrer l'avancée du projet et rendre visible l'ensemble du travail effectué. Une roadmap a également été construite pour fournir une

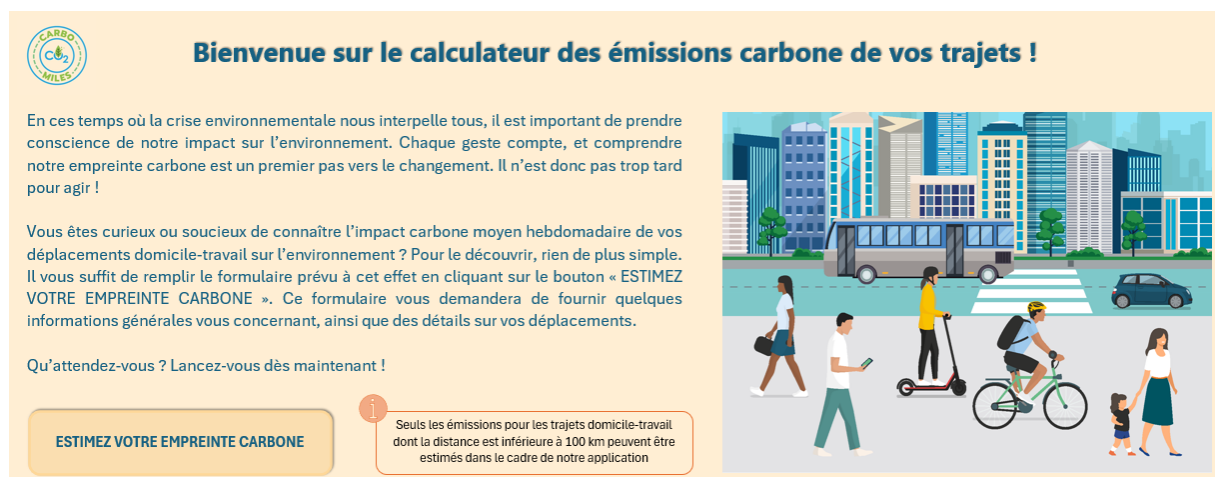
représentation temporelle et incrémentale du produit final découpé en plusieurs releases (Annexe 1). Chaque release correspond à un ensemble de fonctionnalités qui satisfait une partie des besoins de l'utilisateur en vue de récupérer des retours intermédiaires et adapter la roadmap en conséquence. Cela permet ainsi d'atteindre la vision produit globale.

2. Interface d'accueil

Au lancement de l'application CARBOMILES, une page s'affiche, correspondant tout simplement à son interface d'accueil (Figure 1). Cette dernière est composée tout d'abord d'un titre et du logo de notre application. Nous retrouvons ensuite un texte fournissant une brève mise en contexte, ainsi qu'une explication sur le fonctionnement de l'application. Il explique notamment le principe du bouton « ESTIMEZ VOTRE EMPREINTE CARBONE » qui permettra d'afficher un formulaire à remplir par l'utilisateur et qui servira à estimer la quantité de GES émise en moyenne par semaine lors de ses trajets domicile-travail, sur la base des informations qu'il aura fourni.

Par ailleurs, une note d'information est également ajoutée à côté du bouton « ESTIMEZ VOTRE EMPREINTE CARBONE ». Cette note a pour but d'informer l'utilisateur d'une des limites de notre application : les estimations des émissions ne sont limitées qu'aux trajets domicile-travail dont la distance est inférieure à 100 km.

Figure 1 : Page d'accueil de l'application CARBOMILES



Source : réalisé par les membres du projet

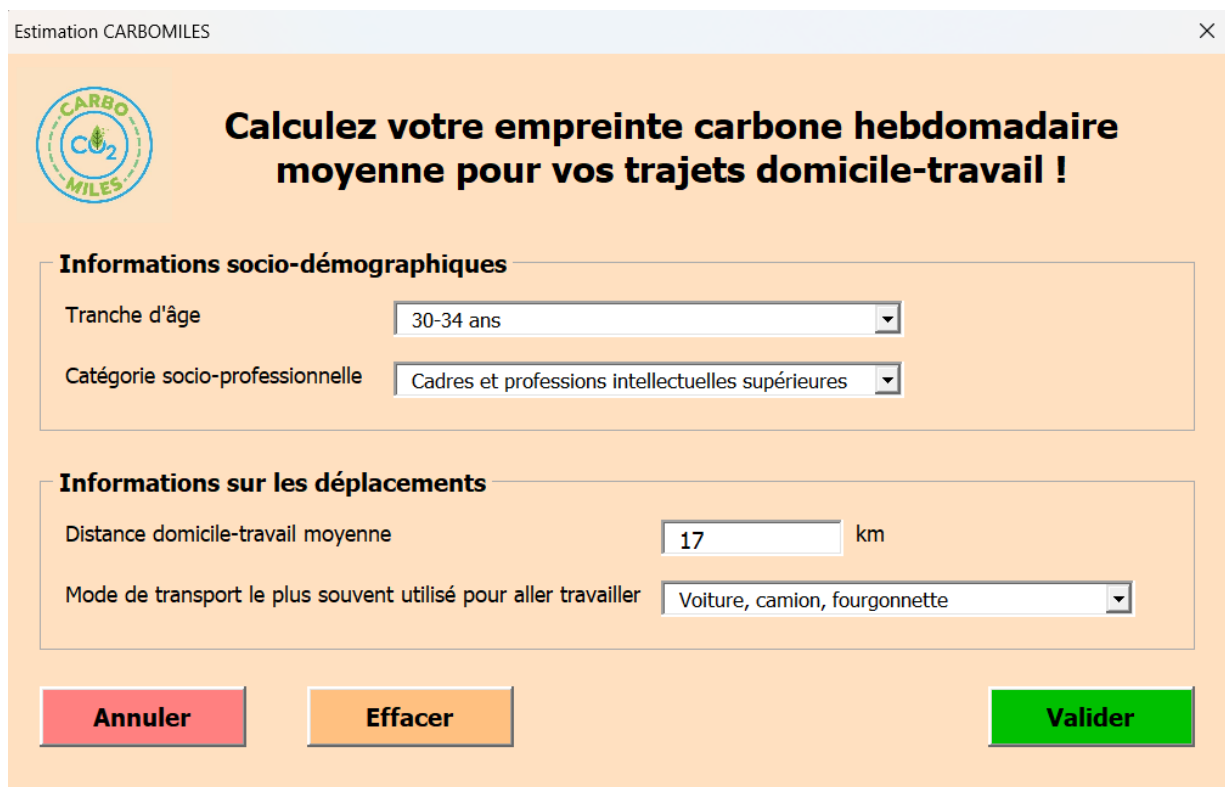
3. Formulaire

À partir du moment où l'utilisateur clique sur le bouton « ESTIMEZ VOTRE EMPREINTE CARBONE » de la page d'accueil, un formulaire s'affiche à l'écran (Figure 2). Ce formulaire est divisé en deux parties afin d'améliorer la visibilité des différentes rubriques pour l'utilisateur.

La première partie du formulaire renvoie aux informations socio-démographiques de l'utilisateur telles que la tranche d'âge à laquelle il appartient et sa catégorie socio-professionnelle, qui seront tous deux présentées sous la forme d'une liste déroulante. Ces informations permettront d'afficher par la suite différents graphiques, afin de lui laisser l'opportunité de comparer ses émissions de GES moyennes hebdomadaires avec toutes les autres personnes actives aux profils similaires ou non. Cela lui permet ainsi d'avoir une vue globale de la situation.

La deuxième partie du formulaire est, quant à elle, consacrée aux informations qui concernent les déplacements qu'il effectue dans le cadre de ses trajets domicile-travail, à savoir le mode de transport principalement utilisé (vélos, voiture, transport en commun, marche à pied, etc.), qui est présenté sous la forme d'une liste déroulante, ou encore la distance moyenne en kilomètres entre son lieu de vie et son lieu de travail.

Figure 2 : Formulaire de saisie à remplir par l'utilisateur



Estimation CARBOMILES

Calculez votre empreinte carbone hebdomadaire moyenne pour vos trajets domicile-travail !

Informations socio-démographiques

Tranche d'âge: 30-34 ans

Catégorie socio-professionnelle: Cadres et professions intellectuelles supérieures

Informations sur les déplacements

Distance domicile-travail moyenne: 17 km

Mode de transport le plus souvent utilisé pour aller travailler: Voiture, camion, fourgonnette

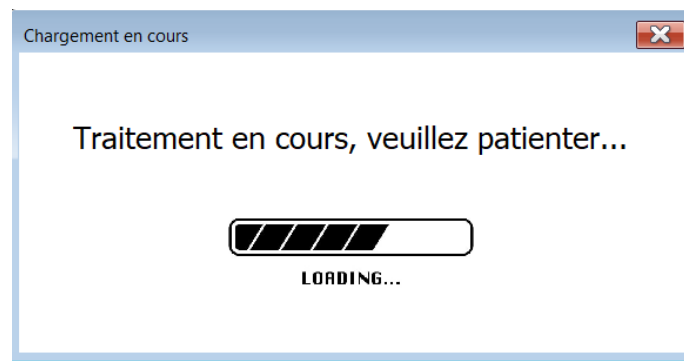
Annuler Effacer Valider

Source : réalisé par les membres du projet

4. Interface de sortie

Une fois que tous les champs sont remplis correctement, le formulaire peut être validé et les informations saisies sont envoyées et traitées par le modèle de prédiction, qui s'exécutera en arrière-plan. Lorsque l'utilisateur valide le formulaire, une fenêtre s'affichera à l'écran afin d'inviter l'utilisateur à patienter le temps du traitement ([Figure 3](#)).

Figure 3 : Fenêtre de chargement



Source : réalisé par les membres du projet

Dès que le chargement est terminé, il sera alors renvoyé vers une interface de sortie ([Figure 4](#)) qui se décompose en plusieurs éléments.

Tout d'abord, un espace est dédié aux éléments qui renvoient à l'empreinte carbone, dont le résultat de l'estimation ainsi qu'un point de comparaison quantifiable humainement pour rendre le résultat compréhensible de tous. Concernant le résultat de l'estimation, ce dernier s'affichera dans une couleur différente selon sa valeur : si l'empreinte carbone moyenne hebdomadaire est faible (inférieure ou égale à 15 000g de CO₂e), le résultat s'affichera en vert ; si elle est moyenne (inférieure ou égale à 30 000g de CO₂e), en orange ; et si elle est élevée (supérieure à 30 000g de CO₂e), en rouge. De plus, pour que l'utilisateur se rend mieux compte de son empreinte carbone moyenne hebdomadaire, nous lui avons mis un équivalent de son impact carbone en termes de quantité de pizzas Margherita fabriquée. Ainsi, en plus, d'aider l'utilisateur à mieux visualiser son impact sur la planète grâce à un exemple parlant, ce point de comparaison a aussi pour but de le faire réagir et de le sensibiliser.

Par ailleurs, le choix de qualifier une empreinte carbone moyenne hebdomadaire comme étant faible, moyenne ou forte selon une certaine valeur est totalement subjectif. Nous avons considéré que cette dernière est faible si elle équivaut à la fabrication de moins de 50 pizzas, moyenne si elle équivaut à la fabrication de 50 à 100 pizzas et forte si elle équivaut à la fabrication de plus de 100 pizzas.

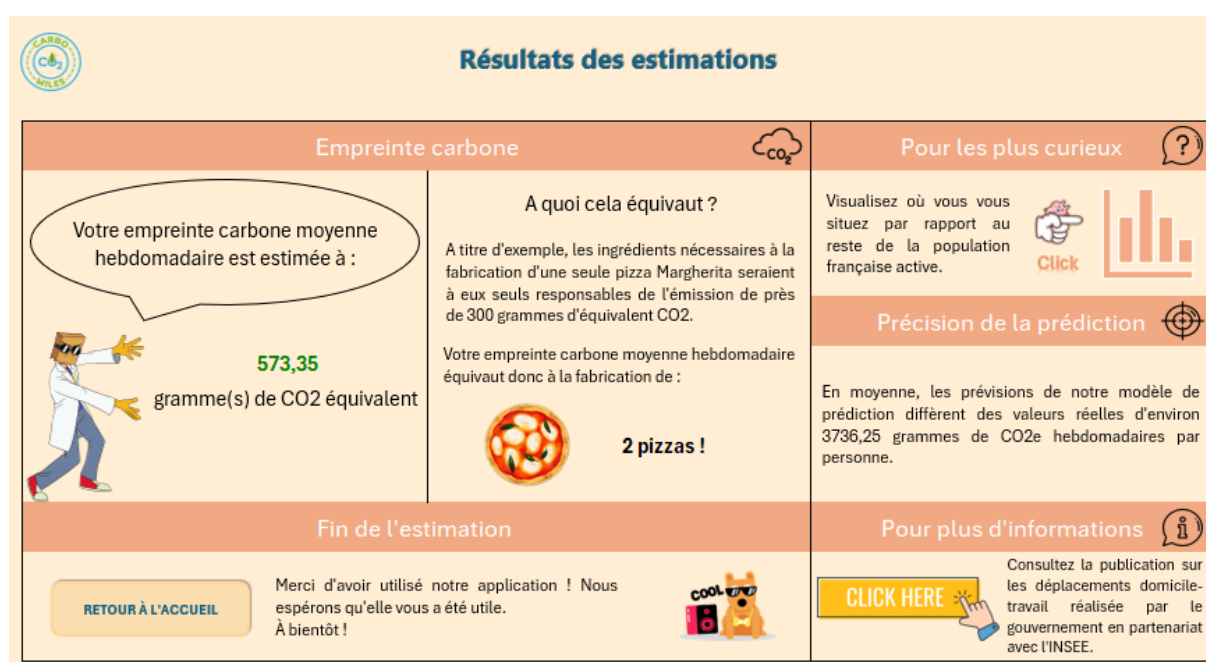
Ensuite, pour les utilisateurs qui souhaitent avoir une vision plus globale de leurs émissions par rapport aux personnes en emploi aux profils similaires ainsi qu'au reste de la population française en activité, il est possible pour eux de consulter des graphiques complémentaires et personnalisés en cliquant sur l'icône correspondant à un graphique. Dès que l'utilisateur clique sur l'icône, la même fenêtre de chargement mentionnée plus haut ([Figure 3](#)) apparaîtra de nouveau, le temps que les graphiques s'affichent à l'écran. Pour améliorer la visibilité et aider l'utilisateur à mieux se repérer, son profil sera affiché dans une couleur distincte sur les graphiques.

Un encart est également réservé à l’affichage de la précision de notre modèle, afin de rassurer l’utilisateur quant à la fiabilité et la précision de l’estimation fournie.

Pour les utilisateurs qui souhaitent se renseigner davantage sur les émissions de GES pour les déplacements domicile-travail, nous leur avons mis à disposition un bouton qui les renvoie directement à un rapport sur le sujet et qui a été publié par le gouvernement en partenariat avec l’INSEE.

Enfin, nous avons créé une zone qui remercie les utilisateurs de notre application, accompagnée d’un bouton « RETOUR À L’ACCUEIL » qui leur permettent de revenir à la page d’accueil de l’application, s’ils le souhaitent.

Figure 4 : Interface de sortie affichant les résultats de l’estimation de l’empreinte carbone de l’utilisateur



Source : réalisé par les membres du projet

II. Les fonctionnalités implémentées dans CARBOMILES

1. Les fonctionnalités générales de l’application

Dans le cadre du développement de notre application CARBOMILES, nous avons ajouté une fonctionnalité pour que l’ouverture de cette dernière soit configurée avec un pourcentage de zoom défini selon la résolution de l’écran. Cette configuration a été notamment appliquée sur la page d’accueil et de résultats pour garantir une expérience utilisateur optimale. Cela évitera ainsi que l’une des deux pages soit rognée à l’ouverture du fichier, surtout pour les écrans

avec une résolution plus basse, obligeant alors l'utilisateur à dézoomer¹. Toutefois, sur certaines résolutions d'écran, les interfaces peuvent sembler trop dézoomées. Nous avons néanmoins privilégié cette approche à celle de pages rognées, car elle offre à l'utilisateur la possibilité d'agrandir la page d'accueil ou de résultats selon ses besoins, tout en garantissant la visibilité de l'intégralité du contenu.

Nous avons également configuré le formulaire de manière à ce que sa taille s'adapte et s'ajuste à toutes les résolutions d'écran possibles. Ainsi, peu importe la résolution de l'écran de l'ordinateur, le formulaire s'ouvrira correctement avec un affichage clair, sans chevauchement ou déformation des éléments ni troncage du contenu.

Pour finir, une feuille a été créée pour permettre l'élaboration des listes déroulantes du formulaire, mais elle est inutile pour l'utilisateur de l'application. C'est pourquoi, nous l'avons masqué de manière à ce qu'elle ne soit jamais visible et accessible pour l'utilisateur à chaque fois que l'application est lancée. La seule manière de l'afficher est d'accéder directement au code VBA. Nous en avons fait de même avec la feuille de résultats qui ne sera visible qu'après validation du formulaire.

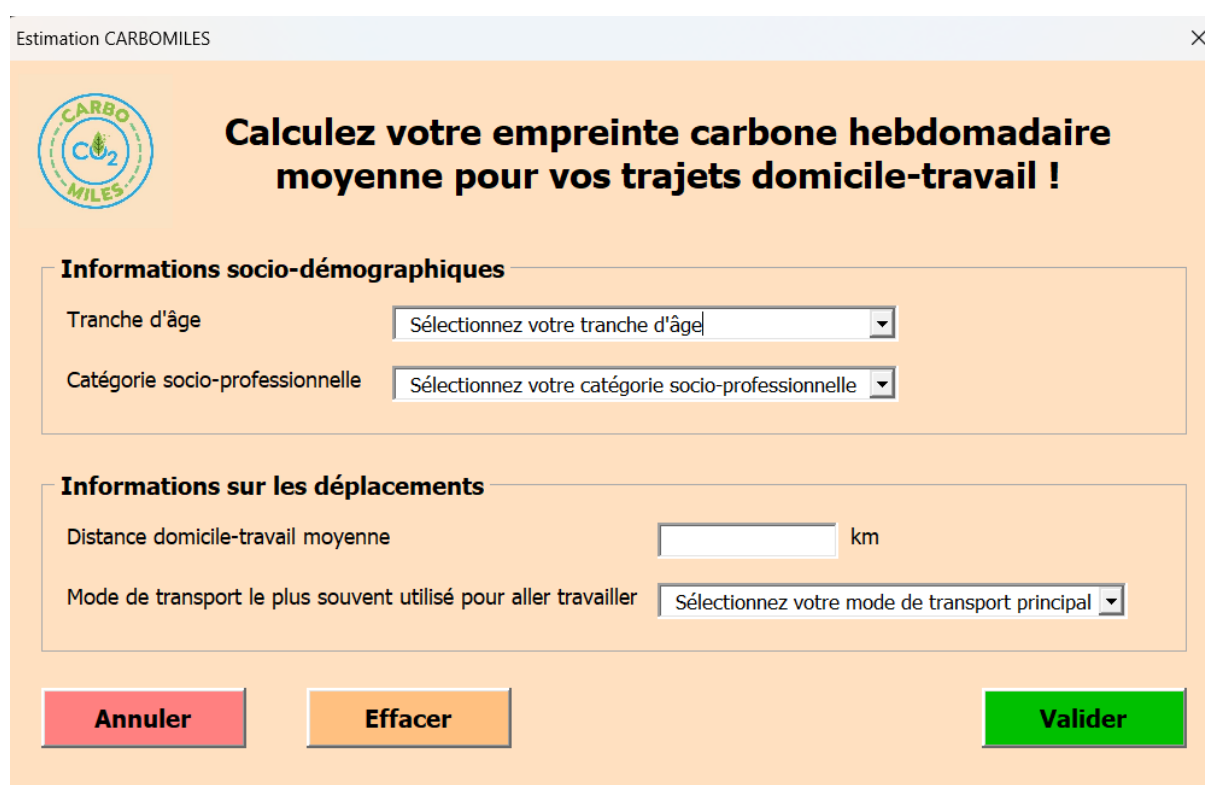
2. Les fonctionnalités du formulaire

Afin de rendre l'utilisation de CARBOMILES plus intuitive et agréable, un ensemble de fonctionnalités a été développé dans le formulaire pour guider au mieux l'utilisateur et sécuriser les informations rentrées en cas de clics ou de saisies involontaires.

Tout d'abord, des messages par défaut ont été insérés dans les champs des différentes listes déroulantes pour indiquer ce que doit faire l'utilisateur (Figure 5). Ce message disparaît dès lors qu'il clique sur l'une des listes déroulantes et réapparaît s'il clique sur un autre champ, dans le cas où il n'aurait rien sélectionné dans cette liste.

¹ Cette fonctionnalité ne fonctionne pas pour les résolutions très basses (1280 x 720, 1024 x 768 et 800 x 600), où le contenu d'au moins une des interfaces est rogné à l'ouverture du fichier.

Figure 5 : Formulaire de saisie contenant des messages par défaut dans les champs



Estimation CARBOMILES

Calculez votre empreinte carbone hebdomadaire moyenne pour vos trajets domicile-travail !

Informations socio-démographiques

Tranche d'âge

Catégorie socio-professionnelle

Informations sur les déplacements

Distance domicile-travail moyenne km

Mode de transport le plus souvent utilisé pour aller travailler

Annuler **Effacer** **Valider**

Source : réalisé par les membres du projet

Des fenêtres d'informations ont également été ajoutées sur chacun des boutons présents dans le formulaire afin d'améliorer l'expérience utilisateur. En effet, elles fournissent des informations contextuelles et renvoient des demandes de confirmation ou des messages d'erreurs, ce qui permet à l'utilisateur de mieux comprendre les actions qu'il effectue et de réduire les erreurs de saisie ou les erreurs dues à des actions involontaires.

Plus précisément, lorsque l'utilisateur appuie sur le bouton « Annuler », une fenêtre s'ouvre afin de confirmer s'il souhaite vraiment annuler le remplissage du formulaire et le fermer. Le bouton « Effacer » renvoie également une demande de confirmation afin de réinitialiser le formulaire si nécessaire, cela effacera alors le contenu de tous les champs. Enfin, quant au bouton « Valider », trois types de messages peuvent s'afficher selon la saisie de l'utilisateur : un premier message peut s'afficher dans le cas où l'utilisateur souhaiterait valider son formulaire, mais qu'au moins un des champs est vide ; un deuxième message peut s'afficher lorsque l'utilisateur saisit une distance non-numérique ou inférieure à 0 ; un troisième message peut s'afficher si la distance saisie est supérieure à 100 km.

Enfin, pour rendre le remplissage plus visuel, nous avons ajouté dans la case de saisie des kilomètres, un fond rouge qui apparaît instantanément lorsque l'utilisateur rentre une valeur non-valide, c'est-à-dire un caractère, un nombre négatif ou un nombre supérieur à 100 (Figure 6). À l'inverse, cette case redevient blanche lorsque la valeur est cette fois-ci correcte. Par ailleurs, en cas de saisie d'un nombre décimal avec pour séparateur décimal une virgule, ce

dernier sera transformé en entier et sera arrondi à l'entier le plus proche afin de réduire les erreurs lors du traitement des données par le modèle de prédiction. En revanche, s'il utilise le point comme séparateur décimal, la case de saisie deviendra rouge car la valeur rentrée ne sera pas considérée comme une valeur numérique.

Figure 6 : Exemple de valeur non-valide dans le champ de la distance

Estimation CARBOMILES

Calculez votre empreinte carbone hebdomadaire moyenne pour vos trajets domicile-travail !

Informations socio-démographiques

Tranche d'âge: 30-34 ans

Catégorie socio-professionnelle: Cadres et professions intellectuelles supérieures

Informations sur les déplacements

Distance domicile-travail moyenne: abc km

Mode de transport le plus souvent utilisé pour aller travailler: Voiture, camion, fourgonnette

Annuler Effacer Valider

Source : réalisé par les membres du projet

3. Construction du pipeline VBA-Python

Nous avons conçu une solution permettant de lier VBA et Python pour automatiser le traitement et la visualisation des données dans Excel. Ce système repose sur une interaction fluide entre ces deux technologies, offrant à l'utilisateur une expérience fluide et intuitive.

Une fois que l'utilisateur a fini de remplir le formulaire et clique sur le bouton « Valider », l'application s'assure que les champs sont correctement remplis et valides. Si les données sont effectivement valides, elles sont ensuite enregistrées dans deux fichiers CSV temporaires : le fichier `donnees_utilisateur` qui stocke les informations relatives à la tranche d'âge et à la CSP et visant à générer les graphiques personnalisés de l'utilisateur, et le fichier `donnees_prediction` qui enregistre la distance domicile-travail moyenne et le mode de transport.

Afin d'afficher le nombre correspondant à la prédiction des émissions de CO2 hebdomadaires moyennes de l'utilisateur, le fichier `donnees_prediction` est utilisé dans Python pour prédire

l'empreinte carbone de l'utilisateur. Le résultat de la prédiction, exprimé en grammes de CO2 équivalent, est ensuite renvoyé à l'interface de sortie. Comme mentionné plus haut dans le rapport, une mise en forme conditionnelle a été appliquée au résultat, permettant de changer la couleur du texte en fonction de la valeur de la prédiction, ce qui garantit ainsi une meilleure visualisation. En effet, cela permet à l'utilisateur d'obtenir une visualisation immédiate et claire de l'impact environnemental de ses choix de transport.

Parallèlement, l'icône représentant un graphique situé dans l'interface de résultats permet à l'utilisateur de générer les graphiques personnalisés à partir des données contenues dans le fichier `donnees_utilisateur`. Si l'utilisateur souhaite visualiser ces graphiques et qu'il clique sur cet icône, les données contenues dans ce fichier, à savoir la tranche d'âge et la CSP sont récupérées et renvoyées au script Python correspondant pour générer les graphiques. Le premier graphique illustre les émissions moyennes de CO2 et les distances domicile-travail moyennes, réparties par catégorie socio-professionnelle, tandis que le second montre ces mêmes indicateurs répartis par tranches d'âge. Ces graphiques, créés par Python et affichés dans une fenêtre distincte, permettent à l'utilisateur de visualiser ses données en comparaison avec les moyennes globales.

Grâce à cette solution, l'utilisateur bénéficie d'une expérience simple et efficace : il saisit ses informations, obtient une prédiction de son empreinte carbone moyenne hebdomadaire, et peut visualiser les résultats et les graphiques en quelques clics seulement.

III. Préparation et analyse des données

1. Présentation de la base de données

Dans le cadre de notre projet de création d'une application pour prédire les émissions de GES moyennes hebdomadaires lors des trajets domicile-travail, nous avons utilisé une base de données issue de la plateforme publique data.gouv.fr. Cette base, fruit d'une collaboration entre l'INSEE et le Service des Données et des Études Statistiques (SDES) du Ministère de la Transition Écologique, fournit une estimation des émissions de GES pour les déplacements domicile-travail de moins de 100 km en France métropolitaine.

La base de données est construite à partir de plusieurs sources, incluant les données du recensement de la population, enrichies par des informations complémentaires telles que les distances calculées par le distancier Metric-OSRM et les fréquences de déplacement issues de l'Enquête Mobilité des Personnes (EMP). Nous avons plus de 6 millions d'observations et 46 variables.

Les variables qui composent cette base sont classées en trois grandes catégories, qui permettent de couvrir différents aspects des déplacements et de leur impact environnemental.

Tout d’abord, nous retrouvons les variables du recensement de la population. Cette première catégorie regroupe les informations issues du recensement de la population. Elle fournit des données sur les caractéristiques individuelles et professionnelles des actifs, ainsi que sur leur situation géographique. Ces variables permettent de définir les profils sociodémographiques des utilisateurs et de relier les émissions de CO2 aux spécificités de chaque région et lieu de travail.

Nous distinguons ensuite les indicateurs de déplacement et les estimations des émissions. Cette seconde catégorie comprend des indicateurs relatifs aux déplacements des actifs et à leurs impacts environnementaux. Ces variables incluent des données telles que les distances parcourues, la durée des trajets, ainsi que les émissions hebdomadaires de CO2 estimées pour chaque actif en fonction de ses déplacements. Ces informations sont calculées ou imputées pour donner une vision détaillée des émissions générées spécifiquement par les trajets domicile-travail.

Enfin, la base de données nous fournit des informations géographiques complémentaires. Cette dernière catégorie rassemble des éléments géographiques additionnels permettant de contextualiser les déplacements. Ces données fournissent des informations sur les aires urbaines, les infrastructures de transport et les connexions intercommunales, contribuant ainsi à une meilleure compréhension de l’accessibilité et de la mobilité dans chaque zone. Elles enrichissent les analyses en tenant compte de la géographie locale et des dynamiques régionales.

Ces trois catégories de variables offrent une vue d’ensemble des comportements de mobilité domicile-travail, facilitant les analyses et les modélisations précises dans notre application pour la prédiction des émissions de CO2.

Parmi ces variables, notre variable cible est la variable `CO2_HEBDO`, qui est une variable de type numérique représentant les émissions de CO2 moyennes par personne et par semaine pour les déplacements domicile-travail (CO2e en g). Cette variable sera donc au cœur de notre analyse et notre but sera alors de prédire la valeur de cette variable en fonction des valeurs des autres variables fournies par l’utilisateur de l’application.

Il est par ailleurs important de noter que dans cette variable cible, tous les GES sont inclus et les émissions sont exprimées en équivalent CO2. Toutefois, les émissions de GES ont été simplifiées dans la documentation par « émissions de CO2 » par convention ou facilité, surtout dans des contextes où le CO2 représente la majorité des émissions étudiées. C’est pourquoi,

nous emploierons également, dans la suite de notre rapport, le terme « émissions de CO2 » pour désigner en réalité les émissions de GES par souci de simplicité.

Le champ des données comprend les actifs en emploi résidant en France métropolitaine et qui effectuent des trajets domicile-travail inférieurs à 100 km pour les modes motorisés, à 30 km pour le vélo, et à 10 km pour les trajets à pied. Les individus qui ne se déplacent pas pour aller travailler ou ceux dont la distance domicile-travail dépasse les seuils prédéfinis sont exclus du calcul des émissions. Elles incluent les émissions directes des trajets (du « réservoir à la roue »), sans comptabiliser les émissions amont (comme celles liées à la production d'énergie ou à la construction des véhicules). Ces données sont pondérées par la variable **IPONDI**, qui représente le nombre d'actifs correspondant à chaque profil sociodémographique, afin de garantir leur représentativité.

Le détail complet sur les variables de la base se trouve en Annexe 2 de ce rapport, où chaque élément est expliqué pour faciliter leur intégration dans notre modèle prédictif et garantir une interprétation correcte des données.

2. Nettoyage de la base de données

Le processus de nettoyage des données que nous avons appliqué repose sur plusieurs étapes complémentaires afin de transformer la base de données brute en un ensemble de données utilisable et cohérent.

Tout d'abord, nous avons procédé à une uniformisation des décimales. Plus précisément, les valeurs numériques dans les colonnes **DIST_HEBDO**, **CARBU_HEBDO**, **IPONDI**, **DIST**, **DUREE** et **CO2_HEBDO** utilisaient des virgules comme séparateurs décimaux, ce qui rendait leur traitement difficile. Pour remédier à cela, nous avons remplacé ces virgules par des points et converti les valeurs en format numérique. Cette transformation nous a permis d'assurer une cohérence dans les données et de faciliter les opérations de calcul.

Nous avons ensuite procédé à la création d'une nouvelle colonne binaire nommée « **va_abs** » pour identifier les lignes comportant des valeurs manquantes dans les colonnes de la distance, la durée et les émissions de CO2 hebdomadaires. Cet indicateur prend la valeur 1 lorsqu'une de ces valeurs est effectivement manquante, et 0 si elles sont toutes présentes. Ainsi, cet ajout a facilité la localisation des lignes incomplètes dans notre base de données.

Une fois cet indicateur de valeurs manquantes créé, nous avons identifié le nombre de lignes avec et sans valeurs manquantes, puis calculé la proportion de valeurs manquantes. Le nombre de lignes comportant des valeurs manquantes (**va_abs = 1**) s'élève à 605 691, soit 8,85% de l'ensemble des données. En revanche, le nombre de lignes sans valeurs manquantes (**va_abs = 0**) est de 6 241 018, représentant ainsi 91,15 % du total. Ces pourcentages

fournissent alors une indication sur la qualité globale de nos données : la majorité des lignes (91,15 %) sont complètes contre seulement moins de 9% de lignes avec des valeurs manquantes. Dans la continuité de ce travail, nous avons appliqué un filtre ne conservant que les observations sans valeurs manquantes afin d'obtenir un jeu de données pertinent pour l'analyse.

Par ailleurs, un détail a suscité notre curiosité : nous avons identifié des individus de bien plus de 70 ans travaillant encore et émettant donc des émissions de CO₂. Mais dans le cadre de notre projet, afin de ne pas fausser nos analyses et nos prédictions par ces individus, que nous considérons comme « atypiques », nous avons décidé de les supprimer car ce type de profil relève de l'exception. En effet, ils ne représentent qu'environ 0,5% de notre jeu de données, ce qui est très peu représentatif et confirme notre idée. Ce filtrage nous garantit ainsi que les données finales sont complètes, exploitables, et prêtes pour l'analyse.

Pour clore ce processus de nettoyage de données, nous avons supprimé les colonnes `va_abs` (utilisée uniquement pour identifier les valeurs manquantes) et `CARBU_HEBDO`, présentant plus de 40% des valeurs manquantes. Cette étape finale nous a permis de simplifier la structure du jeu de données et d'éliminer les éléments superflus.

Ainsi, à l'issue de cette étape de nettoyage, nous obtenons une base de données propre, composée uniquement de lignes complètes et pertinentes pour l'analyse, avec des valeurs prêtes pour les visualisations et les modèles de prédiction.

3. Exploration des données

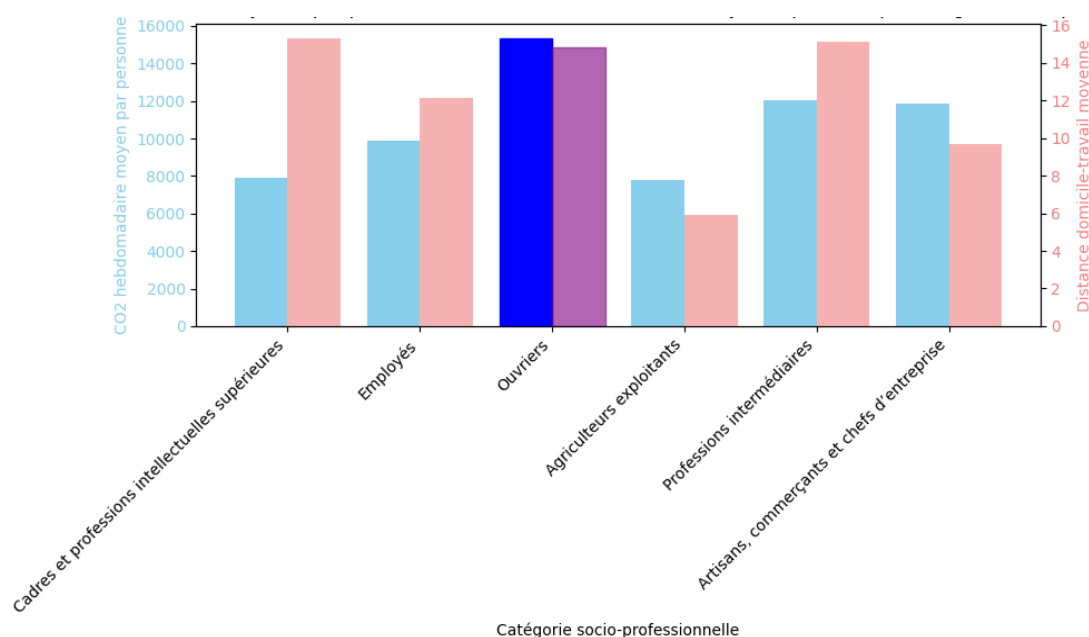
Une fois que les données ont été nettoyées, nous avons réalisé une première analyse des statistiques descriptives des variables pour observer s'il y avait des données aberrantes, telles que des modalités prises par les variables qualitatives qui ne seraient pas dans notre dictionnaire de variables. Nous avons alors observé que ce n'était pas le cas et que nos données semblent être prêtes à être utilisées et analysées.

De plus, nous avons également réalisé des statistiques descriptives sur notre variable cible, qui représente pour rappel les émissions de CO₂ moyennes par personne et par semaine pour les déplacements domicile-travail (CO₂e en g) ([Annexe 3](#)). Nous avons alors constaté que les émissions de CO₂ hebdomadaires moyennes sont autour de 11 269g de CO₂e par personne pour plus de 6 millions d'individus. Ces valeurs varient cependant largement : certaines personnes n'émettent presque pas, tandis que d'autres peuvent atteindre des niveaux d'émissions très élevés, au-delà de 160 000g de CO₂e par semaine. Ces écarts importants peuvent notamment refléter les différences de distances et de modes de transport utilisés par les individus pour leurs trajets quotidiens. Ces statistiques descriptives nous donnent ainsi un

aperçu de l'ordre de grandeur des valeurs prises par cette variable, ce qui nous aide à anticiper les niveaux d'émissions de CO2 que nos futures prédictions pourraient atteindre.

Dans cette phase d'exploration des données, nous avons également réalisé un ensemble de graphiques pour mieux comprendre comment sont réparties les émissions de CO2 hebdomadaires moyennes en fonction des autres variables comme la catégorie socio-professionnelle (CSP) ([Figure 7](#)) des personnes ou leur tranche d'âge ([Figure 8](#)). De plus, ils nous permettent de détecter rapidement de potentielles anomalies ([Figure 9](#)), telles que le problème des personnes âgées dans la base, cité plus haut dans le rapport, que nous avons supprimé.

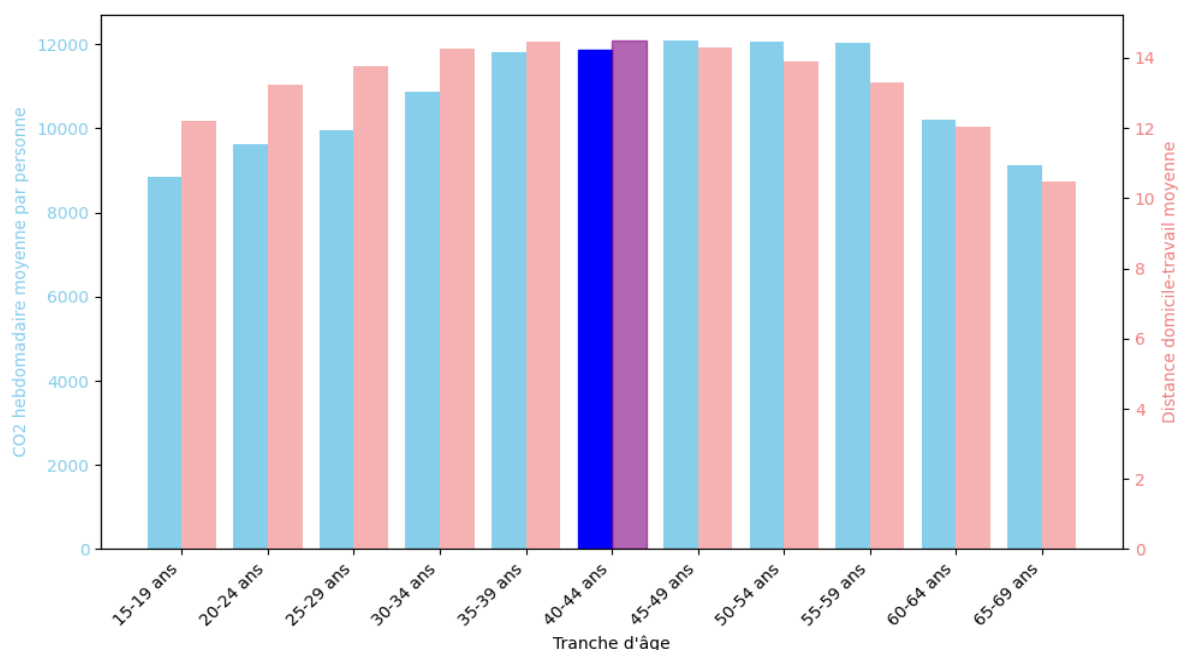
***Figure 7 :** Emissions de CO2 hebdomadaires moyennes par personne et distance domicile-travail moyenne pour chaque catégorie socio-professionnelle*



***Source :** réalisé par les membres du projet*

***Grille de lecture :** les ouvriers émettent en moyenne environ 15 000g de CO2e par personne et par semaine, pour une distance domicile-travail moyenne d'environ 15 km.*

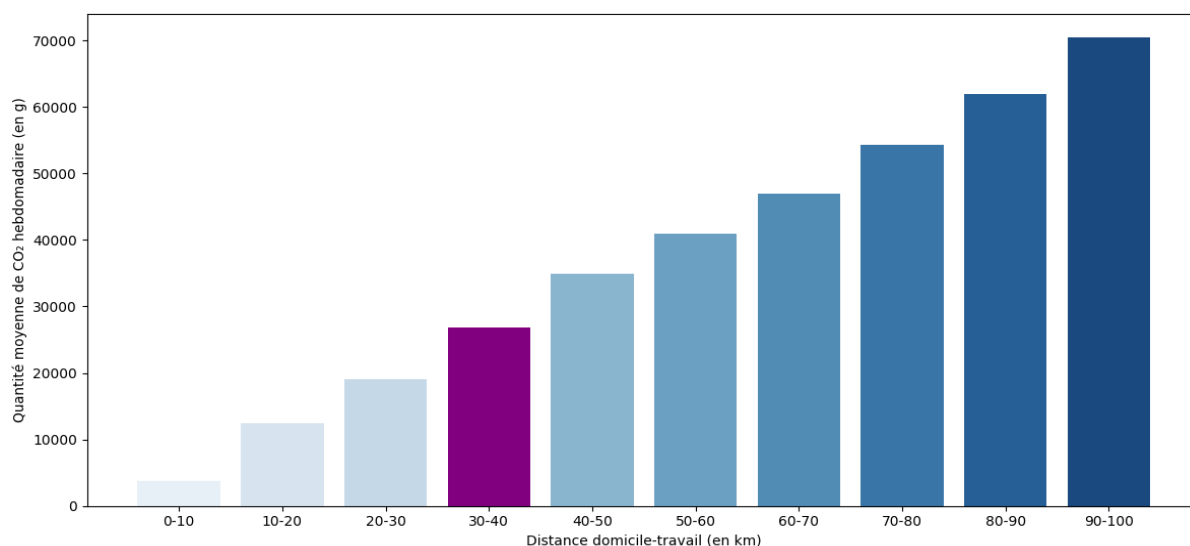
Figure 8 : Emissions de CO2 hebdomadaires moyennes par personne et distance domicile-travail moyenne par tranches d'âge



Source : réalisé par les membres du projet

Grille de lecture : Les personnes ayant entre 40 et 44 ans émettent en moyenne environ 12 000g de CO2e par personne et par semaine, pour une distance domicile-travail moyenne d'environ 14 km.

Figure 9 : Quantité moyenne de CO2 hebdomadaire par personne selon la distance domicile-travail



Source : réalisé par les membres du projet

Grille de lecture : Les personnes se situant entre 30 et 40 km de leur lieu de travail émettent en moyenne environ 29 000g de CO2e par personne et par semaine.

Enfin, des fonctions ont été développées pour personnaliser les graphiques pour chaque utilisateur (représenté par une couleur différente sur les graphiques) selon certains critères tels que la CSP, la tranche d'âge ou encore la distance domicile-travail. Par le biais d'une liaison entre l'interface utilisateur réalisée sur Excel/VBA et le code Python, ces graphiques seront

par la suite affichés dans une interface de sortie dédiée, accompagnés du résultat de la prédiction des émissions de CO2 hebdomadaires moyennes. Cela permettra alors à l'utilisateur de comparer non seulement ses résultats avec des personnes en emploi aux profils similaires (même CSP, même tranche d'âge), mais également avec le reste de la population française.

IV. Elaboration d'un modèle de prédiction

1. Sélection des variables

Avant de chercher un modèle de prédiction, nous avons tenté de réduire, dans un premier temps, notre base de données pour ne conserver que les variables significatives pour la prédiction de l'empreinte carbone moyenne hebdomadaire, et nous avons ensuite supprimé les variables fortement corrélées entre elles. Une première sélection a donc été réalisée en écartant d'abord la variable `CHAMP_CO2`, car elle affichait toujours la même valeur (True) parmi les données restantes. Les variables catégorielles contenant plus de 15 modalités ont également été exclues de cette analyse, car leur traitement nécessite trop d'espace mémoire pour que nos machines puissent les gérer correctement. Ces variables ne sont cependant pas forcément toutes écartées de notre futur modèle prédictif, mais leur analyse devra se faire de manière différente. Les variables concernées sont les suivantes : `LIEU_RESID`, `LIEU_TRAV`, `AGEREVQ`, `DEP_RESID`, `EPCI_EPT_RESID`, `ZE2020_RESID`, `AAV20_RESID`, `DEP_TRAV`, `EPCI_EPT_TRAV`, `ZE2020_TRAV` et `AAV20_TRAV`.

Une fois ce tri effectué, nous avons commencé par analyser les corrélations entre nos variables quantitatives explicatives ([Annexe 4](#)), ainsi que les corrélations entre ces mêmes variables avec notre variable cible `CO2_HEBDO` ([Annexe 5](#)).

Concernant les corrélations entre les variables explicatives, nous observons de très fortes corrélations entre nos variables `DIST`, `DIST_HEBDO` et `DUREE`, ainsi qu'une corrélation quasiment nulle avec la variable `IPONDI` ([Annexe 4](#)). Le second graphique, quant à lui, montre également une forte corrélation entre notre variable cible `CO2_HEBDO` et les variables `DUREE`, `DIST` et `DIST_HEBDO`, et une corrélation proche de 0 avec `IPONDI` ([Annexe 5](#)). Étant donné que les variables `DUREE`, `DIST` et `DIST_HEBDO` sont fortement corrélées entre elles et avec `CO2_HEBDO`, les conserver toutes les trois introduirait un problème de multicolinéarité dans notre modèle. Nous ne garderons donc que la variable `DIST`, qui indique la distance moyenne entre le lieu de travail et le domicile, car nous jugeons qu'il sera plus facile pour un utilisateur de renseigner cette information. Quant à la variable `IPONDI`, cette dernière ne sera pas intégrée dans notre modèle prédictif puisqu'elle n'apporte que très peu d'informations sur notre variable cible et se réfère au ménage de l'individu, et non à l'individu lui-même.

Nous avons ensuite effectué les mêmes analyses pour les variables catégorielles ([Annexe 6](#)). Nous avons pu en déduire qu'aucune des variables catégorielles ne semble être corrélée entre elles. En examinant les corrélations entre nos variables catégorielles et notre variable cible CO2_HEBDO, seules trois variables semblent avoir des modalités corrélées avec notre variable cible. Ces variables sont MODTRANS, ILT et ILTUU.

Dans la poursuite de notre travail de sélection de variables, nous avons analysé la variance de variables par rapport à notre variable cible CO2_HEBDO. Pour cela, nous avons réalisé un test ANOVA sur nos variables catégorielles avec une p-valeur fixée à 0,01. L'idée était de vérifier si certaines variables catégorielles conservaient une moyenne similaire pour toutes leurs modalités, afin de pouvoir les écarter du modèle. Le test s'est révélé significatif pour l'ensemble de nos variables, ce qui ne nous permet donc pas d'en écarter certaines.

Enfin, nous avons également voulu vérifier si certaines variables catégorielles étaient associées entre elles. En effet, l'association de variables posant également un problème de multicolinéarité, l'objectif était donc d'identifier les variables potentiellement associées et les écarter. Nous avons pour cela calculé tous les tableaux de contingence pour chaque paire de variables catégorielles possibles, puis réalisé un test du Chi deux sur chacun d'eux avec une p-valeur fixée à 0,05. Le test s'est révélé significatif à chaque fois, ce qui implique que toutes nos variables catégorielles sont associées entre elles.

Sur la base de nos analyses, nous en avons déduit qu'il semble difficile d'utiliser un modèle de régression simple pour estimer notre variable cible, en raison de la forte multicolinéarité présente dans nos données. Il nous faudra donc choisir un modèle capable de supporter cette multicolinéarité. Parmi les modèles disponibles, nous avons la régression de type Elastic Net, ainsi que des modèles de type forêt ou boosting.

2. Méthodes de prédiction testées et sélection du modèle

Parmi les modèles de prédiction élaborés dans le cadre de notre projet, nous en avons élaboré plusieurs afin de sélectionner celui qui s'avère le plus fiable et précis ([Annexe 7](#)). Pour cela, nous avons testé dans un premier temps le modèle de type Lasso. L'avantage de ce modèle est qu'il va tester différents modèles avec différentes variables et sélectionner celui qui minimise l'erreur. Pour cette méthode, nous avons écarté les variables DIST_HEBDO et DUREE, car elles sont fortement corrélées avec DIST, et de plus, demander la distance entre le lieu de résidence et de travail à l'utilisateur semble plus facile que de lui demander la durée de son trajet ou la distance parcourue chaque semaine. Nous avons également écarté les variables qui n'influencent pas directement les émissions de CO2, comme par exemple le sexe ou le niveau de diplôme. Les variables écartées sont : IPONDI, CS1, DIPL, EMPL, IMMI, INAT, INEEM, INPOM, MOCO, NA5, NPERR, SEXE, STAT, STOCD, TP, TYPL, TYPMR, VOIT, TAAV2017_RESID, TAAV2017_TRAV, ILT, CATEAAV20_RESID, CATEAAV20_TRAV, REG_RESID, REG_TRAV.

Finalement, les seules variables conservées sont la durée du trajet domicile-travail sans congestion (DUREE), le mode de transport principal le plus souvent utilisé pour aller travailler (MODTRANS), l'indicateur urbain du lieu de travail (ILTUU), la grille de densité du lieu de résidence (GRILLE_DENSITE_RESID) et la grille de densité du lieu de travail (GRILLE_DENSITE_TRAV).

Le modèle créé a une erreur moyenne de 5 965,88g de CO₂e hebdomadaires, ce qui est relativement élevée. Etant donné que ce modèle de prédiction possède une erreur non négligeable, nous avons testé un autre modèle afin d'obtenir une erreur plus faible. Nous avons donc choisi de construire comme second modèle de prédiction un arbre de régression. Comme pour la régression Lasso, l'avantage de ce modèle est qu'il reste robuste même lorsque plusieurs variables sont corrélées. Pour ce modèle, nous avons écarté les mêmes variables que pour le modèle Lasso, mais nous avons également écarté les variables ILTUU, GRILLE_DENSITE_RESID et GRILLE_DENSITE_TRAV. La raison de cette exclusion est que ces variables n'avaient que trop peu d'impact sur la prédiction, l'arbre ne prenant en compte que le mode de transport et la distance parcourue pour les trajets domicile-travail.

L'erreur moyenne obtenue pour ce modèle est d'environ 3 736,25g de CO₂e hebdomadaires, ce qui est beaucoup plus faible que l'erreur du modèle Lasso. Nous avons donc décidé de conserver ce modèle pour notre application.

3. Les limites du modèle sélectionné

Bien que nous sommes parvenus à améliorer le modèle de prédiction, plusieurs limites doivent être soulignées pour bien comprendre ses implications.

Tout d'abord, sur les 46 variables initialement disponibles, seules deux (la distance domicile-travail moyenne et le mode de transport utilisé) se sont révélées significatives pour expliquer les variations des émissions. Cela simplifie le modèle, mais limite sa capacité à capturer des facteurs plus complexes ou contextuels, comme les conditions de circulation, l'efficacité énergétique des véhicules, ou les habitudes de déplacement des usagers. Il est également important de souligner que notre base de données ne prend en compte que les trajets domicile-travail inférieurs à 100 km. Ainsi, notre application n'est pas en mesure d'estimer une empreinte carbone moyenne hebdomadaire pour les déplacements dont la distance domicile-travail moyenne est supérieure à 100 km, puisque le modèle n'a pas pu être entraîné sur de telles données.

De plus, les données utilisées, bien qu'importantes en volume (plus de 6 millions d'observations), pourraient contenir des biais structurels tels qu'une sous-représentation de certains modes de transport comme le covoiturage ou les trajets combinant plusieurs moyens. Ces biais pourraient affecter la généralisation du modèle à d'autres populations ou contextes.

Nous pouvons également ajouter qu'une erreur moyenne de 3 736,25g de CO₂e pourrait être perçue comme significative dans certaines plages d'émissions, notamment pour des trajets courts ou utilisant des modes de transport faiblement émetteurs. Ce niveau de précision pourrait ne pas suffire si le modèle est utilisé pour des recommandations ou des politiques nécessitant des estimations très précises.

Conclusion

Pour conclure, nous sommes parvenus à développer une application fonctionnelle permettant de prédire les émissions de GES moyennes hebdomadaires liées aux déplacements domicile-travail. Le modèle de prédiction élaboré présente des performances satisfaisantes, avec une erreur moyenne de 3 736,25g de CO₂e, marquant une nette amélioration par rapport au modèle de type Lasso (5 965,88g de CO₂e).

Toutefois, un certain nombre de limites en lien avec le modèle de prédiction sélectionné ont été identifiées dans le cadre de notre projet, telles que la présence de potentiels biais ou alors le faible nombre de variables explicatives significatifs. Malgré ces limites, le modèle constitue un outil prometteur pour fournir une première estimation des émissions liées aux trajets domicile-travail, utile dans des analyses globales ou des scénarios stratégiques. L'objectif initial reste atteint dans la mesure où nous avons réussi à construire une application fonctionnelle et intuitive, capable de fournir à l'utilisateur une estimation moyenne hebdomadaire de son empreinte carbone dans le cadre de ses déplacements domicile-travail.

Pour aller plus loin

Afin de perfectionner notre application, nous avons identifié plusieurs axes d'amélioration. Tout d'abord, il serait bénéfique d'intégrer des variables supplémentaires provenant d'autres bases de données, telles que les conditions de circulation (trafic, météo, etc.), les comportements individuels ou les habitudes de déplacement (trajets réguliers, covoiturage, etc.). Cela afin de compléter notre base de données actuelle et renforcer la pertinence de notre modèle de prédiction.

Il est également envisageable d'améliorer la représentativité des données. Pour ce faire, il faudrait collecter des données plus équilibrées qui incluent des modes de transport sous-représentés, comme le covoiturage ou encore les trajets multimodaux. Enfin, nous pouvons aussi envisager d'étendre la base de données pour inclure des trajets dépassant 100 km, afin d'améliorer la généralisation du modèle pour les déplacements longue distance.

Bibliographie

Estimation des émissions individuelles de gaz à effet de serre lors des déplacements domicile-travail - data.gouv.fr [en ligne]. [s. d.]. [Consulté le 14 octobre 2024]. Disponible à l'adresse : <https://www.data.gouv.fr/fr/datasets/estimation-des-emissions-individuelles-de-gaz-a-effet-de-serre-lors-des-deplacements-domicile-travail/>

The raw materials in a pizza / alimentarium [en ligne]. [s. d.]. [Consulté le 13 décembre 2024]. Disponible à l'adresse : <https://www.alimentarium.org/fr/learn-play/academy/kid/ecology-and-food-economy/advanced/cycle-of-manufactured-goods/460/1>

Annexes

Annexe 1 : Roadmap

Dans cette annexe se trouve la roadmap détaillant les différentes fonctionnalités prévues et introduites dans notre projet, avec une prise en compte des retours utilisateurs permettant d'adapter la roadmap.

ROADMAP					
	09/10 au 23/10	23/10 au 08/10	08/11 au 21/11	21/11 au 09/12	09/12 au 19/12
Algorithme de prédiction	Construire la roadmap produit découpant les étapes en différents lots de valeur Valider la prise en main du setup de développeur	Etablir un premier modèle de prédiction	Tester des modèles de régression : élastique/Ridge/Lasso, RandomForest et XGBoost	Etablir le modèle de prédiction final	Documenter les limites du modèle
Data préparation	Construire un premier jeu de données permettant de traiter la problématique	Poursuite de la préparation des données Documenter les variables	Sélectionner les variables d'intérêt Etablir une liaison VBA/Python pour afficher les graphiques S'assurer du bon fonctionnement du code	Etablir une liaison VBA/Python pour afficher les résultats de la prédiction et les graphiques sur une même page S'assurer du bon fonctionnement du code et de sa qualité	Effectuer des contrôles qualité pour valider l'intégrité du pipeline Organiser les fichiers et les documenter pour une utilisation future
Interface Excel/VBA	Concevoir un produit analytics en réponse à un besoin client identifié	Poursuivre la réalisation de la maquette d'interface VBA	Faire le formulaire complet et améliorer l'interface de sortie	Finaliser l'interface de sortie afin d'améliorer l'expérience utilisateur et la rendre ergonomique	Finaliser l'interface de sortie et réviser le design pour la rendre plus intuitive (amélioration des couleurs, disposition des boutons, etc.)

Source : réalisé par les membres du projet

Dans le cadre de ce projet, quatre sprints reviews ont été organisés toutes les deux semaines afin de faire part de nos avancées auprès des utilisateurs.

❖ Sprint 1

Lors de la première quinzaine, nous avons tenté d'identifier un besoin client afin de construire un produit permettant d'y répondre. Suite à cette identification du besoin, un jeu de données a été construit et nettoyé de manière à le rendre exploitable pour la suite de notre projet. De plus, des analyses exploratoires et des graphiques ont été réalisés à partir de notre base de données dans le but d'identifier d'éventuelles tendances significatives susceptibles d'être mises en évidence. Du côté de l'interface utilisateur, une première interface d'accueil a été construite avec un bouton permettant d'ouvrir un formulaire. Ce dernier comporte différentes rubriques, quelques questions avec des listes déroulantes opérationnelles ainsi que des boutons fonctionnels permettant de réinitialiser les champs, de quitter le formulaire ou de valider celui-ci.

❖ Sprint 2

Au cours du deuxième sprint, nous avons documenté les variables et poursuivi le nettoyage de notre jeu de données en reformatant les variables sous un type approprié. Nous avons également tenté de supprimer les variables non pertinentes ou redondantes et de sélectionner uniquement les variables significatives dans le cadre de notre application, en procédant notamment à la réduction de dimensions. En ce qui concerne l'interface utilisateur, nous avons implémenté une fonctionnalité visant à enregistrer les informations saisies par l'utilisateur dans le formulaire après sa validation. Ces données avaient notamment pour but d'être renvoyées au modèle de prédiction. De plus, une première interface de sortie a été élaborée dans laquelle les résultats de la prédiction et les graphiques personnalisés sont censés y figurer. Un bouton opérationnel a également été ajouté sur cette interface permettant simplement de retourner à la page d'accueil si besoin. Néanmoins, nous avons dû faire face à plusieurs obstacles au cours de ce sprint. En effet, un premier modèle de prédiction était prévu, mais en raison des difficultés rencontrées dans la manipulation de la base de données, la sélection des variables pertinentes et l'absence d'un membre de l'équipe, nous ne sommes pas parvenus à atteindre cet objectif. Il en va de même pour l'objectif que nous nous sommes fixé concernant la construction d'une première liaison VBA-Python que nous n'avons pas réussi pour des raisons techniques cette fois-ci.

❖ Sprint 3

Pour faire suite à ce deuxième sprint, nous avons d'une part testé lors de cette nouvelle quinzaine plusieurs modèles de prédiction (régression élastique, random forest, XGBoost) de manière à obtenir un premier résultat de prédiction. Dans l'élaboration de ce modèle de prédiction, cinq variables ont été sélectionnées, mais les résultats se sont montrés peu concluants. D'autre part, un formulaire complet est proposé à l'utilisateur selon les variables sélectionnées. Initialement, il était prévu qu'après validation de ce formulaire, l'utilisateur est renvoyé sur l'interface de sortie affichant des graphiques contenant des informations sur ses émissions de CO2 par rapport à des individus partageant des caractéristiques similaires et le reste de la population française. Toutefois, nous n'avons pas réussi à afficher ces graphiques sur cette interface, mais nous avons tout de même réussi à faire une première liaison VBA-Python dans la mesure où nous avons réussi à les faire apparaître dans l'application via un bouton « Graphiques » présent dans le formulaire.

❖ Sprint 4

Lors de ce sprint 4, nous avons pour objectif de poursuivre la création de modèles de prédiction et de ne sélectionner que celui avec la prédiction la plus précise. Une fois choisi, nous sommes parvenus à afficher les résultats de la prédiction dans l'interface de sortie, à laquelle l'utilisateur est renvoyé, grâce à une liaison nouvellement créée entre les données

saisies par ce dernier et le modèle de prédiction. La liaison déjà mise en place au sprint dernier pour afficher les graphiques personnalisés selon les données de l'utilisateur a été revue pour les faire apparaître cette fois-ci dans la même interface que les résultats de la prédiction, à savoir l'interface de sortie. Par ailleurs, nous nous sommes également assurés que l'application s'adapte à tout type de résolution d'écran lorsqu'elle est lancée pour garantir une expérience utilisateur optimale.

❖ **Sprint 5**

Enfin, dans le cadre de notre dernier sprint, il s'agissait essentiellement de peaufiner notre application. Pour cela, nous avons poursuivi l'amélioration visuelle de l'interface de sortie en insérant une fenêtre de chargement pour une meilleure expérience utilisateur. Nous avons également ajouté un élément de comparaison avec l'empreinte carbone moyenne hebdomadaire. Cet ajout vise à rendre le résultat plus significatif pour l'utilisateur, en lui permettant de mieux comprendre son impact. Nous avons notamment envisagé de comparer son empreinte carbone hebdomadaire moyenne à l'équivalent en fabrication de pizzas Margherita. De plus, nous avons pour but d'intégrer un chiffre indiquant la précision du modèle dans l'interface de sortie, afin de rassurer l'utilisateur quant à la qualité et à la fiabilité de l'estimation. Cependant, en raison de l'absence de deux membres de l'équipe lors de ce dernier sprint et de la charge de travail élevée en parallèle, nous n'avons pas pu l'ajouter, mais il sera bien inclus dans la présentation finale du produit.

Annexe 2 : Dictionnaire des variables

La base de données sur laquelle nous avons travaillé dans le cadre de notre projet est composée de 46 variables. Pour une meilleure compréhension de cette base de données, nous avons élaboré une description de chacune des variables.

Le dictionnaire des variables conçu ci-dessous s'est basé sur le travail de l'INSEE disponible à l'adresse suivante : <https://www.data.gouv.fr/fr/datasets/estimation-des-emissions-individuelles-de-gaz-a-effet-de-serre-lors-des-deplacements-domicile-travail/>

❖ **La variable cible**

CO2_HEBDO : Émissions de CO2 moyennes par personne et par semaine pour les déplacements domicile-travail (CO2e en g)

Type : numérique

❖ Les variables explicatives

➤ Variables du recensement de la population

IPONDI : Nombre d'actifs en emploi

Type : numérique

LIEU_RESID : Lieu de résidence

Commune ou arrondissement municipale dans le code officiel géographique 2021

Type : caractère

Code Insee sur 5 positions

LIEU_TRAV : Lieu de travail

Commune ou arrondissement municipal en France dans le code officiel géographique 2021, commune frontalière à l'étranger

Type : caractère

Code Insee sur 5 positions

AGEREVQ : Âge quinquennal en années révolues

Type : caractère

Valeur dans la base	Libellé
15	15 à 19 ans
20	20 à 24 ans
25	25 à 29 ans
30	30 à 34 ans
35	35 à 39 ans
40	40 à 44 ans
45	45 à 49 ans
50	50 à 54 ans
55	55 à 59 ans
60	60 à 64 ans
65	65 à 69 ans
70	70 à 74 ans
75	75 à 79 ans
80	80 à 84 ans
85	85 à 89 ans
90	90 à 94 ans
95	95 à 99 ans
100	100 à 104 ans

105	105 à 109 ans
110	110 à 114 ans
115	115 à 119 ans
120	120 à 124 ans

CS1 : Catégorie socioprofessionnelle sur un caractère

Type : caractère

Valeur dans la base	Libellé
1	Agriculteurs exploitants
2	Artisans, commerçants et chefs d'entreprise
3	Cadres et professions intellectuelles supérieures
4	Professions intermédiaires
5	Employés
6	Ouvriers

DIPL : Diplôme le plus élevé

Type : caractère

Valeur dans la base	Libellé
1	Pas de scolarité ou arrêt avant la fin du primaire
2	Aucun diplôme et scolarité interrompue à la fin du primaire ou avant la fin du collège
3	Aucun diplôme et scolarité jusqu'à la fin du collège ou au-delà
11	CEP (certificat d'études primaires)
12	BEPC, brevet élémentaire, brevet des collèges, DNB
13	CAP, BEP ou diplôme de niveau équivalent
14	Baccalauréat général ou technologique, brevet supérieur, capacité en droit, DAEU, ESEU
15	Baccalauréat professionnel, brevet professionnel, de technicien ou d'enseignement, diplôme équivalent

16	BTS, DUT, Deug, Deust, diplôme de la santé ou du social de niveau bac+2, diplôme Équivalent
17	Licence, licence pro, maîtrise, diplôme équivalent de niveau bac+3 ou bac+4
18	Master, DEA, DESS, diplôme grande école niveau bac+5, doctorat de santé
19	Doctorat de recherche (hors santé)

EMPL : Condition d'emploi

Type : caractère

Valeur dans la base	Libellé
11	En contrat d'apprentissage ou de professionnalisation
12	Placés par une agence d'intérim
13	Emplois aidés (contrat unique d'insertion, d'initiative emploi, d'accompagnement dans l'emploi, avenir, etc.)
14	Stagiaires rémunérés en entreprise
15	Autres emplois à durée limitée, CDD, contrat court, saisonnier, vacataire, etc.
16	Emplois sans limite de durée, CDI, titulaire de la fonction publique
21	Non-salariés : indépendants
22	Non-salariés : employeurs
23	Non-salariés : aides familiaux

ILT : Indicateur du lieu de travail

Type : caractère

Valeur dans la base	Libellé
1	Dans la commune de résidence actuelle
2	Dans une autre commune du département de résidence
3	Dans un autre département de la région de résidence
4	Hors de la région de résidence actuelle : en métropole

5	Hors de la région de résidence actuelle : dans un DOM
6	Hors de la région de résidence actuelle : dans une COM
7	À l'étranger

ILTUU : Indicateur urbain du lieu de travail

Type : caractère

Valeur dans la base	Libellé
1	Réside dans une commune rurale et travaille dans la même commune
2	Réside dans une commune rurale et travaille hors de la commune
3	Réside dans une commune urbaine et travaille dans la même commune
4	Réside dans une commune urbaine et travaille dans une autre commune de la même unité urbaine
5	Réside dans une commune urbaine et travaille en dehors de l'unité urbaine

IMMI : Situation quant à l'immigration

Type : caractère

Valeur dans la base	Libellé
1	Immigrés
2	Non immigrés

INATC : Indicateur de nationalité condensé (Français/Étranger)

Type : caractère

Valeur dans la base	Libellé
1	Français
2	Etranger

INEEM : Nombre d'élèves, étudiants ou stagiaires âgés de 14 ans ou plus du ménage

Type : caractère

Valeur dans la base	Libellé
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
Y	Hors résidence principale
Z	Hors logement ordinaire

INPOM : Nombre de personnes actives ayant un emploi du ménage

Type : caractère

Valeur dans la base	Libellé
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16

17	17
18	18
Y	Hors résidence principale
Z	Hors logement ordinaire

MOCO : Mode de cohabitation

Type : caractère

Valeur dans la base	Libellé
11	Enfants d'un couple
12	Enfants d'une famille monoparentale
21	Adultes d'un couple sans enfant
22	Adultes d'un couple avec enfant(s)
23	Adultes d'une famille monoparentale
31	Hors famille dans ménage de plusieurs personnes
32	Personnes vivant seules
40	Personnes vivant hors ménage

MODTRANS : Mode de transport principal le plus souvent utilisé pour aller travailler

Type : caractère

Valeur dans la base	Libellé
2	Marche à pied (ou rollers, patinette)
3	Vélo (y compris à assistance électrique)
4	Deux-roues motorisé
5	Voiture, camion, fourgonnette
6	Transports en commun

NA5 : Secteur d'activité (en 5 postes)

Type : caractère

Valeur dans la base	Libellé
AZ	Agriculture
BE	Industrie
FZ	Construction
GU	Commerce, transports et services divers
OQ	Administration publique, enseignement, santé humaine et action sociale

NPERR : Nombre de personnes du ménage (regroupé)

Type : caractère

Valeur dans la base	Libellé
1	Une personne
2	2 personnes
3	3 personnes
4	4 personnes
5	5 personnes
6	6 personnes
Z	Hors logement ordinaire

SEXE : Sexe

Type : caractère

Valeur dans la base	Libellé
1	Homme
2	Femme

STAT : Statut professionnel

Type : caractère

Valeur dans la base	Libellé
10	Salariés
21	Non-salariés : indépendants
22	Non-salariés : employeurs
23	Non-salariés : aides familiaux

STOCD : Statut d'occupation détaillé du logement

Type : caractère

Valeur dans la base	Libellé
0	Logement ordinaire inoccupé
10	Propriétaire
21	Locataire ou sous-locataire d'un logement loué vide non HLM
22	Locataire ou sous-locataire d'un logement loué vide HLM
23	Locataire ou sous-locataire d'un logement loué meublé ou d'une chambre d'hôtel

30	Logé gratuitement
ZZ	Hors logement ordinaire

TP : Temps de travail

Type : caractère

Valeur dans la base	Libellé
1	Temps complet
2	Temps partiel

TYPL : Type de logement

Type : caractère

Valeur dans la base	Libellé
1	Maison
2	Appartement
3	Logement-foyer
4	Chambre d'hôtel
5	Habitation de fortune
6	Pièce indépendante (ayant sa propre entrée)
Z	Hors logement ordinaire

TYPMR : Type de ménage regroupé (en 9 postes)

Type : caractère

Valeur dans la base	Libellé
11	Homme vivant seul
12	Femme vivant seule
20	Ménage de plusieurs personnes sans famille
31	Ménage dont la famille principale est monoparentale (un homme avec enfant(s))
32	Ménage dont la famille principale est monoparentale (une femme avec enfant(s))
41	Ménage dont la famille principale est un couple de deux actifs occupés
42	Ménage dont la famille principale est un couple composé d'un homme actif occupé et d'un conjoint 'autre'

43	Ménage dont la famille principale est un couple composé d'une femme active occupée et d'un conjoint 'autre'
44	Ménage dont la famille principale est un couple de deux personnes 'autres'
ZZ	Hors logement ordinaire

VOIT : Nombre de voitures du ménage

Type : caractère

Valeur dans la base	Libellé
0	Aucune voiture
1	Une seule voiture
2	Deux voitures
3	Trois voitures ou plus
X	Logement ordinaire inoccupé
Z	Hors logement ordinaire

➤ Indicateurs sur les déplacements et les émissions de GES

CARBU_HEBDO : Carburant moyen consommé par personne et par semaine pour les déplacements domicile-travail en voiture (en litre)

Type : numérique

CHAMP_CO2 : Champ pour le calcul du CO2 (vrai ou faux)

Type : booléen

DIST : Distance domicile-travail moyenne (en km)

Type : numérique

DIST_HEBDO : Distance moyenne parcourue par personne et par semaine, compte-tenu de la fréquence, des détours et des raccourcis (en km)

Type : numérique

DUREE : Durée du trajet domicile-travail - sans congestion - (en minutes)

Type : numérique

➤ Variables géographiques complémentaires

AAV20_RESID : Aire d'attraction des villes du lieu de résidence

Type : caractère

Code Insee sur 3 positions

AAV20_TRAV : Aire d'attraction des villes du lieu de travail

Type : caractère

Code Insee sur 3 positions

CATEAAV20_RESID : Catégorie de la commune dans le zonage en aires d'attraction des villes du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
11	Commune-centre
12	Autre commune du pôle principal
13	Commune d'un pôle secondaire
20	Commune de la couronne
30	Commune hors attraction des villes
99	Étranger
ZZ	Collectivité d'outre-mer

CATEAAV20_TRAV : Catégorie de la commune dans le zonage en aires d'attraction des villes du lieu de travail

Type : caractère

Valeur dans la base	Libellé
11	Commune-centre
12	Autre commune du pôle principal
13	Commune d'un pôle secondaire
20	Commune de la couronne
30	Commune hors attraction des villes
99	Étranger
ZZ	Collectivité d'outre-mer

DEP_RESID : Département du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
Numéro du département (France Métropolitaine)	Nom du département
971	Guadeloupe
972	Martinique

973	Guyane
974	La Réunion
976	Mayotte

DEP_TRAV : Département du lieu de travail

Type : caractère

Valeur dans la base	Libellé
Numéro du département (France Métropolitaine)	Nom du département
971	Guadeloupe
972	Martinique
973	Guyane
974	La Réunion
976	Mayotte
999	Etranger
ZZZ	Collectivité d’Outre-mer

EPCI_EPT_RESID : EPCI du lieu de résidence (détail EPT pour la Métropole du Grand Paris)

Type : caractère

Code Insee sur 9 positions

EPCI EPT_TRAV : EPCI du lieu de travail (détail EPT pour la Métropole du Grand Paris)

Type : caractère

Code Insee sur 9 positions

GRILLE_DENSITE_RESID : Grille de densité en 7 niveaux du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
1	Grands centres urbains
2	Centres urbains intermédiaires
3	Petites villes
4	Ceintures urbaines
5	Bourgs ruraux
6	Rural à habitat dispersé
7	Rural à habitat très dispersé
9	Étranger
Z	Mayotte

GRILLE_DENSITE_TRAV : Grille de densité en 7 niveaux du lieu de travail

Type : caractère

Valeur dans la base	Libellé
1	Grands centres urbains
2	Centres urbains intermédiaires
3	Petites villes
4	Ceintures urbaines
5	Bourgs ruraux
6	Rural à habitat dispersé
7	Rural à habitat très dispersé
9	Étranger
Z	Mayotte

REG_RESID : Région du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
01	Guadeloupe
02	Martinique
03	Guyane
04	La Réunion
06	Mayotte
11	Île-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Nouvelle-Aquitaine
76	Occitanie
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur
94	Corse

REG_TRAV : Région du lieu de travail

Type : caractère

Valeur dans la base	Libellé
01	Guadeloupe
02	Martinique
03	Guyane
04	La Réunion
06	Mayotte
11	Île-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Nouvelle-Aquitaine
76	Occitanie
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur
94	Corse
99	Etranger
ZZ	Collectivité d'Outre-mer

TAAV2017_RESID : Tranche de taille dans le zonage en aires d'attraction des villes du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
0	Commune hors attraction des pôles
1	Aire de moins de 50 000 habitants
2	Aire de 50 000 à moins de 200 000 habitants
3	Aire de 200 000 à moins de 700 000 habitants
4	Aire de 700 000 habitants ou plus (hors Paris)
5	Aire de Paris
9	Etranger

Z	Collectivité d’Outre-mer
---	--------------------------

TAAV2017_TRAV : Tranche de taille dans le zonage en aires d’attraction des villes du lieu de travail

Type : caractère

Valeur dans la base	Libellé
0	Commune hors attraction des pôles
1	Aire de moins de 50 000 habitants
2	Aire de 50 000 à moins de 200 000 habitants
3	Aire de 200 000 à moins de 700 000 habitants
4	Aire de 700 000 habitants ou plus (hors Paris)
5	Aire de Paris
9	Etranger
Z	Collectivité d’Outre-mer

URBAIN_RURAL_RESID : Typologie urbain / rural en 6 classes du lieu de résidence

Type : caractère

Valeur dans la base	Libellé
1	Rural autonome très peu dense
2	Rural autonome peu dense
3	Rural sous faible influence d’un pôle
4	Rural sous forte influence d’un pôle
5	Urbain densité intermédiaire
6	Urbain dense
9	Étranger
Y	Mayotte
Z	Collectivité d’outre-mer

URBAIN_RURAL_TRAV : Typologie urbain / rural en 6 classes du lieu de travail

Type : caractère

Valeur dans la base	Libellé
1	Rural autonome très peu dense
2	Rural autonome peu dense
3	Rural sous faible influence d’un pôle

4	Rural sous forte influence d'un pôle
5	Urbain densité intermédiaire
6	Urbain dense
9	Étranger
Y	Mayotte
Z	Collectivité d'outre-mer

ZE2020_RESID : Zone d'emploi 2020 du lieu de résidence

Type : caractère

Code Insee sur 3 positions

ZE2020_TRAV : Zone d'emploi 2020 du lieu de travail

Type : caractère

Code Insee sur 3 positions

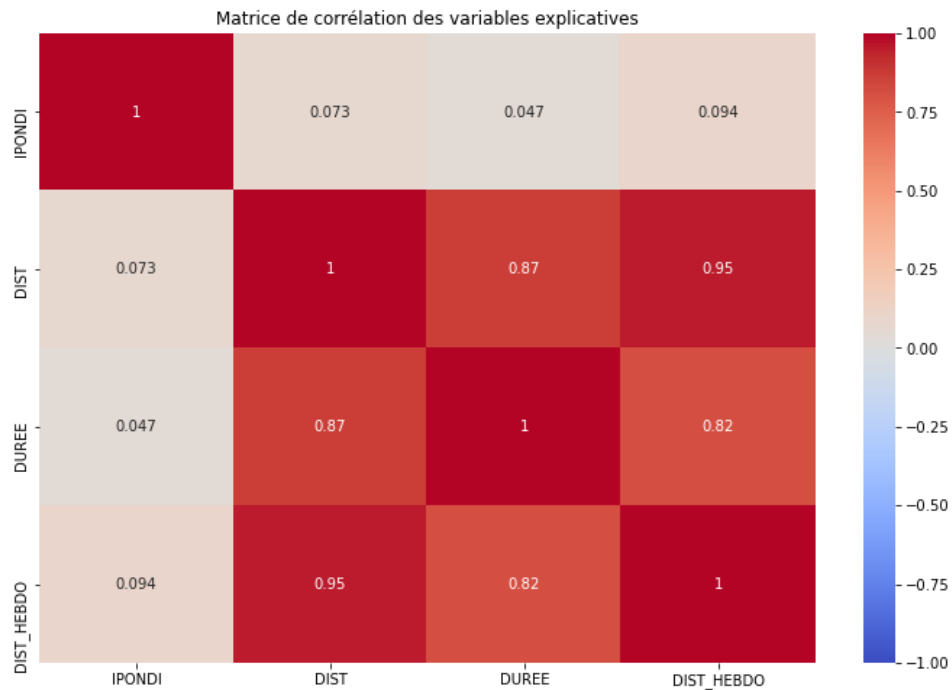
Annexe 3 : Tableau des statistiques descriptives de la variable cible

Voici les principales statistiques descriptives de la variables cible :

Statistiques descriptives	
Nombre d'observations	6 205 730
Moyenne	11 269.3
Médiane	6 244.7
Ecart-type	14 201.8
Valeur minimale	0
Valeur maximale	160315.3

Ces statistiques montrent une dispersion importante autour de la moyenne, signe d'une grande variabilité dans les comportements de déplacement domicile-travail. La médiane, plus basse que la moyenne, indique que la majorité des valeurs sont concentrées en dessous de la moyenne, tandis que quelques valeurs très élevées influencent celle-ci à la hausse. Cette répartition pourrait être due à des différences notables dans les distances domicile-travail et les moyens de transport utilisés, certains individus effectuant de longs trajets ou utilisant des véhicules à forte émission.

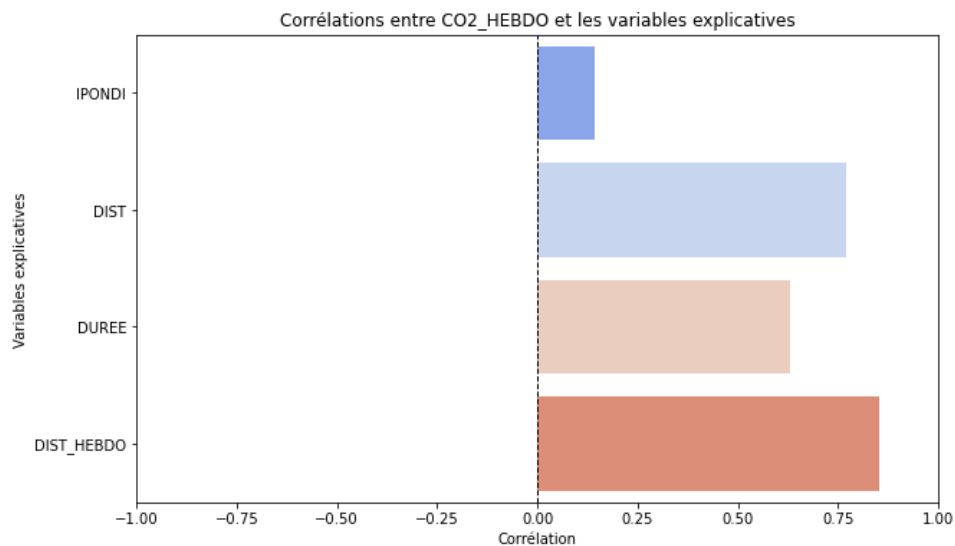
Annexe 4 : Analyse des corrélations entre les variables quantitatives explicatives



Source : réalisé par les membres du projet

Grille de lecture : Le niveau de corrélation entre les variables DIST et DUREE est estimé à 0.87

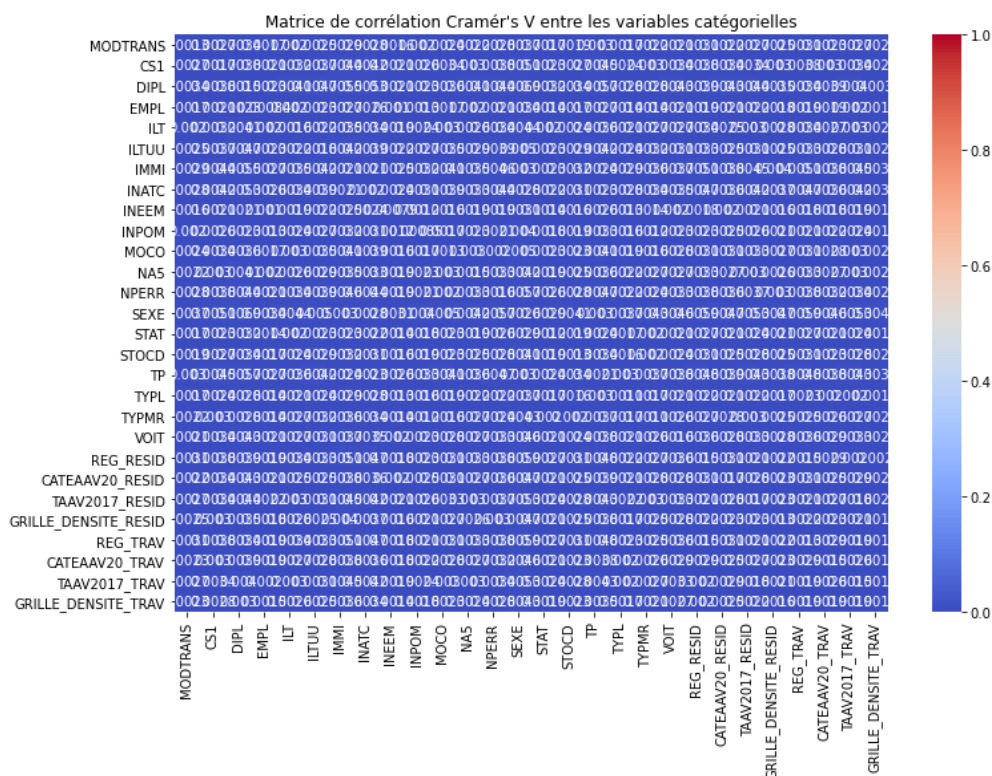
Annexe 5 : Analyse des corrélations entre les variables quantitatives explicatives et la variable cible



Source : réalisé par les membres du projet

Grille de lecture : La corrélation entre la variable DIST_HEBDO et CO2_HEBDO est estimé entre 0.75 et 1.00

Annexe 6 : Analyse des corrélations entre les variables catégorielles explicatives



Annexe 7 : Comparaison des erreurs MSE (Mean Squared Error) et MAE (Mean Absolute Error) du modèle lasso et de l'arbre de régression en grammes de CO2e hebdomadaire

	Lasso	Arbre de régression
MSE	35 593 211,05	13 959 390,97
MAE	5 965,88	3 736,25

Le MSE correspond à la moyenne des carrés des erreurs. Plus cette valeur est faible, plus le modèle est précis, mais elle amplifie l'impact des grandes erreurs. Le MSE est moins intuitif à interpréter directement en raison de l'unité : elle est exprimée en grammes de CO2e², ce qui la rend peu exploitable sans une mise en contexte.

Le MAE, quand à lui, représente la moyenne des erreurs absolues. Cette métrique indique la distance moyenne entre les prédictions et les valeurs réelles. Elle est moins influencée par les grandes erreurs que le MSE. De plus, le MAE est plus facile à interpréter que le MSE puisqu'il

est exprimé dans les mêmes unités que la variable cible (en grammes de CO₂e), ce qui en fait une mesure directe de la précision du modèle.

Dans notre cas, le modèle d'arbre de régression est clairement plus performant que le modèle Lasso selon les deux métriques (MSE et MAE). En effet, il est capable de minimiser à la fois les grandes erreurs (faible MSE) et les écarts moyens (faible MAE), ce qui en fait un choix préférable pour notre objectif de prédiction des émissions de CO₂ hebdomadaires moyennes par personne.