Rapport Projet Ingénieur

Machine Learning et Statistiques pour des Villes Durables
Modélisation des Préférences de Déplacements.

Génie Informatique et Statistiques 5
Salma Hrouma

Tuteur Projet:
DR. Mihaly Petrecszky

Tuteur Polytech:
DR. Serge Petiton

Année universitaire 2024/2025

Résumé

Nous analysons la transition vers des modes de transport plus écologiques en étudiant les comportements de mobilité des citoyens dans la région Hauts-de-France, à partir des données de l'Enquête Ménage Déplacement (EMD) 2016. Cette étude vise à comprendre l'influence des facteurs socio-économiques (âge, sexe, occupation professionnelle, possession d'abonnement de transport en commun...) et comportementaux (fréquence d'utilisation du vélo, voiture, transports en commun) sur les choix de transport.
Afin d'expliquer ces choix, une analyse exploratoire des données a permis d'identifier les variables pertinentes, utilisées pour entraîner des modèles de machine learning, tels que la régression logistique et les forêts aléatoires.Les modèles ont été entraînés et supervisés à l'aide de Mlflow, et une interface interactive sous Streamlit a été développée pour la visualisation des résultats, permettant une exploration facile des prédictions sur des choix de transport.
Les analyses ont révélé un engagement écologique moyen de 34 % chez les habitants. Cette étude apporte une meilleure compréhension des comportements de mobilité, offrant des perspectives pour encourager des stratégies de transport plus respectueuses de l'environnement.En outre, elle permet de mieux cibler les actions nécessaires pour favoriser un passage à des modes de transport plus durables.

Mots-clés : Transition écologique, choix de transport, comportements de mobilité, machine learning, régression logistique, forêts aléatoires

Abstract

We analyze the transition towards more eco-friendly modes of transport by studying the mobility behaviors of citizens in the Hauts-de-France region, using data from the 2016 Household Travel Survey (Enquête Ménage Déplacement, EMD). This study aims to understand the influence of socio-economic factors (age, gender, occupation, public transport subscription ownership) and behavioral factors (frequency of bicycle, car, and public transport use) on transport choices.

To explain these choices, an exploratory data analysis was conducted to identify relevant variables, which were then used to train machine learning models, such as logistic regression and random forests. The models were trained and supervised using Mlflow, and an interactive interface was developed under Streamlit for the visualization of results, allowing easy exploration of predictions related to transport choices.

The analyses revealed an average ecological engagement of 34% among the inhabitants. This study provides a better understanding of mobility behaviors, offering insights to encourage more environmentally friendly transportation strategies. Furthermore, it helps to better target actions needed to promote the shift towards more sustainable modes of transport.

Keywords: Ecological transition, transportation choices, mobility behaviors, machine learning, logistic regression, random forests.

Table des figures

Liste des tableaux

Table des matières

1. Introduction

1.1 Contexte

Dans un contexte de transition énergétique et de lutte contre le changement climatique, la mobilité durable s'impose comme une priorité stratégique pour les territoires. La métropole lilloise, en raison de sa position géographique et de sa dynamique démographique, fait face à des enjeux majeurs en matière de transport. La mobilité contribue de manière significative aux émissions de gaz à effet de serre, et une part importante des déplacements repose encore sur l'usage prédominant de la voiture individuelle, peu respectueuse de l'environnement. Cette situation engendre non seulement une pression écologique, mais également une congestion croissante des infrastructures et une détérioration de la qualité de vie des habitants.

Dans ce cadre, il devient essentiel de comprendre les facteurs qui influencent les choix de transport des citoyens. Ces choix ne sont pas seulement conditionnés par l'accessibilité des modes de transport ou leur coût, mais également par une combinaison des facteurs socio-économiques ,démographiques et de déplacement, tels que le revenu, l'âge, la situation professionnelle, et le type de ménage.
Cependant, malgré l'importance de ces facteurs, il existe une lacune dans la compréhension approfondie de leur impact respectif sur les comportements de mobilité.

1.2 Problématique

La diversité des profils socio-démographiques et des configurations géographiques dans la région Hauts-de-France rend complexe la compréhension des choix de transport des habitants. Des facteurs tels que l'âge, la situation professionnelle ou les habitudes de déplacement influencent ces décisions, ce qui complique la mise en place de stratégies de mobilité plus durables.
Comment ces facteurs déterminent-ils les choix de transport, et comment peut-on prédire ces comportements à l'aide de techniques de machine learning pour orienter des solutions de mobilité plus écologiques ?

1.3 Objectifs de l’étude

Cette étude a pour objectif d'identifier les facteurs les facteurs socio-économiques et démographiques qui influencent les choix de transport des habitants de la région Hauts-de-France. En utilisant des techniques de machine learning, nous cherchons à extraire des insights permettant de mieux comprendre les comportements de mobilité. À partir de ces résultats, nous proposerons des recommandations pour encourager l'adoption de modes de transport plus durables.

1.4 Démarche

Dans cette section, nous présentons la démarche adoptée pour analyser les facteurs influençant les choix de transport des habitants de la métropole lilloise.

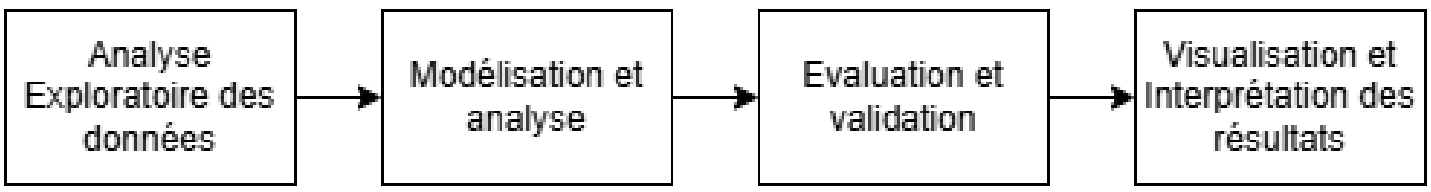

Figure 1: Processus de modélisation et d'analyse des données

1.4.1 Exploration des données

La première étape consiste à explorer les données pour comprendre leur structure, détecter les anomalies et traiter les valeurs manquantes. Les variables socio-économiques et démographiques ont été analysées pour identifier celles influençant les choix de transport.

1.4.2 Modélisation et analyse

Des techniques de modélisation, comme la régression logistique et les arbres de décision, ont été utilisées pour analyser l'impact des facteurs sur les choix de transport. Cela a permis de construire un modèle pour prédire les comportements de mobilité.

1.4.3 Evaluation et validation

Les modèles de régression logistique et d'arbres de décision ont été évalués à l'aide de métriques comme la précision, pour mesurer leur performance.

1.4.4 Interprétation et visualisation des résultats

Les résultats ont été interprétés pour identifier les facteurs influençant les choix de transport et présentés via Streamlit, une interface interactive permettant de visualiser dynamiquement les prédictions des modèles.

1.5 Plan du mémoire

Ce mémoire s'articule autour de 6 sections principales. Après une introduction qui contextualise la problématique et expose les objectifs de l'étude :

Section 1 : Contexte
Cette section présente le cadre général du projet, en abordant les objectifs définis par le cahier des charges et la structure qui a conduit à l'étude.

Section 2 : Cadre théorique et revue de la littérature
Cette section décrit les principes scientifiques utilisés dans le projet, ainsi qu'un compte-rendu détaillé de l'état des connaissances sur la problématique étudiée.

Section 3 : Méthodologie et solutions retenues
Cette partie expose les objectifs spécifiques du projet ainsi que les méthodes employées pour traiter le problème, en mettant l'accent sur la modélisation et l'analyse des données.

Section 4 : Résultats obtenus.
Cette section permet d'exposer clairement les effets des solutions retenues et de mettre en lumière les points forts et les limites des résultats obtenus.

Section 5 : Analyse critique et discussion des résultats
Les résultats obtenus sont confrontés aux objectifs initiaux et discutés en tenant compte des choix méthodologiques effectués. Les limitations identifiées ainsi que les pistes d'amélioration sont également abordées.

Section 6 : Conclusion et perspectives
Le mémoire se clôt par une synthèse des principales contributions du projet. Suivie de recommandations pour des développements futurs.

Conclusion

Ce premier chapitre a permis de poser les bases de l'étude en présentant son contexte, sa problématique, et ses objectifs. Dans un environnement marqué par des défis de transition écologique, comprendre les choix de transport des habitants de la métropole lilloise apparaît comme une nécessité stratégique. La démarche méthodologique adoptée, allant de l'exploration des données à l'évaluation des modèles prédictifs, vise à dégager des recommandations pertinentes pour promouvoir une mobilité plus durable. Les sections suivantes détailleront les fondements théoriques, les méthodes employées, et les résultats obtenus pour répondre aux enjeux identifiés.

2.Contexte du projet

2.1. Présentation de la structure à l'origine du sujet

Le projet a été initié en collaboration avec le Laboratoire Cristal (Centre de Recherche en Informatique, Signal et Automatique de Lille), une unité de recherche reconnue pour son expertise en informatique, data science et intelligence artificielle. Spécialisé dans l'analyse de données complexes et la modélisation algorithmique, le laboratoire Cristal s'intéresse particulièrement aux applications des technologies numériques dans des domaines tels que la mobilité durable, l'optimisation des systèmes et la prise de décision. Dans le cadre de ce projet, le laboratoire Cristal a proposé d'étudier les comportements de mobilité dans la région des Hauts-de-France en utilisant des techniques avancées de machine learning et d'analyse statistique. L'objectif est de contribuer à une meilleure compréhension des facteurs influençant les choix de transport, afin de soutenir la transition vers des modes de déplacement plus écologiques.

2.2. Cahier des charges.

Objectifs :

- Analyser les données de l'Enquête Ménage Déplacement (EMD) 2016 pour comprendre les comportements de mobilité dans la région des Hauts-de-France.
- Identifier les facteurs socio-démographiques et économiques influençant les choix de transport.
- Développer des modèles prédictifs (régression logistique, forêts aléatoires) pour anticiper les préférences de mobilité.
- Proposer des recommandations pour encourager l'adoption de modes de transport plus durables.

Livrables :

- Un rapport détaillé présentant les résultats de l'analyse et les modèles développés.
- Une interface interactive sous Streamlit pour visualiser les prédictions des modèles.
- Le code source des modèles et de l'analyse des données, hébergé sur GitHub pour un accès transparent et reproductible.

Conclusion

Ce projet s'inscrit dans une démarche visant à mieux comprendre les comportements de mobilité dans la région des Hauts-de-France à travers des méthodes d'analyse de données et de machine learning. L'objectif principal étant d'identifier les facteurs socio-économiques et démographiques influençant les choix de transport, afin de favoriser l'adoption de solutions plus durables. Le cahier des charges définit clairement les attentes et les livrables, en mettant l'accent sur des résultats concrets et une accessibilité aux outils et modèles développés. Cette démarche se veut non seulement innovante mais aussi pratique, en offrant des outils visuels et des recommandations stratégiques pour soutenir la transition vers une mobilité plus écologique et

3. Revue de la littérature

3.1. Principes scientifiques utilisés dans le projet

Dans ce projet, nous avons sélectionné un ensemble de features après une analyse exploratoire approfondie, en veillant à choisir des variables ayant une relation directe avec notre variable cible.

Deux principales approches ont été utilisées :
Régression Logistique : Cette méthode est employée pour prédire une variable binaire déterminant si le mode de transport choisi est écologique ou non. Elle permet de modéliser une relation linéaire entre les features explicatives et la probabilité d'appartenance à une classe cible.

Forêt Aléatoire : Ce modèle est utilisé pour prédire le mode de transport préféré par les utilisateurs. La forêt aléatoire présente une capacité essentielle à capturer des relations complexes et non linéaires entre les différentes features, ce qui en fait une solution adaptée aux problématiques de classification nécessitant une flexibilité analytique accrue.
Ces deux approches complémentaires permettent d'exploiter à la fois la simplicité d'interprétation de la régression logistique et la puissance prédictive des forêts aléatoires.

3.2. L'état des connaissances actuelles sur la problématique étudiée

Les études dans la littérature mettent en évidence deux approches principales pour modéliser les choix de mode de transport : les modèles de choix discrets, tels que la régression logistique, et les modèles de machine learning. Selon Mohd. Ali et al. (2021) dans leur étude intitulée *Travel Mode Choice Modeling: Predictive Efficacy between Machine Learning Models and Discrete Choice Model*, les modèles de choix discrets ont historiquement dominé ce domaine, mais les progrès en matière de techniques computationnelles ont conduit à une adoption croissante des modèles de machine learning.

L'étude menée à Kuantan City (Malaisie) a comparé plusieurs modèles de machine learning, notamment les réseaux de neurones, les forêts aléatoires et les arbres de décision, avec la régression logistique binaire pour prédire les choix de mode de transport des utilisateurs. Les principales conclusions de l'étude sont les suivantes :
Les modèles de machine learning offrent une flexibilité supérieure pour capturer des relations complexes dans les données.
L'utilisation de techniques d'importance des features est essentielle pour identifier les variables significatives influençant le choix du mode de transport.
Bien que les modèles de machine learning soient meilleurs pour les prédictions, la régression logistique conserve un avantage pour interpréter les relations entre les variables grâce à une structure mathématique élégante.

L'étude conclut que les modèles de machine learning, tels que la forêt aléatoire, sont particulièrement adaptés aux problématiques où les relations entre variables sont complexes et non linéaires, tandis que les modèles discrets restent précieux pour une meilleure compréhension des relations inférentielles entre variables.

Cette étude nous a servi d'inspiration pour notre projet. En effet, nous avons adopté une approche similaire en utilisant la régression logistique pour prédire une variable binaire déterminant si le moyen de transport est écologique ou non, et la forêt aléatoire pour prédire le mode de transport utilisé. Comme dans l'étude de Nur Fahriza Mohd. Ali et al., nous avons également veillé à choisir des features pertinentes après une analyse exploratoire approfondie, afin d'améliorer la performance de nos modèles.
De plus, nous partageons l'idée selon laquelle les modèles de Machine Learning, tels que la forêt aléatoire, sont mieux adaptés pour capturer des relations complexes et non linéaires entre les variables explicatives. Cette démarche a permis de développer une solution performante pour la modélisation des choix de transport, tout en tenant compte des enseignements de cette étude scientifique.

Conclusion

En conclusion, cette revue de la littérature met en lumière les principes scientifiques fondamentaux utilisés pour modéliser les choix de transport. Elle permet également de situer notre approche dans le contexte des recherches existantes, soulignant l'évolution des modèles de machine learning qui sont de plus en plus privilégiés pour leur capacité à modéliser des relations complexes. L'étude d'inspiration a guidé notre sélection de modèles et de variables, tout en mettant en évidence l'importance de choisir des features pertinentes pour améliorer les performances des modèles. Ces bases théoriques et méthodologiques constituent le fondement sur lequel repose notre projet et orientent les solutions proposées dans les sections suivantes.

4.Méthodologie et moyens mis en œuvre

Dans cette section, nous allons exposer l'ensemble des outils et moyens utilisés pour mener à bien ce projet, ainsi que la démarche suivie pour l'analyse des données.

4.1. Choix des outils et technologies

Le travail s'est principalement basé sur **Python**, un langage de programmation réputé pour sa puissance et sa flexibilité en data science. Pour faciliter l'exécution des analyses et bénéficier de ressources de calcul élevées, **Google Colab** a été utilisé comme environnement de travail. Cette plateforme cloud permet une collaboration en temps réel et évite les contraintes liées à la configuration locale.

Les bibliothèques suivantes ont été utilisées pour répondre aux besoins spécifiques du projet :

Pandas : Pour la manipulation et l'analyse des données sous forme de DataFrame, facilitant ainsi les opérations de nettoyage et de transformation.

Matplotlib.pyplot et Seaborn : Pour la visualisation graphique des données et des résultats, permettant de créer des graphiques clairs et informatifs pour l'analyse exploratoire.

Numpy et Scipy : Pour les calculs numériques et les opérations sur les matrices, en support aux analyses statistiques.

Sklearn : Pour les algorithmes de machine learning, y compris les modèles de régression logistique et de forêts aléatoires, ainsi que pour les outils d'évaluation des performances des modèles (accuracy, confusion_matrix, classification_report).

Statsmodels.api : Pour effectuer des analyses statistiques avancées et des tests d'hypothèses.

Streamlit : Pour le développement d'une interface interactive, facilitant la visualisation des résultats et des prédictions des modèles de manière dynamique et accessible.

MLflow :

MLflow est une plateforme open-source conçue pour la gestion des expérimentations de machine learning. Elle permet de suivre, enregistrer et organiser les résultats des différentes expérimentations tout au long du projet. Grâce à MLflow, il a été possible de collecter et comparer les métriques de performance des modèles, telles que la précision, la matrice de confusion, et d'autres indicateurs clés. Cette plateforme facilite également la gestion des versions des modèles, permettant ainsi d'optimiser et de choisir les modèles les plus performants de manière plus systématique et transparente

4.2. Description des moyens.

Le principal moyen mis à disposition pour ce projet est la base de données de l'Enquête Ménage Déplacement (EMD) 2016. Cette base de données constitue une source précieuse d'informations pour analyser les comportements de mobilité dans la région des Hauts-de-France. Elle fournit une variété de données socio-économiques , démographiques et de déplacement, comprenant notamment :

- Caractéristiques des individus : Âge, sexe, niveau d'instruction, occupation principale (actif, étudiant, retraité, etc.) et statut professionnel.
- Informations économiques : Possession d'un abonnement de transport en commun, accès à des véhicules motorisés au sein du ménage.
- Caractéristiques des ménages : Taille du ménage, composition (nombre d'adultes et d'enfants), localisation géographique.
- Habitudes de déplacement : Fréquences d'utilisation des différents types de transport (vélo, voiture, transports en commun) et motifs de déplacements.

Dans le cadre de ce projet, une attention particulière est portée aux données socio-économiques des individus, car elles permettent d'étudier les interactions entre les caractéristiques personnelles et les préférences en matière de mobilité. Les données comportementales, telles que les fréquences d'utilisation des moyens de transport (vélo, voiture, transport public), offrent également des perspectives intéressantes pour affiner la compréhension des choix de déplacement.

4.3. Stratégie d’analyse des données et modélisation

Pour répondre aux objectifs du projet, une stratégie en trois étapes a été mise en place : prétraitement des données, traitement des données, et post-traitement.

1.**Prétraitement des données**

Dans cette première étape, l'objectif est de préparer les données brutes pour l'analyse. Cela inclut:

Nettoyage des données : gestion des valeurs manquantes et détection des anomalies.

Encodage des variables catégorielles : Transformation des variables non numériques en format adapté aux algorithmes de machine Learning (par exemple, encodage des catégories de transport).

2.**Traitement des données**

À ce stade, les données prétraitées sont utilisées pour construire des modèles et en tirer des conclusions :

- Analyse exploratoire des données (EDA) :Exploration des relations entre les variables explicatives et la variable cible.
- Sélection des features : Choix des variables les plus pertinentes pour les modèles de machine learning.
- Entraînement des modèles : Application des techniques de machine learning, comme la régression logistique et les forêts aléatoires, pour prédire le mode de transport choisi en fonction des facteurs socio-économiques et comportementaux.

3. **Post-traitement**

Cette étape permet d'évaluer et d'affiner les modèles construits :

- Évaluation des performances des modèles : Utilisation de métriques telles que l'exactitude, la matrice de confusion et le rapport de classification pour mesurer la performance des modèles.
- Optimisation des hyperparamètres : Ajustement des paramètres des modèles pour améliorer leur précision.
- Interprétation des résultats : Analyse des résultats pour tirer des conclusions et proposer des recommandations basées sur les facteurs les plus influents dans le choix des moyens de transport.

Conclusion

En résumé, cette section a détaillé les outils, technologies et méthodes utilisés pour analyser les comportements de mobilité à partir des données de l'Enquête Ménage Déplacement (EMD) 2016. Grâce à l'utilisation de Python et de bibliothèques adaptées, ainsi qu'à la plateforme MLflow pour le suivi des expérimentations, nous avons pu préparer et traiter les données de manière efficace, en vue de développer des modèles prédictifs fiables pour étudier les choix de transport des individus.

5. Travaux expérimentaux

5.1. Description des expérimentations réalisées

Dans cette section, nous présentons les travaux expérimentaux réalisés à partir du jeu de données initial, en suivant la méthodologie décrite précédemment. Les données ont été traitées de manière à les rendre exploitables pour l'analyse.

Le tableau suivant montre la structure des données:

Variable	Description
P3	Lien avec la personne de référence
P4	Âge
P4A	Possession de téléphone portable
P5	Possession du permis de conduire
P6	Dernier établissement scolaire fréquenté
P7	Occupation principale
P8	Occupation secondaire
P9	Profession (PCS)
P10	Possession d'un abonnement de transport en commun
P12	Travail ou études à domicile
P13	Lieu de travail ou études (base régionale)
P13L	Lieu de travail ou études (enquête urbaine)
P17	Problème de stationnement en général
P17A	Difficulté de stationnement (lieu de travail/études)
P19	Situation de la personne la veille
P20	Fréquence d'utilisation de la bicyclette
P21	Fréquence d'utilisation de deux-roues à moteur
P23	Fréquence d'utilisation d'une voiture en tant que conducteur
P24	Fréquence d'utilisation d'une voiture en tant que passager
P25	Fréquence d'utilisation du réseau urbain
PGRP	Groupe âge à comportement homogène
PNPC	Taille du ménage

Tableau 1: La structure des données

Prétraitement

La figure ci-dessous (*Figure2*) illustre la répartition des données manquantes. Les variables sont disposées sur l'axe des abscisses, tandis que les observations sont placées sur l'axe des ordonnées. Chaque ligne verticale met en évidence la proportion de valeurs disponibles (en violet) et de valeurs manquantes (en jaune) pour chaque variable. Cette représentation permet de repérer rapidement les variables problématiques nécessitant des décisions de traitement spécifiques.

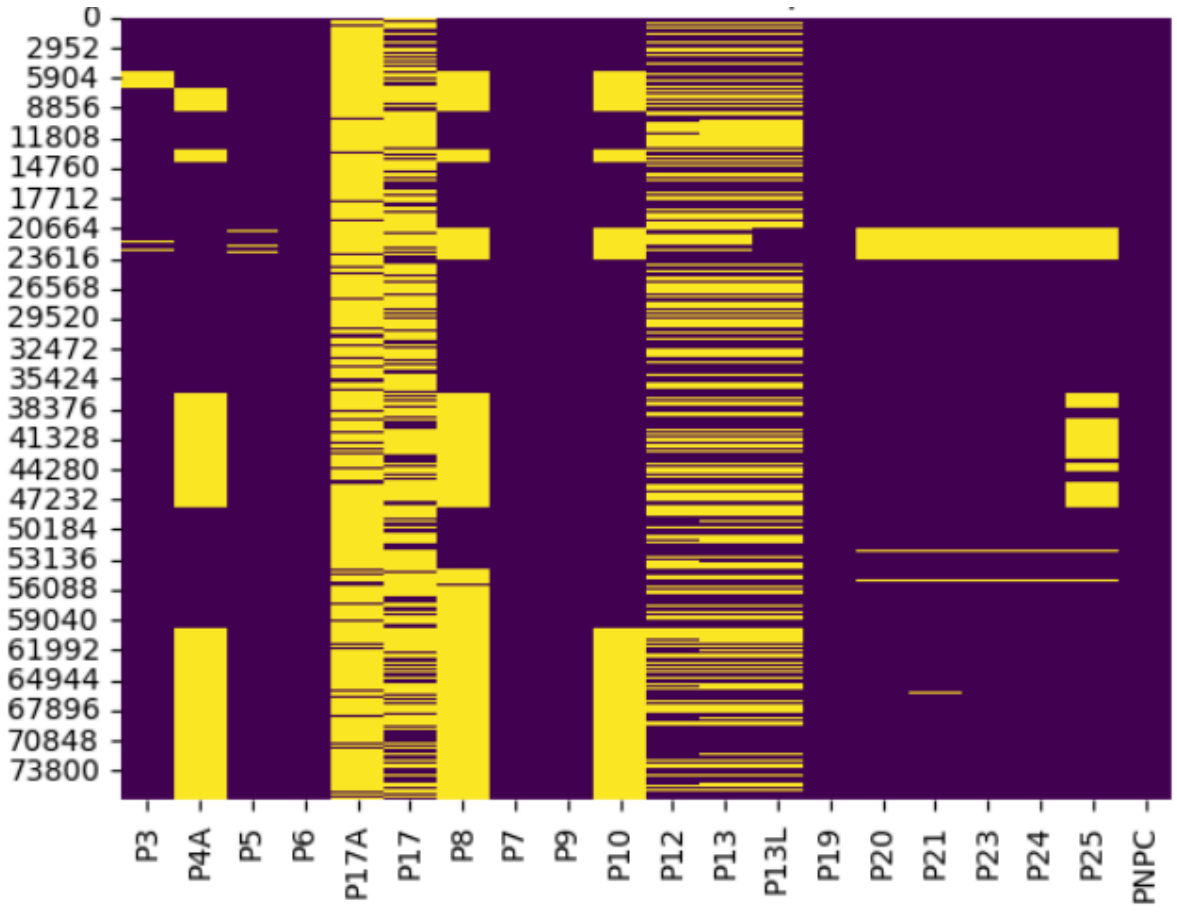

Figure2: Visualisation des Données Manquantes à l'aide d'une Heatmap

Certaines variables ont été supprimées en raison d'un taux de valeurs manquantes supérieur à 50 %. Celles-ci comprennent P17A et P17 , ainsi que P8 (Occupation secondaire), ces variables ayant un taux trop élevé de valeurs manquantes pour être considérées fiables pour l'analyse.

D'autres variables présentant un taux de valeurs manquantes compris entre 30 % et 50 % ont été exclues en raison de la dispersion aléatoire des valeurs manquantes sans relation claire avec d'autres variables. Ces variables incluent P4A , P12 , ainsi que P13 et P13L .

Enfin, pour les variables restantes, dont le taux de valeurs manquantes était faible, une imputation par le mode a été réalisée afin de préserver la cohérence de la distribution des données tout en minimisant la perte d'information.

Encodage des variables

L'encodage des variables a été réalisé afin de transformer les données catégorielles en formats adaptés aux modèles d'analyse. Par exemple, les variables représentant des fréquences d'utilisation des modes de transport (P20, P21, P23, P24, P25) ont été converties en valeurs numériques. Les modalités de ces variables, qui initialement étaient exprimées sous forme de catégories qualitatives, ont été codées de manière numérique pour représenter différents niveaux de fréquence d'utilisation.

De plus, la variable P9, qui décrit les professions des répondants, a été transformée en catégories agrégées. Cette transformation a permis de regrouper les différentes professions en catégories plus larges, facilitant ainsi leur utilisation dans les modèles. Les modalités agrégées ont ensuite été encodées sous forme numérique, permettant de les intégrer efficacement dans le processus d'analyse.

Traitement des données

Cette section concerne le traitement des données utilisées pour construire les modèles prédictifs. Après avoir effectué les étapes de nettoyage, d'encodage et de préparation des données, nous avons analysé les relations entre les variables explicatives et la variable cible.

Pour pouvoir accéder à la variable cible, qui est présente dans un autre fichier de déplacements, nous avons utilisé une clé primaire composée de la combinaison des variables suivantes :

- PEMD (Numéro d'enquête)
- ECH (Code ménage)
- PECH (Code personne)

Cette combinaison de variables nous a permis de lier les différents fichiers de manière cohérente et de relier les informations de manière unique.

La variable cible ECO, une variable binaire indiquant si le mode de transport est écologique (1) ou non (0), a ensuite été étudiée en relation avec l'ensemble des autres variables du dataset. Pour analyser cette association, nous avons utilisé deux tests statistiques : le test du Chi2 et le V de Cramér.

Le test du Chi2 est un test statistique qui permet de vérifier l'indépendance entre deux variables catégorielles. Il compare les distributions observées des catégories de chaque variable avec les distributions attendues sous l'hypothèse d'indépendance.
Le test repose sur deux hypothèses :

- H0 (Hypothèse nulle) : Les deux variables sont indépendantes, c'est-à-dire qu'il n'existe pas d'association significative entre elles.
- H1 (Hypothèse alternative) : Les deux variables sont dépendantes, c'est-à-dire qu'il existe une association significative entre elles.

Pour déterminer si les résultats suggèrent une association entre les variables, on calcule la p-value associée à la statistique Chi2. Si la p-value est inférieure à un seuil de significativité (généralement 0,05), l'hypothèse nulle H0 est rejetée au profit de l'hypothèse alternative H1, indiquant qu'il existe une relation significative entre les variables. En revanche, si la p-value est supérieure à 0,05, on ne peut pas rejeter H0, ce qui suggère qu'il n'y a pas d'association significative entre les variables.

Le V de Cramér, quant à lui, quantifie l'intensité de l'association entre deux variables catégorielles. Il varie de 0 (pas d'association) à 1 (association parfaite). Une valeur élevée du V de Cramér indique une forte relation entre les variables, tandis qu'une valeur proche de 0 indique une association faible.
Ces tests nous ont permis d'évaluer la pertinence des variables et d'identifier celles qui présentent une relation significative avec la variable cible, afin d'orienter la sélection des features pour les modèles de machine learning.

Les variables suivantes ont été retenues pour leur association significative avec la variable cible:

Variable	Chi-square	V de Cramer
P7	3319.360	0.317940
P10	3260.201	0.179315
P20	748.458	0.156544
P23	7877.984	0.507877
P24	451.051	0.121525
P25	3854.460	0.370886
PGRP	2582.812	0.280456
PNPC	636.614	0.139238

Tableau 2: Variables sélectionnées pour la classification avec leurs statistiques associées

Choix des modèles

Dans un premier temps, nous avons choisi d'utiliser la régression logistique binomiale pour prédire la variable cible. L'équation théorique de la régression logistique est la suivante :

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

où P(y=1 | X) est la probabilité que la variable cible prenne la valeur 1, β0 est l'ordonnée à l'origine, et β1,β2,...,βn sont les coefficients associés aux variables explicatives X1,X2,...,Xn.

Justification de ce choix:

Dans un contexte où l'interprétabilité des résultats est importante, la régression logistique reste une solution privilégiée, car :

Coefficients interprétables : Chaque coefficient représente la contribution d'une variable à la probabilité de la classe cible, facilitant l'analyse de l'impact des variables explicatives.

Analyse statistique : Les tests statistiques sur les coefficients permettent de valider leur significativité.

Avant l'entraînement du modèle, les données ont été divisées en deux ensembles : 80% pour l'entraînement et 20% pour les tests. Cette séparation a permis d'évaluer les performances du modèle de manière plus fiable, en évitant le surapprentissage et en assurant une évaluation réaliste sur des données non vues par le modèle.

Les performances du modèle montrent une précision de 78%, avec une matrice de confusion indiquant que le modèle détecte bien les cas non écologiques (88% de rappel pour la classe 0) mais est moins performant pour identifier les cas écologiques (58% de rappel pour la classe 1). La précision pour la classe écologique est de 71%, et l'F1-score de 0.63 (moyenne harmonique entre la précision et le rappel) suggère qu'il y a de la place pour améliorer la détection des clients écologiques.

Après avoir ajouté les variables de déplacement, les performances du modèle ont nettement progressé, atteignant une précision de 91%. Le rappel pour la classe écologique (1) est désormais de 89%, avec un F1-score de 0.87, indiquant une meilleure capacité à détecter les clients écologiques. Ces résultats montrent que la variable cible s'explique mieux en prenant en compte ces nouvelles variables.

Pour approfondir la prédiction et déterminer non seulement si le mode de transport est écologique ou non, mais aussi quel type de transport (voiture, vélo, transport public (pt), marche), nous avons utilisé la variable MODP, qui représente le mode de transport d'origine (encodée pour contenir ces quatre catégories).

Afin de prédire ces types de transport, nous avons évalué deux modèles d'apprentissage automatique : Forêt Aléatoire (Random Forest) et Gradient Boosting. Les deux modèles ont été entraînés en utilisant les caractéristiques socio-économiques et les habitudes de déplacement sélectionnées dans l'ensemble de données.

Comparaison des modèles

Pour comparer les performances de ces deux modèles, nous avons utilisé la validation croisée. Cette méthode permet de diviser l'ensemble des données en plusieurs sous-ensembles, afin de tester les modèles sur des jeux de données différents et d'éviter tout surapprentissage.

Après avoir effectué la validation croisée, nous avons obtenu les résultats suivants :
Random Forest :

- Scores de validation croisée : [0.9236, 0.9431, 0.9506, 0.9484, 0.9242]
- Moyenne des scores : 0.94

Gradient Boosting :

- Scores de validation croisée : [0.9258, 0.9409, 0.9536, 0.9474, 0.9238]
- Moyenne des scores : 0.94

Les deux modèles ont montré des performances très similaires en termes de validation croisée, avec une moyenne des scores de 0.94 dans les deux cas.

Choix du modèle

Bien que les performances des deux modèles soient proches, Random Forest a été choisi en raison de sa robustesse dans l'interprétation des résultats. Random Forest est plus stable dans l'agrégation des prédictions et moins susceptible d'être affecté par des erreurs ou des fluctuations des données. Il fournit également une interprétation plus claire et fiable de l'importance des variables, ce qui est essentiel pour comprendre l'impact de chaque caractéristique dans la prédiction.
En résumé, bien que les deux modèles aient des performances similaires, Random Forest a été privilégié pour sa meilleure capacité à interpréter et à expliquer les résultats.

5.2. Solutions retenues

Pour maximiser la précision des prédictions tout en garantissant une compréhension détaillée des facteurs influençant les choix de transport, nous avons opté pour une approche combinée. Dans un premier temps, la régression logistique est utilisée pour prédire si le mode de transport est écologique ou non. Cette étape permet une interprétation claire des variables explicatives liées à l'adoption de modes de transport durables. Ensuite, pour affiner l'analyse et déterminer précisément le type de mode de transport utilisé, le modèle Random Forest est déployé. Cette combinaison tire parti de l'interprétabilité de la régression logistique et de la robustesse des forêts aléatoires face à des données complexes, offrant ainsi une solution performante et analytiquement riche.

Conclusion

En conclusion, les travaux expérimentaux menés dans cette section ont permis de traiter et d'analyser efficacement le jeu de données, en mettant l'accent sur la gestion des valeurs manquantes et l'encodage des variables. L'utilisation de tests statistiques a permis d'identifier les variables pertinentes pour prédire la variable cible liée à l'écologie des modes de transport. Les modèles de régression logistique et de forêts aléatoires ont été comparés et combinés de manière judicieuse pour maximiser la précision des prédictions. Cette approche hybride, alliant l’interprétabilité de la régression logistique à la robustesse du modèle Random Forest, offre une solution fiable et performante pour l'analyse des choix de transport.

6.Résultats

Régression logistique:

Les résultats de la régression logistique montrent que, en moyenne, les clients affichent un niveau d'engagement écologique d'environ 34%, ce qui suggère une implication modérée dans l'adoption de comportements écologiques. Ce taux d'engagement peut être interprété comme une tendance générale à adopter des pratiques plus écologiques, mais avec encore une marge de progression.

Pour mesurer l'impact des différentes variables sur la probabilité qu'une personne adopte des modes de transport écologiques, tels que le vélo ou les transports publics, les coefficients de régression sont utilisés. Ces coefficients indiquent l'influence de chaque variable sur cette probabilité : un coefficient positif signifie qu'une augmentation de la variable accroît la probabilité d'adopter un mode de transport écologique, tandis qu'un coefficient négatif indique une diminution de cette probabilité. De plus, les p-values associées à chaque coefficient permettent d'évaluer la significativité statistique des variables. Une p-value inférieure à 0,05 suggère que l'effet de la variable est statistiquement significatif.

Variable	Coefficient	P-value
P10 (Abonnement TC)	0.4898	0.000
bike_freq (Fréquence vélo)	1.0719	0.000
public_freq (Transports publics)	1.0833	0.000
car_freq (Fréquence voiture)	-2.0753	0.000
AgeGroup (Groupe d'âge)	-0.8203	0.000
PNPC (taille de menage)	-0.0371	0.034

Tableau 3: Coefficients de la régression logistique

Variables avec impact positif :
P10 (Abonnement TC), bike_freq (Fréquence vélo) et public_freq (Fréquence transports publics) : Une utilisation plus fréquente de ces modes de transport augmente la probabilité d'adopter des comportements écologiques. Par exemple, un abonnement aux transports publics ou une fréquence élevée d'utilisation du vélo favorise l'adoption de modes de transport écologiques.

Variables avec impact négatif :
AgeGroup (Groupe d'âge) : les personnes âgées tendent à ne pas adopter des modes de transport écologiques, préférant la voiture.
PNPC (Taille de ménage) : Une taille de ménage plus grande réduit légèrement la probabilité d'utiliser des modes de transport écologiques.

Forêt Aléatoire

D'après la Figure 3, les résultats obtenus avec le modèle Random Forest indiquent que la fréquence d'utilisation du vélo (bike_freq) et des transports publics (public_freq) sont les variables les plus influentes pour prédire le mode de transport, suivies de la variable P7 qui indique l'occupation principale et des caractéristiques socio-économiques comme P10 et PNPC. L'âge (AgeGroup) a un impact relativement faible sur le choix du mode de transport, ce qui suggère que les habitudes de déplacement et les facteurs socio-économiques sont les déterminants principaux dans ce modèle.

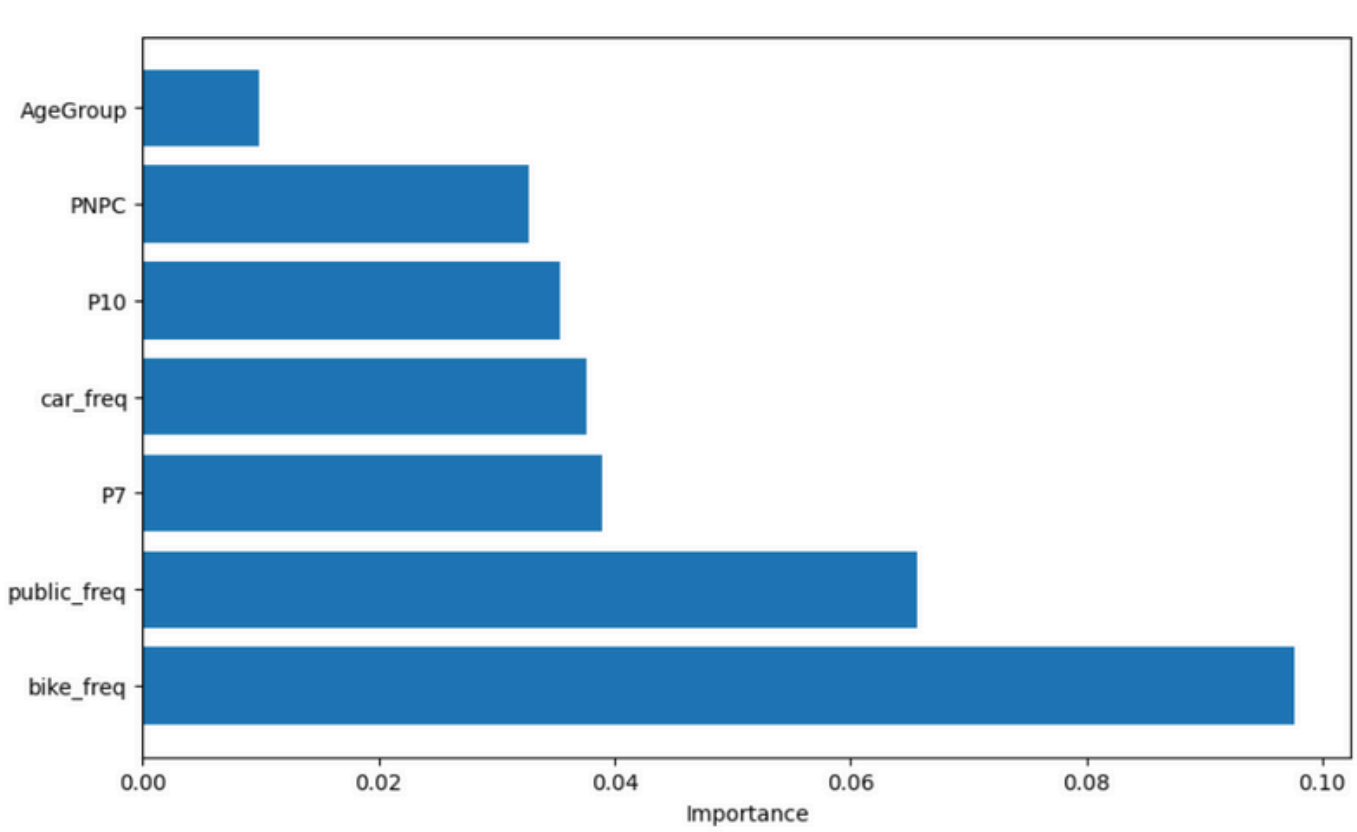

Figure 3 :Importance de chaque variable dans le modèle forêt aléatoire

7.Discussion

7.2. Limites et biais de l’étude

Cette étude présente certaines limites qui peuvent influencer l’interprétation des résultats :

- Données manquantes : La base de données contient de nombreuses valeurs manquantes, ce qui empêche d’avoir une vision complète et précise de certains comportements de transport. Cette situation peut biaiser les résultats en réduisant la diversité des observations.
- Surreprésentation des données concernant la voiture :Une part significative des données (plus de 60%) concerne l'utilisation de la voiture, ce qui est nettement supérieur à celles sur les autres modes de transport tels que le vélo et les transports publics. Cette inégalité dans les données peut fausser les conclusions en accentuant l'influence des variables liées à la voiture.
- Données non récentes : La base de données date de 2016, ce qui ne permet pas de refléter fidèlement les tendances actuelles en matière de mobilité, notamment l’adoption croissante des solutions de transport écologiques.

Malgré ces limitations, l'étude offre des insights précieux sur les comportements de mobilité, notamment en identifiant les facteurs favorisant ou freinant l'adoption des modes de transport écologiques.

8.Conclusion

Ce projet avait pour objectif d'analyser les facteurs socio-économiques, démographiques et de déplacement influençant les choix de transport dans la métropole lilloise, dans un contexte de transition énergétique. En combinant des approches statistiques, notamment la régression logistique, nous avons cherché à identifier les variables déterminantes pour l'adoption de modes de transport écologiques.

Les résultats obtenus montrent que certaines variables ont un impact significatif sur les comportements de mobilité. En particulier, l'âge et la taille du ménage apparaissent comme des facteurs réduisant la probabilité d'adopter des modes de transport durables. En revanche, la fréquence d'utilisation du vélo et des transports publics, ainsi qu'un abonnement aux transports publics, favorisent ces comportements écologiques.

Cependant, l'analyse présente certaines limitations : des données manquantes limitant une vision complète, une surreprésentation des déplacements en voiture faussant l'équilibre des analyses, et des données non récentes qui ne reflètent pas les tendances actuelles en matière de mobilité durable.

Ainsi, cette étude met en évidence l'importance de certains facteurs dans les choix de transport durable et souligne la nécessité d'adapter les politiques publiques en fonction des spécificités démographiques et comportementales pour promouvoir des solutions de mobilité plus écologiques et durables.

Perspectives et recommandations

À partir de ces résultats, plusieurs perspectives et recommandations peuvent être envisagées pour encourager l'adoption de modes de transport durables :

Pour les Jeunes :

Proposer des forfaits ou des réductions pour les transports publics et des services de location de vélos spécifiquement pour les étudiants. Il serait également pertinent d'améliorer les infrastructures cyclables et les stations de vélos dans les zones universitaires afin de faciliter l'accès à ces modes de transport.

Encouragement du vélo :

Il est essentiel de créer des infrastructures sécurisées, telles que des pistes cyclables dédiées et des parkings pour vélos. De plus, des programmes de location de vélos à faible coût pourraient être mis en place pour rendre cette option plus accessible à un large public.

Pour les personnes âgées :
Il est important de proposer des solutions de transport adaptées pour les personnes âgées, telles que des navettes ou des bus à faible coût. Par ailleurs, il convient de promouvoir l'usage de véhicules écologiques partagés, tout en garantissant confort et sécurité. Des campagnes de sensibilisation sur les avantages des modes de transport écologiques pour cette tranche de la population pourraient également contribuer à cette transition.

Pour les Ménages nombreux (PNPC) :
Pour les familles nombreuses, il serait pertinent d'offrir des réductions ou des forfaits familiaux pour les transports publics. Promouvoir l'utilisation de véhicules écologiques partagés pourrait également réduire l'impact écologique tout en répondant aux besoins de mobilité de ces foyers.

9.Positionnement par rapport aux compétences de la formation IS

Identification et mobilisation des connaissances scientifiques et techniques pointues :
Dans le cadre de ce projet, j'ai mobilisé des connaissances avancées en statistiques, machine learning, et analyse de données pour aborder les problématiques socio-économiques. J'ai utilisé des techniques de classification et d'apprentissage automatique via Python, telles que la régression logistique et Random Forest, afin d'analyser les données socio-économiques. Ces méthodes ont permis d'identifier des tendances clés et de prédire les comportements de mobilité en fonction des facteurs socio-économiques.

Compréhension de la structure des données et prétraitement :
Avant de procéder à l'analyse, j'ai effectué une exploration approfondie des données. Cela a impliqué des analyses univariées et bivariées afin de mieux comprendre la structure des données : identification des unités statistiques, des variables observées, des types de variables, ainsi que la gestion des données manquantes. Cette phase a permis de choisir la modélisation statistique la plus appropriée et d'assurer la qualité des données avant de les utiliser pour la modélisation.

Description des données et sélection des variables :
J'ai ensuite procédé à la description des données à l'aide de statistiques descriptives, en utilisant des indicateurs numériques et des graphiques pour résumer et visualiser les données. En parallèle, j'ai analysé les relations entre la variable cible et les variables explicatives par des analyses de corrélation. Cette étape a été cruciale pour sélectionner les variables les plus pertinentes et éviter la surabondance de données inutiles.

Modélisation et apprentissage automatique :
Enfin, j'ai construit et interprété des modèles d'apprentissage automatique, tels que la régression logistique et Random Forest. Cette étape a consisté à comprendre les dépendances présentes dans les données, et à valider les hypothèses du modèle. Ces modèles ont permis de prédire avec précision les comportements de mobilité et d'apporter des insights significatifs pour répondre à la problématique posée par le projet.

Références bibliographiques

Ali, N. F. M., Sadullah, A. F. M., Abdul Majeed, A. P. P., Mohd Razman, M. A., Zakaria, M. A., & Ab. Nasir, A. F. (2021). Travel mode choice modeling: Predictive efficacy between machine learning models and discrete choice model. The Open Transportation Journal, 15(1), 24-41.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. Statistical Methods in Medical Research, 20(1), 1–23.