

Rapport Projet Ingénieur

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

GENIE INFORMATIQUE STATISTIQUE

HROUMA Samia

Tuteur Projet:
Dr. Mihaly Petreczky

Tuteur Polytech:
Dr. Serge Petiton

Année universitaire 2024/2025

Résumé

Ce projet porte sur la modélisation des choix de transport des habitants de la Métropole Européenne de Lille (MEL), à l'aide de techniques de machine learning et de méthodes statistiques, en exploitant les données de l'Enquête Ménage Déplacement (EMD) 2016. L'objectif principal est de mieux comprendre les comportements et motivations de déplacement des individus afin de favoriser une transition vers des moyens de transport plus durables.

Pour ce faire, une analyse statistique exploratoire a été menée pour extraire des informations pertinentes des données, Ensuite, des modèles de classification, notamment la régression logistique et les forêts aléatoires, ont été implémentés en Python. Ces modèles permettent de modéliser les préférences individuelles en matière de transport. L'entraînement des modèles a été supervisé avec Mlflow, et une interface interactive basée sur Streamlit a été développée pour visualiser les résultats.

L'analyse révèle que le niveau moyen d'engagement écologique des habitants, calculé à partir des prédictions, est de seulement 32 % en ce qui concerne les choix de transport. Ce travail met en lumière des pistes d'amélioration pour promouvoir des pratiques de déplacement plus respectueuses de l'environnement.

Keywords : machine learning, statistiques, transport durable, classification, régression logistique, forêt aléatoire.

Abstract

This project focuses on modeling the transportation choices of residents in the Lille European Metropolis (MEL) using machine learning techniques and statistical methods, leveraging data from the survey "Enquête Ménage éplacement - EMD" 2016 . The primary goal is to better understand individuals' travel behaviors and motivations in order to promote a shift toward more sustainable modes of transportation.

To achieve this, an exploratory statistical analysis was conducted to extract relevant information from the data. Then, classification models, including logistic regression and random forests, were implemented in Python. These models help to model individual transportation preferences. The training of the models was supervised using Mlflow, and an interactive interface was developed using Streamlit to visualize the results.

The analysis reveals that the average ecological engagement level among residents, as estimated from the predictions, is only 32% concerning transportation choices. This work highlights areas for improvement to encourage more environmentally friendly travel practices.

Keywords: machine learning, statistics, sustainable transportation, classification, logistic regression, random forests.

Liste des figures

Figure 1 : <i>Schéma de la démarche suivie dans le projet</i>	1
Figure 2: <i>Interface de prédiction des choix de transports</i>	22
Figure 2 : <i>Importance des Variables dans le Choix d'un Mode de Transport</i>	26

Liste des tableaux

Tableau 1 : <i>Description des Variables du Jeu de Données</i>	15
Tableau 2 : <i>Résultats du test du Chi carré pour les variables catégorielles</i>	18
Tableau 3 : <i>Résultats du test t de Student pour les variables continues</i>	19
Tableau 4 : <i>Coefficients de la régression logistique</i>	24

Table des matières

Résumé	1
Abstract	2
Liste des figures	3
Liste des Tableaux	3
Table des matières	4
1. Introduction	5
1.1 Contexte général	5
1.2 Problématique	5
1.3 Objectifs	6
1.4 Conduite du projet	6
1.5 Plan du mémoire	7
2. Contexte du projet	9
2.1 Présentation de la structure	9
2.2 Cahier des charges	9
3. Revue de la littérature	10
3.1 Principes scientifiques utilisés dans le projet	10
3.2 État des connaissances actuelles sur la problématique étudiée	10
3.3 Contribution de cette étude à notre problématique	11
4. Méthodologie et Moyens mis en œuvre	12
4.1 Choix des outils et technologies	12
4.2 Base de données utilisée	13
4.3 Stratégie	13
5. Travaux expérimentaux	15
5.1 Préparation des données	16
5.2 Modélisation	20
5.3 Visualisation des résultats	22
5.4 Solutions retenues	23
6. Résultats	24
7. Discussion	27
8. Conclusion et perspectives	28
9. Positionnement par rapport aux compétences de la formation IS	29
10 Références	30

Introduction

1.1 Contexte général

Face aux défis liés au changement climatique, la transition vers des modes de vie plus durables constitue un enjeu majeur pour les sociétés modernes. Cette transformation est indispensable pour limiter les impacts environnementaux, préserver les ressources naturelles et garantir un avenir viable pour les générations futures. Parmi les secteurs clés à transformer, les transports occupent une place centrale en raison de leur contribution significative aux émissions de gaz à effet de serre, à la pollution atmosphérique et à la consommation énergétique.

Dans ce contexte, il devient essentiel de comprendre les comportements de mobilité des habitants, afin de mieux cibler les actions en matière de transport durable. Ce projet se concentre sur la Métropole Européenne de Lille (MEL), une région urbaine dynamique regroupant plus d'un million d'habitants. Il s'agit d'analyser et de modéliser les comportements de déplacement des individus, en identifiant les facteurs qui influencent leurs choix de modes de transport.

Une meilleure compréhension de ces comportements nous conduit à l'élaboration de politiques publiques adaptées, visant à encourager l'adoption de pratiques de transport plus respectueuses de l'environnement.

1.2 Problématique

Les choix de transport des habitants de la Métropole Européenne de Lille sont influencés par une multitude de facteurs complexes et interdépendants, tels que les motifs du déplacement, et les spécificités des déplacements (durée, distance parcourue, nombre de trajets).

Dans ce cadre, la problématique posée est la suivante :

Quels sont les facteurs de déplacement qui influencent les choix de transport des habitants de la MEL, et comment peut-on prédire ces choix à l'aide de modèles de machine learning ?

1.3 Objectifs

L'objectif est de modéliser les choix de transport des habitants de la MEL en utilisant des modèles de classification, notamment la régression logistique et les forêts aléatoires, basés sur les données de l'enquête Ménage Déplacement (EMD) de 2016. Il s'agit d'identifier les facteurs de déplacement, tels que la distance parcourue, la durée du trajet, qui influencent ces choix. Ensuite, des recommandations seront élaborées pour promouvoir des solutions de mobilité durable.

1.4 Conduite du projet

Pour atteindre ces objectifs, nous avons suivi une approche méthodologique structurée en plusieurs étapes :

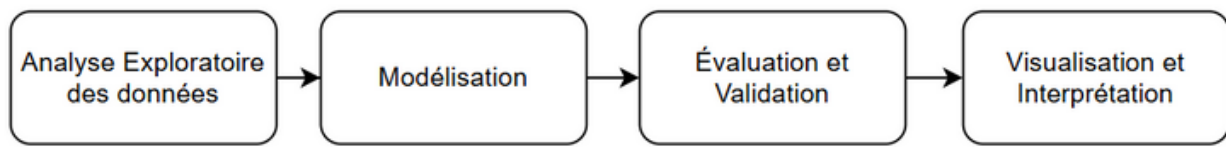


Figure 1 : Schéma de la démarche suivie dans le projet

- Analyse exploratoire : L'analyse exploratoire s'appuiera sur les données de l'EMD 2016 pour identifier les tendances principales et les variables clés via des visualisations et des tests statistiques.
- Modélisation : Implémentation et optimisation de deux modèles de classification (régression logistique et forêts aléatoires).
- Évaluation et validation : Analyse des performances des modèles à travers des métriques spécifiques, telles que la précision, le rappel et le F1-score.
- Visualisation : Création d'une interface interactive avec Streamlit pour visualiser les résultats.
- Interprétation : Analyse des résultats pour proposer des politiques de mobilité durable.

1.5 Plan du mémoire

Ce mémoire est structuré en 9 parties :

Introduction :

Situe le contexte du projet et présente clairement la problématique ainsi que la démarche adoptée pour traiter le sujet.

Section 2 : Contexte du projet

Dans cette section, nous présentons la structure à l'origine du sujet ainsi que le cahier des charges du projet.

Section 3 : Revue de la littérature

Cette section explore les principes scientifiques utilisés dans le projet. Elle propose également une revue de la littérature pour situer l'état des connaissances actuelles sur la problématique étudiée.

Section 4: Méthodologie et moyens mis en œuvre

Les méthodes employées pour résoudre la problématique sont détaillées dans cette section, ainsi que les moyens de mis en œuvre.

Section 5: Travaux Pratiques

Dans cette section, nous détaillons les différentes étapes pratiques pour réaliser le projet.

Section 6 : Résultats

Cette partie est consacrée à la présentation des résultats obtenus à travers les travaux menés. Ces résultats sont analysés et interprétés en lien avec les objectifs du projet.

Section 7 : Discussion

La discussion confronte les résultats obtenus aux objectifs initiaux, et met en lumière les difficultés rencontrées.

Section 8: Conclusion et perspectives

Cette dernière section récapitule les principaux résultats du projet et ouvre sur des pistes de réflexion ou d'amélioration pour des travaux futurs.

Section 9: Positionnement par rapport aux compétences de la formation IS

Conclusion

Dans ce premier chapitre, nous avons présenté le contexte général du projet en mettant en lumière les enjeux liés à la mobilité durable au sein de la Métropole Européenne de Lille. Nous avons défini la problématique et détaillé les objectifs principaux, ainsi que la conduite adoptée pour mener à bien ce travail, notamment l'analyse exploratoire, la modélisation et la visualisation des résultats.

Section 2 : Contexte du projet

2.1 Présentation de la structure à l'origine du sujet

Le projet est dirigé par CRISAL (Centre de Recherche en Informatique, Signal et Automatique de Lille), un laboratoire de recherche affilié au CNRS (Centre National de la Recherche Scientifique). Ce centre est reconnu pour ses recherches en intelligence artificielle, en traitement des données, en machine learning et en statistiques. CRISAL collabore avec de nombreux acteurs académiques, industriels et institutionnels

2.1 Cahier des charges

Le projet vise à modéliser les préférences de déplacement des habitants de la Métropole Européenne de Lille (MEL) en utilisant des techniques de machine learning et des méthodes statistiques. Les objectifs spécifiques sont les suivants:

1. Comprendre les comportements de mobilité des habitants de la MEL.
2. Identifier les facteurs clés influençant les choix de transport.
3. Prédire les préférences de transport à l'aide de modèles de classification (régression logistique, forêts aléatoires).
4. Développer une interface interactive pour visualiser les résultats.
5. Proposer des recommandations pour encourager l'adoption de modes de transport plus durables.

Livrables Attendus

- 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.

Section 3: Revue de la littérature

3.1 Principes scientifiques utilisés dans le projet

Dans le cadre de la modélisation des choix de transport, deux approches principales sont souvent employées :

- **Les modèles économétriques traditionnels**, comme la régression logistique, qui permettent de modéliser les préférences individuelles tout en offrant une interprétation causale des résultats.
- **Les modèles de machine learning**, comme les forêts aléatoires, qui sont adaptés à la prédiction et permettent de capturer des relations complexes et non linéaires dans les données.

Dans ce projet, nous avons combiné ces deux approches pour exploiter leurs forces respectives. La régression logistique a été utilisée pour son interprétabilité et sa capacité à expliquer les préférences individuelles, tandis que le modèle des forêts aléatoires a été implémenté pour maximiser la précision des prédictions.

3.2 État des connaissances actuelles sur la problématique étudiée

De nombreuses études se sont intéressées à la modélisation des choix de transport, à l'aide de techniques allant des modèles économétriques traditionnels aux méthodes modernes de machine learning. Une étude notable est celle de **Gusarov (2024)** intitulée *"Performance of econometric and Machine Learning models for the analysis of discrete choice data"*. Cette étude compare les performances des modèles économétriques et des techniques de machine learning dans l'analyse des choix individuels.

L'étude montre que les modèles de machine learning, tels que les forêts aléatoires, surpassent souvent les modèles économétriques traditionnels en termes de précision prédictive, notamment dans les contextes où les relations entre les variables explicatives et la variable cible sont complexes ou non linéaires. Cependant, les modèles économétriques conservent un avantage en termes d'interprétabilité et restent utiles pour des analyses nécessitant une compréhension causale des choix des individus.

Ces résultats trouvent un écho direct dans notre projet. En effet, lors de la comparaison des performances entre la régression logistique et les forêts aléatoires, nous avons constaté que ces dernières offraient une meilleure performance prédictive dans notre contexte, avec les données issues de l'Enquête Ménage Déplacement (EMD) 2016 de la Métropole Européenne de Lille. Cette observation confirme les conclusions de Gusarov, tout en les appliquant à un cas d'étude concret centré sur une métropole française.

3.3 Contribution de cette étude à notre problématique

L'étude de Gusarov (2024) a également mis en lumière l'intérêt de combiner les forces des deux approches. Dans notre projet, cela s'est traduit par une double utilisation des modèles :

- La régression logistique a servi à identifier et comprendre les facteurs déterminants dans les choix de transport des habitants, mettant en évidence un faible niveau d'engagement écologique (32 %).
- Les forêts aléatoires, quant à elles, ont permis d'obtenir des prédictions précises et fiables.

En combinant ces approches, notre travail s'aligne sur les recherches actuelles tout en apportant une contribution spécifique à l'étude des comportements de transport dans la Métropole Européenne de Lille.

Conclusion

La revue de la littérature a mis en lumière les deux principales approches pour modéliser les choix de transport : les modèles économétriques (comme la régression logistique), appréciés pour leur interprétabilité, et les modèles de machine learning (tels que les forêts aléatoires), reconnus pour leur précision dans la prédiction de relations complexes. Les travaux de Gusarov (2024) ont souligné l'intérêt de combiner ces approches pour exploiter leurs atouts respectifs.

À la suite de cette revue de la littérature, nous détaillerons dans la section suivante la méthodologie et les moyens mis en œuvre pour mener à bien cette étude. Cela inclura une présentation des données utilisées, des techniques d'analyse employées, ainsi que des justifications des choix méthodologiques opérés pour répondre à notre problématique.

Section 4 : Méthodologie et moyens mis en œuvre

4.1 Choix des outils et technologies

Pour réaliser ce projet, nous avons opté pour des outils et technologies adaptés à l'analyse des données et à la modélisation des choix de transport.

1. Language de programmation : Python

Python a été choisi pour sa flexibilité et la richesse de ses bibliothèques dédiées à l'analyse de données et au machine learning.

Les bibliothèques utilisées sont:

- **Pandas** : pour la manipulation et le prétraitement des données
- **NumPy** : pour les calculs numériques
- **Scikit-Learn** : pour la construction des modèles de classification
- **Matplotlib et Seaborn** : pour la visualisation des données et les résultats
- **Streamlit** : pour créer une interface utilisateur interactive permettant de visualiser les résultats des modèles et d'explorer les prédictions.

2. MLflow : Gestion des expériences de machine learning

MLflow est une plateforme open-source qui permet de gérer le cycle de vie des modèles de machine learning, de leur création à leur déploiement. Nous l'avons utilisé pour suivre les expérimentations, gérer les modèles et optimiser les performances des différents modèles de classification.

3. Environnement de développement : Google Colab

Google Colab est un environnement de développement cloud-based qui permet d'écrire et d'exécuter du code Python dans un notebook interactif. Nous l'avons utilisé pour plusieurs avantages, notamment l'accès gratuit à des ressources matérielles puissantes, telles que les GPU, et la facilité de partage et de collaboration

4.2 Base de données utilisée

Le jeu de données utilisé dans ce projet, "Enquête Ménage Déplacement (EMD) 2016", fournit une vue détaillée des comportements de mobilité dans la Métropole Lilloise en 2016.

Il comprend une variété de données concernant les trajets effectués, telles que la distance parcourue, la durée des déplacements, ainsi que les modes de transport utilisés. Ces données permettent une analyse approfondie des choix de transport des habitants de la région.

Dans le cadre de cette étude, l'accent est mis sur le traitement des données issues des fichiers Déplacement et Trajet.

4.3 Stratégie

Notre démarche s'articule en trois phases principales : **préparation des données, modélisation et interprétation**. Cette structure permet une progression logique.

I. Préparation des données

Cette phase vise à transformer les données brutes de l'enquête Ménage Déplacement (EMD) 2016 en un format exploitable pour la modélisation. Elle comprend trois étapes clés :

1) Nettoyage des données :

- gestion des valeurs manquantes (par imputation ou suppression).
- Détection et traitement des valeurs aberrantes (outliers) pour éviter les biais dans les modèles.

2) Transformation des données :

- Encodage des variables catégorielles.
- Normalisation des variables numériques : Standardisation des variables telles que la distance parcourue pour garantir une échelle comparable.
- Etude des corrélations

3) Division des données:

- Division des données en ensemble d'entraînement et ensemble de test pour évaluer la généralisation des modèles.

II. Modélisation

Cette phase a pour objectif de construire et d'optimiser des modèles de classification afin de prédire le choix du mode de transport. Plusieurs modèles ont été envisagés, notamment la régression logistique et les forêts aléatoires, chacun présentant des avantages spécifiques selon les caractéristiques des données. L'objectif est de sélectionner les modèles les plus performants, capables de capturer efficacement les relations entre les variables explicatives et la variable cible.

III. Visualisation

Cette phase est consacrée au développement d'une interface interactive avec Streamlit, permettant aux utilisateurs d'explorer dynamiquement les résultats des modèles de classification et de visualiser les prédictions de manière intuitive.

Conclusion

Notre méthodologie, structurée en trois phases (préparation des données, modélisation, visualisation), permet de transformer les données de l'EMD 2016 en insights stratégiques pour la mobilité durable. Grâce à des outils modernes (Python, Scikit-learn, Streamlit, MLflow), on peut nettoyer, modéliser et analyser les choix de transport, en mettant en avant des modèles interprétables et performants. Cette approche rigoureuse et reproductible offre une base solide pour des recommandations concrètes adaptées aux besoins de la MEL.

Section 5 : Travaux expérimentaux

Cette section décrit les actions concrètes réalisées dans le cadre du projet, en suivant la méthodologie présentée précédemment. Elle détaille les étapes de préparation des données, de modélisation et d'interprétation.

Le tableau suivant présente les variables du jeu de données utilisé:

Variable	Description	Type
D5AA	Motif à la destination du déplacement	Catégorique
D2AA	Motif à l'origine du déplacement	Catégorique
DGRP	Groupes à comportement homogène d'âge	Catégorique
Speed	Vitesse du déplacement en km/h	Continue
D8C	Durée de déplacement en minutes	Continue
DIST	Distance parcourue en mètres	Continue
D4A	Heure du départ	Continue
MODP	Mode de transport (car, bike, walk, pt)	Catégorique

Tableau 1 : Description des Variables du Jeu de Données

5.1 Préparation des données

Traitement des valeurs aberrantes:

Lors de l'analyse du jeu de données, nous avons identifié des valeurs aberrantes dans plusieurs variables, notamment Durée de déplacement (D8C), Distance parcourue (DIST) et Heure de départ (D4A).

Pour détecter et éliminer ces valeurs aberrantes, nous avons utilisé l'Intervalle Interquartile (IQR).

L'IQR mesure la dispersion des données en prenant la différence entre le troisième quartile (Q3) et le premier quartile (Q1). Il permet ainsi de définir un intervalle dans lequel la majorité des données se situent. L'écart interquartile (IQR) est calculé comme suit : $IQR = Q3 - Q1$

Les valeurs considérées comme aberrantes sont celles qui se situent en dehors de l'intervalle défini par : $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$

Cette méthode a été choisie car elle présente plusieurs avantages:

- Robustesse face aux valeurs aberrantes:

Contrairement à des mesures telles que la moyenne ou l'écart-type, l'IQR repose uniquement sur les quartiles (Q1 et Q3) et n'est pas influencé par les valeurs extrêmes. Cela permet une identification plus fiable des anomalies sans être biaisée par des valeurs exceptionnelles.

- Adaptation à différentes distributions de données:

L'IQR ne nécessite pas que les données suivent une distribution spécifique, comme une distribution normale. En étant non paramétrique, il s'applique à tous types de distributions.

Encodage des variables catégorielles

L'encodage des variables catégorielles est une étape essentielle dans le pré-traitement des données avant l'entraînement des modèles de machine learning. Les modèles, comme les forêts aléatoires ou la régression logistique, ne peuvent généralement pas traiter directement les données sous forme de texte ou de catégories non numériques.

La variable DGRP, qui représente les groupes d'âge, a été encodée en 3 catégories distinctes :

- Inactifs : 0 , Retraités : 1, Actifs : 2

D5AA et D2AA (Motifs du déplacement à destination et à l'origine) : Ces deux variables catégorielles ont été encodées en 5 catégories chacune :

- home : 1 , work : 2 , education : 3 , shop : 4 , leisure : 5

L'encodage de ces catégories permet de transformer les différents motifs de déplacement en valeurs numériques qui peuvent être utilisées pour prédire les choix de transport.

Normalisation des variables continues

Pour les variables continues, telles que Speed, D8C, DIST, et D4A, nous avons utilisé la méthode de normalisation par MinMaxScaler. Cette méthode permet de transformer les valeurs des variables en un intervalle compris entre 0 et 1, ce qui facilite leur traitement par les modèles de machine learning en les mettant sur une échelle comparable.

La normalisation a été réalisée de manière à préserver la distribution relative des données tout en assurant qu'aucune variable n'influence de manière disproportionnée le modèle en raison de son échelle initiale.

Variable Cible et Explicatives

Dans le cadre de cette modélisation, l'objectif est de prédire le choix du moyen de transport. La variable cible est ECO, qui a été construite à partir de MODP. Ainsi, ECO est une variable binaire où la valeur 0 représente l'usage de la voiture ("car") et la valeur 1 représente l'utilisation des autres modes de transport, à savoir le vélo ("bike"), la marche ("walk") et les transports publics ("pt"). Les autres variables du jeu de données servent de variables explicatives, utilisées pour expliquer et prédire la valeur de la variable cible ECO.

Etude des corrélations:

L'analyse des corrélations entre les variables explicatives et la variable cible (ECO) est une étape clé pour mieux comprendre les relations dans le jeu de données. Pour ce faire, des tests statistiques appropriés ont été réalisés en fonction du type de variable.

Variables Catégorielles:

Le test du Chi carré a été utilisé pour évaluer la corrélation entre les variables catégorielles explicatives et la variable cible. Ce test permet de mesurer l'association entre deux variables qualitatives afin de déterminer si elles sont indépendantes ou statistiquement liées.

Il compare les fréquences observées dans les données aux fréquences attendues sous l'hypothèse d'indépendance entre les variables. Si les fréquences observées diffèrent significativement des fréquences attendues, on rejette l'hypothèse d'indépendance et on conclut qu'il existe une association entre les variables.

L'hypothèse nulle (H0) du test stipule qu'il n'existe aucune relation entre les variables, tandis que l'hypothèse alternative (H1) postule une association significative.

Toutefois, le test du Chi carré ne permet pas d'évaluer la force de l'association détectée. Pour pallier cette limite, le V de Cramer a été calculé.

Les résultats obtenus sont présentés dans le tableau suivant :

Variable	P-value	V de Cramer	Association
D5AA	< 0.0001	0.179	+
D2AA	< 0.0001	0.185	+
AgeGroup	< 0.0001	0.099	+

Tableau 2 : Résultats du test du Chi carré pour les variables catégorielles

L'analyse des résultats montre que toutes les p-values sont inférieures à 0.05, indiquant que toutes les variables catégorielles étudiées sont significativement associées à la variable cible

En termes d'intensité de l'association, les variables D5AA et D2AA présentent des associations modérées avec le mode de transport, tandis que DGRP présente une association plus faible.

Variables Continues:

Pour les variables continues, nous avons utilisé le test t de Student afin d'évaluer l'existence de différences significatives entre les groupes définis par la variable cible.

Le test t est une méthode statistique utilisée pour comparer les moyennes d'une variable quantitative entre deux groupes indépendants. Il permet de déterminer si une différence observée entre les moyennes est statistiquement significative ou simplement due au hasard.

L'hypothèse nulle (H_0) du test stipule que les moyennes des deux groupes sont égales, tandis que l'hypothèse alternative (H_1) postule une différence significative entre ces moyennes.

Les résultats obtenus sont présentés dans le tableau suivant :

Variable	t-stat	P-value	Interprétation
Speed	366.37	< 0.0001	Différence significative
D8C	-5.23	< 0.0001	Différence significative
DIST	260.02	< 0.0001	Différence significative
D4A	34.10	< 0.0001	Différence significative

Tableau 3 : Résultats du test t de Student pour les variables continues

Les résultats du test t de Student indiquent des différences significatives entre les groupes de la variable cible pour toutes les variables continues analysées ($p < 0.05$).

Ces résultats suggèrent que des facteurs tels que la vitesse moyenne, la distance parcourue, la durée du déplacement et l'heure de départ jouent un rôle important dans le choix du mode de transport. Ces observations seront essentielles pour l'optimisation des modèles de classification et l'interprétation des facteurs influençant les préférences de transport.

Division des données

Avant d'entamer la partie modélisation, les données ont été divisées en deux ensembles distincts : 80% pour l'entraînement et 20% pour le test. Cette division permet d'entraîner les modèles sur une grande proportion des données tout en préservant un échantillon indépendant, utilisé pour évaluer la performance des modèles et leur capacité à généraliser sur de nouvelles données.

5.2 Modélisation

Cette phase consiste à construire et optimiser des modèles de classification pour prédire le choix du mode de transport. Deux algorithmes ont été retenus, chacun pour des raisons spécifiques liées à la nature des données et à l'objectif de modélisation.

1 Régression Logistique

La régression logistique a été choisie principalement en raison de son **interprétabilité**. Elle permet de quantifier l'impact de chaque variable explicative sur la probabilité d'appartenance à l'une des deux classes de ECO (0 ou 1), ce qui facilite l'interprétation des résultats. Ce modèle est particulièrement adapté lorsque l'on cherche à comprendre la relation directe entre les variables indépendantes et la variable cible. Il offre également une bonne performance lorsque les relations sont relativement simples et que les données sont bien séparées, avec peu de bruit.

2 Forêts aléatoires

Les forêts aléatoires ont été sélectionnées pour leur capacité à gérer des relations non linéaires complexes et des interactions multiples entre les variables. Ce modèle est particulièrement efficace pour capturer des patterns difficiles à appréhender avec une régression logistique simple. De plus, il présente l'avantage d'être très robuste aux déséquilibres de classes, un problème fréquent dans de nombreux jeux de données réels. Les forêts aléatoires fonctionnent bien même avec des données hétérogènes et peuvent gérer un grand nombre de variables explicatives sans surajuster le modèle.

Les forêts aléatoires utilisent la variable cible MODP pour prédire le mode de transport, prenant ainsi en compte l'ensemble des choix possibles de transport.

Métriques d'Évaluation

Les modèles ont été évalués en utilisant des métriques de performance telles que précision, rappel, et F1-score. Ces métriques permettent de mesurer la capacité des modèles à prédire correctement le mode de transport choisi en fonction des variables explicatives.

Les résultats obtenus montrent des performances satisfaisantes pour les deux modèles :

- Pour la regression logistique:

Précision (0.89) : Le modèle est fiable pour ses prédictions positives, car 89% des prédictions positives sont correctes.

Rappel (0.88) : Le modèle détecte correctement 88% des cas où l'instance est effectivement positive.

F1-score (0.87) : Le modèle atteint un bon compromis entre précision et rappel, ce qui est particulièrement utile lorsque les classes sont déséquilibrées

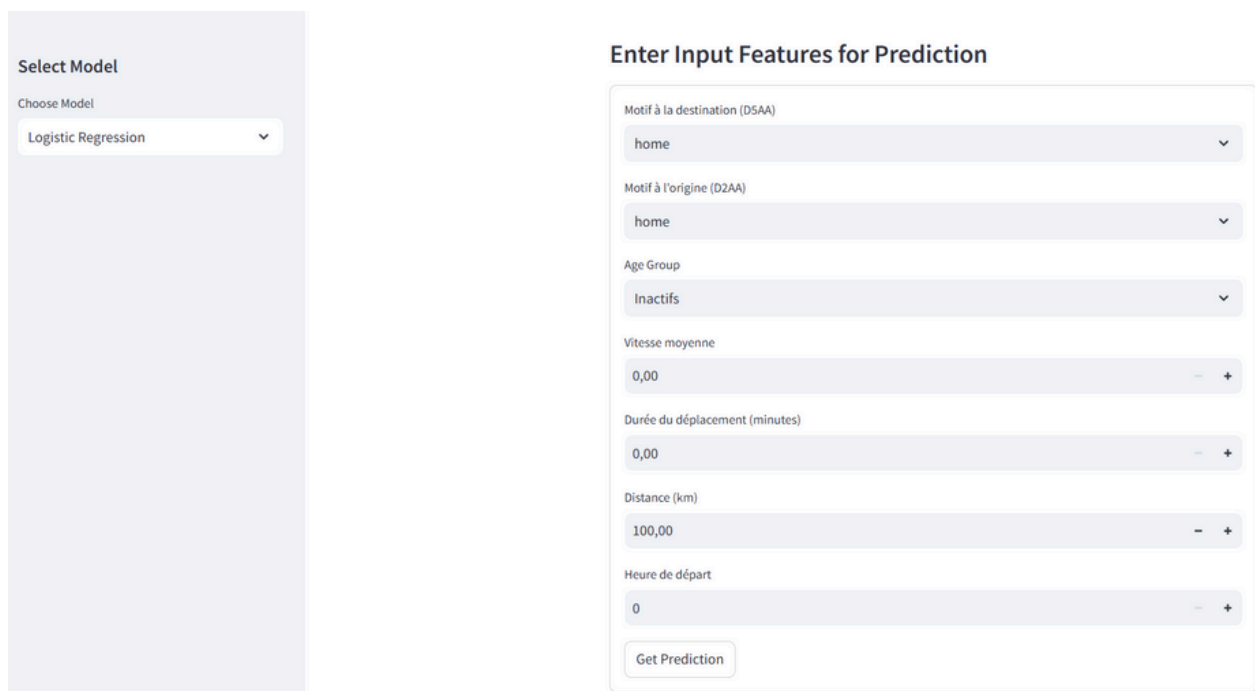
- Pour Random Forest:

Précision (0.92) : Le modèle des forêts aléatoires est fiable lorsqu'il prédit la classe positive, avec seulement 8% d'erreurs de classification positives.

Rappel (0.94) : Le modèle est très performant pour détecter presque toutes les instances positives, avec une légère marge d'erreur.

F1-score (0.93) : Cela montre que le modèle atteint un bon compromis entre les faux positifs et les faux négatifs, ce qui est crucial, notamment en cas de classes déséquilibrées.

5.3 Visualisation



The image shows a Streamlit web application interface for predicting transport choices. It is divided into two main sections. The left section, titled 'Select Model', contains a 'Choose Model' dropdown menu with 'Logistic Regression' selected. The right section, titled 'Enter Input Features for Prediction', contains several input fields: 'Motif à la destination (DSAA)' and 'Motif à l'origine (D2AA)' are dropdown menus both set to 'home'; 'Age Group' is a dropdown menu set to 'Inactifs'; 'Vitesse moyenne', 'Durée du déplacement (minutes)', 'Distance (km)', and 'Heure de départ' are numeric input fields with values 0,00, 0,00, 100,00, and 0 respectively. Each numeric field has minus and plus buttons for adjustment. At the bottom of the right section is a 'Get Prediction' button.

Figure 3: Interface de prédiction des choix de transports

Cette interface interactive, développée avec Streamlit, permet d'explorer et de comparer deux modèles de classification : la Régression Logistique et la Forêt Aléatoire (Random Forest). Son objectif est de faciliter la sélection de modèles, la configuration des variables explicatives, et l'obtention de prédictions tout en offrant une visualisation claire des résultats.

5.4 Solutions retenues

À l'issue des expérimentations, la comparaison des modèles a montré que les forêts aléatoires surpassent la régression logistique en termes de performance prédictive. Cependant, la régression logistique reste utile pour l'interprétation des facteurs influençant les choix de transport, permettant d'identifier les variables les plus significatives. Ainsi, nous allons utiliser les deux modèles de manière complémentaire afin de répondre à notre problématique : les forêts aléatoires pour obtenir des prédictions précises et la régression logistique pour analyser et interpréter les facteurs déterminants dans le choix du mode de transport écologiques.

Conclusion

Cette section a permis de détailler l'ensemble des étapes mises en œuvre pour modéliser les choix de transport des habitants. De la préparation des données à l'entraînement des modèles, en passant par l'optimisation et l'évaluation des performances, chaque étape a été essentielle pour garantir la robustesse des résultats.

Section 6 : Résultats

Dans cette section, nous présentons les résultats obtenus à travers les modèles de classification développés dans le cadre de ce projet. Nous débutons par l'analyse des coefficients de la régression logistique, qui permet d'interpréter l'impact des différentes variables explicatives sur le choix du mode de transport. Ces coefficients reflètent l'effet de chaque variable sur la probabilité que l'observation appartienne à l'une des deux classes cibles, ici le choix entre un mode de transport écologique (1) et non écologique (0)

Variable	Coefficient
Constante	3.7591
D5AA (Motif à destination)	-0.1232
D2AA (Motif à l'origine)	-0.1273
AgeGroup (Groupe d'âge)	-0.0944
speed (Vitesse du déplacement)	-311.1727
D8C (Durée du déplacement)	-2.2303
DIST_km (Distance parcourue)	7.5893
D4A (Heure de départ)	-0.8425

Tableau 4 : Coefficients de la régression logistique

D'après les résultats présentés dans le Tableau 4, nous pouvons distinguer deux catégories de facteurs influençant le choix d'un mode de transport écologique : ceux qui réduisent son utilisation et ceux qui le favorisent.

Parmi les facteurs qui **diminuent** la probabilité d'adopter un transport écologique, nous retrouvons principalement:

- La vitesse du déplacement (speed), qui a le coefficient négatif le plus élevé, Cela indique que les trajets effectués à des vitesses élevées sont fortement associés aux modes de transport non écologiques (voiture).
- La durée du déplacement (D8C) a également un impact négatif, suggérant que les trajets longs en temps incitent à privilégier des moyens de transport plus rapides et souvent moins durables.

- L'heure de départ (D4A) joue un rôle défavorable : plus l'heure de départ est tardive, moins les individus optent pour un mode de transport écologique.
- L'âge (AgeGroup) joue aussi un rôle négatif : les personnes plus âgées ont moins tendance à adopter des modes de transport écologiques, peut-être par confort.

À l'inverse, d'autres facteurs **favorisent** l'utilisation d'un mode de transport écologique. notamment la Distance parcourue (DIST_km), dont le coefficient positif indique que les trajets plus longs en distance sont plus susceptibles d'être réalisés avec un mode de transport durable, comme les transports en commun.

En résumé, les transports écologiques sont privilégiés pour les longs trajets en distance, mais moins adoptés pour les déplacements rapides, tardifs et de longue durée.

Dans le cadre de la régression logistique, la probabilité d'un événement (ici, le choix écologique des usagers) est calculée à partir des coefficients estimés du modèle selon la formule de la fonction logistique, aussi appelée sigmoïde.

La formule de la probabilité est la suivante :

$$p = \frac{1}{1 + e^{-z}}$$

z est la somme des produits des variables explicatives par leurs coefficients.

Dans notre analyse, après avoir estimé les coefficients du modèle à partir des données, nous avons calculé les probabilités de choisir un mode de transport écologique pour chaque individu. La probabilité moyenne est de **0.32**. Cela signifie qu'en moyenne, les habitants de la MEL ont 32% de chances de choisir un mode de transport écologique, en fonction des facteurs analysés.

Ce résultat met en lumière un niveau d'adoption relativement faible des modes de transport durables dans la région. Cela suggère qu'il existe un potentiel important pour encourager un changement vers des pratiques de déplacement plus écologiques.

L'interprétation des résultats de **Random Forest** repose sur l'analyse des importances des variables, qui montrent quelles caractéristiques ont le plus d'influence dans la prédiction du choix du mode de transport.

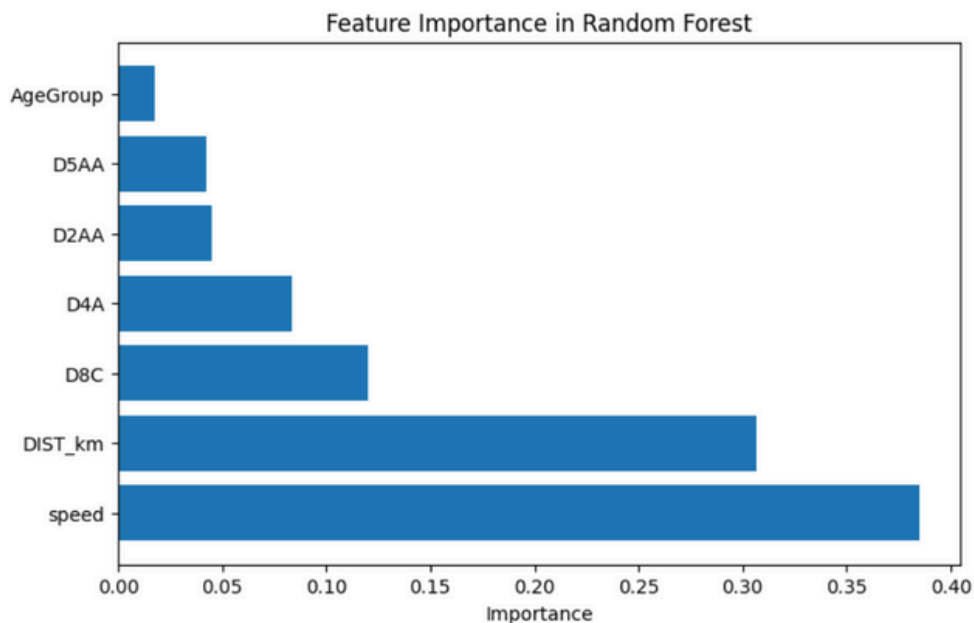


Figure 3 : Importance des Variables dans le Choix d'un Mode de Transport

Les résultats du Figure 2 montrent que les variables les plus influentes sur le choix du mode de transport écologique sont la vitesse et la distance parcourue et la durée, avec une influence plus faible des autres facteurs comme , l'âge et les motifs de déplacement. Ces résultats confirment les conclusions tirées de la régression logistique.

Conclusion

L'analyse des modèles de classification montre que la vitesse, la durée et l'heure de départ réduisent l'adoption des transports écologiques, tandis que la distance parcourue la favorise. Le score moyen de 32 % reflète un faible recours aux modes durables dans la MEL. Les résultats de Random Forest confirment ces tendances, soulignant l'importance de la vitesse et de la distance. Ces conclusions mettent en évidence la nécessité de mesures pour encourager des alternatives de transport plus durables.

Section 7 : Discussion

Au cours de ce projet, plusieurs défis ont été identifiés, influençant à la fois la qualité des résultats et la généralisation des modèles de classification.

1 Données Anciennes:

La base de données utilisée pour entraîner et évaluer les modèles provient de l'enquête de 2016 sur les déplacements. Bien que ces données fournissent une base solide pour l'analyse, elles ne reflètent pas nécessairement les tendances actuelles en matière de mobilité. Par exemple, les habitudes de déplacement ont pu évoluer en raison de facteurs tels que le télétravail, les changements dans les infrastructures de transport, ou encore les préoccupations environnementales accrues. Cette limitation temporelle peut affecter la pertinence des prédictions, notamment si les modèles sont appliqués à des contextes plus récents. Pour pallier ce problème, il serait essentiel de mettre à jour la base de données avec des informations plus récentes.

2 Déséquilibre des Classes:

Un autre défi réside dans le déséquilibre des classes au sein du jeu de données. En effet, la catégorie "voiture" représente près de 70 % des observations, tandis que les autres modes de transport (ex. transports en commun, vélo, marche) sont sous-représentés. Ce déséquilibre peut biaiser les performances des modèles de classification, car ceux-ci ont tendance à privilégier la classe majoritaire pour maximiser la précision globale, au détriment des classes minoritaires.

Section 8 : Conclusion et perspectives

Au cœur de la Métropole Européenne de Lille, notre exploration de la mobilité durable a dévoilé des insights précieux. En combinant la clarté des modèles économétriques avec la puissance prédictive du machine learning, nous avons pu identifier les facteurs qui influencent les choix de transport des habitants.

Nos analyses révèlent que la vitesse, la durée du trajet et l'heure de départ sont des freins majeurs à l'adoption des modes de transport écologiques. Cependant, une distance parcourue plus longue encourage le recours à ces alternatives durables. Ce constat met en lumière un défi important : avec seulement 32 % d'utilisation des modes de transport durables.

Grâce à une méthodologie rigoureuse et des outils innovants comme Python et Scikit-learn, nous avons conçu des modèles interprétables et performants qui peuvent guider des actions concrètes pour promouvoir l'utilisation des modes de transport plus respectueuses de l'environnement.

Ainsi, pour encourager l'adoption des modes de transport durables au sein de la Métropole Européenne de Lille, plusieurs mesures concrètes peuvent être mises en place :

1. Améliorer la compétitivité des modes écologiques : Optimiser les réseaux de transports publics en proposant des lignes plus rapides et plus directes afin de rendre ces modes de transport plus attractifs.
2. Développer l'usage du vélo : Déployer des vélos en libre-service dans des zones stratégiques pour encourager les habitants à utiliser cette alternative écologique.
3. Adapter les horaires des transports écologiques : Étendre les heures de fonctionnement des transports publics, notamment en zones urbaines et rurales, pour mieux répondre aux besoins des usagers.

Section 9 : Positionnement par rapport aux compétences de la formation IS

Compétences Métiers

Je me positionne en tant que Data Analyst car mon travail se concentre principalement sur :

IS4.1 : Identifier et mobiliser des connaissances scientifiques et techniques

- Utilisation de méthodes statistiques avancées et de programmation (Python, pandas, scikit-learn) pour modéliser les données de déplacements.

IS4.2 : Comprendre la structure des données à analyser

- Analyse de la structure des données de déplacements.
- Identification des variables pertinentes (vitesse, motifs, groupes d'âge, etc.).

IS4.3 : Décrire les données avec des statistiques descriptives et analyser en grande dimension

- Utilisation d'indicateurs numériques (moyennes, écarts-types) et de visualisations (matrices de corrélation, histogrammes) pour décrire les données.

IS4.4 : Construire et interpréter un modèle d'apprentissage

- Construction d'un modèle de régression logistique pour prédire la variable cible (ECO).
- Développement d'un modèle de Random Forest pour améliorer la précision des prédictions

Références:

- Gusarov, N. (2024). *Performance of econometric and machine learning models for the economic study of discrete consumer choices*. Economics and Finance, Université Grenoble Alpes.
- Hunault, G. (n.d.). *Pratique de la régression logistique*.
- Genuer, R., & Poggi, J.-M. (n.d.). *Les forêts aléatoires avec R*. Presses Universitaires de Rennes.
- Foucart, T. (1991). *Introduction aux tests statistiques*.