



Titre: Optimizing Machine Learning Techniques to Better Understand Sustainability in Transportation Planning
Title:

Auteur: Hamed Naseri
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Naseri, H. (2023). Optimizing Machine Learning Techniques to Better Understand Sustainability in Transportation Planning [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. <https://publications.polymtl.ca/54849/>
Citation:

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/54849/>
PolyPublie URL:

Directeurs de recherche: Owen Waygood, & Zachary Patterson
Advisors:

Programme: Génie civil
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Optimization in Machine Learning Techniques to Better Understand
Sustainability in Transportation Planning**

HAMED NASERI

Département des génies civil, géologique et des mines

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie civil

Août 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Optimizing Machine Learning Techniques to Better Understand Sustainability in Transportation Planning

présentée par **Hamed NASERI**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Martin TREPANIER, président

Owen WAYGOOD, membre et directeur de recherche

Zachary PATTERSON, membre et codirecteur de recherche

Francesco CIARI, membre

Nizar BOUGUILA, membre externe

ACKNOWLEDGEMENTS

This dissertation was supported by:

- (1): Trottier Institute of Energy, Ph.D. Excellence Scholarship.
- (2): FRQNT funding for the GHG project (grant number 322727).
- (3) SSHRC funding for the children's research (grant number 435-2020-1292).

RÉSUMÉ

Le secteur des transports a une incidence importante sur le développement social, environnemental et économique. Ce secteur est une composante de la lutte contre le changement climatique puisqu'il représente environ 25 % de la consommation d'énergies fossiles, un tiers de la consommation mondiale d'énergie et 23 % des émissions mondiales de gaz à effet de serre (GES). Par conséquent, il est essentiel d'étudier des méthodes capables d'améliorer la durabilité dans le secteur des transports. Par le passé, l'analyse des enquêtes de déplacement a principalement été réalisée à l'aide de méthodes statistiques conventionnelles. Cependant, de nouvelles approches d'intelligence artificielle (IA) permettent aujourd'hui d'améliorer la précision de ces analyses. Néanmoins, ces approches de modélisations et d'analyse par IA présentent certaines lacunes :

- Leur précision est sensible aux variables inappropriées utilisées dans la modélisation.
- Elles sont de type boîte noire, ce qui rend leur interprétation difficile, voire impossible, pour les décideurs et les planificateurs.
- Elles sont très sensibles aux « hyperparamètres » qui doivent être choisis par l'analyste/modélisateur. Les techniques conventionnelles de réglage des « hyperparamètres » (c'est-à-dire les techniques de régularisation) ne peuvent pas optimiser les hyperparamètres, et il n'y a aucune garantie qu'elles puissent détecter les meilleures valeurs pour les hyperparamètres.
- Elles ont besoin d'observations de données supplémentaires, appelées données de test, pour évaluer la puissance de prédiction du modèle, mais le pourcentage optimal de données de test à utiliser n'est pas clair.

Pour résoudre ces problèmes, cette thèse propose des méthodes pour surmonter ces faiblesses dans le contexte des comportements de mobilité des individus et de leur lien avec la durabilité.

Une autre question abordée dans la recherche sur les comportements de mobilité est le grand nombre de variables explicatives potentielles (ou caractéristiques) qui sont, peuvent ou ont été prises en compte dans l'analyse des comportements de déplacement. Une technique couramment utilisée dans d'autres domaines de recherche dans de telles circonstances est l'application de techniques de sélection de variables. Cependant, l'application de ces techniques n'a pas reçu beaucoup d'attention dans l'étude des comportements de transport. De plus, les techniques conventionnelles de sélection de variables ne peuvent pas obtenir la précision de prédiction la plus

élevée, et la plupart d'entre elles ne peuvent pas présenter un ensemble de variables optimales. Pour pallier ces lacunes, une nouvelle méthode de sélection des variables est développée pour optimiser la sélection de celles employées dans les modèles de prédiction. Ces processus peuvent augmenter considérablement la précision du modèle et éviter de présenter des solutions inexactes.

Ensuite, les techniques d'apprentissage automatique contiennent certains paramètres, appelés hyperparamètres, qui doivent être réglés. Les performances des techniques d'apprentissage automatique sont très sensibles à ces paramètres. Par conséquent, une nouvelle méthode est introduite pour ajuster les hyperparamètres en tenant compte de plusieurs indicateurs de performance à l'aide d'un cadre d'optimisation métaheuristique multiobjectif. Par la suite, une question pratique liée à la meilleure façon de diviser les données de formation et de test est examinée. Une nouvelle technique basée sur un algorithme d'optimisation métaheuristique est développée qui détermine automatiquement la division qui conduit à la plus grande précision de prédiction.

Ces techniques avancées sont appliquées à différentes questions de comportement de déplacement durable. En tant que tel, le stade de changement du changement climatique individuel (une variable ordinale) est évalué en fonction de différentes caractéristiques du comportement lié au transport. Les résultats suggèrent que l'application de techniques de sélection de caractéristiques est essentielle, en particulier lorsque le nombre de variables indépendantes est élevé. De plus, l'application du comportement de transport pour prédire le stade de changement du changement climatique sont des questions liées au transport. Ensuite, un autre modèle est développé pour prédire les choix de mode de déplacement des enfants (une variable catégorique) et détecter comment différentes variables influencent les choix de mode des enfants. Les résultats montrent que le réglage des hyperparamètres avec des techniques multi-objectifs surpasse les techniques conventionnelles pour régler les hyperparamètres lors de la comparaison de la précision de la prédiction, en particulier dans les ensembles de données déséquilibrés (c'est-à-dire le choix du mode enfant). De plus, la distance parcourue, le score de marche, l'âge, le score de vélo, le revenu du ménage et l'accessibilité aux écoles secondaires sont les variables les mieux classées dans le choix du mode enfant. Enfin, un modèle de prédiction est développé pour prédire qui préférera les véhicules électriques aux véhicules à moteur à combustion interne (une variable binaire) et pour savoir ce qui influence le choix d'un véhicule électrique par rapport à un véhicule

à moteur à combustion interne traditionnel. Les résultats ont montré que les personnes qui se sont identifiées comme étant au sommet du changement climatique-étape du changement (CC-SoC4 & 5), les Québécois français, les personnes âgées de plus de 58 ans et celles qui avaient obtenu un doctorat (par exemple, Ph. .D.), ou un diplôme en médecine, en dentisterie, en médecine vétérinaire ou en optométrie avaient la plus haute intention d'acheter des véhicules électriques.

ABSTRACT

The transportation sector significantly affects social, environmental, and economic development. This sector is a component in trying to combat climate change as it represents roughly 25% of fossil fuel consumption, a third of global energy consumption, and 23% of worldwide greenhouse gas (GHG) emissions. Hence, it is vital to investigate methods capable of enhancing sustainability in the transportation sector. Previous investigations on travel behavior have mostly used conventional statistical analysis. However, new Artificial Intelligence (AI) approaches exist that could significantly improve the accuracy of such analysis. Nonetheless, these AI techniques have some deficiencies:

- Their accuracy is sensitive to inappropriate variables being used in the modeling;
- They are often black-box tools and hard to interpret, making it difficult or impossible for policymakers and planners to interpret them;
- They are highly sensitive to “hyperparameters” that must be chosen by the analyst/modeler. The conventional techniques for tuning “hyperparameters” (i.e., regularization techniques) can not optimize hyperparameters, and there is no guarantee that they can detect the best values for hyperparameters;
- They need additional data observations, called testing data, to evaluate the prediction power of the model, but it is not clear what is the optimal percentage of testing data.

To address these issues, this thesis proposes methods to overcome these weaknesses in the context of individual travel behavior choices and how they relate to sustainability.

To address the problem of inappropriate variables, a common technique used outside of transportation research in such circumstances is the application of feature selection techniques. However, the application of feature selection techniques has not received much attention in transportation studies. Conventional feature selection techniques cannot obtain the highest prediction accuracy, and most of them cannot present a set of optimal features. Accordingly, first, a new feature selection method is developed to optimize the features employed in prediction models. The results of this study showed that these processes increased model accuracy considerably and avoid presenting inaccurate predictions.

Then, machine learning techniques contain some parameters, called hyperparameters, that should be tuned. The performance of machine learning techniques is highly sensitive to these parameters. Therefore, a new method is introduced to tune hyperparameters considering multiple performance indicators using a multi-objective metaheuristic optimization framework. Subsequently, a practical question related to how best to split the training and testing data is examined. A new technique based on a metaheuristic optimization algorithm is developed that automatically determines the split that leads to the greatest prediction accuracy.

These advanced techniques are applied to different sustainable travel behavior questions. As such, individual Climate Change Stage of Change (an ordinal variable) is evaluated based on different features of transport-related behavior. The results suggest that the application of feature selection techniques is vital, particularly when the number of independent variables is high. Moreover, the application of transportation behavior to predict the Change Stage of Change is vital since 32.6% of optimal variables to predict the Climate Change Stage of Change are transportation-related questions. Then, another model is developed to predict children's travel mode choices (a categorical variable) and detect how different variables influence children's mode choices. The results show that tuning hyperparameters with multi-objective techniques outperforms the conventional techniques to tune hyperparameters when comparing the prediction accuracy, particularly in imbalanced datasets (i.e., child mode choice). With respect to predicting children's mode choice, trip distance, walk score, age, bike score, household income, and accessibility to secondary schools are the top-ranked variables in child mode choice. Finally, a prediction model is developed to predict who prefers electric vehicles over internal combustion engine vehicles (a binary variable) and to know what influences the choice of an electric vehicle over a traditional internal combustion engine vehicle. The results showed that individuals who self-identified as being at the top of the Climate Change-Stage of Change (CC-SoC4 & 5), French Quebecers, those older than 58, and those who had completed a doctoral degree (e.g., Ph.D.), or a degree in medicine, dentistry, veterinary medicine, or optometry had the highest intention to buy EVs.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	III
RÉSUMÉ.....	IV
ABSTRACT	VII
TABLE OF CONTENTS	IX
LIST OF TABLES	XII
LIST OF FIGURES.....	XIII
LIST OF SYMBOLS AND ABBREVIATIONS.....	XV
LIST OF APPENDICES	XVII
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 CRITICAL OVERVIEW OF THE LITERATURE AND RESEARCH NEEDS	3
CHAPTER 3 OVERALL RESEARCH APPROACH	9
CHAPTER 4 ARTICLE 1: A NOVEL FEATURE SELECTION TECHNIQUE TO BETTER PREDICT CLIMATE CHANGE STAGE OF CHANGE	13
4.1 Introduction	15
4.2 Literature review.....	17
4.2.1 Research contributions	20
4.3 Methodology.....	21
4.3.1 Data preparation	22
4.3.2 Classification Techniques	26
4.3.3 Feature selection process.....	28
4.3.4 Sensitivity Analysis.....	36
4.4 Results and discussions	36

4.4.1	Optimal number of features.....	36
4.4.2	Feature selection technique performance.....	38
4.4.3	Classifiers' accuracy	41
4.4.4	The most important features.....	41
4.4.5	Comparing the results with previous studies.....	46
4.5	Conclusions	47
CHAPTER 5 ARTICLE 2 : APPLICATION OF MACHINE LEARNING TO CHILD MODE CHOICE WITH A NOVEL TECHNIQUE TO OPTIMIZE HYPERPARAMETERS		49
5.1	Introduction	51
5.2	Methods	57
5.2.1	Datasets and variables	59
5.2.2	Modeling	62
5.2.3	Tuning hyperparameters.....	63
5.2.4	Results interpretation.....	67
5.3	Results and Discussions.....	67
5.3.1	The performance of hyperparameter tuning techniques.....	67
5.3.2	The relative influence of variables on children's TMC	73
5.3.3	Analyzing the influence direction of top variables	74
5.3.4	Managerial implications.....	77
5.4	Conclusions	78
CHAPTER 6 ARTICLE 3 : INTERPRETABLE MACHINE LEARNING APPROACH TO PREDICT ELECTRIC VEHICLES BUYING DECISIONS		79
6.1	Introduction	80
6.2	Methods	83
6.2.1	Data and survey	85

6.2.2	Optimizing the percentage of training and testing datasets.....	88
6.2.3	Classification Process.....	90
6.2.4	Interpreting black-box machine learning techniques	92
6.3	Results and discussions	92
6.3.1	Optimal split ratio.....	92
6.3.2	Contribution of variables to increase the intention to buy EVs	94
6.3.3	Interpretation of machine learning results.....	96
6.4	Conclusions	102
CHAPTER 7 GENERAL DISCUSSION		104
7.1	Research gaps	104
7.2	Contributions	106
7.2.1	Finding optimal variables for CC-SoC prediction	106
7.2.2	Using a multi-objective approach to tune hyperparameters applied to child TMC.....	107
7.2.3	Optimizing the split ratio applied to vehicle engine preferences	108
7.2.4	Interpreting the results of machine learning techniques.....	109
7.3	Limitations.....	109
CHAPTER 8 CONCLUSIONS AND RECOMMANDATIONS FOR FUTURE STUDIES.....		112
8.1	Conclusions	112
8.2	Recommendations for future studies	114
8.3	Final summary	115
REFERENCES.....		117
APPENDIX A		140

LIST OF TABLES

Table 3.1 The analyzed variables in this thesis.	11
Table 4.1 Selected socio-demographic characteristics of the respondents.	24
Table 4.2 The accuracy of classifiers based on various feature selection technique's optimal features.	39
Table 4.3 The optimal features for CC-SoC prediction and their corresponding feature groups.	43
Table 5.1 A summary of recent studies on the application of machine learning to travel mode choice prediction.	54
Table 5.2 Attribute of selected variables.	61
Table 5.3 Set of hyperparameters for various hyperparameter tuning techniques.	66
Table 5.4 The optimal values of hyperparameters.	68
Table 5.5 Results of Multiple Logistic Regression (variables are organised by relative influence from most to least).	75
Table 6.1 Correlation matrix of socio-demographic information and environmental attitudes of respondents.	86
Table 6.2 Discrete choice experiment attributes and their levels.	87
Table 8.1 A summary of this dissertation	115

LIST OF FIGURES

Figure 3.1 The flowchart of this dissertation.	11
Figure 4.1 The methodology flowchart.	22
Figure 4.2 The average validation data accuracy of classifiers by considering different values as the optimal number of features.	37
Figure 4.3 Whole cycle running time of feature selection techniques.	41
Figure 4.4 The average testing data accuracy of different classification techniques.	41
Figure 5.1 The flowchart of the methodology.	59
Figure 5.2 The pseudo-code of MOHT.	66
Figure 5.3 The testing data accuracy of hyperparameter tuning techniques.	70
Figure 5.4 The testing data accuracy of hyperparameter tuning techniques.	71
Figure 5.5 The Area Under the Curve of Receiver Operating Characteristic, (a): MOHT , (b): random search, (c): Hyperopt, and (d): grid search; SB: school bus, AP: automobile as a passenger, BU: bus, RA: rail transit, CY: cycling, WA: walking.	73
Figure 5.6 The testing data accuracy of hyperparameter tuning techniques.	73
Figure 5.7 The relative influence of variables on children's TMC.	74
Figure 6.1 Flowchart of the methodology.	85
Figure 6.2 New-vehicle labels with different framing techniques designed to analyze the impacts of different GHG framings on individual's vehicle choices.	88
Figure 6.3 Results of the conventional model for split ratio optimization.	93
Figure 6.4 The relative influence of socio-demographic, environmental concern, and vehicle attributes on individuals' intention to buy EVs.	95
Figure 6.5 The relative influence of variables' interactions on individuals' intention to buy EVs.	96
Figure 6.6 Effect of CC-SoC on EV purchase likelihood where: 1 = not concerned; 2 = concerned, but don't plan to reduce emissions; 3 = concerned, but don't know what to do; 4 = concerned,	

and planning to reduce emissions; 5 = concerned, and have significantly reduced emissions.	98
Figure 6.7 Effect of EV to ICEV purchase price ratio on EV purchase likelihood.	99
Figure 6.8 Effect of age on EV purchase likelihood.	100
Figure 6.9 Effect of GHG framing on EV purchase likelihood. Frame 1 = current NRCan label; 2 = Societal goal with color; 3 = Societal goal as pressure gauge; 4 = Societal goal with new emojicons; 5 = Steps towards societal goal with patriotism; 6 = Societal goal as dirty air scale; 7 = Societal goal as Tree health; 8 = Societal goal with thumbs up/down	101
Figure 6.10 Effect of the interaction of CC-SoC and EV to ICEV purchase price ratio on EV purchase likelihood.	102

LIST OF SYMBOLS AND ABBREVIATIONS

AdaBoost	AB
Artificial Intelligence	AI
Accumulated Local Effects	ALE
Artificial Neural Network	ANN
Area Under the Curve	AUC
Battery Electric Vehicle	BEV
climate change stage of change	CC-SoC
Coyote Optimization Algorithm	COA
Coyote Optimization Algorithm-Quadratic Discriminant Analysis	COA-QDA
Coyote Optimization Algorithm - eXtreme Gradient Boosting	COA-XGBoost
Discrete Choice Experiment	DCE
Decision Tree	DT
Elastic Net	EN
Extra Trees Feature Selection	ETFS
Electric Vehicle	EV
General Ecological Behavior	GEB
Greenhouse Gas	GHG
Internal Combustion Engine Vehicle	ICEV
Identically and Independently Distributed	IID
K-Nearest Neighbor	KNN
Light Gradient Boosting Machine	LGBM

Local Interpretable Model-Agnostic Explanations	LIME
Logistic Regression	LR
Machine Learning	ML
Multi-Layered Perceptron	MLP
Multi-Objective HyperParameter Tuning	MOHPT
Multinomial Logit	MNL
Gaussian Naïve Bayes	NB
New Environmental Paradigm	NEP
Natural Resources Canada	NRCan
Non-dominated Sorting Genetic Algorithm-III	NSGA-III
Principal Component Analysis	PCA
Partial Dependence Plot	PDP
Plug-in Hybrid Electric Vehicle	PHEV
Random Forest	RF
Random Forest Feature Selection	RFFS
Receiver Operating Characteristic	ROC
SHapley Additive exPlanations	SHAP
Sport Utility Vehicle	SUV
Support Vector Machine	SVM
Travel Mode Choice	TMC
eXtreme Gradient Boosting	XGBoosting

LIST OF APPENDICES

APPENDIX A The socio-demographic information and environmental attitudes of respondents and vehicle labels in chapter 6	140
---	-----

CHAPTER 1 Introduction

In this section, general information and general descriptions of the current project are first presented:

Transportation is a fundamental part of people's lives, allowing them to access destinations and activities. The transportation sector significantly affects environmental, social, and economic development (Augeri, Greco, & Nicolosi, 2019). It is also a crucial industry to consider when trying to combat climate change and air pollution as it represents roughly 25% of fossil fuel consumption, a third of global energy consumption (Jahanbakhsh, Karimi, Naseri, & Nejad, 2020), and 23% of worldwide Greenhouse Gas (GHG) emissions (Climate Watch, 2018). In Canada, the transportation sector contributes to 27% of GHG emissions, and the GHG emissions of the transportation sector increased by 53% from 1990 to 2018 in Canada (Environment and Climate Change Canada, 2020).

Along with GHG emissions, the transportation sector contributes to a third of global air pollution (Jahanbakhsh et al., 2020). Mortality from air pollution is higher than that from road collisions, tuberculosis, and AIDS/HIV combined (OECD, 2016). That is, global air pollution contributes to 6.5 million annual deaths, of which 3 million deaths occur due to outdoor air pollution, and this value is estimated to increase to 4.5 million by 2040 (OECD, 2016). Unfortunately, in several cities, the transportation sector is the highest contributor to air pollution. For instance, transportation vehicles contributed to 85% of air pollution generated in Tehran, Iran, in 2013 (Panahi, Dehghani, Kinder, & Ezeji, 2019).

Contrary to some other industries, the transportation sector's emissions and energy consumption are increasing due to many reasons. For example, average vehicle size has increased, due in part to an increase in the use of Sport Utility Vehicles (SUVs) (Brazeau, Denoncourt, Gruber, & Lachapelle, 2021). Moreover, vehicle growth is considerably higher than population growth in Canada, leading to increasing cars per capita (Canada Statistics, 2021; Statistics Canada, 2021b). Hence, it is vital to investigate the methods capable of reducing this significant amount of GHG emissions, air pollution, and energy consumption in the transportation sector.

If societies want to reduce transportation-sector emissions, there is a need to better understand the variables affecting transportation-related choices. To this end, statistical analyses have been widely

applied. Due to the complexity of travel behavior, conventional statistical methods may not be able to analyze and predict travel behavior with the highest possible accuracy. In this case, machine learning techniques can be applied to model and predict travel behavior with higher accuracy (S. Wang, Mo, Hess, & Zhao, 2021).

Machine learning techniques generally contain some parameters that should be tuned, called hyperparameters. For example, the number of estimators in the Random Forest technique is a hyperparameter that should be tuned. The number of estimators implies the number of weak learners that should be generated to form the final Random Forest prediction model. These hyperparameters can significantly influence the performance of the prediction model in terms of prediction accuracy. That is, considering inappropriate values for hyperparameters can significantly reduce the prediction accuracy.

To model a prediction technique (e.g., machine learning), data sets should be divided into training, validation, and testing sets. Training data is used to educate the model. Validation data is used to tune hyperparameters. Then, unseen data (testing data) is used to evaluate the performance of the prediction techniques (Bishop, 2006). It is worth noting that it is possible not to separate training and validation data by using the K-fold cross-validation technique. The K-fold cross-validation technique applies the training set to tune hyperparameters. Finding the optimal percentage of testing data is called determining the optimal split. Determining the split ratio is vital to maximizing the accuracy of machine learning techniques.

CHAPTER 2 Critical overview of the literature and research needs

Analyzing travel behavior is vital for transportation planning and for improving sustainability in transportation systems. However, travel behavior is complex and hard to analyze (Riggs, 2017). Researchers generally employ statistical methods and discrete choice models to understand, predict, and analyze travel behavior. For example, the Multinomial Logit (MNL) has been the dominant discrete choice model in Travel Mode Choice (TMC) analysis since it is well understood, relatively easy to estimate thanks to the straightforward closed form formula and easy to interpret (E. J. Kim, 2021a). Consequently, MNL has been applied to different TMC problems, such as shopping trips (Andrade, Uchida, & Kagaya, 2006), comparing the determinants of TMC in different countries (Buehler, 2011), commuter TMC in the presence of a new public transit infrastructure (Hensher & Rose, 2007), individual daily TMC (Viegas de Lima, Danaf, Akkinapally, De Azevedo, & Ben-Akiva, 2018), and analyzing home to school/work commute patterns (Eluru, Chakour, & El-Geneidy, 2012).

With MNL models, it is assumed that the error terms are Identically and Independently Distributed (IID). This assumption can reduce the model's accuracy. Therefore, other discrete choice models that can relax the IID hypothesis have been used to model TMC. As such, Nested Logit, Multinomial Probit, Heteroscedastic Extreme Value, and Mixed Multinomial Logit have been applied to TMC analyses (Aboutaleb, Danaf, Xie, & Ben-Akiva, 2021; Bhat, 2011; X. Chen, Liu, & Li, 2013). The Latent Class, Hybrid Choice, and Joint models have also been used to investigate travel behavior. For instance, Morikawa et al. (2002) used a Latent Class model to analyze travel behavior. Similarly, Shen (2009) applied a Latent Class model with a Mixed Logit model to two different TMC datasets to analyze travel behavior.

Due to the complexity of travel behavior, however, conventional statistical methods and DCMs may not be able to analyze and predict travel behavior with the highest possible accuracy. Moreover, most of the conventional statistical methods tend to only consider a limited number of variables in the modeling process, and hence, the influence of only a limited number of variables on travel behavior can be captured. As well, many conventional statistical methods assume linear relationships between the explanatory and dependent variables which is not necessarily the case. As another issue, in DCMs, utility functions should be defined which might be a complicated task to optimize their structure. Machine learning algorithms are robust tools in big data analysis and

pattern recognition and researchers have begun applying them to predict travel behavior (Koushik, Manoj, & Nezamuddin, 2020).

Machine learning techniques can address some of the problems associated with conventional statistical methods. Powerful machine learning techniques exist that have the potential to significantly improve the accuracy of such analyses, and they can more easily include more variables in the modeling process (Hillel, Bierlaire, Elshafie, & Jin, 2021; S. Wang et al., 2021). Machine learning methods can capture non-linear relationships between model features and response variables. Further, machine learning techniques are flexible in structure (i.e., no need for defining utility functions), leading to a higher prediction accuracy than the traditional methods to model TMC (Hillel et al., 2021).

Researchers have examined ML techniques against discrete choice models. Chapleau et al. (2019) used the Random Forest and a Multinomial Logit to predict the use of different travel modes in Montreal. The results showed that the Random Forest classifier was more accurate than the Multinomial Logit for TMC prediction. Similarly, Gradient Boosting (Pineda-Jaramillo & Arbeláez-Arenas, 2022), Softmax Regression, Support Vector Machine (Mi, Wang, Shao, Zhang, & Chen, 2021), and Neural Network (Pham, Jiang, & Zhang, 2022) were found to be significantly more accurate than the Multinomial Logit regarding the prediction accuracy. Richards and Zill (2019) used four machine learning techniques, including Random Forest, Neural Network, Gradient Boosting, Logistic Regression, and a nested logit model (discrete choice model) to predict TMC. The results indicated that machine learning techniques performed better than the discrete choice model based on accuracy. Finally, the Cross-Nested Logit model and Nested Logit model were compared with the Neural Network classifier, and the Neural Network significantly outperformed the mentioned discrete choice models regarding the prediction accuracy (Nam & Cho, 2020). From an accuracy perspective, ML techniques would be the preferred approach.

However, there are some problems associated with robust machine learning techniques. First, these methods are highly sensitive to variables (i.e., features) applied in the modeling (J. Yang & Ma, 2019). There might be many variables that might influence travel behavior, but it is not clear which of them should be applied in the modeling to reach the highest accuracy. That is, model prediction accuracy may not be improved simply by increasing the number of features. Including a high number of features in prediction models may reduce the accuracy, lead to overfitting, and increase

the computation time (Chang et al., 2019). Therefore, the optimal variables that should be applied in the modeling should be detected, and it is a complicated task when a large number of variables are available in the data set (J. Yang & Ma, 2019).

In this case, using powerful feature selection techniques to detect the optimal features can be vital (J. Yang & Ma, 2019). Feature selection techniques identify the most important variables allowing the number of features to be reduced in the modelling process. Therefore, the computational complexity (e.g., run time) can be reduced, accuracy may be increased, and the likelihood of overfitting is decreased (Chang et al., 2019). However, the application of feature selection techniques has not received much attention in transportation studies. Further, some feature selection techniques can not present the optimal number of features, and they can only present the importance weight of features. In this case, considering an importance weight threshold for eliminating features is a complicated task. Further, many feature selection techniques (e.g., Lasso and Elastic Net) are linear techniques, and they may not be appropriate for detecting optimal features in travel behavior modeling since travel behavior generally does not follow a linear trend (e.g., the amount driven might increase with age to a certain point and then start to decrease. In this case, developing a new non-linear feature selection technique that can detect optimal features that lead to the highest prediction accuracy is essential when applying machine learning techniques to travel behavior analyses.

The second consideration is that machine learning techniques are quite sensitive to “hyperparameters” that need to be set by users. Regularization (tuning algorithms’ hyperparameters) is one of the most important processes to increase the prediction accuracy and prevent the model from over-fitting or under-fitting (Montesinos López, Montesinos López, & Crossa, 2022). However, a few studies applied conventional regulation processes to tune classifiers’ hyperparameters in transportation-based prediction studies. Grid search and random search are the conventional methods to tune algorithms’ hyperparameters. Grid search is a brute force (comprehensive search) technique in which all the combinations of hyperparameters in a predefined set are investigated to tune these parameters. Random search is a more efficient method that tries to find the optimal value of hyperparameters using a random selection technique. Nonetheless, grid search and random search are not efficient techniques (Liashchynskyi & Liashchynskyi, 2019), particularly for tuning continuous hyperparameters. That is, it is nearly impossible to find the optimal value of a continuous variable by checking a finite number of

samples in their ranges (Naseri, Jahanbakhsh, Hosseini, & Moghadas Nejad, 2020). Moreover, conventional tuning techniques (such as Grid Search or Random Search) generally consider a single performance indicator (e.g., accuracy for classification or mean absolute error for regression) to determine the optimal value for hyperparameters. However, in some cases, there might be a need to consider multiple performance indicators (e.g., considering accuracy and F1-score simultaneously) to tune the hyperparameters of machine learning techniques when using imbalanced data sets. As a result, proposing formalized regularization techniques to tune hyperparameters is vital to increasing the accuracy of travel behavior prediction models.

Moreover, when using machine learning approaches, datasets are traditionally divided into training and testing datasets. Training data is used to train the model. Then, testing data is used to evaluate the performance of the prediction techniques (Sanchez, Repello, Carrega, & Bolla, 2021). Nonetheless, determining the optimal split for training and testing datasets has been of immense concern. One possible approach to find the optimal share of the mentioned datasets is to check a limited number of randomly selected percentages for training data (e.g., 90%, 80%, 70%) and consider the shares that led to the highest testing data accuracy as the optimal percentage of datasets (Nguyen et al., 2021). However, finding the optimal value of a continuous-ranged parameter by checking a limited number of possible options is essentially impossible (Naseri, Jahanbakhsh, et al., 2020). Hence, developing an efficient method for optimizing the split ratio is essential to maximize the prediction accuracy of machine learning techniques.

Finally, most of the powerful machine learning techniques are black-box tools, and hard to interpret. Hence, the outcomes of machine learning techniques are difficult or impossible for policymakers and planners to apply them. Although there are white-box machine learning techniques (e.g., regression models), their accuracy is typically much lower, and their application leads to inaccurate interpretation (Naseri, Jahanbakhsh, et al., 2020; Naseri, Jahanbakhsh, & Moghadas Nejad, 2019). In this regard, the application of novel approaches to make black-box prediction methods interpretable and apply them to set new transportation policies remains another significant concern.

The above discussed the technical research gaps that this thesis addresses. This is done through the development of techniques that will be tested on three different sustainable transportation questions: travel behavior and level of concern about climate change; children's travel mode

choice; electric vehicle choice versus internal combustion engine vehicle choice. Each of these questions was chosen as they are important questions in themselves, but they are also representative questions related to sustainable transportation. Moreover, the data sets address different travel behavior-related questions that have not received enough attention in previous studies. Finally, each has a different type of dependent variable. The three datasets and studies are summarized below.

Travel behavior influences climate change. However, it is unknown whether travel behavior is a determining factor when it comes to an individual's climate change attitude and efforts to reduce personal emissions. To make it clear, analyzing individuals' Climate Change Stage of Change (CC-SoC) can be useful. CC-SoC was demonstrated to be an important indicator to estimate the influence of climate change attitudes on vehicle choice (E. O. D. Waygood, Wang, Daziano, Patterson, & Kohlová, 2021). However, the role of transport-related behavior in predicting CC-SoC is not well known. Perhaps it is not a behavior that people consider when they self-assess their climate change attitudes and behavior. Further, how a multitude of general environmental behaviors, attitudes, and sociodemographic characteristics related to the CC-SoC is not well known.

Additionally, although research has been conducted on adult Travel Mode Choice (TMC), child TMC has not received enough attention, and this group is often excluded from transportation studies. That is, most previous travel mode choice studies have focused on adults while analyzing TMC in children has received less attention. Likely related to this, most transportation planning is based on adults' needs, and children are more and more reliant on adults for their transportation needs (Freeman & Tranter, 2012). Children are members of society, and travel mode choice has been found to relate to their overall wellbeing (Waygood, Friman, Olsson, & Taniguchi, 2017). Further, children's travels can influence parents' travel behavior (Zwerts, Janssens, & Wets, 2007). Also, children become adults, and their childhood behaviors can impact their adult behaviors. Therefore, it is important to examine children's TMC for all trips (i.e., not only school-related trips) and how different parameters impact healthier and more sustainable mode choices.

One proposed solution to reduce GHG emissions, oil dependency, and air pollution is to replace Internal Combustion Engine Vehicles (ICEVs) with Electric Vehicles (EVs) (Kucukoglu, Dewil, & Cattrysse, 2021). Although many governments have implemented EV promotion policies such

as EV subsidies (Y. Dong, 2022), fuel taxation (Barros, Cruz, Júdice, & Sarmiento, 2021), and improving charging infrastructure (Funke, Sprei, Gnann, & Plötz, 2019), the EV share in international vehicle stock is low, only nearing 1% in 2020 (Austmann, 2021). Hence, it is essential to investigate the factors affecting EV preference/choice. Although previous studies investigated the intention to buy EVs or the willingness to pay for EVs, only a few prediction models have been generated to determine who prefers EVs over ICEVs. Further, the influence of GHG information framing on individuals' willingness to buy EVs has not been tested in EV buyer prediction models.

CHAPTER 3 Overall Research Approach

The aim of this thesis is to contribute to better predict individual travel behavior through improved machine learning approaches. As mentioned, the application of machine learning methods and the comparison of their performance to determine the most accurate prediction algorithms have not received much attention in the transportation literature. Thus, the performance of different powerful Machine Learning prediction algorithms will be compared for various travel behavior prediction problems.

Additionally, large numbers of variables are included in the prediction models to obtain the highest possible accuracy and detect which of them influence travel behavior the most. To do so, feature selection methods will be used. Moreover, a new feature selection technique is proposed combining a machine learning technique and an optimization algorithm. The proposed feature selection techniques should be capable of determining the optimal number of features, and accordingly, they should present the optimal feature set. Consequently, the performance of the proposed feature selection methods and conventional feature selection techniques will be compared to detect the best feature selection algorithms that would be appropriate for travel behavior estimation problems.

Hyperparameters of artificial intelligence methods need to be calibrated to increase the prediction accuracy and make the outcomes reliable. Thus, this investigation attempts to propose novel formalized regularization techniques to tune hyperparameters with lower computational complexity. These proposed regularization techniques are compared with conventional regularization techniques based on computational cost and prediction accuracy. As a result, appropriate regularization techniques for sustainable transportation behavior problems are identified. Using appropriate prediction methods, proper feature selection, and suitable regularization techniques can lead to the highest accuracy.

In traditional machine learning, data sets are divided into training data and testing data to generate a prediction model. Training data is used to tune the prediction model, and testing data is used to assess the prediction accuracy of the model. The common approach to finding the optimal share of the mentioned data sets is to check a limited number of randomly selected percentages for training data and consider the shares that led to the highest testing data accuracy as the optimal split ratio (Nguyen et al., 2021). However, finding the optimal split ratio by checking a limited number of possibilities may be impossible. Hence, in this study, a new method is introduced to determine the

optimal share of training and testing data sets. This method applies an optimization method to obtain the mentioned objective, and it can maximize the prediction accuracy.

Although machine learning techniques are accurate, most of them are black-box methods, limiting interpretability. Hence, another contribution of this project is to apply different methods to make the results of black-box models interpretable. To do so, the parameters affecting sustainable travel behavior are prioritized based on their importance. Accordingly, policymakers and urban planners can concentrate on the most important optimal features and make optimal decisions to reduce private car use and enhance sustainability in the transportation sector.

This chapter will present the overall approach to the research and how the articles relate to the overall objective. To address the issues presented in the previous chapter, the overall objectives of this thesis are to:

1. Propose a new feature selection technique (chapter 4);
2. Develop a novel regularization technique (i.e., a technique for tuning hyperparameters) (chapter 5);
3. Introduce a novel method to find the optimal share of training and testing data (chapter 6).

An overview of how each article relates to the overall objective and each sub-objective is shown in Figure 3.1.

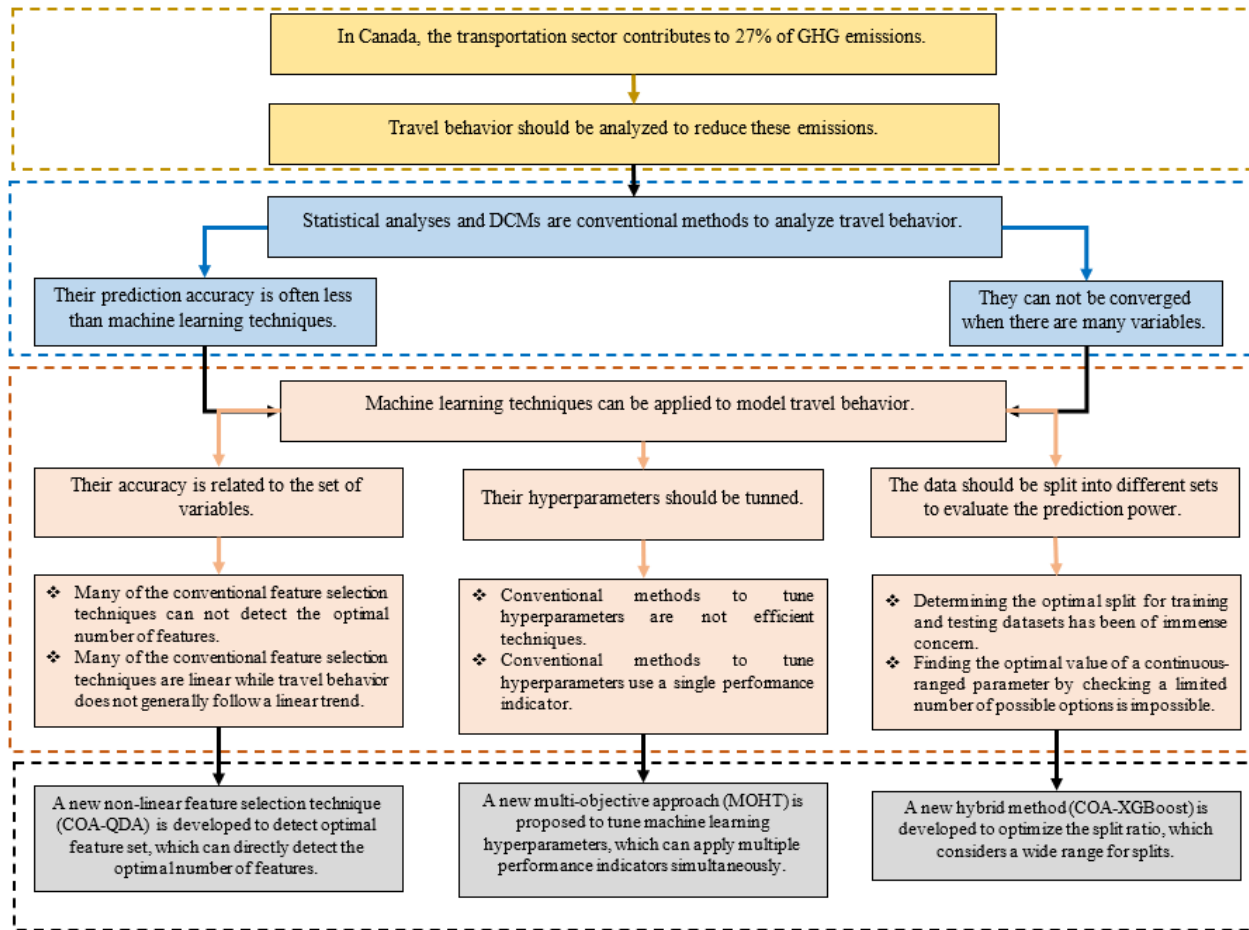


Figure 3.1 The flowchart of this dissertation.

Each sub-objective considers a different type of variable. The dependent variables applied in this thesis, their type, and their relation to sustainable travel behavior are shown in Table 3.1.

Table 3.1 The analyzed variables in this thesis.

Chapter	Dependent variable	Type of variable	Relation to sustainable travel behavior
4	CC-SoC	Ordinal	CC-SoC can explain the environmental motivations in transportation choices
5	Child TMC	Categorical	TMC directly influences the generated GHG emission during the trips
6	Engine type (EV vs. ICEV)	Binary	The GHG emission of EVs is less than ICEV

The following chapters include three accepted research articles that are presented to investigate the overall objectives of this dissertation. In Chapter 4, a new feature selection technique, named the Coyote Optimization Algorithm-Quadratic Discriminant Analysis (COA-QDA), is proposed to find the optimal features to predict individuals' level of concern about climate change. In Chapter 5, a novel technique, called Multi-Objective HyperParameter Tuning (MOHT), is proposed to select hyperparameters using a multi-objective metaheuristic optimization framework. In Chapter 6, a new method is introduced to optimize the share of training and testing data to reach the highest prediction accuracy.

CHAPTER 4 Article 1: A Novel Feature Selection Technique to Better Predict Climate Change Stage of Change

In this chapter, the first objective of this dissertation is investigated. That is, a new feature selection technique is developed to find the optimal feature that should be applied in the modeling to maximize prediction accuracy as well as reduce the computational cost. This chapter was accepted for presentation in the “Transportation Research Board 101st Annual Meeting”, located in Washington DC, United States, in 2022. Then, it was published in “Sustainability” journal on 21/12/2021.

The conference version of this paper can be found here: <https://trid.trb.org/view/1909388>.

The journal version of this research article can be found here: <https://doi.org/10.3390/su14010040>.

The co-authors of this study are:

Owen Waygood, Professor at Polytechnique Montreal, Canada.

Bobin Wang, Professor at Laval University, Quebec City, Canada.

Zachary Patterson, Professor at Concordia University, Montreal, Canada.

Ricardo Daziano, Professor at Cornell University, Ithaca, The United States.

Abstract

Indications of people's environmental concern are linked to transport decisions and can provide great support for policymaking on climate change. This study aims to better predict individual climate change stage of change (CC-SoC) based on different features of transport-related behavior, General Ecological Behavior, New Environmental Paradigm, and socio-demographic characteristics. Together these sources result in over 100 possible features that indicate someone's level of environmental concern. Such a large number of features may create several analytical problems, such as overfitting, accuracy reduction, and high computational costs. To this end, a new feature selection technique, named the Coyote Optimization Algorithm-Quadratic Discriminant Analysis (COA-QDA), is first proposed to find the optimal features to predict CC-SoC with the highest accuracy. Different conventional feature selection methods (Lasso, Elastic Net, Random Forest Feature Selection, Extra Trees, and Principal Component Analysis Feature Selection) are employed to compare with the COA-QDA. Afterward, eight classification techniques are applied to solve the prediction problem. Finally, a sensitivity analysis is performed to determine the most important features affecting the prediction of CC-SoC. The results indicate that COA-QDA outperforms conventional feature selection methods by increasing average testing data accuracy from 0.7% to 5.6%. Logistic Regression surpasses other classifiers with the highest prediction accuracy.

Keywords: climate change stage of change; feature selection; transport-related behavior; optimization; classification.

4.1 Introduction

Governments around the world are trying to reduce transportation-related greenhouse gas (GHG) emissions in response to concerns about climate change. An important aspect of trying to reduce emissions is individual attitudes towards climate change (McCright, Marquart-Pyatt, Shwom, Brechin, & Allen, 2016). Awareness plays a crucial role in minimizing the negative impacts on climate change. It was demonstrated that individuals' environmental awareness could affect their behaviors aiming to protect the environment and reduce adverse effects on the environment (M. X. Yang, Tang, Cheung, & Zhang, 2021). Various research within the field of transport has demonstrated that environmental attitudes can also help explain travel behavior (e.g., Anable (2005); Susilo et al. (2012); Gaker and Walker (2013); Waygood and Avineri (2016)). It is important to understand the links between attitudes and transport behavior as it is a major challenge when it comes to personal emissions (Wynes & Nicholas, 2017).

A few common measures of environmental behavior and attitudes exist. One of the most established measures is the General Ecological Behavior (GEB) tool which includes roughly 50 questions on various behaviors, including a few on transport. Another more general "world view" measure for the environment is the New Environmental Paradigm (NEP) tool that includes fifteen questions related to attitudes towards the environment. A simpler measure is the Climate Change Stage of Change (CC-SoC), which was developed to quickly capture attitudes and behavior with respect to personal climate-emissions (E. O. D. Waygood et al., 2021). CC-SoC was developed based on the Transtheoretical Model (e.g., Prochaska et al. (2015)) where individuals are presumed to go through stages with respect to a problematic behavior. Essentially, the process starts from whether or not an individual believes there's a problem (precontemplation), moves through stages of motivation to do some behavior to address the problem (contemplation, preparation), taking action, maintaining it, and then establishing a habit (termination). Detailed descriptions of these stages can be found in Prochaska and Velicer (1997) and Prochaska and Colleen (2015). The CC-SoC was first proposed and used to examine differences in response strength to information on climate change emissions in the Carbon Aware Travel Choices (CATCH) research project by Waygood and Avineri (2011) and subsequently used in various studies (e.g., Waygood and Avineri (2016), Daziano et al. (2017), Daziano et al. (2021), Wang et al. (2021)). It has been demonstrated

that the simpler CC-SoC measure can replace the more complex measures of GEB and NEP with a good assessment of people's environmental motivations (E. O. D. Waygood et al., 2021). Thus, it is worthy of being predicted if possible.

In contrast to other environmental behaviors like recycling or heating/cooling practices, transport is essential for conducting many daily activities, and a disconnect may exist between its use and climate change. As demonstrated, common environmental behaviors such as recycling are not strong predictors of climate change behavior (Wynes & Nicholas, 2017), and people may do these as “token” environmental behavior. Behaviors such as recycling may be so commonplace that it might not be a good measure of whether a person has strong climate change attitudes or behaviors, though an individual may see themselves as environmental for having done such token behaviors. Knowing what environmental and transport behaviors and attitudes are associated with stronger *climate change* attitudes and behaviors can help create proxies for such measures to better estimate how individuals might respond to climate change policies.

In previous work (E. O. D. Waygood, Wang, Daziano, Patterson, & Kohlová, 2020), a variable attrition approach was used to analyze what behaviors and attitudes related to the CC-SoC. In this regard, an ordered logistic regression was performed to model and predict CC-SoC. In the modeling process, 89 variables were employed, and the model reached the Pseudo R^2 of 0.1364. However, Artificial Intelligence (AI) methods have the potential to improve the accuracy of the predictions, as well as the selection of the most important predictive variables. At the same time, when dealing with large numbers of variables, such as from the General Environmental Behavior questions (50), it is difficult to determine which combination of variables will provide the most accurate prediction model. Therefore, feature selection techniques are needed to select the most important predictors. Several research questions will be investigated here:

- a. Can the prediction accuracy of belonging to a CC-SoC be improved considerably by applying AI techniques such as Machine Learning (ML) or deep learning?
- b. Many ML methods exist, but which might be the most accurate for this type of measure (non-linear nominal variable)?
- c. When dealing with large numbers of variables, can using all variables in the prediction model maximize the prediction accuracy?

4.2 Literature review

Analyzing individual levels of concern about the environment has been investigated in various studies. For example, Zha et al. (2020) attempted to examine customers' environmental level of concern while purchasing electrical appliances, such as washing machines and refrigerators. The authors used the appliance's energy label as a proxy measure for individuals' environmental concerns. A mixed logit model was used to consider the effects of various parameters, including energy label, power consumption, performance, price, and brand, on customers' choices. The results showed that energy labels, power consumption, price, and brand significantly affected customers.

Bedard and Tolmie (2018) investigated the effects of online interpersonal and social media usage on sustainable behavior in terms of purchasing. The relation between the cultural dimensions and green purchase intentions in their study. The dataset came from the Mechanical Turk service of Amazon, and only millennial was considered the target group. Subsequently, a linear regression was applied for the modeling process. The results indicated that the impacts of online interpersonal and social media usage on green purchase intentions were significant. However, the influences of individualism were insignificant.

Cheung et al. (2021) investigated the role of consumer-brand interaction and consumer-consumer interaction in driving the consumer-brand engagement's cognitive, behavioral, and emotional dimensions. Furthermore, the influences of consumer-brand interaction and consumer-consumer interaction on consumers' behavioral intentions were examined considering ongoing search behavior and repurchase intention. A case study including 316 customers was applied, and the Partial Least Square – Structural Equation Modelling was used for the modeling process. The results indicated that consumer participation influenced ongoing search behavior, and behavioral and emotional engagements significantly impacted repurchase intention.

Likewise, environmental concern has been considered in making transport-based decisions. For example, Liu and Cirillo (2018) modeled vehicle purchase behavior and predicted future preferences using a generalized dynamic discrete choice approach. Impacts of different scenarios, including changes in vehicle purchase prices, vehicle characteristic improvements, and fuel price changes, on environmental behavior were taken into account. The results indicated that all the

mentioned scenarios influenced environmental behavior and could significantly affect the adoption of electric vehicles.

Although discrete choice models are easily interpretable methods and powerful models to scrutinize variables, it has been recognized that they generally have lower prediction accuracy than machine learning techniques. Moreover, discrete choice models have longer computational time than machine learning techniques (S. Wang et al., 2021). Although some ML techniques are black-box, sensitivity-analysis can be applied to know the influence strength of different features. Hence, researchers have begun to apply AI classification techniques to predict environmental behaviors. Researchers have applied different classification techniques to predict environmental behaviors.

Lee et al. (2017) applied three prediction methods: a deep learning neural network, an ordinary artificial neural network, and least square regression to predict environmental consumption levels in different regions. Six features, i.e., health expenditure, pre-primary education, pro-environmental consumption index, past orientation, and two features related to the gross domestic product, were used in the classification modeling. The results indicated that deep learning neural performed better than other prediction methods based on the prediction accuracy.

Amasyali and El-Gohary (2021) proposed an approach to predict the energy consumption of cooling in office buildings. Five sets of parameters, including window status, occupancy density, cooling setpoint, the power density of electric equipment, and density of lighting power, were considered as the model's input variables. Decision tree, deep neural network, artificial neural network, and ensemble bagging tree were used for the classification process. The results showed that the proposed approach could predict energy consumption as an environmental behavior. Furthermore, the deep neural network was the most accurate classification method. Aiming to predict whether people adopted green electricity policies, Lee et al. (2016) applied a machine learning approach to information on anti-environmental and pro-environmental attitudes. The outcomes of the mentioned study revealed that environmental attitudes had a significant role in adopting green electricity policies.

In a transport-related study, the prediction of fuel consumption was examined by Ping et al. (2019). To this objective, trip route, vehicle type, weather condition, and traffic conditions were used as the prediction model's features. A deep learning network method was modeled for classification

purposes. The proposed deep learning method could effectively detect the relationship between fuel consumption and driving behavior.

Given the many variables now available and considered in real-life prediction problems, feature selection techniques are increasingly used and can increase the prediction accuracy. Feature selection techniques can make a prediction model easier to interpret, increase the model's generalization capability, and remove noisy features (Chang et al., 2019). Chang et al. (2019) proposed a model to predict individual behavior in terms of transportation mode choice and detect the most important features. The travel history of 162 households over six years, comprising roughly 52,000 trips, was considered for the dataset. Twenty-three parameters relating to individual characteristics, household characteristics, and trip properties were considered in the initial feature set. A feature selection technique was employed, and the 14 features with the highest importance weights were retained. Subsequently, a set of feature selection techniques were utilized, and the results revealed that Random Forest was the most accurate prediction method.

Wade et al. (2017) compared the performance of two feature selection methods, Random Forest Feature Selection and LASSO, on a subcortical brain surface morphometry prediction problem. Three machine learning algorithms, including Random Forest, Naïve Bayes, and Support Vector Machine, were used for classification. The results indicated that Random Forest feature selection outperformed LASSO based on the prediction accuracy. On the other hand, LASSO was the better alternative for minimizing running time.

Sanchez-Pinto et al. (2018) compared the performance of various feature selection methods on two datasets. Four regression-based feature selection methods, including LASSO, Elastic Net, stepwise backward selection, Akaike information criterion, and four tree-based feature selection methods, including Regularized Random Forest Feature Selection, Random Forest Feature Selection, Gradient Boosted Feature Selection, and Boruta, were considered in their comparison. The results showed that regression-based methods obtained better parsimony in the smaller dataset, while tree-based methods achieved better parsimony in the larger dataset. The regression-based feature selection methods had better (or equal) performance than the model without feature selection. However, some performance loss was reported for tree-based methods.

CC-SoC was demonstrated to be an important indicator to estimate the influence of climate change attitudes vehicle choice (E. O. D. Waygood et al., 2021). To the best of the authors' knowledge,

although environmental behavior prediction has been investigated in some studies, the prediction of individual CC-SoC has not received enough attention considering the crisis at hand. The transport industry generates a third of global air pollution (Jahanbakhsh et al., 2020) and 22.7% of global GHG emissions (Climate Watch, 2018), and understanding how transport-related behavior relates to CC-SoC is essential to address the crisis. However, the role of transport-related behavior in predicting CC-SoC is not well known. Perhaps it is not a behavior that people consider when they self-assess their climate change attitudes and behavior. Further, how a multitude of general environmental behaviors, attitudes, and socio-demographic characteristics related to the CC-SoC is not well known.

Although there are a number of features to predict the CC-SoC, such as transport-related behavior, GEB, NEP, and socio-demographic characteristics, model prediction accuracy may not be improved simply by increasing the number of features. To this end, using robust feature selection techniques to detect the optimal features can be vital. However, detecting the optimal features for environmental behavior prediction has rarely been taken into account. As well, comparing the performance of several AI techniques, to obtain the highest accuracy is essential and is often overlooked in environmental behavior predictions. Furthermore, prioritizing the model's features and detecting the most important parameters can be critical for policymakers. Nonetheless, detecting the features' importance and ranking may be neglected in the mentioned classification problem.

4.2.1 Research contributions

To address the mentioned concerns, in this study, a new approach is proposed to predict individuals' environmental attitudes and behaviors (i.e., CC-SoC). Due to the significant effects of transportation on generating harmful emissions, transport-related behavior is taken into account as the model's variable as well as socio-demographic characteristics and environmental behaviors (GEB) and attitudes (NEP). Large number of variables increases the model's computational complexity and may reduce the prediction accuracy (Chang et al., 2019). Thus, a new feature selection technique is introduced, capable of finding the optimal number of features and the optimal feature set to maximize the prediction accuracy. Moreover, different common feature selection techniques are implemented and compared, and the new approach improves model performance in the context of the CC-SoC prediction problem. Similarly, various AI prediction methods are used

to detect the best prediction algorithms for the CC-SoC prediction problem. Finally, a sensitivity analysis is performed to prioritize the optimal features and determine the effectiveness of each variable on prediction accuracy increment.

4.3 Methodology

This study proposes a methodology to predict individual CC-SoC using a number of different types of variables, including socio-demographic characteristics, the 50 questions from the GEB, and the 15 questions from the New Ecological Paradigm (NEP) indices. Moreover, it aims to detect which variables have the greatest effect on people's CC-SoC. With this objective in mind, eight classification techniques are applied as prediction tools. Hence, one of the primary objectives of this study is to compare different prediction methods and detect the most accurate classifiers to solve the mentioned prediction problem. Subsequently, a new feature selection technique, named Coyote Optimization Algorithm-Quadratic Discriminant Analysis (COA-QDA), is introduced to determine the optimal features and the optimal number of features to obtain the highest prediction accuracy. The COA-QDA is compared with five conventional feature selection techniques based on the average accuracy of classification methods to assess their effectiveness and determine the most valuable feature selection technique. Finally, a sensitivity analysis is proposed to rank the features based on their importance on CC-SoC prediction accuracy.

The methodology flowchart is illustrated in Figure 4.1. As can be seen, the first step of this research is data preparation. Afterward, the proposed feature selection technique (COA-QDA) is developed. Then, different feature selection techniques are applied, and their performance is improved using classifier average prediction accuracy. A model without applying feature selection (i.e., using all features) is used to evaluate the effectiveness of feature selection methods on prediction accuracy. In the next step, the variables resulting from the different feature selection methods are employed to predict CC-SoC using eight classification techniques. Accordingly, the performance of feature selection techniques and classifiers are compared. The best combination of feature selection and classification techniques is determined, and its optimal features are applied in a proposed sensitivity analysis to prioritize the optimal features.

In this section, the data preparation process is first described. Then, the classification techniques applied in this study are presented. Following that, feature selection methods are explained. Finally, the sensitivity analysis is presented.

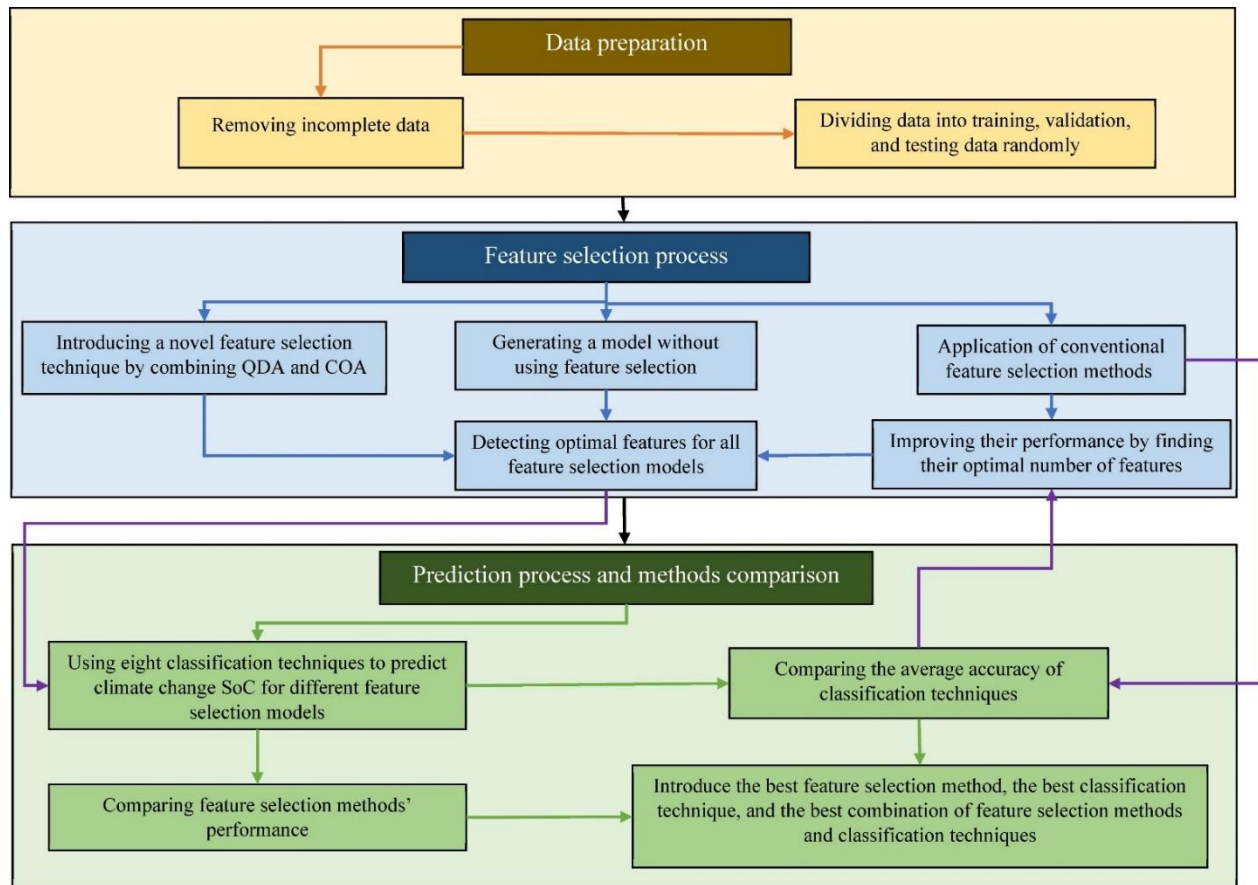


Figure 4.1 The methodology flowchart.

4.3.1 Data preparation

The data comes from a project on framing CO₂ emissions to predict individual willingness to pay for emissions (R. A. Daziano et al., 2017). An online survey was conducted between December 2015 and March, 2016 in Boston and Philadelphia, USA. As the original project was focused on vehicle purchases, the survey was restricted to only car owners. As such, the transport questions in this survey were predominantly car focused. A total of 1,580 complete responses were collected through the recruitment agency Qualtrics. Some selected socio-demographic information for the survey participants is displayed in Table 4.1.

The survey included questions on attitudes towards the environment including the NEP and GEB questions, attitudes towards various relevant government policies, a CC-SoC question (see below), and various transport-related questions. Additional information about GEB and NEP questions was

presented by Kaiser and Wilson (2004) and Dunlap et al. (2000), respectively. All questions in the survey were quantitative, and as a result, all input variables in the problem are categorical. The prediction model's input variables (features) can be divided into five groups, including socio-demographic (18 features), GEB (53 features¹), NEP (15 features), transport-related features (14 features), and extra features (11 features). The extra features category includes some questions on policy support for emission reduction and climate change attitudes. Hence, the prediction problem includes 111 features.

After collecting data, incomplete responses and responses where individuals failed “trap questions” (i.e., questions that are used to identify whether or not the respondent is paying attention) were eliminated from the initial dataset. The final dataset includes 1,536 samples. The final data were divided into three groups: training data, testing data, and validation data. Training data is applied to educate the prediction models. Validation data is employed to tune hyperparameters. Testing data is used to assess and compare the prediction ability of soft computing methods. The portion of training, testing, and validation data is considered 70%, 15%, and 15% (Majidifard et al., 2021). The model attempts to predict classes (categories) of respondent reported Climate Change Stage of Change (CC-SoC). The label of classes is based on the responses to the question “Please choose the phrase that most corresponds to you for reducing greenhouse gases”. The possible responses are as follows:

- 1) I am not concerned;
- 2) I would like to reduce my emissions, but I don't know how;
- 3) I would like to reduce my emissions, and will do so in the future;
- 4) I have already reduced my emissions significantly.

¹ Small changes were made in the GEB questions such as separating cycling and public transport.

Table 4.1 Selected socio-demographic characteristics of the respondents.

Socio-demographic variables		Frequency	Percent	Mode
Gender	Male	794	50.2	☒
	Female	787	49.8	
Household cars	1	616	39.0	
	2	739	46.7	☒
	3	155	9.8	
	4 or more	71	4.5	
Residence location	Greater Philadelphia	926	58.6	☒
	Greater Boston	655	41.4	
Education	Professional or Doctorate degree	80	5.1	
	Master's degree	229	14.5	
	Bachelor's degree	610	38.6	☒
	Associate degree	145	9.2	
	Some college, no degree	314	19.9	
	High School Graduate (Diploma or equivalent GED)	192	12.1	
	1-12th grade	11	0.7	
Household income	Less than \$30,000	105	6.7	

Table 4.1 Selected socio-demographic characteristics of the respondents (continued).

Socio-demographic variables		Frequency	Percent	Mode
Household income	\$30,000-\$39,999	105	6.6	
	\$40,000-\$49,999	134	8.5	
	\$50,000-\$59,999	163	10.3	
	\$60,000-\$74,999	247	15.6	☒
	\$75,000-\$84,999	133	8.4	
	\$85,000-\$99,999	165	10.4	
	\$100,000-\$124,999	174	11.0	
	\$125,000-\$149,999	109	6.9	
	\$150,000-\$174,999	66	4.2	
	More than \$175,000	103	6.5	
	I prefer not to answer	77	4.9	
Hispanic	Yes	104	6.6	
	No	1477	93.4	☒
Political	Strongly conservative	110	7.0	
	Moderately conservative	364	23.0	
	Independent	633	40.0	☒
	Moderately liberal	320	20.2	
	Strongly liberal	154	9.7	

4.3.2 Classification Techniques

Eight classification techniques, including Multi-Layered Perceptron (MLP), Gaussian Naïve Bayes (NB), Logistic Regression (LR), Decision Tree classifier (DT), K-Nearest Neighbor classifier (KNN), Random Forest classifier (RF), Support Vector Machine classifier (SVM), and AdaBoost (AB) are applied to model and predict the CC-SoC. Moreover, these methods are employed to compare the performance of different classifiers and obtain the highest possible accuracy. The classifiers are briefly explained in this section.

MLP is a deep Artificial Neural Network (ANN) containing more than one hidden layer. ANNs can be employed to model complicated problems in a short time. They are good at nonlinear prediction problems in a reasonable amount of time (Naseri, Jahanbakhsh, Khezri, & Shirzadi Javid, 2021). An MLP generally includes an input layer, some hidden layers, and an output layer. There are some processing units in each layer, called neurons. All neurons are connected to other neurons by various connection weights (unidirectional connections). The input layer receives the row information, adjusts them, and transfers them to the first hidden layer. The function of the hidden layers is to allocate different weights to each neuron. Then, activation functions are applied to change data representation, and the combination of neurons information and their corresponding weights are transferred to the next hidden layer. Finally, the output layer receives information from the last hidden layer and presents the prediction values or labels (Hasan, Alam, Das, Hossain, & Hasan, 2020).

Gaussian Naïve Bayes (NB) is one of the fastest and most straightforward classification methods. In NB, each sample's posterior probability is maximized during the labels' allocation. NB assumes that the voxel contributions follow a Gaussian distribution, and they are conditionally independent. NB applies a discriminant function for each category. The mentioned function is based on the summation of the squared distances to each classes' centroid weighted by its variance. Then, Bayes' rules are used to calculate the logarithm of the priori probability to train the model. Ultimately, for each testing data sample, the discriminant function is calculated for all classes, and the testing data sample is assigned to a class including the maximum discriminant function value (Ontivero-Ortega, Lage-Castellanos, Valente, Goebel, & Valdes-Sosa, 2017).

Logistic Regression (LR) is a powerful statistical modeling method that has been applied to solve classification problems. LR considers an explanatory variables' set to assess the dichotomous

outcome event probability (L. Dong, Wesseloo, Potvin, & Li, 2016). Dichotomous variables generally denote the occurrence or not of some events. Generally, LR assumes the relationship between the explanatory variables is linear. Thus, LR applies linear decision boundaries while using a non-linear model (Hajmeer & Basheer, 2003).

The Decision Tree classifier (DT) was inspired by the shape of trees and their nodes and leaves. DT is easy to understand and interpret. Furthermore, DT easily supports adding new scenarios if introduced, it can work as a white-box method, and it can be efficient while using an enormous volume of data. Classification rules are mainly modeled based on a set of selections in DT. DT is constituted of decision rules according to optimal feature cut-off thresholds. These thresholds divide each feature into different groups in every leaf node. Then, this process is continued in a hierarchical manner, and at each level, the available samples are divided into different groups based on the splitting criterion (Suresh, Udendhran, & Balamurgan, 2020). At each step, the current node's branching condition is assessed by splitting criteria. All the mentioned processes are called DT construction. Subsequently, the pruning process is performed. Pruning is a back forward process that eliminates the additional branches to reduce the computational costs and improve the algorithm's efficiency (Rau et al., 2017).

K-Nearest Neighbor classifier (KNN) is a black-box classification technique, which has been applied for statistical analysis since the 1970s. KNN is a non-parametric prediction algorithm, and it predicts a sample's label based on the labels of similar samples (Duca, Bacciu, & Marchetti, 2017). KNN plots all samples in a hyper-dimension space based on their features' values. Afterward, a distance function is utilized, and K nearest samples to the test sample are detected. The test sample's label is the most frequent label in the corresponding K nearest neighbor's label set. Considering a large value for K leads to high running time. Moreover, KNN cannot perform well in the circumstances where more than one frequent label is detected in the K nearest neighbor's label set (Thanh Noi & Kappas, 2017).

Random Forest (RF) is a prediction technique employed for solving regression or classification problems. RF is an ensemble method that combines different DTs to improve prediction accuracy. A particular number of DTs are modeled in the modeling process, and each tree is generated from a random vector. Subsequently, all DT models are run, and the label is determined by considering all DTs' results (Dogru & Subasi, 2018). Different DT models are run in RF simultaneously, and

the majority of class votes determine the predicted label. Research in transport has shown that RF is a powerful method when the problem is large-scale such as an origin-destination survey (Chapleau et al., 2019).

Support Vector Machine (SVM) is a powerful method used for classification, estimation, and pattern recognition. A set of kernel-based functions are generally applied by SVM to predict class labels in classification problems. Low-dimensional data are converted to high-dimensional vector spaces by nonlinear mapping functions in SVM. As SVM utilizes the theory of structural risk minimization, the over-fitting probability of the problem is reduced (Fan, Wang, Zhang, Ma, & Wu, 2020). Furthermore, nonlinear complex models can be transformed into simple linear form problems by SVM. Accordingly, SVM can apply linear regression function in a high dimensional space. Consequently, SVM allocates different values of bias and various weights to the model. The SVM model is replaced with a mathematical optimization problem using the principle of structural risk minimization. Afterward, slack variables are added to the new model, and the ultimate prediction model is generated considering fitting error. Ultimately, the optimal solution to the optimization problem is presented as the final classification model (L. L. Li, Zhao, Tseng, & Tan, 2020).

AdaBoost (AB) is an ensemble prediction method that works iteratively. AB combines different weak classifiers in a model to generate an accurate classification method. First, some weak classifiers (sub-classifiers) are generated, and equal weights are assigned to them. Subsequently, the sub-classifiers are trained, and their corresponding error is calculated. Then, the assigned weights are updated based on sub-classifiers' errors, and the updated weights are allocated to sub-classifiers in the next iteration. This iterative process is continued, and ultimately, the class labels are predicted using the results of sub-classifiers and their corresponding weight in the last iteration (Hu, Yin, Wan, Zhang, & Fang, 2020).

4.3.3 Feature selection process

This study aims to introduce an accurate model to predict an individual's CC-SoC. One approach to generate a precise model and obtain the highest accuracy is to detect optimal features that should be applied as the classifiers' inputs. In this regard, a new feature selection technique capable of finding the optimal number of features is introduced in the current study. In other words, the proposed technique can detect the optimal number of features and optimal features simultaneously

based on an optimization approach. Moreover, different conventional feature selection methods – Lasso, Elastic Net, Random Forest Feature Selection, Extra Trees, and Principal component analysis feature selection – are applied. Their structure is improved to enhance their performance. Hence, the other objective of this study is to compare the performance of the introduced feature selection technique with the improved version of some conventional feature selection techniques to detect the best set of variables that leads to the maximum possible accuracy. In this section, the introduced feature selection technique is presented. Afterward, the conventional feature selection technique and the method applied to improve their performance are described.

COA-QDA feature selection

As mentioned, a new feature selection technique is introduced in this study to find the optimal features leading to the highest accuracy. COA-QDA is developed with a combination of the Coyote Optimization Algorithm (COA), as a metaheuristic optimization algorithm, and Quadratic Discriminant Analysis (QDA), as a robust and fast machine learning technique. In this section, COA and QDA are described respectively, and afterward, the modeling of COA-QDA is presented.

COA is a metaheuristic optimization algorithm introduced by Pierezan and Coelho (2018). COA is a swarm intelligence algorithm inspired by the interactions and social behavior of *Canis Latrans* (coyotes). This algorithm applies a particular number of solution vectors, called coyotes, to investigate the problems' feasible regions and find optimal solutions. In the metaheuristic optimization process, each solution vector includes one value for each optimization problem's dependent variable. The set of independent variable values for each solution vector (coyote) is called the coyote social behavior in COA, as presented in Equation 4.1.

$$soc_c^{h,iter} = \vec{x} = (x_1, x_2, \dots, x_D) \quad (4.1)$$

Where $soc_c^{h,iter}$ signifies the social behavior of coyote c in herd h at the iteration of $iter$. Meanwhile, x_i and D imply the value of independent variable i and the optimization problem's dimension (number of independent variables), respectively.

Initially, various solution vectors are generated by assigning random values to each independent variable. The assigned values should be between the lower and upper bounds of independent variables. Subsequently, all coyote social condition (fitness value) is determined using the

problem's objective function. Then, coyotes are divided into different groups (herds). In other words, solution vectors are classified in order to investigate different parts of the problem's feasible region simultaneously. The coyotes are ranked based on their fitness value in their herds, and the coyote with the highest fitness value (i.e., the least objective function value in minimization optimization problems) is called alpha in each herd. That is to say, alpha coyotes are the best solution vectors in their groups. Equation 4.2 is applied to spot the alpha in each herd at each iteration (Pierezan, Maidl, Massashi Yamao, dos Santos Coelho, & Cocco Mariani, 2019).

$$\alpha^{h,iter} = \{soc_c^{h,iter} \mid \arg_{c=\{1,2,\dots,N_c\}} \min f(soc_c^{h,iter})\} \quad (4.2)$$

Where $\alpha^{h,iter}$ is the alpha in herd h at the iteration of $iter$.

Consequently, “culture” is transferred within each herd. Each coyote moves toward its groupmates and alpha in the feasible region in the culture transfer operation. The gravity of each groupmate to attract a coyote depends on the social condition, and the solution vectors with higher fitness values generate more attraction (gravity). Similarly, each coyote is transferred to the nearest point to the group alpha (Pierezan, Coelho, Mariani, & Lebensztajn, 2019). Therefore, the capable regions can be investigated meticulously by attracting more solution vectors. Some coyotes are transferred between herds, and this process is called culture transfer. The culture transfer operator avoids remaining in the local optimal solutions by scattering some solution vectors across the problem's feasible region. The death and birth process is another operator improving algorithm performance by removing the weakest coyotes and generating new coyotes. In each iteration, the solution vectors with the lowest fitness values are removed from the society (through death), and new solution vectors are generated randomly to investigate unseen areas (Naseri, Ehsani, Golroo, & Moghadas Nejad, 2021). The mentioned operators are run until the termination criteria are met. Ultimately, the solution vector with the highest fitness value is introduced as the optimal solution to the problem. More details about the algorithm's pseudo-codes and the algorithm process are provided by Pierezan and Coelho (2018) and Pierezan et al. (2019).

QDA is a supervised classification technique. QDA applies a Gaussian distribution to model each category likelihood. Consequently, posterior distributions are employed to predict the labels for testing data samples. The Gaussian parameters for all categories can be predicted using maximum likelihood estimation and training data samples (Srivastava, Gupta, & Frigiyik, 2007). In QDA, it

is assumed that the feature vector is multivariate normally distributed in the group with a given mean vector in a particular group and a specific covariance matrix. Hence, non-linear decision boundaries are used in the classification process (K. S. Kim, Choi, Moon, & Mun, 2011).

The COA-QDA aims to maximize the prediction accuracy by selecting the optimal features. That is, maximizing the prediction accuracy is an optimization problem that should be solved by an optimization algorithm. Since the type of the mentioned problem is Integer Programming, and the number of decision variables is high, the problem is non-deterministic polynomial-time (NP-hard). Exact optimization algorithms (e.g., branch and bound) cannot solve NP-hard problems (Naseri, Shokoohi, Jahanbakhsh, Golroo, & Gandomi, 2021). Moreover, exact optimization cannot be synced with machine learning techniques. Therefore, a metaheuristic optimization algorithm should be employed to solve the mentioned problem (Naseri et al., 2019). As a result, as a robust metaheuristic algorithm, COA is applied for optimization purposes.

Moreover, a powerful and fast classifier is required to predict the labels for each solution vector in COA and calculate the accuracy. Hence, QDA is used as the classifier in the proposed method. The modeling of the COA-QDA is as follows:

$$\text{Maximize } z = (\alpha_1 \times \text{Accuracy}_{\text{training}}) + (\alpha_2 \times \text{Accuracy}_{\text{validation}}) \quad (4.3)$$

Subject to:

$$\alpha_1 \in \{1, 2, \dots, \alpha_{\max}\} \quad (4.4)$$

$$\alpha_2 \in \{1, 2, \dots, \alpha_{\max}\} \quad (4.5)$$

$$N_{OPT} = n \quad n \in \{1, 2, \dots, N - 1\} \quad (4.6)$$

$$x_i \in \{fea_1, fea_2, \dots, fea_N\}, \quad \forall i \in \{1, 2, \dots, N_{OPT}\} \quad (4.7)$$

$$i \leq N_{OPT} \quad (4.8)$$

$$\text{If } x_j = x_k \Rightarrow fea_j = fea_k \quad \forall j, k \in \{1, 2, \dots, N_{OPT}\} \quad (4.9)$$

Where α_1 and α_2 are the calibration weights. $Accuracy_{training}$ and $Accuracy_{validation}$ signify the accuracy of QDA for predicting training data and validation data, respectively. α_{max} denotes the maximum value of calibration weights. N_{OPT} and N imply the optimal number of features and the number of features in the initial features set. x_i and fea_i are the optimal feature i and the feature i in the initial features set.

In the proposed optimization process, Equation 4.3 is the problem's objective function. This equation maximizes the model's training and prediction accuracy. Considering validation data accuracy is necessary to avoid over-fitting in the feature selection process and selecting the optimal features that increase the model's prediction power. Moreover, calibration weights are applied to investigate the optimal calibration weights according to the details provided by Naseri et al. (2020). After running the model and obtaining the solutions, the testing data is applied to determine the calibration weight optimal value. That is to say, the calibration weights leading to the highest testing data accuracy are considered the optimal calibration weights. Equation 4.4 and Equation 4.5 guarantee that the calibration weights are selected from the given range. α_{max} is considered to be 3 based on Naseri et al. (2020). Equation 4.6 is another constraint that prevents the model from selecting the optimal number of features higher or equal to the number of features in the initial dataset. This constraint is applied due to not limiting the model from selecting each feature at most once. That is to say, the model can select one feature as an optimal feature more than once if the feature's duplication improves the model's performance. Additionally, the approach is to reduce the input's dimension, and the number of features should be reduced.

Equation 4.7 guarantees that exactly one feature is assigned to each optimal feature. Meanwhile, Equation 4.8 forces the model to select exactly N_{OPT} features, which is the optimal number of features. Based on Equation 4.9, only one feature from the initial feature set should be assigned to each optimal feature. After running the model, the x_i set related to the optimal solution is considered as the optimal feature set.

Lasso

Lasso is a soft computation technique proposed by Tibshirani (1996). Lasso has been extensively applied to feature selection and regularization processes. Lasso shrinks the model's input size by minimizing the summation of the coefficients' absolute value (L1-penalty function) using conventional least squares regression. The L1-penalty function is utilized to avoid overfitting and detect the selected features. That is to say, the penalty parameter prevents the model from selecting significant values for coefficients (Muthukrishnan & Rohini, 2017). Hence, the coefficient of unimportant features becomes zero automatically. The features with the assigned coefficient of zero are removed from the model. On the other hand, the parameters with the corresponding non-zero coefficients are considered the selected features (Fonti, 2017).

Elastic net

Elastic Net (EN) is another feature selection technique applied to improve the performance of prediction models influenced by multicollinearity. In the cases that the data is affected by multicollinearity, the model's variance is significant while least squares predictions are unbiased. Accordingly, the model estimation can be inaccurate. EN is a conventional least squares regression modified with two penalty parameters, including the L1-penalty function and L2-penalty function (Cui, Bai, Wang, Jin, & Hancock, 2021). In other words, EN is the combination of lasso regression and ridge regression. EN minimizes all coefficients' absolute values by adding the summation of coefficients' absolute value and summation of coefficients' square to the least-squares function. Moreover, each penalty function is multiplied by a tuning parameter that controls the shrinkage amount. Ultimately, the features with the coefficients of zero are eliminated from the input sets, and the other features are taken into account selected features (Zou & Hastie, 2005).

Random Forest Feature Selection

Random Forest Feature Selection (RFFS) is a robust feature selection reducing the number of features based on the features' importance score. It has been proved that RFFS is efficient on dimensionality reduction when the model includes hundreds of features (Jaiswal & Samikannu, 2017). RFFS is an ensemble technique that generates several decision trees by choosing random observations and random variables and combines them. Then, the votes generated by each decision tree are aggregated, and hence, the variables' predicted likelihood and features' importance score

are calculated. The features with the highest importance scores are generally considered the chosen features, and the other features are overlooked (Yamauchi, 2013). Nonetheless, there is not a particular threshold for features' importance score, and it is a complicated task to detect the number of optimal features in RFFS.

Extra Trees Feature Selection

Extra Trees Feature Selection (ETFS) is an ensemble method that has been used for feature selection. ETFS is a variant of RFFS with higher randomization for selecting decision boundaries at all steps. The generated trees in ETFS have more leaf nodes compared with RFFS, and the computational efficiency of ETFS can be higher than RFFS. Meanwhile, the variance-bias trade-off in ETFS may be higher than that of RFFS due to a higher level of randomization. However, more randomization may lead to a reduction in the model's accuracy. ETFS combines different decision trees, and the aggregated votes are presented as the features' importance factor (Sharma, Giri, Granmo, & Goodwin, 2019). Like RFFS, ETFS cannot detect the optimal number of features that should be selected to obtain the highest classification accuracy.

New feature selection-based Principal component analysis

Principal Component Analysis (PCA) is a powerful technique in data structure investigation. PCA generates new variables (principal components or latent variables) by data variance maximization. Hence, PCA application reduces the problem's dimensionality. Although PCA reduces the dimensionality, the number of original features is not reduced as all original features can be applied to generate principal components (Guo, Wu, Massart, Boucon, & De Jong, 2002). In the current investigation, the PCA is converted to PCA feature selection based on the details provided by Song et al. (Song, Guo, & Mei, 2010). The weight of each feature to generate all principal components are summed, and the obtained value is considered the importance weight of the corresponding feature. Moreover, the PCA model is run $N-2$ times by considering the number of principal components equal to 2, 3, ..., $N-1$. Where N represents the number of original features in the initial features set. Consequently, the average value of importance weights over $N-2$ runs is calculated for all features, which is called the ultimate importance weight. Finally, the features are ranked based on their ultimate importance weight, and the feature with the highest ultimate importance weight is the most important feature, followed by the features with the next rankings.

Finding the optimal number of features for conventional feature selection techniques

One of the primary drawbacks of most feature selection techniques is not presenting the optimal number of features. RFFS, ETFS, and PCA prioritize the features based on their importance weights. However, there may not be a practical rule in order to define a threshold for importance weights and remove features from the data set. Hence, it may be impossible to realize the optimal number of features based on importance weights. On the other hand, Lasso and EN can present the optimal number of features by removing unimportant features. Nevertheless, there may be some features with very small coefficients in Lasso and EN, and similarly, there may not be a standard threshold for selecting or not selecting features with small coefficients. Thus, there is a need to improve the performance of these feature selection techniques. In this regard, Equation 4.10 is used for finding the optimal number of features for conventional feature selection techniques. Initially, the features are ranked based on their importance weights. Then, all classification techniques are run by considering the first and the second most important features, and the average value of validation data for all classifiers is calculated. Subsequently, all classifiers are run considering the first, second, and third most important features, and the average value of validation data for all classifiers is calculated. Then, the four most important features are applied, and validation data average accuracy is assessed. This process is continued until the most important $N - 1$ features are employed in the model. Consequently, different combinations of features are compared based on validation data average accuracy, and the optimal number of features is determined for each feature selection technique. Finally, the optimal feature set is used to train all classifiers, and the average value of testing data accuracy is applied to compare the performance of different feature selections.

$$N_{opt}^i = \text{Arg max}\{Accuracy_{val}FS_n^i\} \quad \forall n \in \{2, 3, \dots, N - 1\} \quad (4.10)$$

Where N_{opt}^i is the optimal number of features for conventional feature selection i . $Accuracy_{val}FS_n^i$ represents the validation data accuracy of feature selection i run by considering n features with the highest importance weights in the model.

4.3.4 Sensitivity Analysis

After detecting the best optimal feature set leading to the highest prediction accuracy, a sensitivity analysis is performed to prioritize the optimal features. Initially, one optimal feature is removed from the optimal feature set. Then, all classifiers are run, and their average testing data accuracy is calculated. Afterward, the average testing data accuracy reduction for all classifiers is recorded. This process is performed for all optimal features. The features are ranked based on their average testing data accuracy reduction. Accordingly, the feature with the highest average testing data accuracy reduction is considered the most important feature (first rank) and so on.

4.4 Results and discussions

As mentioned, this research proposes an approach for predicting CC-SoC. In addition to conventional environmental indexes (GEB and NEP) and socio-demographic variables, transport-related features are considered to generate a robust prediction model. Different feature selection techniques are applied to select optimal features. Various classifiers are used to obtain the highest accuracy and spot the best classifier that fits the problem. The results of this investigation are presented here. First, the results of improving conventional feature selection techniques and their optimal number of features are presented. Then, the performance of different feature selection methods is scrutinized. Classification technique performance is then analyzed, and accuracy results are presented. Finally, the results of the proposed sensitivity analysis for the most accurate feature set are presented.

4.4.1 Optimal number of features

Initially, conventional feature selection techniques are run, and feature importance weights are obtained. Then, the optimal number of features are tested incrementally from 2 to 110 by decreasing value of importance weight. All classifiers are run, and the average value of validation data accuracy is calculated for each possible optimal number of features and for each conventional feature selection technique. The results of this analysis are shown in Figure 4.2. As can be seen, the optimal number of features for RFFS, ETFS, Lasso, EN, and PCA is 18, 19, 16, 35, and 17, respectively. A more detailed look at this graph reveals that the applied method enhances the performance of feature selection techniques, even for EN and Lasso that already determine the optimal number of features. The average accuracy of classifiers for EN and Lasso features can be

increased by 2.8% and 0.8% respectively by considering the introduced improvement to find the optimal number of features. Therefore, it can be inferred that the conventional versions of Lasso and EN do not present the optimal number of features if the introduced improvement technique is overlooked in their process. Additionally, these versions of RFFS, ETFS, and PCA can present the optimal number of features. It should be noted that there is not a direct correlation between increasing the number of features and an increase in the prediction accuracy. By increasing the number of features, the accuracy is increased until a threshold, and afterward, it reduces for all feature selection techniques. Thus, applying the improved versions to find the optimal number of features for conventional feature selection techniques in problems with a high number of features can be vital.

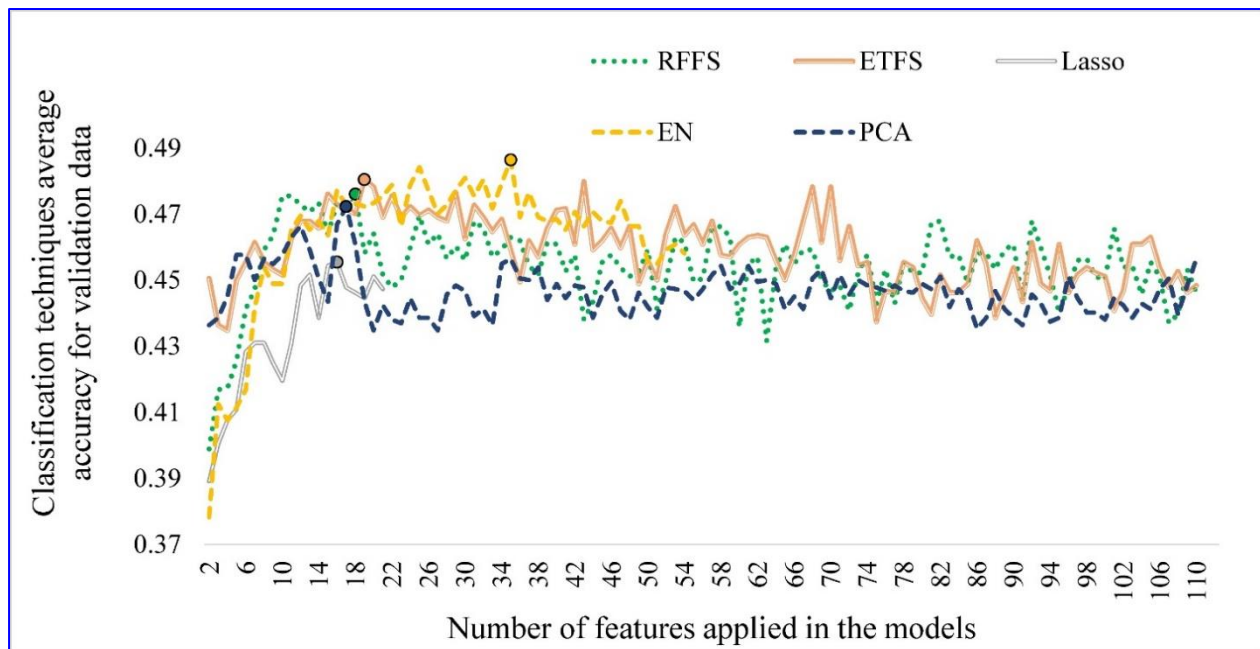


Figure 4.2 The average validation data accuracy of classifiers by considering different values as the optimal number of features.

The optimal feature sets for all feature selection techniques are used to train all classifiers, and then, the accuracy of the testing data is calculated to compare their performance. COA-QDA directly obtains the optimal number of features. The optimal number of features is determined to be 46 by COA-QDA. Furthermore, the optimal value of α_1 and α_2 is 1 and 2, respectively.

4.4.2 Feature selection technique performance

The training and testing data accuracy of all classifiers for different feature selection techniques is shown in Table 4.2. According to the results presented in Table 4.2, COA-QDA provides the highest average testing data accuracy, followed by ETFS, EN, RFFS, all features, Lasso, and PCA. That is to say, the average testing data accuracy of COA-QDA is 0.7%, 0.9%, 2.2%, 3.8%, 4.8%, and 5.6% higher than that of ETFS, EN, RFFS, all features, Lasso, and PCA, considering all classifiers. Thus, it can be inferred that COA-QDA is better at detecting the optimal features for CC-SoC prediction. Meanwhile, applying COA-QDA, ETFS, EN, RFFS can improve the average prediction accuracy compared with a model without using any feature selection. On the other hand, the average testing data accuracy of Lasso and PCA is lower than the all-features model, so the application of these feature selection techniques is not recommended for the CC-SoC prediction problem.

Drawing on the results presented in Table 4.2, the highest accuracy is obtained by COA-QDA, with a value of 53.7%. The maximum accuracy achieved by EN, RFFS, all features, ETFS, Lasso, and PCA is 1.3%, 2.6%, 3%, 3.9%, 5.6%, and 6.1% lower than COA-QDA. Hence, it can be proposed that COA-QDA outperforms other feature selection techniques based on obtaining the highest accuracy. The performance of EN and RFFS are also desirable as their maximum accuracy is higher than that of the all-features model. However, ETFS, Lasso, and PCA cannot improve the accuracy if they are replaced with the model without using any feature selection.

Another purpose of the current study is to find the best combination of feature selection techniques and classifiers to achieve the highest prediction accuracy. For the column of Maximum accuracy in Table 4.2, the highest testing data accuracy is related to COA-QDA optimal features trained by logistic regression (COA-QDA/LR). The combination of COA-QDA and LR leads to the highest testing data accuracy of 53.7%, followed by EN/LR, RFFS/SVM, COA-QDA/NB, COA-QDA/RF, and all features/SVM, with the values of 52.4%, 51.1%, 50.6%, 50.6%, and 50.6%, respectively.

Table 4.2 The accuracy of classifiers based on various feature selection technique's optimal features.

Feature selection techniques	Data type	MLP	NB	LR	DT	KNN	RF	SVM	AB	Average accuracy	Maximum accuracy	Accuracy standard deviation
All features	Training	0.461	0.528	0.623	0.570	0.438	0.979	0.838	0.519	0.620	0.979	0.179
	Testing	0.338	0.481	0.502	0.351	0.390	0.502	0.506	0.459	0.441	0.506	0.066
ETFS	Training	0.571	0.521	0.554	0.557	0.514	0.900	0.687	0.503	0.601	0.900	0.125
	Testing	0.498	0.476	0.494	0.446	0.407	0.494	0.468	0.494	0.472	0.498	0.030
EN	Training	0.735	0.505	0.561	0.554	0.497	0.935	0.754	0.515	0.632	0.935	0.149
	Testing	0.476	0.485	0.524	0.398	0.420	0.489	0.498	0.472	0.470	0.524	0.039
Lasso	Training	0.620	0.473	0.507	0.532	0.477	0.887	0.647	0.474	0.577	0.887	0.133
	Testing	0.437	0.433	0.455	0.416	0.394	0.481	0.437	0.398	0.431	0.481	0.027
RFFS	Training	0.556	0.491	0.517	0.559	0.473	0.914	0.659	0.479	0.581	0.914	0.138
	Testing	0.459	0.481	0.494	0.394	0.381	0.494	0.511	0.446	0.457	0.511	0.045
PCA	Training	0.620	0.408	0.470	0.548	0.473	0.927	0.627	0.477	0.569	0.927	0.153
	Testing	0.398	0.381	0.459	0.420	0.407	0.476	0.455	0.394	0.424	0.476	0.033
COA-QDA	Training	0.529	0.500	0.553	0.569	0.428	0.946	0.768	0.493	0.598	0.946	0.161
	Testing	0.494	0.506	0.537	0.433	0.385	0.506	0.502	0.472	0.479	0.537	0.046

Computational complexity is a vital criterion to compare different soft computing techniques. Running time is a straightforward method that is generally taken into consideration to compare different methods. To this end, the running time of feature selection techniques is evaluated and presented in Figure 4.3. The running time considers the whole-cycle running time, including running the method and the time spent on finding the optimal number of features. As can be seen from Figure 4.3, Lasso requires the minimum time to find the optimal feature set. Lasso's short running time may be due to removing a significant portion of features in the first step. Hence, in the second step, the number of runs for different classifiers is reduced considerably. EN is the second fastest feature selection technique. Hence, considering the average accuracy, maximum accuracy, and running time, EN is the best option among the conventional feature selection techniques. COA-QDA is the third fastest feature selection technique. Thus, the performance of COA-QDA is highly attractive considering average testing data accuracy, highest testing data accuracy, and running time. Therefore, COA-QDA is found to be a competent approach to the CC-SoC prediction problem. PCA, RFFS, and ETFS are the fourth, fifth, and sixth algorithms based on running time ranking.

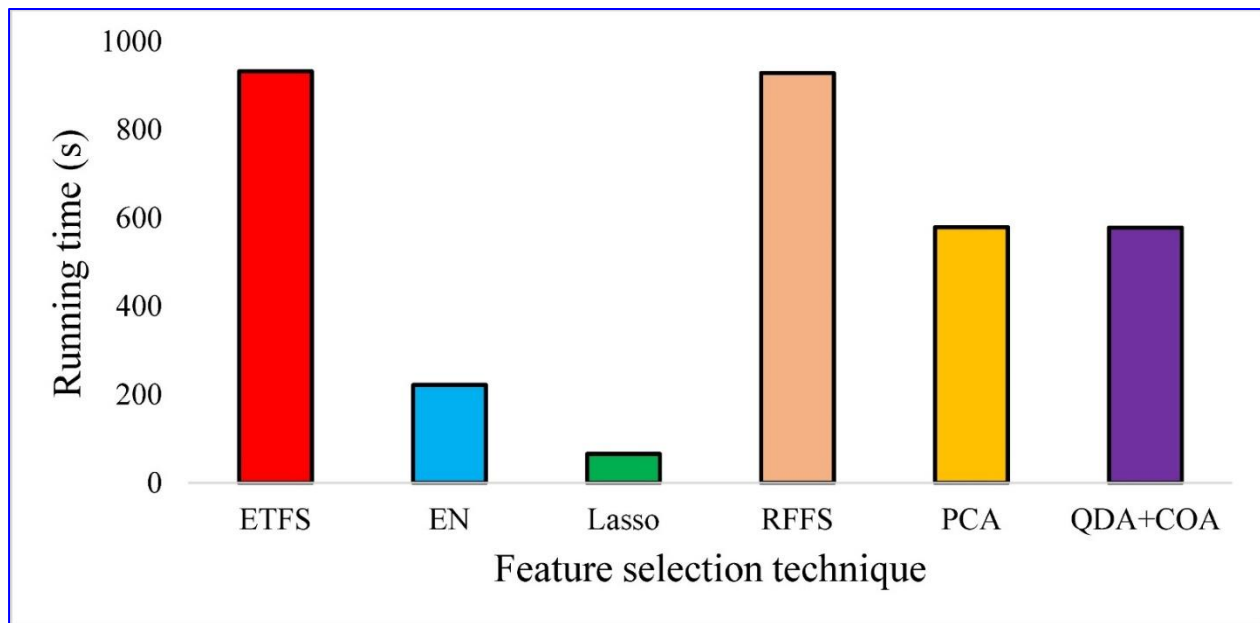


Figure 4.3 Whole cycle running time of feature selection techniques.

4.4.3 Classifiers' accuracy

The average testing data accuracy of different classifiers over different datasets, generated by various feature selection techniques, is presented in Figure 4.4. As can be seen, LR provides the highest average accuracy considering testing data. The average testing data accuracy of LR is 0.1%, 0.76%, 1.57%, 3.36%, 4%, 6.93%, and 7.26% higher than that of RF, SVM, NB, AB, MLP, KNN, and DT, respectively. The average testing data accuracy of all classifiers on all datasets is 42.51%. Considering this value (i.e., 42.51%) as a threshold, LR, RF, SVM, and NB can be considered appropriate classification techniques to predict CC-SoC. On the other hand, the average testing data accuracy of AB, MLP, KNN, and DT was less than the average prediction accuracy of all classifiers. Furthermore, it can be deduced that LR and RF outperform other classifiers based on testing data average accuracy. In contrast, DT and KNN may not be appropriate techniques to predict CC-SoC as they obtain the lowest testing data accuracy.

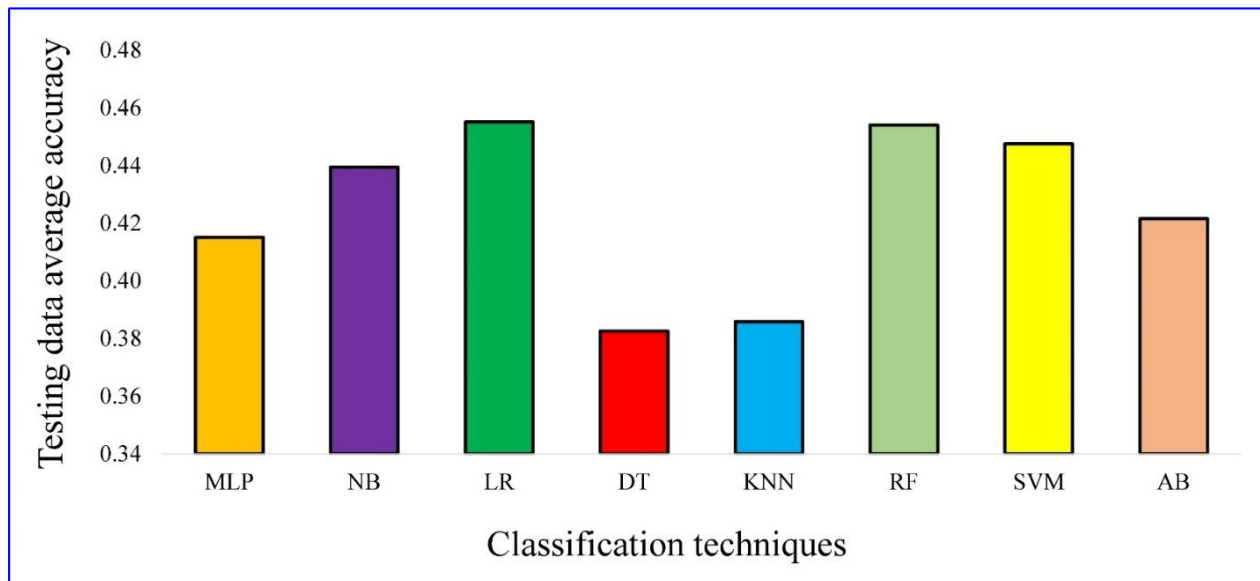


Figure 4.4 The average testing data accuracy of different classification techniques.

4.4.4 The most important features

As mentioned, one of the main purposes of this investigation is to detect the vital features that should be used in classifiers to obtain the highest prediction accuracy. Thus, COA-QDA/LR (LR trained by COA-QDA optimal features), as the most accurate model, is applied in the introduced

sensitivity analysis to prioritize the optimal features. COA-QDA contains 46 features in the optimal features set. Each individual feature is eliminated from the dataset to test for its influence. The model is then run, and the average testing data accuracy reduction of all classifiers is calculated. In other words, the features are ranked based on their effects on the prediction accuracy reduction. The ranking of optimal features is presented in Table 4.3.

As can be seen from Table 4.3, the portion of GEB, transport-related, socio-demographic, NEP, and extra features in the optimal feature set is 45.7%, 19.6%, 15.2%, 13%, and 6.5%. Before highlighting the transportation features, we should point out the sample only contains Americans who own at least one car. In this sample, the production year of the current vehicle is the most important transport-related feature on CC-SoC prediction. Similarly, availability of a car with optional upgrades, expectation time to buy or lease a new car, current car makes, the expected time to keep the next car, annual mileage driving, selecting between purchase or lease, frequency of using a car, and model of the current car are selected in the optimal feature sets, and they should be applied in order to generate an accurate CC-SoC prediction model. Interestingly, six GEB questions in the optimal set are based on transport behavior. Owning a fuel-efficient car, taking a plane for long trips, driving the car into the city, being a member of a carpool, driving in such a way as to keep one's fuel consumption as low as possible, and using public transport for distances up to 20 miles are the transport based GEB questions selecting as the optimal features. Thus, 32.6% of features are related to transport behavior considering transport-related and GEB questions. Therefore, it can be postulated that transport-related behavior can be considered as climate-change related indices, and they should be applied to predict CC-SoC.

Table 4.3 The optimal features for CC-SoC prediction and their corresponding feature groups.

Ranking	Questions (features)	Group
1	What was your total household income before taxes during the past 12 months?	Socio-demographic
2	I buy milk in returnable bottles	GEB
3	How much would you be willing to pay per ton of additional GHG emissions?	Extra features
4	The production year of the current vehicle	Transport-related
5	I talk with friends about problems related to the environment	GEB
6	In summer, I turn the AC off when I leave my home for more than 4 hours.	GEB
7	I own a fuel-efficient automobile	GEB
8	Government rules allow mini-vans, vans, pick-ups, and SUVs to pollute more than passenger cars, for every gallon of gas used	Extra features
9	For long trips (more than 6 hours), I take an airplane.	GEB
10	Do you have the base model or do you have a model with optional upgrades?	Transport-related
11	Age	Socio-demographic
12	When do you expect to purchase (or lease) your next car?	Transport-related
13	I buy convenience foods	GEB

Table 4.3. The optimal features for CC-SoC prediction and their corresponding feature groups (continued)

Ranking	Questions (features)	Group
14	Please select the make of your car	Transport-related
15	What is your relationship status?	Socio-demographic
16	I reuse my shopping bags.	GEB
17	I have pointed out unecological behavior to someone.	GEB
18	How many people have driver licenses in your household (including you)?	Socio-demographic
19	What is your gender?	Socio-demographic
20	Human destruction of the natural environment has been greatly exaggerated.	NEP
21	We are approaching the limit of the number of people the earth can support.	NEP
22	How many people are in your household including you?	Socio-demographic
23	I buy beverages in cans	GEB
24	Plants and animals have as much right as humans to exist	NEP
25	I wait until I have a full load before doing my laundry	GEB
26	I put dead batteries in the garbage.	GEB
27	If I am offered a plastic bag in a store, I take it.	GEB

Table 4.3. The optimal features for CC-SoC prediction and their corresponding feature groups (continued)

Ranking	Questions (features)	Group
28	How long would you plan on keeping your next car?	Transport-related
29	Current car annual mileage	Transport-related
30	Will you purchase or lease your next car?	Transport-related
31	After meals, I throw leftovers in the garbage disposal.	GEB
32	How often do you commute by car?	Transport-related
33	Describe your housing type	Socio-demographic
34	I drive my car into the city	GEB
35	All cars, mini-vans, vans, pickups, and SUVs pollute about the same amount for each mile driven.	Extra features
36	In hotels, I have the towels changed daily.	GEB
37	Please select the model of your car	Transport-related
38	When humans interfere with nature it often produces disastrous consequences	NEP
39	I am a member of a carpool.	GEB
40	I bought solar panels to produce energy	GEB
41	I drive in such a way as to keep my fuel consumption as low as possible	GEB

Table 4.3. The optimal features for CC-SoC prediction and their corresponding feature groups (continued)

Ranking	Questions (features)	Group
42	I requested an estimate on having solar power installed	GEB
43	The earth has plenty of natural resources if we just learn how to develop them	NEP
44	If things continue on their present course, we will soon experience a major ecological disaster	NEP
45	For distances up to 20 miles, I use public transport	GEB
46	I bring empty bottles to a recycling bin	GEB

4.4.5 Comparing the results with previous studies

Ramachandran et al. (2018) compared the performance of random forest classifier and logistic regression on predicting an ordinal variable (fall detection in geriatric healthcare systems). Their study showed that logistic regression outperformed random forest classifier based on prediction accuracy on ordinal variable prediction, which is in line with the outcomes of the current study. Meti et al. (2021) applied five machine learning techniques, including Random Forest classifier, Support Vector Machine, K-Nearest Neighbor classifier, Multi-Layered Perceptron, and Naive Bayes, to predict neoadjuvant chemotherapy response in breast cancer. Subsequently, they compared the prediction accuracy of the mentioned classifiers, and the results indicated that the random forest classifier had a better prediction performance than the other machine learning techniques. Hence, their results are in harmony with the results of this study.

In another study, Vanhoenshoven et al. (2017) compared the performance of different classification techniques, including Multi-Layer Perceptron, Naïve Bayes, Decision Trees, k-Nearest Neighbors, Random Forest Classifier, and Support Vector Machines, on a binary classification problem. The

results demonstrated that Random Forest Classifier was the best classifier in terms of prediction accuracy, which is consistent with the results of this investigation.

Ahmad et al. (2016) employed k-Nearest Neighbors, Multi-Layer Perceptron, Naïve Bayes, Random Forest Classifier, and Support Vector Machine to model a gender recognition task problem. Comparing the prediction accuracy of classifiers revealed that Support Vector Machine was the best classifier to predict gender using speech. Therefore, this outcome contradicts the results of the current study that SVM could not perform well. This contradiction is due to the difference in prediction problems' output. That is, the prediction output variable of this study is an ordinal variable, while a binary variable (i.e., gender) was considered the prediction output in Ahmad et al. (2016) study.

4.5 Conclusions

This study proposed a new AI approach that was applied to predict individual CC-SoC. Behaviors such as recycling may be more commonly thought of as environmental, but transport must be considered as it is a major contributor of CO₂ emissions. As such, so transport's role in predicting CC-SoC was examined. Transport-related behaviors, socio-demographic characteristics, General Environmental Behaviors (GEB; established tool for measuring environmental attitudes and behavior), and New Environmental Paradigm (NEP; established tool for measuring environmental attitudes) features were all employed to generate a prediction model. As the model includes several features (variables), a new feature selection technique was introduced to find the optimal number of features and optimal features to obtain the highest accuracy. Different conventional feature selection methods, including Lasso, Elastic Net, Random Forest Feature Selection, Extra Trees, and Principal component analysis feature selection, were used to select the most valuable feature selections. Moreover, a new approach was presented to improve the performance of conventional feature selection techniques and find their optimal feature sets. Consequently, eight different classification techniques were applied to achieve the highest accuracy. Ultimately, a sensitivity analysis was utilized to prioritize and rank the optimal features. The main conclusions are as follows:

- Fifteen optimal features (out of 46) are based on transport behavior: nine from transport-related questions and six from GEB transport-based questions. Hence, 32.6% of optimal features are related to transport behavior. This suggests that the application of transport

behavior to predict CC-SoC is vital. It should be noted that the original survey focus was on vehicle choice and included only car owners. As such, future research should examine a larger array of transport behaviors with a general population sample.

- The introduced improvement method for conventional feature selection models can increase the average prediction accuracy of EN and Lasso by 2.8% and 0.8%, respectively. RFFS, ETFS, and PCA can also determine the optimal number of features using the proposed improvement method.
- The average testing data accuracy of COA-QDA is 0.7%, 0.9%, 2.2%, 4.8%, and 5.6% higher than that of ETFS, EN, RFFS, Lasso, and PCA. Accordingly, COA-QDA outperforms other feature selection techniques in terms of accuracy. Using an appropriate feature selection technique, such as COA-QDA, can increase the average accuracy by 3.8% as compared to not using all features in the model.
- COA-QDA provides the highest testing accuracy, with a value of 53.7%. The highest COA-QDA testing data accuracy is 1.3%, 2.6%, 3.9%, 5.6%, and 6.1% higher than that of EN, RFFS, ETFS, Lasso, and PCA, respectively. Furthermore, using all features in the prediction models results in a model with 3% lower testing data accuracy than COA-QDA.
- Lasso is the fastest feature selection method regarding the average running time, followed by EN, COA-QDA, PCA, RFFS, and ETFS.
- The highest testing data accuracy is obtained by combining COA-QDA and LR (COA-QDA/LR), followed by EN/LR, RFFS/SVM, COA-QDA/NB, COA-QDA/RF, and all features/SVM. The testing data accuracy these methods is equal to 53.7%, 52.4%, 51.1%, 50.6%, 50.6%, and 50.6%, respectively. This may be a result of the type of dependent variable (ordinal).
- The average testing data accuracy of LR, RF, SVM, NB, AB, MLP, KNN, DT is 45.5%, 45.4%, 44.8%, 43.9%, 42.2%, 41.5%, 38.6%, and 38.3%, in the order given. Therefore, in this study LR and RF outperformed other classifiers based on the average prediction accuracy.

CHAPTER 5 Article 2 : Application of Machine Learning to Child Mode Choice with a Novel Technique to Optimize Hyperparameters

In this chapter, the second objective of this dissertation is taken into account. To this end, a novel method is proposed to optimize hyperparameters using a multi-objective metaheuristic optimization framework. The proposed method is used to predict child travel mode choice and determine which variables influence the child mode choice.

This chapter was accepted for presentation in the “Transportation Research Board 102nd Annual Meeting”, located in Washington DC, United States, in 2023. Then, it was published in “International Journal of Environmental Research and Public Health” on 15/12/2022. The journal version of this research article can be found here: <https://doi.org/10.3390/ijerph192416844>.

The co-authors of this study are:

Owen Waygood, Professor at Polytechnique Montreal, Canada.

Bobin Wang, Professor at Laval University, Quebec City, Canada.

Zachary Patterson, Professor at Concordia University, Montreal, Canada.

Abstract

Travel Mode Choice (TMC) prediction is crucial for transportation planning. Most previous studies have focused on TMC in adults, and predicting TMC in children has received less attention. On the other hand, previous children's TMC prediction studies have generally focused on home-to-school TMC. Hence, Light Gradient Boosting Machine (LGBM), as a robust machine learning method, is applied to predict children's TMC and detect its determinants since it can present the relative influence of variables on children's TMC. Nonetheless, the use of Machine Learning introduces its own challenges. First, these methods and their performance are highly dependent on the choice of "hyperparameters." To solve this issue, a novel technique, called Multi-Objective Hyperparameter Tuning (MOHT), is proposed to select hyperparameters using a multi-objective metaheuristic optimization framework. The performance of the proposed technique is compared with conventional hyperparameters tuning methods, including random search, grid search, and "Hyperopt." Second, machine learning methods are black-box tools and hard to interpret. To overcome this deficiency, the most influential parameters on children's TMC are determined by LGBM, and Logistic Regression is employed to investigate how these parameters influence children's TMC. The results suggest that MOHT outperforms conventional methods to tune hyperparameters based on prediction accuracy and computational cost. Trip distance, the "walkability" and "bikeability" of origin location, age, and household income are principal determinants of child mode choice. Further, older children, those who live in walkable and bikeable areas, low-income groups, and short-distance travelers are more likely to travel by sustainable transportation modes.

Keywords: Children's travel mode choice, Multi-objective hyperparameter tuning, Metaheuristic optimization, Light Gradient Boosting Machine.

5.1 Introduction

Predicting travel mode choice is essential for transportation planning. However, most previous travel mode choice studies have focused on adults, while analyzing TMC in children has received less attention. Possibly, as a result, most transport planning is based on adults' needs, and children are more and more reliant on adults for their transport needs (Freeman & Tranter, 2012). Children are members of society, and travel mode choice has been found to relate to their overall wellbeing. Children travels can influence parents' travel behavior (Zwerts et al., 2007). Also, children become adults, and their childhood behaviors can impact their adult behaviors. Therefore, it is important to examine children's TMC and how different parameters impact healthier and more sustainable mode choices.

In real-life travel behavior, individuals choose between different transportation modes. Therefore, the features of different transportation modes (e.g., travel time, travel cost, and availability of different modes), as well as respondent characteristics, are used as factors influencing the choice between available alternatives (Tuerlinckx & Wang, 2004). Traditionally, different statistical methods, such as structural equation modeling (Golob, 2003), multivariate regression (Handy, 1996), bivariate analysis (Klinger & Lanzendorf, 2016), and fuzzy logic (Přibyl & Goulias, 2003), have been used to determine the factors affecting child travel behavior. For TMC, discrete choice models have been considered appropriate methods to analyze these kinds of datasets, and hence, they have been widely employed (Andrade et al., 2006; X. Chen et al., 2013; Eluru et al., 2012).

Multinomial Logit has been the most used discrete choice model to analyze travel behavior due to its simplicity and straightforward interpretation (E. J. Kim, 2021a). To this end, Multinomial Logit has been applied in many settings, such as daily TMC modeling (Viegas de Lima et al., 2018), modeling the influence of new public transportation infrastructures on TMC (Hensher & Rose, 2007), modeling shopping trips (Andrade et al., 2006), and home to work/school commute modeling (Eluru et al., 2012). The use of the Multinomial Logit requires assuming error terms to be Identically and Independently Distributed (IID), which can reduce prediction accuracy. Therefore, researchers have applied other discrete choice models to relax the IID assumption and obtain still more accurate results. In this regard, Heteroscedastic Extreme Value, Nested Logit, Mixed Multinomial Logit, Multinomial Probit (Aboutaleb et al., 2021; Bhat, 2011; X. Chen et al.,

2013), Joint models, Hybrid Choice (Andani, La Paix Puello, & Geurs, 2021), and Latent Class discrete choice models (Shen, 2009) have been used to model individual travel mode choice.

Recently, due to the widespread application of machine learning techniques in various fields, these learning techniques have received growing attention for travel behavior modeling (Zhao, Yan, Yu, & Van Hentenryck, 2020). Machine learning techniques generally outperform discrete choice models when comparing the prediction accuracy (Pham et al., 2022; Pineda-Jaramillo & Arbeláez-Arenas, 2022; Salas, De la Fuente, Astroza, & Carrasco, 2022; S. Wang et al., 2021). Various machine learning techniques have been employed to predict TMC, such as Support Vector Machines (Assi, Shafiullah, Nahiduzzaman, & Mansoor, 2019), Random Forest (Chang et al., 2019), Naïve Bayes (Chao, 2021), eXtreme Gradient Boosting (J. Liu, Wang, & Xiao, 2021), Kernel Logistic Regression (Martín-Baos, García-Ródenas, & Rodríguez-Benitez, 2021), Softmax Regression (Mi et al., 2021), Adaptive-neuro-fuzzy classification (Minal, Sekhar, & Madhu, 2019), k-Nearest Neighbor (Mohd Ali, Mohd Sadullah, P.P. Abdul Majeed, Mohd Razman, & Muazu Musa, 2020), and Gradient Boosting (Pineda-Jaramillo & Arbeláez-Arenas, 2022). Although machine learning techniques have been found to be accurate techniques for prediction problems (e.g., TMC prediction), two challenges are associated with these techniques:

a) Machine learning techniques require the determination of hyperparameters (Bishop, 2006).

b) Most of the powerful machine learning techniques are “black-box,” and as a result, their results are not easily interpretable (Naseri, Hosseini, Jahanbakhsh, Hosseini, & Gandomi, 2022).

While ML techniques require the determination of hyperparameters, their determination is typically done *ad hoc*. Table 5.1 provides a summary of recent studies on TMC prediction using machine learning techniques and the methods applied to tune hyperparameters. As can be seen, over 30% of those studies did not tune hyperparameters at all, and as a result, their models might suffer from over- or under-fitting. When hyperparameter tuning is undertaken, it is normally done by breaking datasets into training-, validation- and sometimes even testing-datasets. Further, the existence of excessive outliers on validation data can lead to selecting non-optimal values for hyperparameters. To overcome these deficiencies, the application of k-fold cross-validation is recommended (Bishop, 2006). However, only 42.4% of studies shown in Table 5.1 employed the k-fold cross-validation process to tune hyperparameters.

The application of a robust method to tune hyperparameters is vital to develop an accurate prediction model. As shown in Table 5.1, trial and error is the most commonly used method in TMC prediction studies. However, the trial and error method has two major problems: it is a time-consuming technique and depends on modeler experience (Hertel, Collado, Sadowski, Ott, & Baldi, 2020). Accordingly, other researchers have applied systematic methods, including grid search, random search, and Hyperopt. Grid search is a brute force method, and it is not computationally efficient. Random search does not guarantee that optimal hyperparameters are found (Liashchynskyi & Liashchynskyi, 2019). Moreover, all of these methods (i.e., trial and error, grid search, random search, and Hyperopt) only apply a single performance indicator (e.g., prediction accuracy) to tune hyperparameters. However, in many real-life prediction problems (e.g., TMC), datasets are not balanced. For instance, there tend to be many more car than bicycle trips (e.g. (Chapleau et al., 2019)). The promotion of such active modes may be a policy objective, but models using only overall accuracy might not adequately predict low-frequency modes.

Thus, rather than simply using accuracy as a single performance indicator, multiple performance indicators (such as accuracy and F1-score together) should be applied to solve the problem of the imbalanced distribution of transportation modes. Hence, developing a new method that can consider multiple performance indicators in hyperparameter tuning is important, but currently overlooked.

Table 5.1 A summary of recent studies on the application of machine learning to travel mode choice prediction.

Reference	Optimizing hyperparameters	Considering k-fold cross-validation	Method for tuning hyperparameters	Children mode choice
Pham et al. (2022)	✓	✓ (Ten-fold)	Trial and error	✗
Pineda-Jaramillo and Arbeláez-Arenas (2022)	✓	✓	Random search	✗
Kashifi et al. (2022)	✗	✓ (Ten-fold)	-	✗
Salas et al. (2022)	✓	✓ (Five-fold)	Hyperopt	✓
Chao (2021)	✗	✗	-	✗
Brenner et al. (2022)	✗	✓ (Five-fold)	-	✗
Mohd Ali et al. (2022)	✓	✗	-	✗
Li et al. (2021)	✗	✗	-	✗
Tariq and Shakeel (2021)	✓	✗	Trial and error	✓
Aschwanden et al. (2021)	✗	✗	-	✗
Buijs et al. (2021)	✓	✗	Random search	✗
Liu et al. (2021)	✓	✓ (Five-fold)	Grid search	✗
Lu et al. (2021)	✓	✗	Trial and error	✗
Sun and Wandelt (2021)	✓	✗	Trial and error	✗

Table 5.1 A summary of recent studies on the application of machine learning to travel mode choice prediction (continued)

Reference	Optimizing hyperparameters	Considering k-fold cross-validation	Method for tuning hyperparameters	Children mode choice
Kim (2021b)	✓	✗	Grid search	✗
Martín-Baos et al. (2021)	✓	✓ (Two-fold)	Random search	✗
Gao et al. (2021)	✓	✓ (Ten-fold)	Trial and error	✗
Qian et al. (2021)	✓	✓ (Ten-fold)	?	✗
Mi et al. (2021)	✓	✓ (Five-fold)	Grid search	✗
Liang et al. (2020)	✗	✗	-	✗
Mohd Ali et al. (2020)	✗	✗	-	✗
Nam and Cho (2020)	✓	✗	Trial and error	✗
Thanh et al. (2020)	✗	✗	-	✗
Zhou et al. (2019)	✓	✓ (Three-fold)	Grid search	✗
Chang et al. (2019)	✓	✓ (Five-fold)	Grid search	✓
Yang and Ma (2019)	✗	✗	-	✗
Pirra and Diana (2019)	✓	✓ (Five-fold)	?	✗
Richards and Zill (2019)	✓	✓ (Ten-fold)	Grid search	✓
Cheng et al. (2019)	✓	✗	Trial and error	✗

Table 5.1 A summary of recent studies on the application of machine learning to travel mode choice prediction (continued)

Reference	Optimizing hyperparameters	Considering k-fold cross-validation	Method for tuning hyperparameters	Children mode choice
Kim (2021b)	✓	✗	Grid search	✗
Chapleau et al. (2019)	✓	✗	Trial and error	✓
Minal et al. (2019)	✓	✗	Trial and error	✗
Assi et al. (2019)	✓	✗	Trial and error	✓

✓: yes; ✗: no; ?: could not find the in the manuscript

As mentioned, the second problem with machine learning techniques is their black-box nature. To address this, many white-box prediction techniques have been developed, such as programming techniques (e.g., Soccer League Competition (Naseri, Jahanbakhsh, et al., 2020), Water Cycle Programming (Naseri et al., 2019), Coyote Optimization Programming (Naseri, Hosseini, et al., 2022), Marine Predator Programming (Naseri, Jahanbakhsh, et al., 2021)) and M5tree (Yong Wang & Witten, 1997). Programming techniques can not be applied to classification problems. Additionally, M5tree can not represent the influence of variables on the response variable considering all respondents. In this regard, researchers have begun to use ensemble machine learning techniques (e.g. Gradient Boosting) for TMC prediction problems since these methods can present the relative influence of each input variable on the response variable (Chang et al., 2019; Chapleau et al., 2019; J. Liu et al., 2021; Pineda-Jaramillo & Arbeláez-Arenas, 2022). Although ensemble techniques can determine the influence of each variable, they can represent the direction of those influences .

Accordingly, after detecting the input variables with the highest relative influence on the response variable, different methods, such as Accumulated Local Effects (Apley & Zhu, 2020), SHapley Additive exPlanations (SHAP) (Ullah, Liu, Yamamoto, Shafiullah, & Jamal, 2022), Partial Dependence Plot (PDP) (Y. Dong et al., 2022), and Local Interpretable Model-Agnostic Explanations (LIME) (Bas, Zou, & Cirillo, 2021), can be applied to represent in which direction (positively, negatively, linearly, quadratically, etc.) top input variables impact the response variable. However, LIME cannot indicate the influence direction of variables for all respondents, and it is a disaggregated technique. Although SHAP, PDP, and ALE can illustrate the influence direction of variables considering all data samples, they cannot represent whether the behavior of different groups is significantly different or not. Hence, Multinomial Logistic Regression (Fei Wang et al., 2018) is often used to determine the influence direction of variables and detect which groups significantly behave differently.

As can be seen from Table 5.1, although research has been conducted on adult TMC, child TMC has not received enough attention, and this group has been excluded in most of the studies. To address this issue, this study develops a model to predict the TMC of children and determine which variables significantly influence child mode choice for all trips (i.e., not only school-related trips) using an ensemble learning approach. Since conventional techniques to tune hyperparameters may not be highly efficient, and they can only optimize a single indicator during the tuning process, a new technique is proposed in this study that can optimize multiple indicators. The proposed technique can be highly effective for imbalanced datasets (e.g., Montreal TMC data) as the F1-score and accuracy can be maximized simultaneously. After detecting the most important variables on children's TMC, Multinomial Logistic Regression is applied to make the results of the black-box prediction technique interpretable. In other words, Multinomial Logistic Regression is used to represent in which direction top-ranked variables can influence child TMC and how these variables can support sustainable transportation.

In the following sections, the datasets used in this study is initially explained. Then, the developed technique and the conventional techniques applied for tuning hyperparameters are described. Afterward, the results are presented and discussed.

5.2 Methods

The main objectives of this studies are:

- Developing a new method to tune hyper parameters
- Predict child mode choices accurately
- Determine which variables influence the child travel mode choice

The flowchart of the methodology is shown in Figure 5.1. As can be seen, initially, different datasets are merged to develop a comprehensive dataset including many variables. Then, a new method is developed to tune hyperparameters. The developed method is compared with conventional regularisation techniques based on prediction accuracy and running time. The most accurate hyperparameter tuning technique is then used to run the final model. Subsequently, the machine learning technique is run, and the relative influence of variables is determined. Finally, Multiple Logistic Regression is used to interpret the results of the machine learning technique.

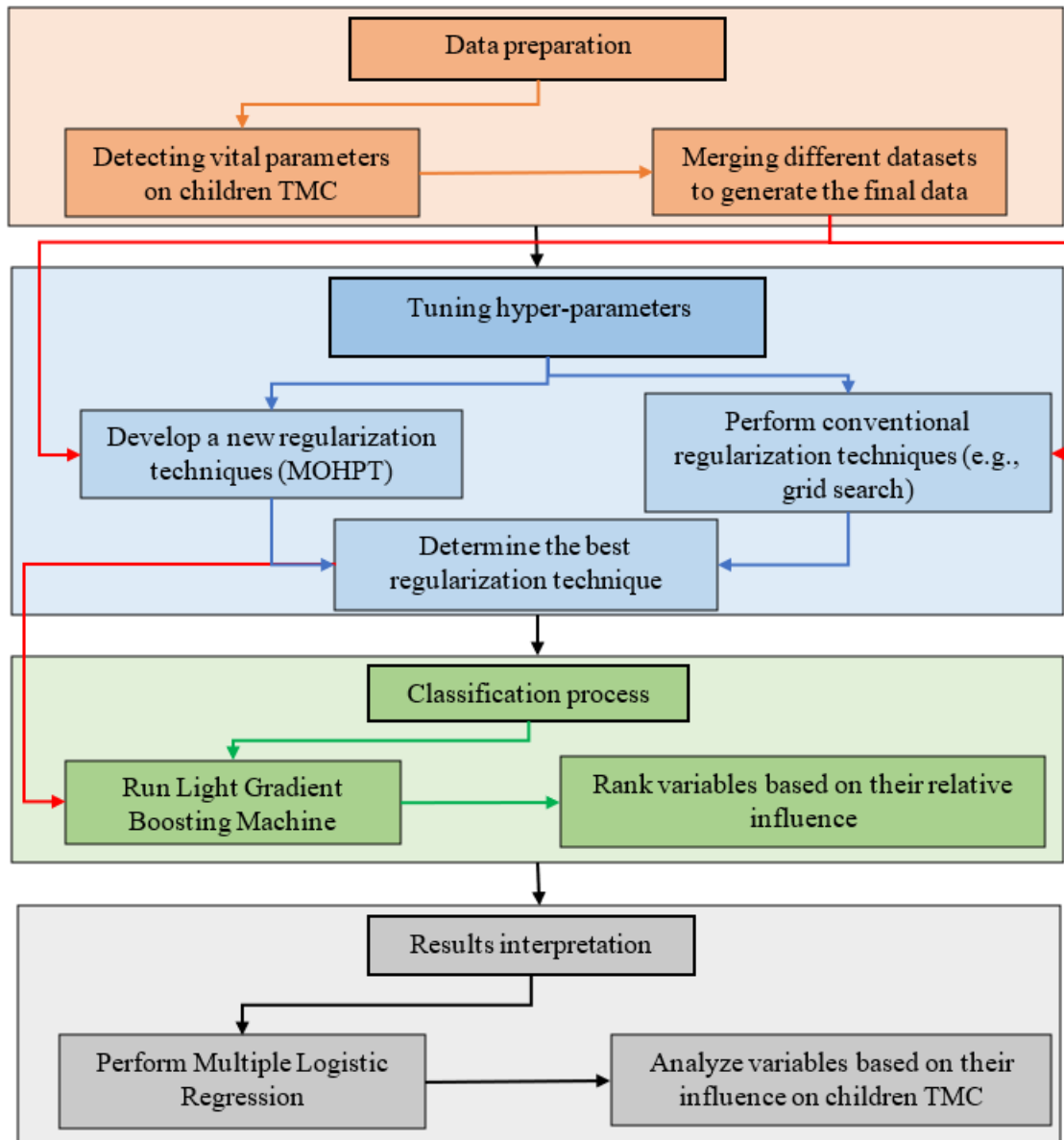


Figure 5.1 The flowchart of the methodology.

5.2.1 Datasets and variables

To the best of the authors' knowledge, most previous studies have used trip details, individual and household characteristics to model TMC. However, in this study, additional variables, such as accessibility, geographic, and land-use variables, are added to the mentioned variables to help explain TMC. To this end, three datasets are taken into account, including the 2018 Montreal OD survey, Walk Score, and Montreal proximity measure data.

The Montreal OD survey was conducted in the fall of 2018, and roughly 400,000 trips were recorded for “an average fall” day. From this survey, 14 variables are considered, including age, gender, availability of a monthly transit pass, disability status, interview language, household income, the presence of people in the household with restrictions in movement, number of members in the household, number of cars in the household, trip distance, start time of the trip, reason for trip, region of origin, and region of destination.

From the Walk Score dataset (Walk Score, 2021), Walk Score, Transit Score, and Bike Score variables are collected. Walk Score measures the walkability of a location according to the distance to different amenities, including schools, parks, restaurants, grocery stores, and coffee shops. Transit Score represents how well a location is served by public transit. The Bike Score indicates how a location is good for biking based on the availability of bike lanes, road connectivity, hilliness, and nearby amenities. These indexes quantify the quality of walking, transit, and biking trips from 0 (worst) to 100 (excellent).

The built environment data is further enriched by using proximity data for Montreal. Ten variables were added including accessibility level to primary school, secondary school, childcare facility, park, library, grocery store, health facility, pharmacy, employment source, and public transit. These indexes measure the closeness of a dissemination block to the mentioned services using a gravity-based accessibility measure. Dissemination blocks are the smallest geographic area bounded on all sides by streets or boundaries of Statistics Canada’s standard geographic areas (Statistics Canada, 2021a). For more information about the mentioned accessibility indexes, please visit: Statistics Canada (Statistics Canada, 2020). In contrast to Walk Score, which provides an overall score, these values are destination specific.

Altogether, 27 variables are applied to explain TMC. The attributes of selected variables are shown in Table 5.2. Since this investigation focused on children’s TMC, the trips of individuals aged from 5 to 17 are taken into consideration (five is the minimum age for trips to be collected on an individual level in Montreal; 18 is considered to be an adult in Canada). The trips where the origin is home are considered because the first trip’s mode restricts the following TMC (Chapleau et al., 2019) and built environment data is collected based on the individual’s residential location. In the final dataset, the number of relevant trips is 9597. These observations are randomly divided into training (80% of total samples) and testing data (20% of total samples). Six transportation modes

were used for the mentioned trips: school bus (18.6%), car as a passenger (33.6%), bus (10.9%), rail transit (6.7%), cycling (2.4%), and walking (27.8%). The share of these modes in the dataset is 18.6%, 33.6%, 10.9%, 6.7%, 2.4%, and 27.8%, respectively. Hence, the share of transportation modes is imbalanced, and it is more appropriate to develop a model that can maximize F1-score as well as accuracy in the hyperparameter tuning process.

Table 5.2 Attribute of selected variables.

Variable	Frequency	Percentage	Variable	Frequency	Percentage
Gender			Household income		
Male	4989	51.98	Less than \$ 30,000	543	5.66
Female	4608	48.02	\$ 30,000 to \$ 59,999	1189	12.39
Availability of a monthly transit pass			\$ 60,000 to \$ 89,999	1654	17.23
Yes	2118	22.07	\$ 90,000 to \$ 119,999	1856	19.34
No	7479	77.93	\$ 120,000 to \$ 149,999	1042	10.86
Disability status			\$ 150,000 to \$ 179,999	603	6.28
Yes	84	0.88	\$ 180,000 to \$ 209,999	376	3.92
No	9513	99.12	\$ 210,000 and more	713	7.43
Number of cars in the household			Refusal	1185	12.35
1	3669	38.23	Don't know	436	4.54
2	4532	47.22	Number of members in the household		
3	686	7.15	1	4	0.04
4	129	1.34	2	299	3.12

Table 5.2. Attribute of selected variables (continued).

Variable	Frequency	Percentage	Variable	Frequency	Percentage
5	31	0.32	3	1484	15.46
6	5	0.05	4	4543	47.34
7	0	0.00	5	2412	25.13
8	1	0.01	6	604	6.29
9	0	0.00	7	170	1.77
10	0	0.00	8	50	0.52
11	0	0.00	9	16	0.17
12	0	0.00	10 and more	15	0.16
13	3	0.03	Interview language		
14 and more	0	0.00	French	8988	93.65
			Other	609	6.35

5.2.2 Modeling

For modeling TMC, an ensemble learning approach is applied due to two reasons. First, the results of recent studies showed that ensemble prediction techniques generally outperform other modeling techniques, such as Naïve Bayes, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Artificial Neural Network, Nested Logit, and Multinomial Logit, in explaining TMC in terms of prediction accuracy (Chapleau et al., 2019; E. J. Kim, 2021a; Mohd Ali et al., 2020; Richards & Zill, 2019). Second, ensemble techniques can prioritize variables based on their relative influence on the response variable (Naseri, Waygood, Wang, Patterson, & Daziano, 2021).

In this study, Light Gradient Boosting Machine (LGBM), a powerful and fast ensemble technique, is employed for the prediction process. LGBM is an updated version of tree-based gradient

boosting developed by Microsoft. Like other ensemble techniques, LGBM combines different weak learners (i.e., decision trees) to form a powerful and robust prediction algorithm (Yang Liu, Lyu, Khadka, Zhang, & Liu, 2020). LGBM is a quick method, and it is highly efficient for large-scale prediction problems. Parallel learning is supported by LGBM, and as a result, memory usage is significantly reduced. A leaf-wise leaf growth strategy is used in LGBM modeling that can limit the depth growth in the splitting process. The mentioned leaf-wise leaf growth strategy splits the same layer of leaves simultaneously. Therefore, LGBM can implement multi-threaded optimization. To this end, the complexity of the model is controlled automatically, and the probability of overfitting is considerably reduced (Z. Zhou, Wang, Huang, Lin, & Lv, 2021).

5.2.3 Tuning hyperparameters

A new technique is proposed to optimize hyperparameters of machine learning techniques considering multiple performance indicators. That is, a new Multi-Objective Hyperparameter Tuning (MOHT) approach is developed in this study. In this regard, Non-dominated Sorting Genetic Algorithm-III (NSGA-III), a multi-objective metaheuristic algorithm, is used as the optimization tool. Genetic algorithms have been widely used to optimize several engineering problems (Naseri, 2019; Naseri, Fani, et al., 2020; Naseri, Shokoohi, et al., 2021). NSGA-III is used for the optimization process since it is a multi-objective metaheuristic optimization technique, and metaheuristic can sync with machine learning techniques (Naseri, Jahanbakhsh, et al., 2022). In this technique, the hyperparameters' values are optimized using an optimization framework. In each iteration of the optimization process, NSGA-III assigns different values to hyperparameters. Then, the machine learning technique is run to evaluate the performance indicators (i.e., accuracy and F1-score) for each of the assigned hyperparameters. Then, the model tries to improve the performance indicators by optimizing the hyperparameters. The optimization modeling of the proposed method is presented in Equations 5.1 to 5.5:

$$Z_1 = \text{maximize Accuracy}^{K-CV} \quad (5.1)$$

$$Z_2 = \text{maximize } F_1^{K-CV} \quad (5.2)$$

$$HP_i^{Int} \in Set_i \quad \forall i \in I \quad (5.3)$$

$$HP_j^{Con} \geq HP_j^{\min} \quad \forall j \in J \quad (5.4)$$

$$HP_j^{Con} \leq HP_j^{\max} \quad \forall j \in J \quad (5.5)$$

Where Z_1 and Z_2 are the objective functions of the proposed optimization model. $Accuracy^{K-CV}$ and F_1^{K-CV} imply the accuracy and F1-score of validation data calculated using K-fold Cross Validation technique. In this study, a 5-fold Cross Validation is used for the tuning process (K=5). HP_i^{Int} and HP_j^{Con} denote integer and continuous-ranged hyperparameters. Set_i is the defined set of integer hyperparameter i . $HP_j^{\min}$ and $HP_j^{\max}$ are the minimum and maximum defined values for continuous-ranged hyperparameters j . I and J represent the number of integer and continuous-ranged hyperparameters, respectively.

Equations 5.1 and 5.2 are the objective function of the proposed technique. That is, in the hyperparameters tuning process, accuracy and F1-score are maximized simultaneously. Equation 5.3 guarantees that the optimal value of integer hyperparameters is selected from their defined set. Equations 5.4 and 5.5 are the constraints that force the model to select the optimal value of continuous-ranged hyperparameters from their allowed range.

As mentioned, NSGA-III is employed to solve the multi-objective optimization problem. NSGA-III is a metaheuristic optimization algorithm that is used for solving multi-objective optimization problems. This algorithm aims to find non-dominated sorting optimal solutions integrating all objective functions rather than converting all objective functions into a single objective function. As a result, NSGA-III presents a Pareto front in which the optimal solutions can not dominate each other based on all objective functions (M. T. Tariq, Hadi, & Saha, 2021). The pseudo-code of MOHT is shown in Figure 5.2.

Three conventional hyperparameter tuning techniques, including grid search, random search, and Hyperopt, are used to evaluate the effectiveness of the proposed hyperparameter tuning approach (i.e., MOHT). Grid search checks all the possible combinations of hyperparameters to find their optimal values. That is, a possible set for each hyperparameter should be defined. Then, all the possible combinations of hyperparameters in the possible set are used to run the model. Finally, the combination that leads to the highest accuracy is considered the optimal value of

hyperparameters. Random search only checks some random possible combinations of hyperparameters and tunes hyperparameters based on the limited number of random combinations (Liashchynskyi & Liashchynskyi, 2019). Hyperopt is an efficient hyperparameter tuning method that applies parallel and serial optimization to efficiently optimize hyperparameters (Bergstra, Yamins, & Cox, 2013).

Grid search is a brute force technique, and it only assigns a limited number of possible hyperparameter values to the hyperparameters' initial set. That is, for hyperparameters with a continuous range, only a few values can be checked, and the optimal value of hyperparameters may not be found. However, grid search is an exact algorithm, and its optimal solution is not changed in different runs. Random search may not find the optimal values for hyperparameters because it assigns random values to hyperparameters. Nonetheless, random search is a quick technique and it is computationally efficient when the number of hyperparameters is significant (Liashchynskyi & Liashchynskyi, 2019). Hyperopt is computationally more efficient than grid search while its running time is generally higher than random search. Furthermore, all these techniques apply a single performance indicator (e.g., accuracy) to tune hyperparameters. To address this issue, this study develops the MOHT.

The defined set for hyperparameters is presented in Table 5.3. As can be seen, grid search cannot cover all the range, and a set with a few possible options should be considered in this technique since it is a comprehensive search method.

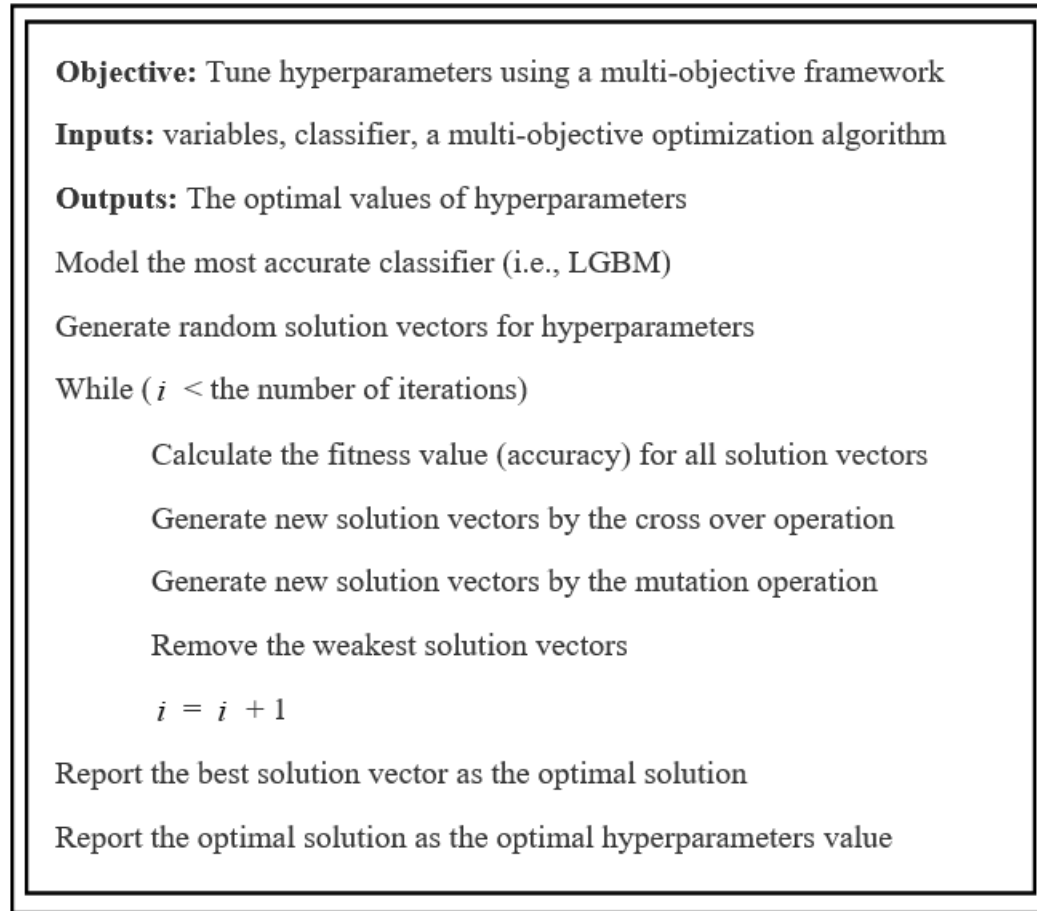


Figure 5.2 The pseudo-code of MOHT.

Table 5.3 Set of hyperparameters for various hyperparameter tuning techniques.

	MOHT	Random search	Hyperopt	Grid search
Number of estimators	{10,11,12,...,510}	{10,11,12,...,510}	{10,11,12,...,510}	{10,50,100,200,300,400,500,600}
Maximum depth	{1,2,3,...,11}	{1,2,3,...,11}	{1,2,3,...,11}	{1,3,5,7,9,11}
Minimum data in leaf	{5,6,7,...,105}	{5,6,7,...,105}	{5,6,7,...,105}	{10,20,30,40,50,60,70}
Learning rate	(0,1]	(0,1]	(0,1]	{0.0001,0.001,0.01,0.1,1}

5.2.4 Results interpretation

Although LGBM can rank variables based on their relative influence on the response variable (i.e., children's TMC), it cannot interpret how each variable (e.g., trip distance) impacts the children's TMC. To solve this issue, after detecting the variables with the highest relative influence on children's TMC, Multinomial Logistic Regression is applied to determine how these top-ranked variables influence the children's TMC. Since Multinomial Logistic Regression can not be converged when the number of variables is significant, the top variables on child TMC is detecting using the relative influence presented by LGBM. Then, those top variables are applied for modeling Multinomial Logistic Regression.

Multinomial Logistic Regression is a robust statistical modeling technique that can be used for classification and interpretation. A set of explanatory variables are used in Multinomial Logistic Regression for assessing the probability of dichotomous outcome events. Dichotomous variables mainly represent whether some events occur or not. In this technique, it is assumed that the relation between the explanatory variables is linear. Therefore, Multinomial Logistic Regression uses linear decision boundaries, but it is a non-linear technique (Naseri, Waygood, et al., 2021). From a sustainable transport perspective, car as a passenger is considered the reference in the Multinomial Logistic Regression to determine how top variables can attract children to use more sustainable transportation modes.

5.3 Results and Discussions

In this section, the results of hyperparameter tuning techniques are initially presented, and the best technique is determined. Then, the ranking of variables based on their relative influence on children' TMC is presented using the most accurate hyperparameter technique and LGBM. Ultimately, the results of a Multinomial Logistic Regression model are represented.

5.3.1 The performance of hyperparameter tuning techniques

The optimal values of hyperparameters using different techniques are shown in Table 5.4. Although MOHT is a multi-objective algorithm and generally provides the users with multiple non-dominated optimal solutions (i.e., a Pareto front), it presents a single optimal solution for the applied case study. If MOHT presents over one optimal solution, it is recommended to apply Gray

Relational Analysis to find the best optimal solution based on the details provided by Naseri et al. (2021).

Table 5.4 The optimal values of hyperparameters.

	Number estimators	of Maximum depth	Minimum data in leaf	Learning rate
Definition hyperparameters	The number of decision trees	The maximum depth of the tree	The minimum data required to be at a leaf node	Convergence magnitude
MOHT	172	10	29	0.319
Random search	50	9	60	0.481
Hyperopt	429	10	54	0.028
Grid search	500	11	10	0.1

The testing data accuracy and testing data F1-score of different hyperparameter tuning techniques are shown in Figure 5.3 and Figure 5.4. As can be seen, the proposed technique in this study (MOHT) obtains the highest testing data accuracy, followed by grid search, Hyperopt, and random search. That is, applying MOHT can increase the prediction accuracy by 1.25%, 2.81%, and 3.59%, respectively, compared to grid search, Hyperopt, and random search. Similarly, MOHT outperforms other techniques in terms of testing data F1-score. The testing data F1-score of MOHT is 1.74%, 3.61%, and 4.89% more than that of the grid search, Hyperopt, and random search, respectively. Therefore, the testing data F1-score improvement of MOHT is more than its prediction accuracy, and it is related to considering both accuracy and F1-score in the objective function of MOHT. Hence, it can be postulated that considering multiple performance indicators in the tuning hyperparameter techniques can improve the overall performance of the model. However, techniques including a single performance indicator can only improve the prediction accuracy and not all vital performance indicators.

The Receiver Operating Characteristic (ROC) of different hyperparameter tuning techniques are indicated in Figure 5.5. Drawing on the results, the highest Area Under the Curve (AUC) of ROC is related to MOHT, followed by grid search, Hyperopt, and random search, with values of 0.81, 0.80, 0.79, and 0.78, respectively. Accordingly, MOHT is the best technique regarding the AUC of ROC. MOHT obtains the highest AUC of ROC for the least frequent mode (cycling) with a value of 0.62, which is 2%, 4%, and 6% more than that of the grid search, Hyperopt, and random search. This improvement results from considering prediction accuracy and F1-score in MOHT, proving that MOHT is highly efficient for modeling this imbalanced TMC dataset. Therefore, considering an optimization framework to tune hyperparameters can even improve the performance indicators that have not been considered in the objective function of the optimization model.

The running time of hyperparameter tuning techniques is presented in Figure 5.6. MOHT can reduce the computational time by 68% and 71% compared to Hyperopt and grid search, indicating MOHT is a highly efficient technique regarding the computational cost. However, the MOHT running time is 2.5 times more than that of the random search. As mentioned, random search checks a limited number of random combinations, and hence, it is the fastest technique. On the other hand, random search is less likely to find optimal values of hyperparameters, and the testing data accuracy and F1-score obtained by the random search are significantly less than MOHT. Therefore, it can be postulated that MOHT outperforms other techniques when considering the testing data accuracy, testing data F1-score, and running time together.

Liashchynskyi and Liashchynskyi (2019) compared the performance of grid search and random search regarding the prediction accuracy and running time. The result suggested that although random search is a faster technique, grid search can obtain higher prediction accuracy. Hence, their results are in line with the findings of this study.

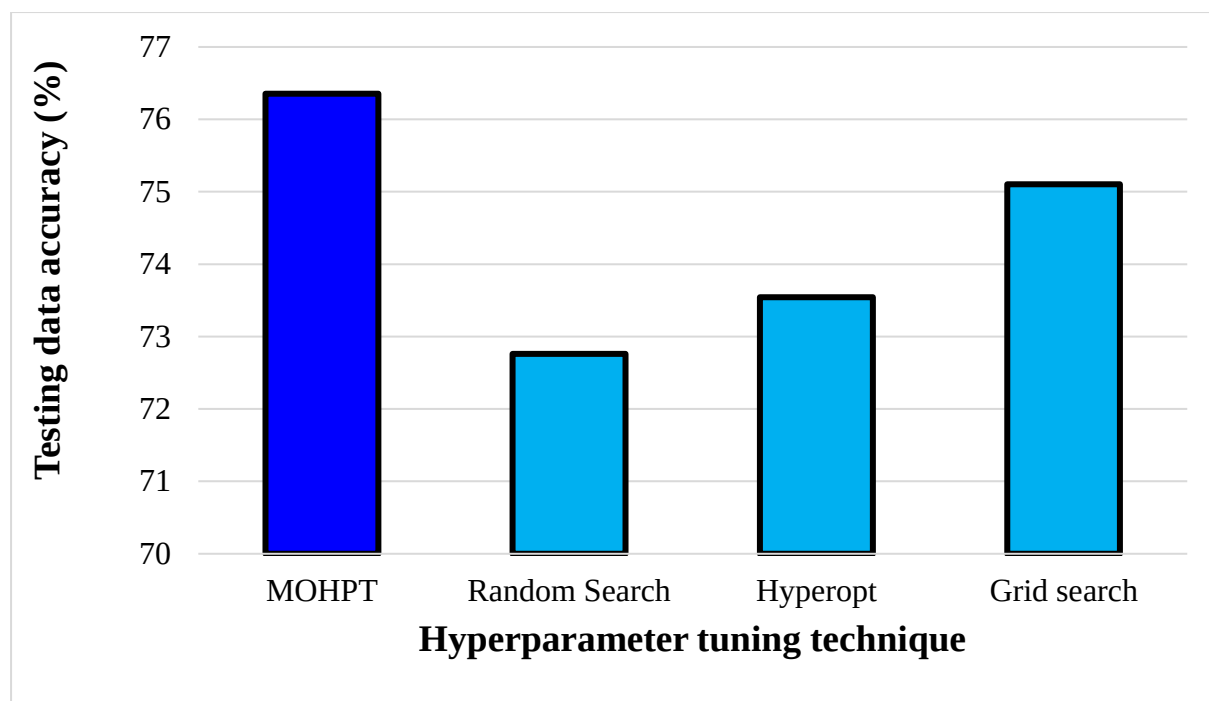


Figure 5.3 The testing data accuracy of hyperparameter tuning techniques.

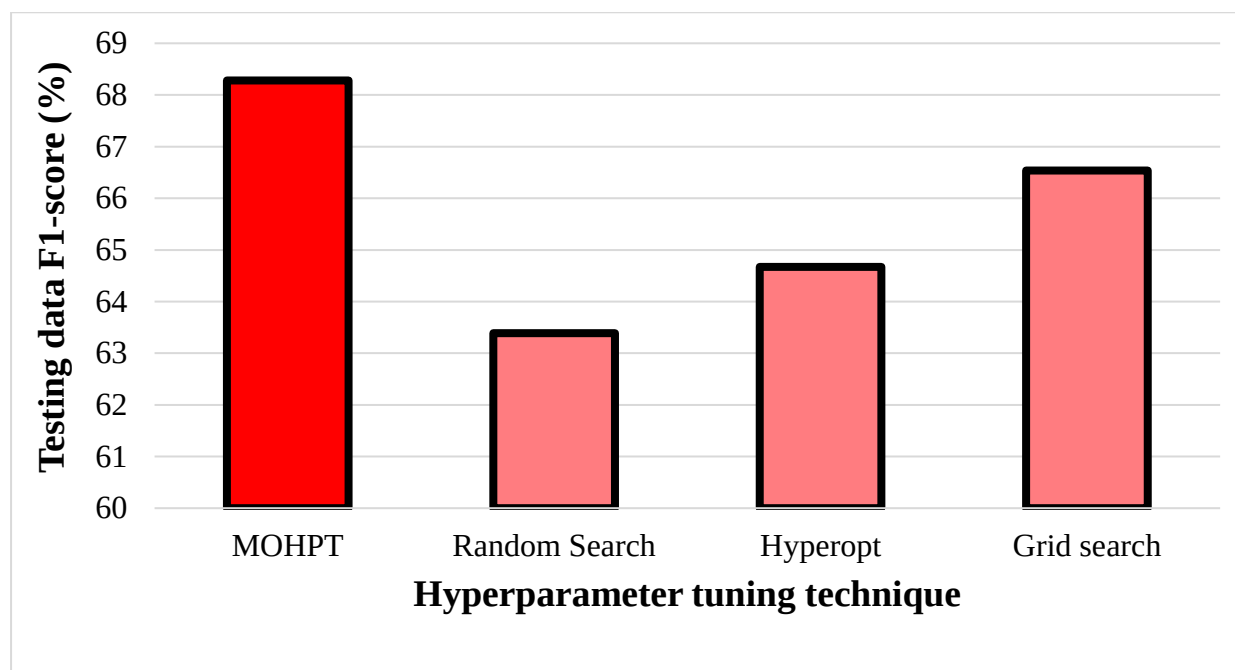
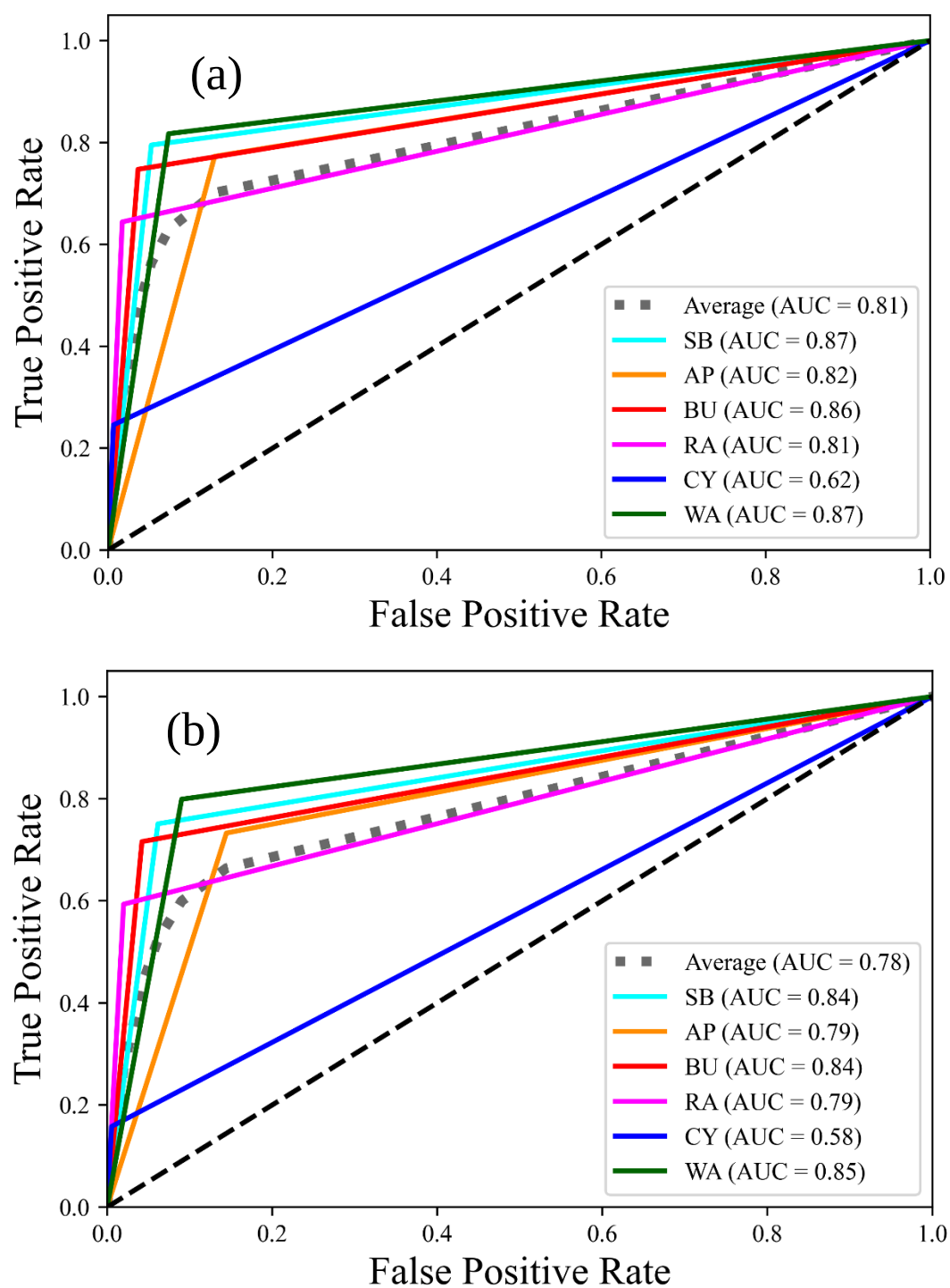


Figure 5.4 The testing data accuracy of hyperparameter tuning techniques.



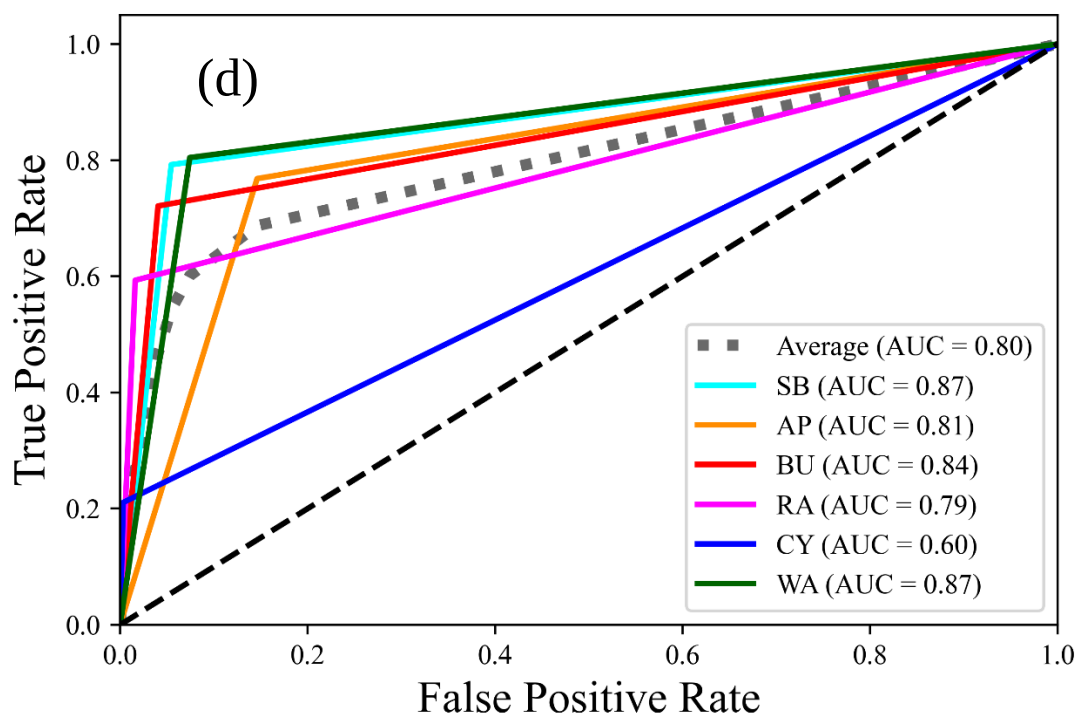
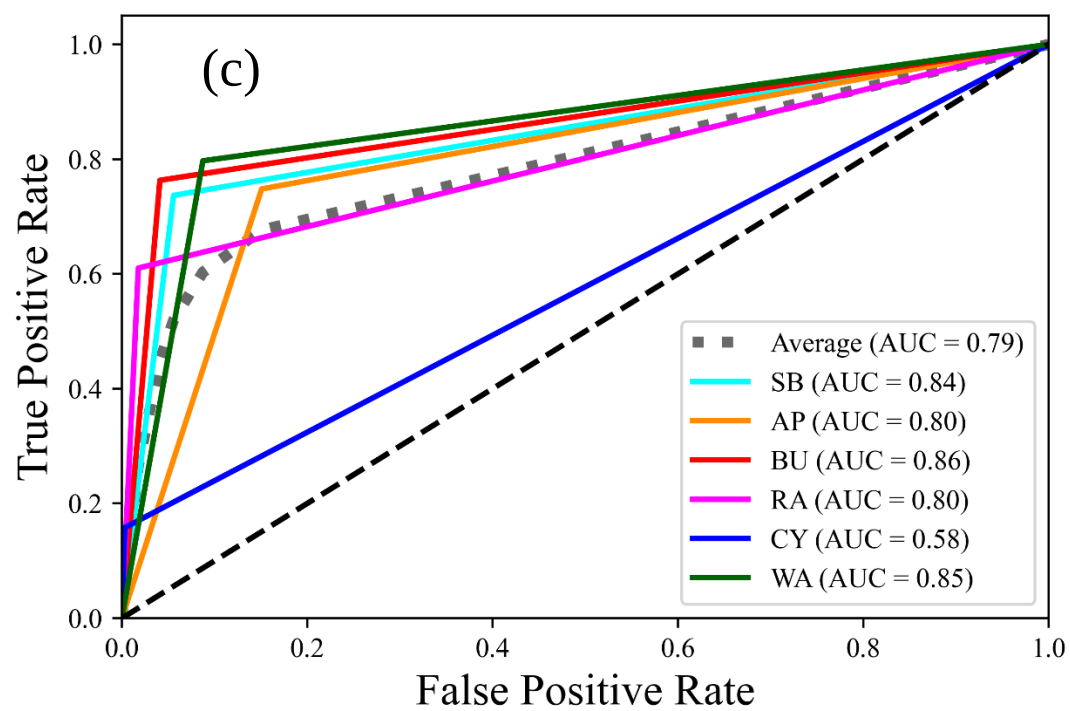


Figure 5.5 The Area Under the Curve of Receiver Operating Characteristic, (a): MOHT , (b): random search, (c): Hyperopt, and (d): grid search; SB: school bus, AP: automobile as a passenger, BU: bus, RA: rail transit, CY: cycling, WA: walking.

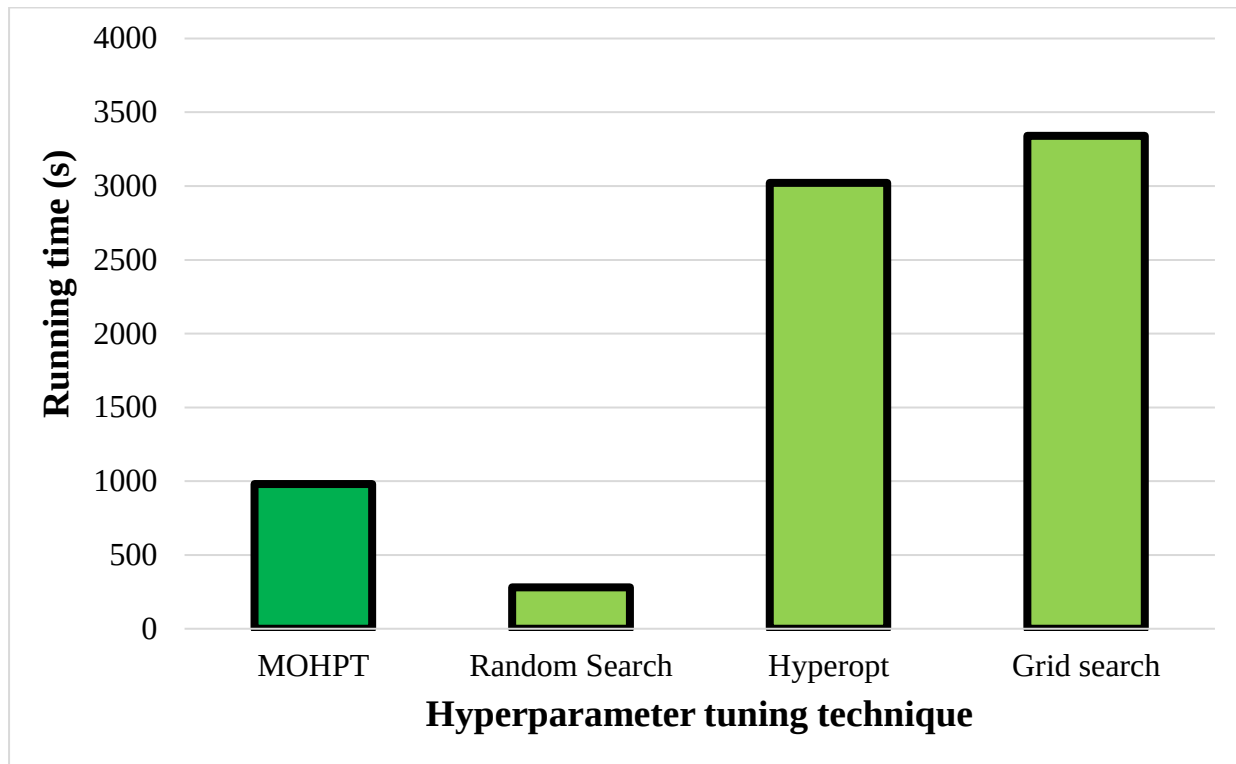


Figure 5.6 The testing data accuracy of hyperparameter tuning techniques.

5.3.2 The relative influence of variables on children's TMC

Since MOHT leads to the highest prediction accuracy, the LGBM is performed using the optimal hyperparameter values found by MOHT. Then, the relative influence of variables is extracted to determine which variables impact the children's TMC the most. The relative influence is illustrated in Figure 5.7. As can be seen, trip distance has by far the highest impact on children's TMC, with a relative influence of 15.5%. Walk Score, age, Bike Score, household income, and accessibility to secondary school are the next variables. The relative influence of other variables is less than 5%. Among accessibility parameters, accessibility to secondary school, accessibility to libraries, and accessibility to grocery stores have the greatest influence on children's TMC.

Wang and Ross (2018) investigated the relative influence of different variables on adults' TMC, and the results suggested that the relative influence of trip distance is significantly higher than the

number of vehicles per capita, population density, and the number of people in the household. Accordingly, their results are in line with the results of the current study.

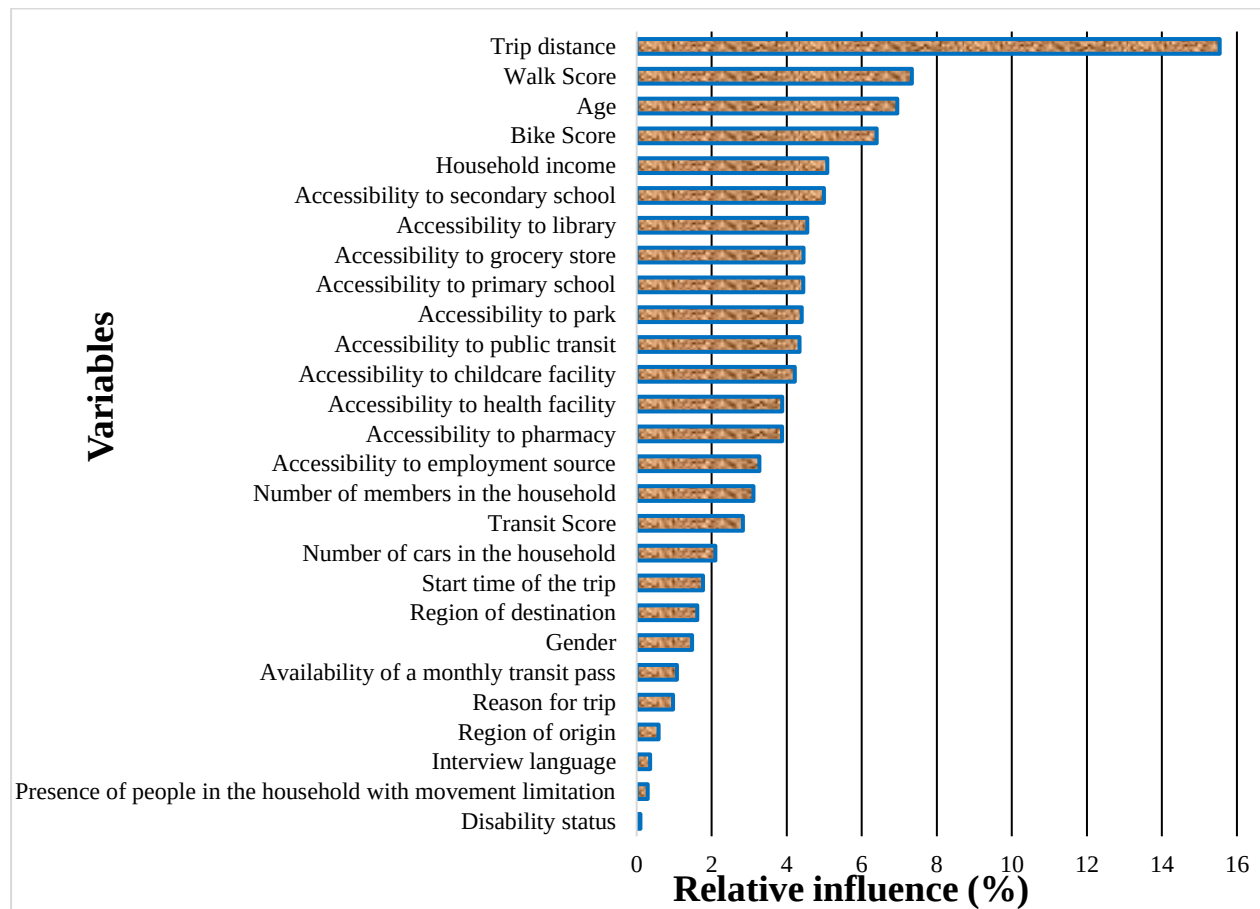


Figure 5.7 The relative influence of variables on children's TMC.

5.3.3 Analyzing the influence direction of top variables

Although LGBM can rank the variables based on their relative influence on the response variable, it cannot determine how changing variables affects the response variable. In this regard, Multinomial Logistic Regression is performed to examine the direction influence of top-ranked variables, and the results are shown in Table 5.5. In the mentioned analysis, the car as a passenger is considered the reference. Based on the results, most of the variables were found to be statistically significant on children's TMC and it may be related to considering top-ranked variables of LGBM in the Multinomial Logistic Regression.

Children are more likely in Montreal to travel by public transit (i.e., rail transit and bus) over the car as a passenger. Nonetheless, they are less likely to travel by school bus or active transportation

(i.e., cycling and walking) compared to car as a passenger. By reducing the trip distance, children are more likely to walk or cycle to their destination, while the probability of traveling by public transit or school bus is reduced. In regions with a lower Walk Score, children are more likely to travel by school bus, and public and active transportation are less likely than car as passengers. Children aged under 12 years are more likely to travel by car as a passenger. In the Kim (2021a) study, age had a considerably higher influence than gender in terms of relative influence on TMC, which is consistent with the results shown in Table 5.5. For those who are aged over 15, rail transit is their preference, followed by cycling, bus, and walking. However, they are not likely to prefer the school bus over the car as a passenger. Bike Score reduction leads to a reduction in the probability of choosing bus, rail transit, cycling, and walking over the car as a passenger. Further, a statistically significant difference is not found between the car as a passenger and the school bus if Bike Score is changed.

Table 5.5 Results of Multiple Logistic Regression (variables are organised by relative influence from most to least).

		School bus		Bus		Rail transit		Cycling		Walking	
		Estimate	SL	Estimate	SL	Estimate	SL	Estimate	SL	Estimate	SL
Constant		-1.121	***	0.489	***	1.509	***	-1.899	***	-2.269	***
Trip distance (km)	0-0.8	-0.918	***	-1.693	***	-4.928	***	2.246	***	5.214	***
	0.8-1.6	-0.195	*	-0.585	***	-2.610	***	1.942	***	3.548	***
	1.6-3.2	0.042		0.033		-1.309	***	1.187	***	1.549	***
	>3.2 (ref)	0		0		0		0		0	
Walk Score	0-46	1.031	***	-1.240	***	-1.768	***	-0.028		-0.773	***
	46-70	0.580	***	-0.685	***	-1.341	***	-0.731	**	-0.603	***
	70-85	0.184	+	-0.247	*	-0.716	***	-0.693	***	-0.412	***
	86-100	0		0		0		0		0	

Table 5.5. Results of Multiple Logistic Regression (variables are organised by relative influence from most to least) (continued).

		School bus		Bus		Rail transit		Cycling		Walking	
		Estimate	SL	Estimate	SL	Estimate	SL	Estimate	SL	Estimate	SL
Age	5-8	-0.800	***	-3.432	***	-3.358	***	-2.147	***	-1.683	***
	9-11	-0.482	***	-2.621	***	-2.107	***	-1.359	***	-1.008	***
	12-15 (ref.)	0		0		0		0		0	
	16-17	-0.503	***	0.570	***	0.913	***	0.662	***	0.415	**
Bike Score	0-59	0.115	+	-0.625	***	-1.310	***	-1.503	***	-0.718	***
	60-71	-0.100	+	-0.349	*	-1.137	***	-1.321	***	-0.726	***
	72-84	0.101	+	-0.179	+	-0.870	***	-0.631	**	-0.614	***
	85-100 (ref)	0		0		0		0		0	
Household income (thousand CAD)	<60	0.771	***	0.623	***	0.487	**	0.170	+	0.073	+
	60 - 120	0.315	***	0.032		0.129	+	-0.030		-0.003	
	>120 (ref)	0		0		0		0		0	
	Prefer not to answer	0.432	***	0.140	+	0.506	***	-0.224	+	-0.102	+
Accessibility to secondary school	<0.048	0.286	**	0.144	+	-0.685	***	-0.255	+	0.103	+
	0.048 - 0.072	-0.222	*	0.140	+	-0.919	***	-0.577	**	0.002	
	0.072 - 0.114	0.170	+	0.091	+	-0.307	*	-0.160	+	0.136	+
	>0.114 (ref)	0		0		0		0		0	

Note: SL= Significance level; +, *, **, and *** imply the significance difference at the levels of 0.1, 0.05, 0.01, and 0.001, respectively.

As compared to high-income households, children in the low-income group (<60 thousand CAD annually) prefer school bus, bus, rail transit, cycling, and walking over the car as a passenger. The middle-income group (60-120 thousand CAD annually) is more likely to travel by school bus as compared to the high-income group, but a significant difference cannot be seen between the car as a passenger and other transportation modes. A reduction in the accessibility to secondary school results in an increment in intention to choose school bus, bus, and walking over the car as a passenger. On the other hand, Reducing the accessibility to secondary school decreases the probability of choosing rail transit and cycling over the car as a passenger.

5.3.4 Managerial implications

Individuals over 15, those who live in regions with higher Walk Score, Bike Score, and accessibility to secondary schools, the low-income group, and short-distance travelers are more likely to travel by active transportation. Moreover, older children (aged over 15), long-distance travelers, residents of regions with higher Walk Score and Bike Score, and the low-income group generally use public transit more than the car as a passenger. Children aged 12 to 15, residents of regions with the lowest level of Walk Score, Bike Score, and accessibility to secondary schools, long-distance travelers, and low and middle-income groups are more likely to travel by school bus than by car as a passenger.

Therefore, improving the Walk Score can increase the share of active and public transportation in child trips. Similarly, the Bike Score needs to be increased if the goal is to promote active transportation in children. Accessibility to schools should be improved if the governments tend to attract children to travel by active transport.

One of the limitations of this study is to only apply NSGA-III algorithm to develop a multi-objective hyperparameter tuning technique. It is recommended to consider other multi-objective optimization algorithms to develop new hyperparameter tuning techniques and compare their accuracy with the method proposed in this study.

5.4 Conclusions

In this study, the travel mode choice of children aged 5 to 17 was investigated using a robust ensemble learning technique, LGBM. To maximize the model's performance, a new multi-objective approach (MOHT) was proposed to tune machine learning techniques hyperparameters. The performance of the proposed technique was compared with the conventional tuning methods. MOHT was demonstrated to be an appropriate technique for tuning hyperparameters of imbalanced datasets (such as travel mode choice) since it can consider multiple machine learning performance indicators in the tuning process. MOHT outperformed other hyperparameter tuning techniques in terms of machine learning performance indicators (e.g., prediction accuracy, F1-score, and AUC). Moreover, this technique could significantly reduce the computational cost compared to grid search and Hyperopt. However, the running time of MOHT was considerably higher than the random search, but it could present more accurate solutions.

The independent variables were ranked based on their relative influence on children's TMC and trip distance. Walk Score, age, Bike Score, household income, and accessibility to secondary schools were the top-ranked variables. Since LGBM cannot represent how these top-ranked variables influence children's TMC, Multiple Logistic Regression was applied to better understand the influence of these variables on children's TMC. In reference to trips by car, the results suggest that as trip distance decreases, active modes are more likely. The built environment, as measured by Walk Score, was positively associated with all sustainable and independent modes as was Bike Score to a lesser degree. As age increases, children use more sustainable and independent modes. Finally, the highest household income was associated with more car as passenger trips, but the relationship with active modes was less strong. The results suggest that policies for a mixed-use development with high-quality public transport networks, such as Singapore's 20-minute towns and the 45-minute city (Manifesty & Park, 2022), can facilitate both local travel and the use of public transport by children.

CHAPTER 6 Article 3 : Interpretable Machine Learning Approach to Predict Electric Vehicles Buying Decisions

In this chapter, the third objective of this dissertation is investigated. In this regard, a new technique based on a metaheuristic optimization algorithm is developed that automatically determines the split that leads to the greatest prediction accuracy. The proposed technique is used to predict individual decisions between Electric Vehicles (EVs) and Internal Combustion Engine Vehicles (ICEVs) and detect what influences the choice of an EV over an ICEV.

This chapter was accepted for presentation in the “Transportation Research Board 102nd Annual Meeting”, located in Washington DC, United States, in 2023. Then, it was published in “Transportation Research Record” on 12/05/2023. The journal version of this research article can be found here: <https://doi.org/10.1177/03611981231169533>.

The co-authors of this study are:

Owen Waygood, Professor at Polytechnique Montreal, Canada.

Bobin Wang, Professor at Laval University, Quebec City, Canada.

Zachary Patterson, Professor at Concordia University, Montreal, Canada.

Abstract

To address the problem of climate change emissions from the transport sector, many countries are promoting electric vehicles (EVs). In order to support such efforts, it is essential to know what influences the choice of an EV over a traditional Internal Combustion Engine Vehicle (ICEV). To study this, a discrete choice experiment was developed, and 2015 valid responses were gathered from Canadian adults with a driver's license. In place of a more traditional analysis, a machine learning approach, XGBoost, was applied. However, two key issues were addressed with respect to its application. First, a practical question related to how to best split the training and testing data was examined. A new technique based on Coyote Optimization Algorithm (COA) is developed that automatically determines the split that leads to the greatest prediction accuracy. The policy-relevant results of the analysis found that an individual's Climate Change-Stage of Change (CC-SoC) and the price ratio of EVs to ICEVs are the most important direct influences. The interaction effect of the first two (CC-SoC and price ratio) is also influential. However, this leads to the second key issue: interpretability. Although high prediction accuracy (87.1%) was achieved, the black-box nature of the approach limits its policy relevance. As such, this research applied a technique, Accumulated Local Effects (ALE), that can determine the strength and direction of influence of the variable. This research demonstrates how machine learning can be applied to a policy-relevant question and provide information that is useful to policy decision-makers.

Keywords: Electric vehicle adoption, eXtreme Gradient Boosting, Accumulated Local Effects, GHG Information framing

6.1 Introduction

The transportation sector is one of the largest sources of greenhouse gas (GHG) emissions around the world (Austmann, 2021). This sector contributes to roughly 24% of CO₂ emissions around the world (Solaymani, 2019) and approximately a fourth of global fossil fuel consumption (Jahanbakhsh et al., 2020). Therefore, governments have been trying to reduce GHG emissions related to this sector (Hardman, 2019). Fossil fuels are responsible for air pollution and polluting the environment (Jalilian, Najafpour, & Khajouei, 2020). One proposed solution to reduce GHG emissions, oil dependency, and air pollution is to replace Internal Combustion Engine Vehicles (ICEVs) with Electric Vehicles (EVs) (Kucukoglu et al., 2021). Although many governments have implemented EV promotion policies such as EV subsidies (Y. Dong, 2022), fuel taxation (Barros

et al., 2021), and improving charging infrastructure (Funke et al., 2019), the EV share in international vehicle stock is low, only nearing 1% in 2020 (Austmann, 2021). Hence, it is essential to investigate the factors affecting EV preference/choice.

Different methods have been used to investigate determinants of EV preferences. Discrete choice models are conventional methods to analyze EV adoptions (Louviere, Hensher, Swait, & Adamowicz, 2000). Although discrete choice models are well-understood and transparent for interpretation, it has been demonstrated that artificial intelligence techniques (e.g., machine learning) often outperform discrete choice models with respect to prediction accuracy (S. Wang et al., 2021). Machine learning techniques can capture non-linear interactions between dependent and independent variables. Moreover, robust machine learning methods are highly flexible in structure and can result in higher accuracies than traditional techniques with predetermined (fixed) structures (Hillel et al., 2021). Further, machine learning techniques can model problems including many features (independent variables), and they are powerful methods in pattern recognition and big data analysis (Koushik et al., 2020).

Recently, machine learning techniques have been applied to predict who prefers electric vehicles and determine which parameters highly contribute to electric vehicle adoption. For instance, Bas et al. (2021) developed a model to predict individuals' electric vehicle adoption. As the survey was done in only one state of the USA, it is not known how generalizable the results are. Different machine learning classifiers were used to solve the prediction problem, such as Gradient Boosting Machine, Deep Learning, and Generalized Linear Model. The results indicated that Gradient Boosting Machine outperformed other classification techniques regarding prediction accuracy with an Area Under the Curve (AUC) of 92.8%. A further analysis using the most accurate classifier (i.e., Gradient Boosting Machine) revealed that county, next vehicle engine type preference, tax deductions, EV price, and EV range were influential parameters of a high willingness to adopt electric vehicles. Attitudes towards EVs and environmental attitudes were also found to have an influence. Subsequently, Local Interpretable Model-Agnostic Explanations (LIME) was applied to interpret the results of machine learning techniques. However, LIME may not be an appropriate method for policy making since it can not represent the influence of variables on intention to buy EVs for all respondents.

Similarly, Bas et al. (2021) proposed a classification model to predict potential electric vehicle purchasers in the United States. The same survey (i.e., the survey used in Bas et al. (2021)) was used as above was taken into account. The results of the leading model found that attitudes towards EVs, household income, and environmental concerns were the top factors. Although the influential parameters of EV adoption could be determined, how these parameters impacted the response variable (EV adoption) was not clear.

Although machine learning techniques are accurate and can capture non-linear relations with features, two major issues are associated with these techniques. The first is a practical use problem: there is no standard procedure to determine the optimal percentage of training and testing datasets for different prediction problems. The second is a policy relevance issue: machine learning techniques are black-box tools, and hence, they are hard to interpret (Naseri, Waygood, Wang, & Patterson, 2022).

Regarding the first issue, datasets should be divided into training and testing datasets. Training data is used to educate the model. Then, unseen data (testing data) is used to evaluate the performance of the prediction techniques (Sanchez et al., 2021). Nonetheless, determining the optimal split for training and testing datasets has been of immense concern. One possible approach to find the optimal share of the mentioned datasets is to check a limited number of randomly selected percentages for training data (e.g., 90%, 80%, 70%) and consider the shares that led to the highest testing data accuracy as the optimal percentage of datasets (Nguyen et al., 2021). However, finding the optimal value of a continuous-ranged parameter by checking a limited number of possible options is impossible (Naseri, Jahanbakhsh, et al., 2020).

The second problem with machine learning techniques is their black-box nature. In other words, machine learning techniques are hard to interpret, and they cannot represent in which direction independent variables affect dependent variables. Researchers generally apply decision tree-based machine learning techniques (e.g., Random Forest and Gradient Boosting Machine) to solve this problem since these techniques can determine the relative influence of each independent variable on the dependent variable. Although decision tree-based techniques can prioritize variables based on their relative influence, they cannot determine how these variables (negatively, positively, linearly, quadratically, etc.) affect the problem's response variable.

To solve this problem, researchers have developed white-box prediction techniques such as metaheuristic programming techniques (e.g., Water Cycle Programming (Naseri et al., 2019), Biogeography-based Programming (Golafshani & Behnood, 2019), Soccer League Competition Algorithm (Naseri, Jahanbakhsh, et al., 2020), Marine Predator Programming (Naseri, Jahanbakhsh, et al., 2021), and Coyote Optimization Programming (Naseri, Hosseini, et al., 2022)) and M5tree (e.g., Wang and Witten (Yong Wang & Witten, 1997)). However, metaheuristic programming techniques can not model classification problems, and they can only be applied for regression problems. Moreover, M5tree divides the data into smaller groups, and hence, it can not represent the direction of influence of the variables for all the respondents. To solve this problem, Accumulated Local Effects (ALE) was developed (Apley & Zhu, 2020) and is increasingly used (K. Kim et al., 2021; Zhang, 2021), though its use in the transportation field has been limited (E. J. Kim, 2021a). This research will apply ALE to demonstrate how it can make the black-box nature of ML results useful from a policy and practice perspective.

In this study, eXtreme Gradient Boosting (XGBoost), a robust tree-based ensemble machine learning technique, was applied to predict who will buy EVs (with respect to both PHEV and BEV). Using the XGBoost, variables could be ranked based on their influence on individuals' intention to buy EVs and the interaction of variables could be examined. Then, a new method is introduced to determine the optimal share of training and testing datasets. After finding the optimal share of datasets, XGBoost was performed to calculate the importance of each parameter on the intention to buy EVs. Moreover, XGBoost was used to understand the interaction of parameters, and the interactions with the highest impact on the likelihood of buying EVs were scrutinized. Ultimately, Accumulated Local Effects (ALE) was employed to make the results interpretable.

6.2 Methods

This study attempted to investigate which parameters influence intention to buy EVs using an ensemble learning approach. The methodology flowchart is illustrated in Figure 6.1. As can be seen, a Discrete Choice Experiment (DCE) was designed and used to collect a dataset. A new method was introduced to maximize the model's accuracy by optimizing the share of training and testing datasets. After optimizing the data split, an ensemble learning method (XGBoost) was performed. Consequently, the variables and their interactions were ranked based on their relative

influence on intention to buy an EV. Ultimately, ALE was used to understand how the model's independent variables and their interactions impact intention to buy EVs.

In this section, the data collection and preparation processes are initially presented. Then, the introduced method and conventional technique to tune the share of training and testing datasets are described. Afterward, the methods used for classification and interpretation processes are explained.

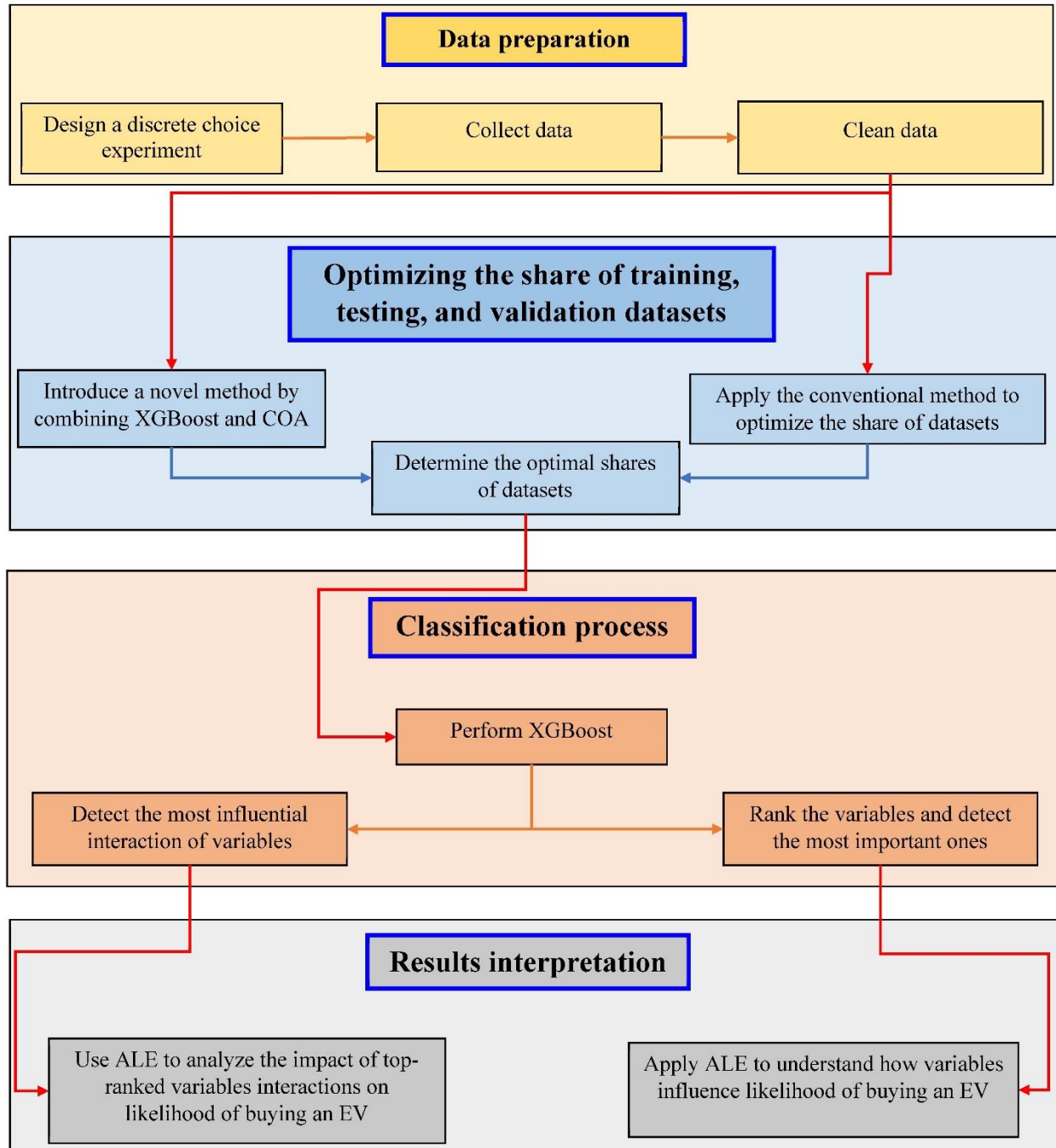


Figure 6.1 Flowchart of the methodology.

6.2.1 Data and survey

A survey was conducted to reach the objectives of this study using a Discrete Choice Experiment. The mentioned survey was administered as an online survey in spring 2021. Only people aged 18 or older (legal adult age in Canada) and driver's license holders could participate in the survey as the survey was related to car purchases. The survey included some trap questions to detect participants who did not pay attention to questions. Such respondents were removed from the dataset. Finally, the survey responses of 2015 participants remained.

Socio-demographic and environmental attitudes

The socio-demographic information and environmental attitudes of respondents are shown in Appendix A (Table A1). Gender, region, age, employment status, education, ethnicity, household income, vehicle ownership, urbanization, and Climate Change Stage of Change (CC-SoC) were the characteristics of respondents collected during the survey. Correlation matrix of socio-demographic information and environmental attitudes of respondents is shown in Table 6.1. As can be seen, the correlation between all variables is not significant, and hence, all variables are applied in the final prediction model. The maximum correlation between variables is related to the correlation between household income and employment status, which is 0.26. CC-SoC is a strong measure to capture behavior and attitudes with respect to personal climate emissions (E. O. D. Waygood et al., 2021). Hence CC-SoC was applied to capture the environmental attitudes of participants. The possible responses for CC-SoC were as follows:

CC-SoC1: I am not concerned about climate change.

CC-SoC2: I am concerned about climate change, but I do not plan to reduce my emissions.

CC-SoC3: I would like to reduce my emissions, but I don't know how.

CC-SoC4: I would like to reduce my emissions, and will do so in the future.

CC-SoC5: I have already reduced my emissions significantly.

Further, responses of Quebec residents (the predominantly francophone province) were classified into two categories to consider the impact of language and culture: French Quebecers (QC-French) and English Quebecers (QC-English).

Table 6.1 Correlation matrix of socio-demographic information and environmental attitudes of respondents.

Correlation	CC- SoC	Region	Gender	Age	Employment status	Education	Ethnicity	Household income	Vehicle ownership	Urbanization
CC-SoC	1.00	0.05	0.06	-0.01	0.05	0.10	0.03	0.02	-0.08	0.08
Region	0.05	1.00	0.01	0.11	0.04	0.00	-0.07	-0.06	0.02	0.00
Gender	0.06	0.01	1.00	-0.07	0.06	-0.02	0.04	-0.04	-0.03	-0.05
Age	-0.01	0.11	-0.07	1.00	0.35	-0.13	-0.22	-0.07	0.18	-0.11
Employment status	0.05	0.04	0.06	0.35	1.00	-0.21	-0.08	-0.26	-0.06	-0.09
Education	0.10	0.00	-0.02	-0.13	-0.21	1.00	0.12	0.25	-0.06	0.13
Ethnicity	0.03	-0.07	0.04	-0.22	-0.08	0.12	1.00	0.10	-0.09	0.13
Household income	0.02	-0.06	-0.04	-0.07	-0.26	0.25	0.10	1.00	0.10	0.03
Vehicle ownership	-0.08	0.02	-0.03	0.18	-0.06	-0.06	-0.09	0.10	1.00	-0.08
Urbanization	0.08	0.00	-0.05	-0.11	-0.09	0.13	0.13	0.03	-0.08	1.00

Discrete choice Experiment

The survey included a discrete choice experiment (DCE) prior to questions on socio-demographics. Half the respondents saw the Climate Change Stage of Change question before the DCE and half afterward to control for any influence of having seen such a question on responses. Discrete choice experiments present survey participants with hypothetical tasks (choice situations). The attributes of alternatives are defined by an experimental design. Participants need to select between alternatives based on their preferences. In this study, the discrete choice experiment was designed according to Wang et al. (2021), which had the same underlying experiment and precise econometric model. Each respondent was asked to select between two alternatives (an EV and an ICEV) in 12 different choice situations (tasks). In each task, the EV could be Battery Electric Vehicle (BEV) or Plug-in Hybrid Electric Vehicle (PHEV).

For the discrete choice experiment design, four attributes were considered in the utility function, including purchase price, fuel costs, electric vehicle range, and CO₂ emissions. Therefore, in each task, the purchase price, annual fuel cost, and annual GHG emission of alternatives varied. The alternatives were labeled as an ICEV, a PHEV, and a BEV. Respondents needed to select between an ICEV label and an EV label, which could be a BEV or a PHEV. For each vehicle, the annual driven distance was assumed to be 20,000 km, following the actual Natural Resources Canada

(NRCan) label. The monthly CO₂ emission and monthly fuel cost were calculated based on the driven distance and the unit CO₂ emission and unit fuel cost of vehicles, reported by NRCan. The current investigation adopted a D-efficient design in Ngene for the discrete choice experiment. The levels used for attributes are summarized in Table 6.2.

Table 6.2 Discrete choice experiment attributes and their levels.

Alternatives	Price (CA\$)	Monthly fuel cost (CA\$)	CO ₂ emissions (kg/month)	EV battery range (km)
ICEV	[17,000; 20,000; 27,000]	[100; 170; 210]	[200; 300; 400]	—
PHEV	[33,000; 38,000]	[55; 62]	[50; 80]	[40; 90]
BEV	[42,000; 48,000]	[30; 40]	[0]	[150; 240; 430]

Vehicle labels were used to present alternatives attributes in the survey. In the designed labels, eight different framings were used to present GHG information (CO₂ emissions) since it was demonstrated that GHG information framing significantly influences individual vehicle engine choices (R. A. Daziano et al., 2017). For further discussion on framing, please see Daziano et al. (2017) or Wang et al. (2021). Each respondent is randomly assigned to one treatment where all the labels use one framing technique for the two-choice tasks. In this experiment, the labels were: current NRCan label, societal-goal framing with color (best performing label from Wang et al. (Wang et al., 2021)), pressure gauge label, new emoticons, goal-oriented with patriotism, dirty air, tree, and thumbs-up thumbs down. For more information about these framings, please see Xun et al. (Xun et al., 2022). Figure 6.2 provides an example for first and second framings. Apart from Figure 6.2(a), all of the emissions are framed with respect to the Government of Canada's objective to reduce GHG emissions by 30% below 2005 levels by 2030, and these framings were designed based on this goal (for more information, please see: Wang et al. (Wang et al., 2021)). An example for other framings is indicated in Appendix A (Figure A1).



(a) frame1: NRCan label and vehicle choice (current label in Canada)



(b) frame2: -30% societal goal with color label and vehicle choice

Figure 6.2 New-vehicle labels with different framing techniques designed to analyze the impacts of different GHG framings on individual's vehicle choices.

6.2.2 Optimizing the percentage of training and testing datasets

The dataset should be divided into training and testing data in a prediction process. In this investigation, a new method, called COA-XGBoost, was developed to optimize the data split ratio. Recently, Nguyen et al. (2021) proposed a new method to optimize the share of training and testing datasets. Therefore, the proposed method by Nguyen et al. (2021) was considered the conventional method.

The conventional method for data split optimization

Split optimization is a research question that was recently addressed by Nguyen et al. (2021). In the current study, different split ratios were taken into account for training, validation, and testing

datasets, including 98/1/1, 90/5/5, 85/7.5/7.5, 80/10/10, 75/12.5/12.5, 70/15/15, 65/17.5/17.5, 60/20/20, 55/22.5/22.5, 50/25/25, 45/27.5/27.5, 40/30/30, and 35/32.5/32.5. In the mentioned split ratios, the percentage of validation and testing datasets was considered equal, according to Majidifard et al. (2021), since validation data acts as testing data in the split ratio optimization. For each split ratio, the model is run ten times using different seeds to randomly split training, testing, and validation data. Then, the split ratio leads to highest average validation data accuracy (over different seeds) is considered the optimal split ratio. Then, the optimal model is run, and the testing data accuracy is calculated.

The introduced method for data split optimization

In this study, a new hybrid method was proposed to optimize the share of training and testing datasets to maximize prediction accuracy. Initially, hyperparameters of XGBoost were tuned using Grid Search and K-fold cross-validation techniques. Then, an optimization framework was developed to maximize the prediction accuracy. The hyperparameters of COA were tuned according to the details provided by Naseri et al. (2022) to tune the hyperparameters of metaheuristic algorithms. Further, the xgboost library in Python was used to perform XGBoost, and COA was coded in Python according to Pierezan and Coelho (2018). Then, an optimization framework was developed to maximize the prediction accuracy. In the introduced hybrid method, XGBoost, a powerful machine learning technique, was applied for classification purposes. Moreover, Coyote Optimization Algorithm (COA), a robust metaheuristic optimization algorithm, was employed to find the optimal solution that resulted in the highest prediction accuracy. The optimization modeling of the proposed approach is presented in Equations 6.1 to 6.5:

$$\max \quad z = Accuracy_{val} \quad (6.1)$$

$$SH_{val} = SH_{test} \quad (6.2)$$

$$SH_{train} \geq SH_{min} \quad (6.3)$$

$$SH_{test} \geq SH_{min} \quad (6.4)$$

$$SH_{train} + SH_{val} + SH_{test} = 1 \quad (6.5)$$

Where $Accuracy_{val}$, SH_{train} , SH_{val} , SH_{test} , and SH_{min} imply the accuracy of validation data, the share of training data, the share of validation data, the share of testing data, and the split threshold, respectively. Equation 6.1 represents the objective function of the proposed optimization model. The objective function aims to maximize the prediction accuracy. In this regard, the model is trained using XGBoost and training data. Then, the accuracy of validation data is calculated, and COA maximizes the validation accuracy by optimizing the split ratio. Equations 6.2 to 6.5 are the optimization model's constraints. Equation 6.2 guarantees that the percentage of validation data and testing data is equal (Majidifard et al., 2021) since validation data acts as testing data in optimizing the split ratio. The training and testing dataset can be null. Hence, Equations 6.3 and 6.4 assure that training and testing percentages are higher than a split threshold. The minimum feasible percentage of training and testing datasets was considered to be 1%, and hence, $SH_{min} = 0.01$. Furthermore, Equation 6.5 guarantees that all the data samples are used in the modeling.

Developed in 2018, COA is a metaheuristic algorithm. Based on *Canis latrans*' social behavior and interactive experience, the COA algorithm was developed. Coyotes are associated with solution vectors, and their social behavior determines their fitness value. The first step is to randomly classify coyotes into different groups. Each group of coyotes is ranked based on its social behavior, and the most valuable coyote is called Alpha. The Alpha of each group and its group mates ultimately influence each coyote. By doing so, solution vectors are moved towards the solution vectors of their group and the most optimal solution in that group. A further method of transferring cultures involves replacing coyotes with those from other groups. In this way, solution vectors are not accumulated in local-minimum areas, but instead are checked in a larger area. In the end, weaker coyotes die, and new generations replace them (Pierezan & Coelho, 2018; Pierezan, Maidl, et al., 2019).

6.2.3 Classification Process

After finding the optimal data split ratio, an ensemble learning technique (i.e., XGBoost) was applied to predict who prefers EV in the choice tasks and investigate which parameters impact EV preference. That is, the individuals' choices (EV or ICEV) were considered the response variable to determine which parameters influence the intention to buy EV. Therefore, the dependent variable was a binary (EV or ICEV) variable. The independent variables were: the ratio of EV to ICEV

price, the ratio of EV to ICEV fuel cost, the ratio of EV to ICEV GHG emission, battery range of EV, EV engine type (BEV or PHEV), GHG framing (treatment), and socio-demographic variables of respondents (i.e., gender, region, age, employment status, education, ethnicity, household income, vehicle ownership, urbanization, and Climate Change Stage of Change (CC-SoC)). Since the survey included 2015 participants, and each participant did twelve choice tasks, the number of observations was $2015 \times 12 = 24180$.

XGBoost was performed to model the classification problem due to four reasons. First, XGBoost is an ensemble learning technique, and hence, it can present the relative influence of variables on the EV purchase likelihood. Second, XGBoost prioritizes the variables' interactions according to the influence they have on the response variables. Third, Problems with high complexity can generally be solved with high accuracy using this technique (T. Chen & Guestrin, 2016). Fourth, XGBoost has outperformed several machine learning techniques in terms of prediction accuracy. For instance, Kim (2021b) compared the prediction accuracy of XGBoost with random forest and artificial neural networks on a travel mode choice prediction problem. The results showed that the accuracy of XGBoost was 2.7% and 8.1% higher than random forest and artificial neural networks, respectively. Similarly, XGBoost was found to be more accurate than logistic regression (Yan Wang & Sherry Ni, 2019), support vector machine (Nguyen-Sy et al., 2020), decision tree (Jamal et al., 2021), etc.

XGBoost is a tree-based prediction technique using boosting modeling in the prediction process. XGBoost is characterized by prediction based on parallel processing and rapid learning (Jeon, Seo, Park, & Choi, 2020). This technique can be used for classification and regression problems, and it generally reaches high prediction accuracy in complicated problems (T. Chen & Guestrin, 2016). XGBoost is more accurate than conventional Gradient Boosting due to using a more accurate approximation (Mahmoud, Abdel-Aty, Cai, & Yuan, 2021). XGBoost generates some weak learners (decision trees) and tries to reach a powerful learner by combining the weak decision trees. An optimization problem is modeled during each iteration of XGBoost to minimize the prediction error. Moreover, the result of the mentioned optimization problem and the residuals are applied to optimize tree structures in each iteration. Residuals are the difference between target values and predicted values. Further, the objective function of the optimization problem contains a regularized term that controls the model's over-fitting. The second-order and first-order gradient statistics are

applied to solve the XGBoost optimization problem and optimize the model's structure (Zhu & Zhu, 2019).

Since XGBoost is a tree-based ensemble technique, it can determine the relative influence of each independent variable on the dependent variable. Moreover, XGBoost can prioritize the variables' interactions based on their influence on the response variable. XGBoost was modeled using the `xgbfir` python library to analyze the interactions of independent variables and detect the most important ones. The library can identify the interactions' importance based on the split points in the XGBoost structure (Hillel, Bierlaire, Elshafie, & Jin, 2019).

6.2.4 Interpreting black-box machine learning techniques

Although XGBoost can present the relative influence of independent variables and their interaction on the response variable, it can not interpret how these variables impact the response variable. To this end, ALE was used in this study to interpret the results of XGBoost. ALE is a plot-based analysis to interpret black-box prediction tools, developed by Apley and Zhu (Apley & Zhu, 2020). ALE divides independent variables by set intervals. Then, the lower and upper bounds of the interval are calculated. The estimated differences are accumulated, and the mean prediction is centered at zero (K. Kim et al., 2021). More details about ALE can be found in Apley and Zhu (Apley & Zhu, 2020).

6.3 Results and discussions

In this section, the optimal split ratio is initially presented. Then, the relative influence of variables and their interactions on the intention to buy EVs are introduced. Finally, ALE results are indicated to understand how variables impact the intention to buy EVs.

6.3.1 Optimal split ratio

As discussed, different values were considered in the conventional model to find the optimal value of testing data. The results of the conventional model are shown in Figure 6.3. In this figure, the range of validation data accuracy and the average validation data accuracy over different seeds are presented. As can be seen, the highest average validation data accuracy was achieved when the percentage of validation data was considered 7.5%. Accordingly, the optimal percentage of testing data was 7.5% in the conventional split ratio optimization model. Therefore, the accuracy of testing

data for the optimal split ratio (7.5%) was calculated, and the model reached the testing data accuracy of 83.6%.

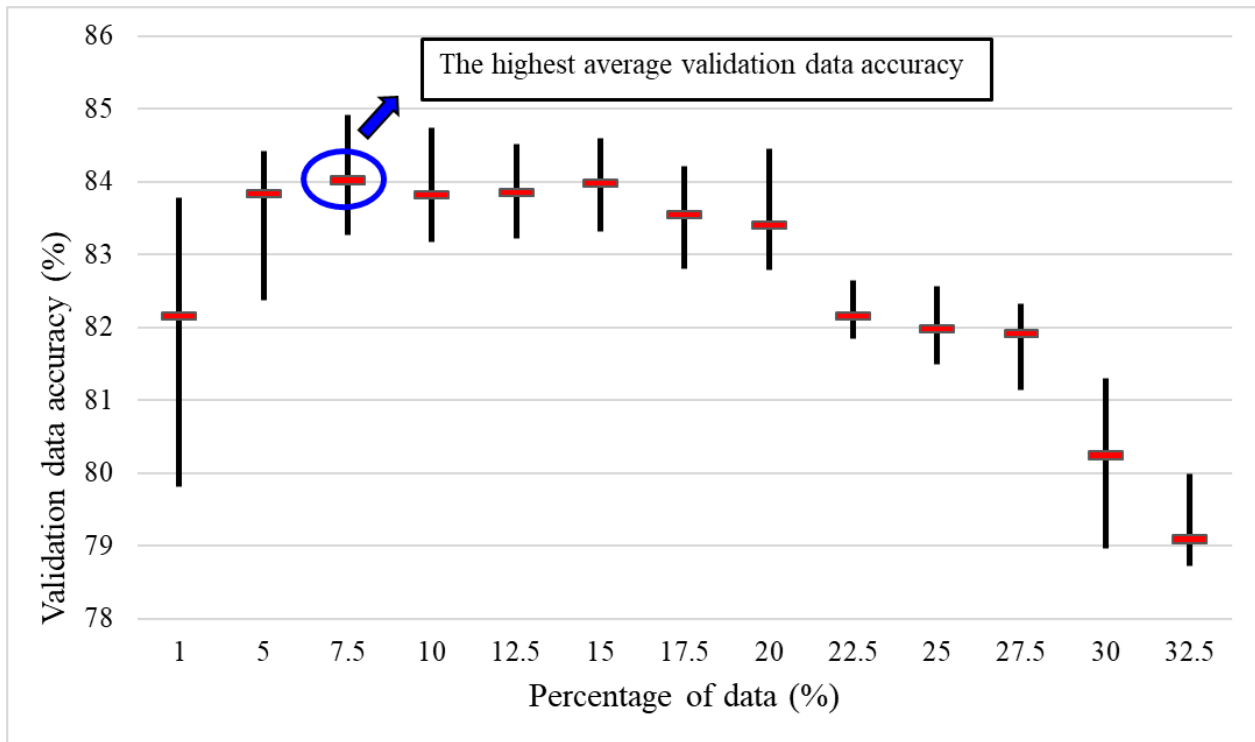


Figure 6.3 Results of the conventional model for split ratio optimization.

The hyperparameters of COA and XGBoost were first tuned. In XGBoost, the number of estimators, minimum data in leaves, maximum depth, and learning rate were 200, 11, 6, and 0.3, respectively. In COA, the optimal number of packs and the number of coyotes in each pack were 10 and 8. Then, the introduced method in this study (i.e., COA-XGBoost) was performed, reaching a validation accuracy of 87.1%. In the optimal solution, the validation data percentage was 3.93%. Hence, the percentage of testing data was considered 3.93%, and then the accuracy of testing data was calculated. In the proposed optimal solution, the testing data accuracy was 87.4%. Accordingly, applying COA-XGBoost to optimize the split ratio increased the testing data accuracy by 3.8% (from 83.6% to 87.4%), and the introduced method is highly qualified to maximize the accuracy of prediction models. Since COA-XGBoost obtained a higher accuracy, it was used to determine the relative influence of variables, and the results are presented in the following part.

6.3.2 Contribution of variables to increase the intention to buy EVs

COA-XGBoost was performed to determine the relative influence of variables on intention to buy an EV, and the results are illustrated in Figure 6.4. As can be seen, CC-SoC, as a strong measure to assess the level of concern about climate change, is by far the most influential variable on vehicle choice. Hence, it can be postulated that concern about climate change plays a crucial role in the intention to buy EVs. Bas et al. (2021) investigated the effects of different variables on electric vehicle adoption, and the results showed that awareness of environmental protection was one of the most important variables in explaining the high willingness to adopt EVs. Moreover, previous studies demonstrated that people with different levels of CC-SoC are statistically different in willingness to buy EVs (Wang et al., 2021; E. O. D. Waygood et al., 2021).

EV to ICEV purchase price ratio is the second significant parameter of vehicle choice since its contribution to the likelihood of buying an EV is roughly 9%. It has been demonstrated that monetary incentives, such as income tax deduction for EVs and vehicle price, significantly influence individuals' choice when selecting between an ICEV and an EV (Bas, Zou, et al., 2021; Wang et al., 2021). The mentioned results are consistent with the outcomes of the current research. This result would suggest that reducing the difference in vehicle costs (either by increasing the costs of ICEV or reducing the costs of EVs) is a policy-relevant lever.

Contextual and individual variables are the next most influential. The region, age, and education are the following variables based on the relative influence on intention to buy an EV. GHG information framing is the sixth influential parameter on vehicle choice, with an importance weight of 6.73%. Interestingly, the impact of GHG information framing is much more than some other vehicles' attributes, such as EV to ICEV fuel price ratio, EV battery range, the ratio of EV to ICEV GHG emission, and EV engine type (BEV or PHEV). Therefore, it can be concluded that it is not sufficient to simply present GHG emissions, but to present them in efficient ways.

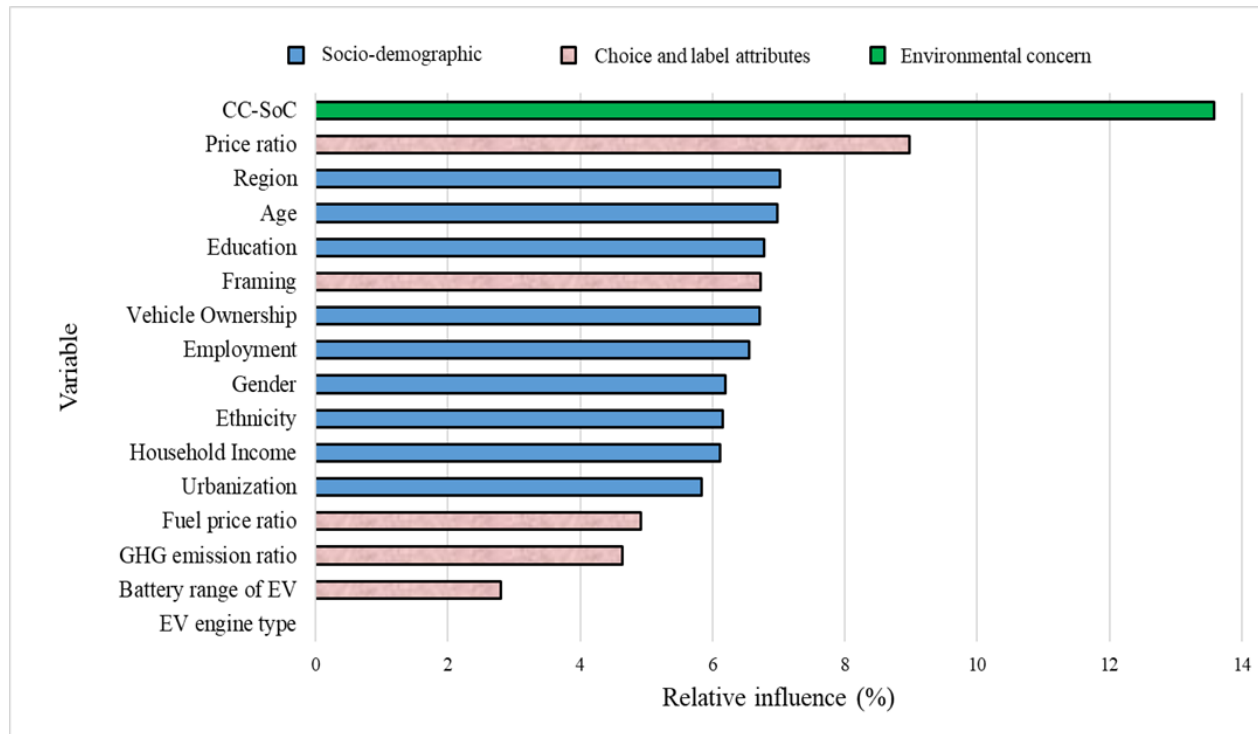


Figure 6.4 The relative influence of socio-demographic, environmental concern, and vehicle attributes on individuals' intention to buy EVs.

Further, the interaction of variables was ranked based on their impact on vehicle choices using XGBoost. The top-ranked variables' interactions are shown in Figure 6.5. The interaction of CC-SoC and EV to ICEV purchase price ratio has the strongest influence on the intention to buy EVs.

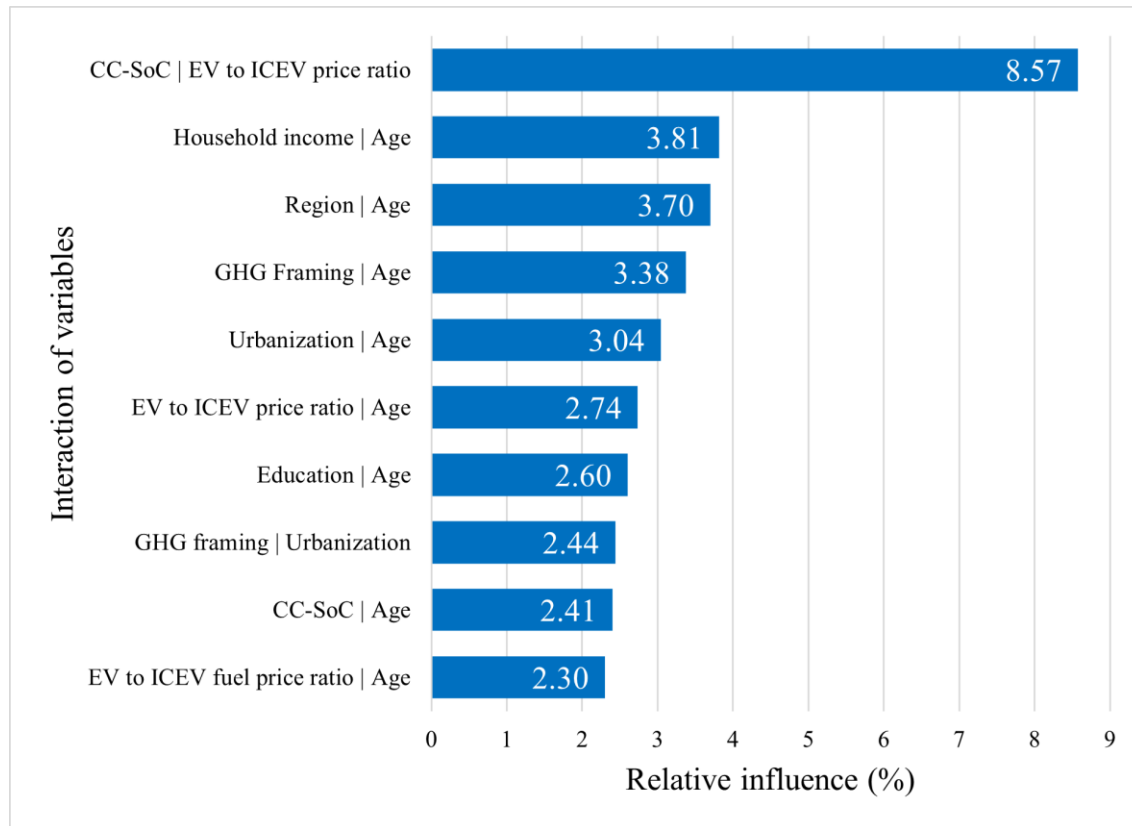


Figure 6.5 The relative influence of variables' interactions on individuals' intention to buy EVs.

6.3.3 Interpretation of machine learning results

In general, machine learning methods are considered black-box tools because it is impossible to obtain the influence direction of inputs on output variable (Naseri, Ehsani, et al., 2021). As a result, the results of machine learning techniques may be misinterpreted (Momenzadeh, Shamsa, & Meyer, 2022). Further, interpretation of machine learning results is essential to provide policy makers the required information (Posada-Quintero, Molano-Vergara, Parra-Hernández, & Posada-Quintero, 2020). That is, the interpreted results of machine learning techniques can be used to set new policies to promote EVs. interpretation of machine learning may be an exciting avenue for researchers to increase the prediction accuracy of their models by exploiting prior knowledge along with any other benefits of interpretation (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). In other words, researchers can detect which variables have the highest influence on the response variable (i.e., the intention to buy EVs) and how these variables impact the intention to buy EVs. Therefore, the top variables can be applied to better investigate inclined EV buyers.

In this section, the results of ALE for variables are first presented. Then, the outcomes of ALE for top-ranked interactions are discussed. It should be mentioned that size in ALE plots represents the number of observations in each group.

ALE results for variables

The effect of CC-SoC (the most influential parameter on intention to buy EVs) on the likelihood of EV purchase is shown in Figure 6.6. In ALE figures (e.g., Figure 5.6), the bar chart implies the number of data samples (size) in the data set, and the bar chart values are shown on the right axis. For example, the number of data samples in CC-SoC1 is 2664 (222 individuals and 12 choice tasks for each for them: $222 \times 12 = 2664$). The trend line indicates the EV purchase likelihood, and the probability values can be seen on the left axis. The EV purchase likelihood of the two groups can be compared by subtraction of their corresponding EV purchase likelihood. For instance, if the EV purchase likelihood of group 1 and group 2 is -2% and 4.2%, the EV purchase likelihood of group 2 is 6.2% ($4.2 - (-2) = 6.2$) higher than that of group 1. Further, the value of zero in the left axis denotes the middle group, who their EV purchase likelihood is the average of all respondents.

As shown in Figure 6.6, individuals who stated “they have reduced emissions (CC-SoC5)” or “they will reduce their emissions (CC-SoC4)” are more likely to purchase EVs. Knowing what percentage of the population is at these stages of change will help policy makers tune their approaches. On the other hand, people who are not concerned about climate change (CC-SoC1) and who are concerned but do not plan to reduce their emissions (CC-SoC2) are significantly less likely to buy EVs. In other words, the likelihood of buying an EV for the CC-SoC1 and CC-SoC2 groups is roughly 26% and 28% less than that of the CC-SoC5 group. Wang et al. (Wang et al., 2021) investigated the influence of CC-SoC on willingness to pay for EVs. Their investigation showed that the CC-SoC5 group was more likely to buy an EV, followed by CC-SoC4, CC-SoC3, CC-SoC1, and CC-SoC2 groups. Moreover, individuals who stated they are in the CC-SoC1, CC-SoC2, and CC-SoC3 stages were found to be statistically different from the CC-SoC4 group when comparing their intention to buy EV. However, the difference between people who stated, “they have reduced emissions significantly (CC-SoC5)” and “they will reduce their emissions (CC-SoC4)” was not statistically different. Therefore, the results of this study are in harmony with the results presented by Wang et al. (2021), where Multinomial Logit was applied.

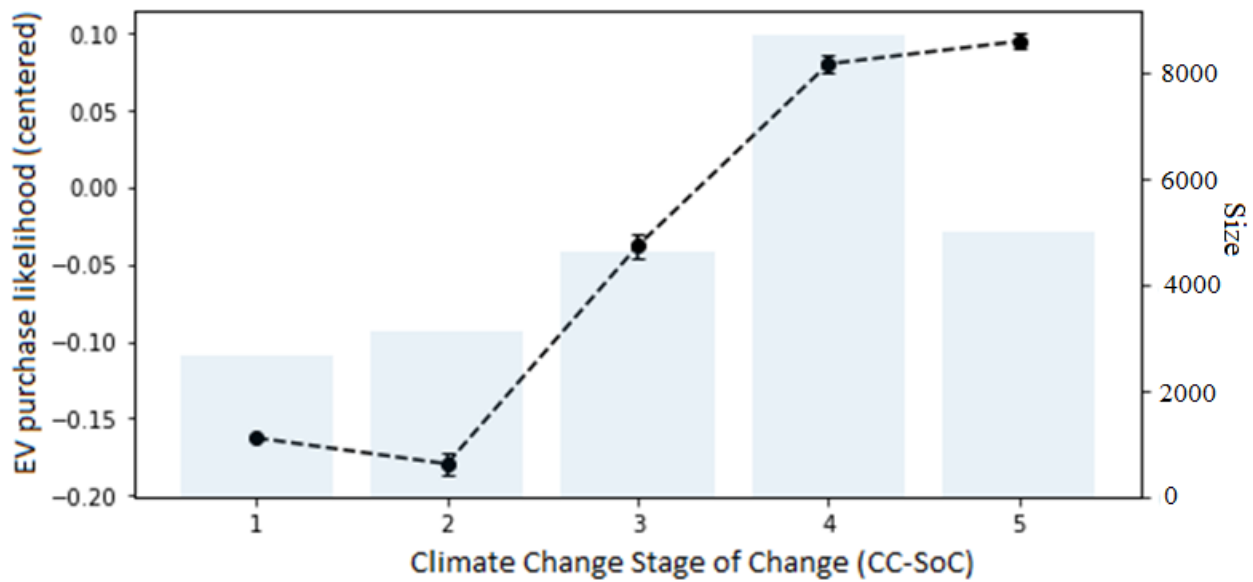


Figure 6.6 Effect of CC-SoC on EV purchase likelihood where: 1 = not concerned; 2 = concerned, but don't plan to reduce emissions; 3 = concerned, but don't know what to do; 4 = concerned, and planning to reduce emissions; 5 = concerned, and have significantly reduced emissions.

The effect of EV to ICEV purchase price on the likelihood of EV purchase is indicated in Figure 6.7. As can be seen, increasing the purchase price ratio of EV to ICEV from 1.22 to 1.56 reduces the intention to buy an EV sharply. By increasing the purchase price ratio from 1.22 to 1.56, the likelihood of buying an EV is reduced by approximately 20%. After this level (the ratio of 1.56), the intention to buy EVs is reduced slightly by EV to ICEV purchase price increment. The probability of EV preference reaches its lowest level when the purchase price ratio is 2.82. The likelihood of buying an EV is over 30% decreased if the EV to ICEV purchase price is increased from 1.22 to 2.82. A more detailed look at Figure 6.7 reveals that the maximum EV to ICEV purchase price ratio should be 1.78 if the aim is to increase EV adoption. Because the effect of price ratio on intention to buy EV is zero when the ratio is 1.78 and increasing the ratio to more than 1.78 results in a reduction in intention to buy EVs.

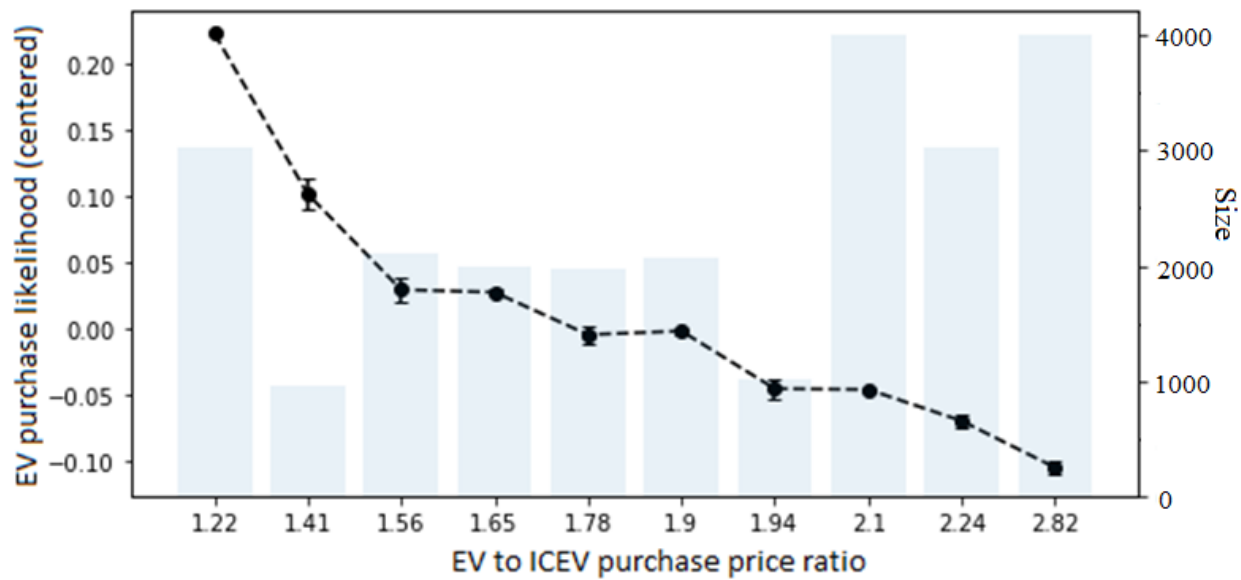


Figure 6.7 Effect of EV to ICEV purchase price ratio on EV purchase likelihood.

The effect of age on the likelihood of EV purchase is displayed in Figure 6.8. The intention to buy EVs does not change significantly for those younger than 33 years old. After 38 years of age, the intention to choose EVs is reduced, and it reaches its minimum level for people aged between 50 and 53 years. On the flip side, the willingness to buy an EV is gradually increased by increasing the age from 58 to 87 years. The oldest group is expected to be the wealthiest group, and it might be much easier for them to afford to buy EVs, which are more expensive than ICEVs. The previous study demonstrated that people aged 50 to 59 have the least willing to pay for EVs, and those under 40 or more than 60 are more willing to pay for EVs in Canada (Wang et al., 2021). Therefore, their results are in line with the results obtained in this study. The additional benefit here is that the non-linear relationships can be seen on a finer level.

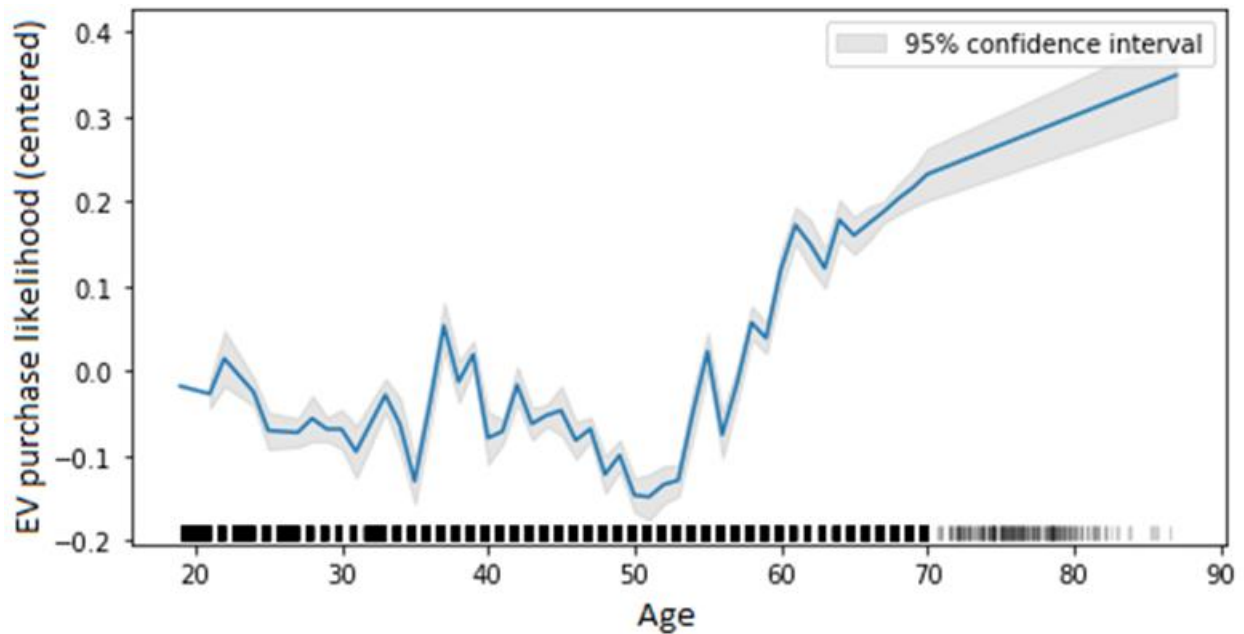


Figure 6.8 Effect of age on EV purchase likelihood.

The impact of GHG information framing on the likelihood of EV purchase is presented in Figure 6.9. Regarding GHG information framings, presenting the emissions with respect to a national goal along with the new emojis developed (frame 4 which integrated color, injunctive norms, and air quality) is the most effective, followed by highlighting the information with only color (frame 2; red for high emissions, blue for low emissions) and they can increase the probability of selecting EV by 12% and 9% compared to the current mock-up (i.e., frame1; NRCan label with CO₂ emissions g/km). Therefore, it can be deduced that GHG information framings (e.g., frame 4 and frame 2) on labels play a critical role in motivating individuals to buy EVs. Further, frame 2 (societal goal with color label) was the most effective framing in the previous study (Wang et al., 2021). Nonetheless, frame 4 (integrating color, injunctive norms, and air quality), which is developed in this study, can increase the intention to buy EVs by 4% compared to frame 2.

Frame 3 (pressure gauge label) and frame 5 (patriotic goal label) are the third and fifth effective treatments to enhance EV adoption, and they increase the EV selection likelihood by nearly 6% and 4% compared to the current label in Canada (frame 1). Frame 1 (current label in Canada) is the weakest frame for EV adoption. Similarly, frame 6 (dirty air label), frame 7 (tree scale label), and frame 8 (thumbs up/down label) are not effective treatments to increase the intention to buy EVs.

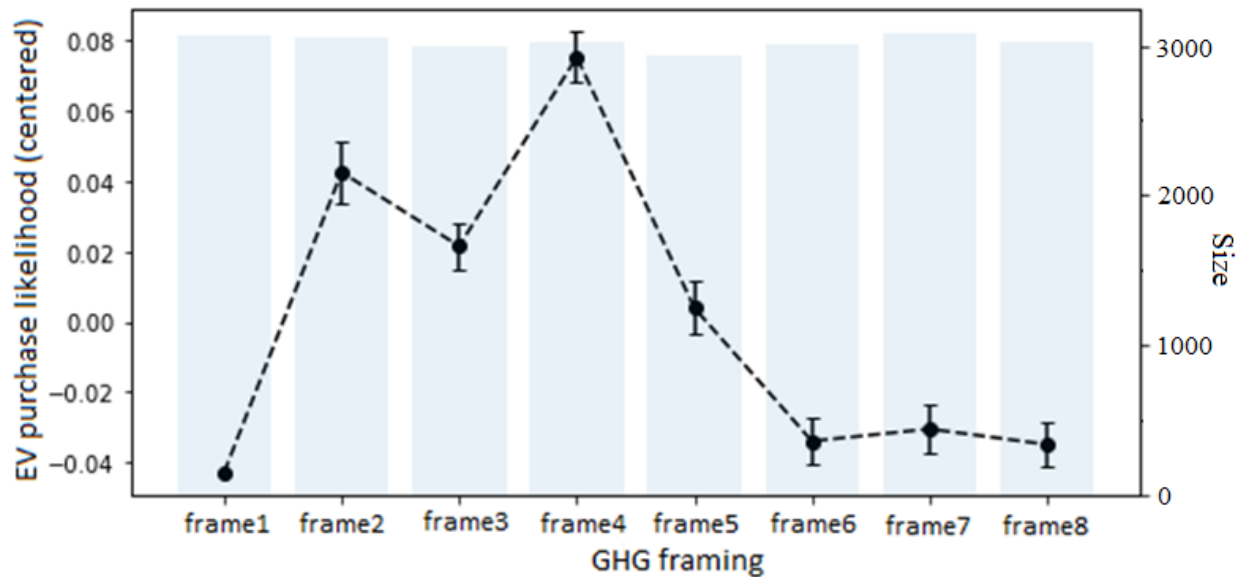


Figure 6.9 Effect of GHG framing on EV purchase likelihood. Frame 1 = current NRCan label; 2 = Societal goal with color; 3 = Societal goal as pressure gauge; 4 = Societal goal with new emoticons; 5 = Steps towards societal goal with patriotism; 6 = Societal goal as dirty air scale; 7 = Societal goal as Tree health; 8 = Societal goal with thumbs up/down

ALE results for variable interactions

To better understand the direction of variables on the EV adoption, the first level interaction of variables was extracted using XGBoost (shown in Figure 6.5). Then, ALE was used to better understand how variables' interactions influence the likelihood of buying EVs. In this section, ALE results for first-rank variables' interactions are presented.

The impact of the most influential interaction (CC-SoC/EV to ICEV purchase price ratio) on the probability of choosing EVs is indicated in Figure 6.10. As can be seen, the purchase price of EV to ICEV does not considerably affect the choices of those not concerned about climate change (CC-SoC1) since they are more likely to buy ICEVs. Likewise, decisions of the CC-SoC2 group (concerned but not planning to reduce their emissions) are not significantly influenced by the EV to ICEV purchase price ratio. The CC-SoC2 group is slightly more likely to buy an EV when the EV price is approximately the same as that of the ICEVs. The CC-SoC3 group prefer EVs when the EV to ICEV purchase price ratio has the lowest level (1.21). Therefore, the CC-SoC3 group prefers PHEVs to BEVs, and they are not inclined to pay a considerable amount of money for reducing their emissions. The CC-SoC4 group is more likely to buy EVs when the EV to ICEV purchase price ratio is between 1.78 and 1.9. The ratio of 1.78 implies the task in which the EV

option is BEV with a price of \$48000 and the ICEV price is \$27000. Moreover, the ratio of 1.9 denotes the task with a PHEV and an ICEV with purchase prices of \$38000 and \$20000, respectively. Therefore, this group prefers both PHEVs and BEVs to ICEVs, but when the price ratio is not higher than 1.9. The people who stated “have reduced their emissions (CC-SoC5)” are more likely to buy BEVs. In other words, the CC-SoC5 group prefers BEVs to both PHEVs and ICEVs. Interestingly, the CC-SoC5 group is willing to pay 1.82 times more than ICEVs for GHG emissions reduction.

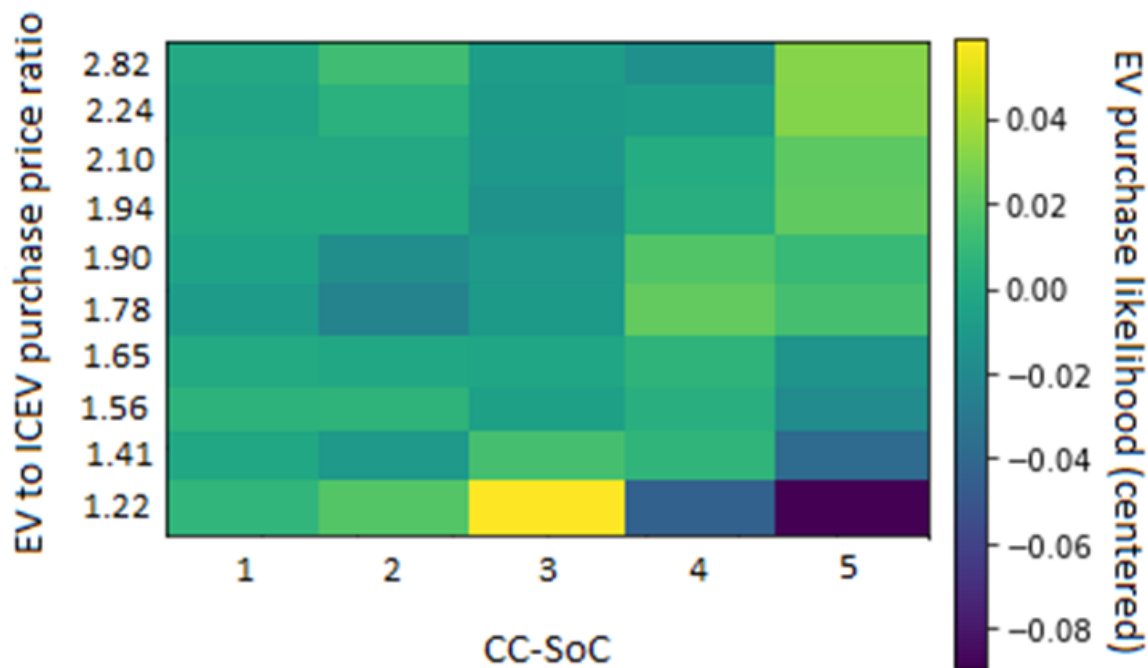


Figure 6.10 Effect of the interaction of CC-SoC and EV to ICEV purchase price ratio on EV purchase likelihood.

6.4 Conclusions

The study aimed to predict who will buy EVs and detect the determinants of EV adoption. To this end, a discrete choice experiment was designed, and a dataset of 24180 observations was collected from 2015 respondents. Applying a new hybrid method (COA-XGBoost), the results showed that the climate change stage of change (CC-SoC) was the most important parameter on individuals' decisions. Regarding variables' interactions, CC-SoC/EV to ICEV price ratio was the most important variables' interaction based on the relative influence on the likelihood of buying an EV.

To address the question of data splitting (optimizing the share of training and testing datasets), a new hybrid method (COA-XGBoost) was introduced to maximize the model's accuracy. COA-XGBoost found the optimal percentage of testing data to be 3.92%, while the optimal percentage of testing data was 7.5% in the conventional split ratio optimization model. Using COA-XGBoost, the mode reached the testing data accuracy of 87.4, which was 3.8% higher than the conventional method to optimize the split ratio. Therefore, the proposed technique could enhance the prediction power significantly by finding the optimal data split.

Since machine learning techniques are black-box tools, ALE was applied to interpret the results and determine in which direction variables and variables' interactions influence the likelihood of EV preference. ALE results showed that individuals who self-identified as being at the top of the climate change stage of change (CC-SoC4 & 5), French Quebecers, those older than 58 years, and those who completed a doctoral degree (e.g., Ph.D.), or a degree in medicine, dentistry, veterinary medicine, or optometry had the highest intention to buy EVs. As well, individuals who do not own a vehicle, part-time workers, females, people of Oceania origins, and individuals with household incomes of over \$200000 are more likely to prefer EVs to ICEVs. The results suggested that the purchase price of EVs should not be over 78% higher than ICEVs since EV to ICEV purchase price ratios of higher than 1.78 result in a reduction in intention to buy EVs. In terms of GHG framing, frame 4 (integrating color, injunctive norms, and air quality) is the most effective treatment to increase the likelihood of buying EVs.

In terms of policy, the government can more concentrate on those who are more likely to buy EVs (e.g., CC-SoC4 & 5, females, and French Quebecers) to increase the share of EVs in the Canada stock market. Moreover, the current mock-up (NRCan label) should be replaced with more effective labels to attract more car purchasers to buy EVs. The results of this study suggest that the integration of color, pressure gauge, emoticon, patriotic goal, dirty air figures, trees, and thumbs up/down can better present GHG information and attract customers to choose more sustainable options. Further, emoticon label maximizes the intention to buy EVs, followed by color labels, pressure gauge, and patriotic goal. Therefore, replacing these labels with the current labels (NRCan label) can considerably increase the EV purchase likelihood.

CHAPTER 7 General Discussion

This chapter discusses the contributions of this Ph.D. project concerning the research gaps in the literature. In this chapter, the research gaps mentioned in the previous chapters are first summarized. Then, the contributions of this study are briefly discussed. Then, the limitations of this dissertation are presented.

7.1 Research gaps

Statistical methods have been generally applied to analyze and predict travel behavior. Since travel behavior is complex to analyze, conventional statistical methods may not result in the highest possible accuracy of transportation phenomena. Further, traditional statistical methods may not easily converge when many explanatory variables are included in models so that often only relatively few variables are considered. At the same time, powerful machine learning techniques exist that can increase accuracy compared to conventional techniques as well as including larger numbers of variables.

Nonetheless, there are some challenges when using machine learning techniques. Although these techniques can include many features (i.e., variables) in the modeling, considering all the available features may not lead the highest prediction accuracy (Naseri, Shokoohi, Jahanbakhsh, Karimi, & Waygood, 2023). Travel behavior surveys generally include many variables, but an optimal feature set should be applied in the modeling to maximize the model's accuracy. However, detecting the optimal feature set may be a complicated task when many variables are available in the data set. In this circumstance, powerful feature selection techniques are required.

The second challenge is to tune the hyperparameters of machine learning techniques since these techniques are quite sensitive to them. If hyperparameters are not tuned well, models may be over- or underfitted and their results may not be accurate. To address this, some studies have applied conventional tuning processes to tune classifier hyperparameters in travel behavior studies. Grid search and random search are the conventional methods to tune algorithm hyperparameters. Nonetheless, grid search and random search are not efficient techniques (Liashchynskyi & Liashchynskyi, 2019), particularly for tuning continuous hyperparameters. Further, grid search and random search generally consider a single performance indicator (e.g., accuracy for classification or mean absolute error for regression) to determine the optimal value for hyperparameters.

However, in some cases, there might be a need to consider multiple performance indicators (e.g., considering accuracy and F1-score simultaneously) to tune the hyperparameters of machine learning techniques when using imbalanced data sets. As a result, developing a multi-objective regularization technique to tune hyperparameters is essential to increasing the accuracy of travel behavior prediction models.

Third, datasets need be divided into training and testing datasets. However, how to determine the optimal split for training and testing datasets has been a question of debate. One possible approach to finding the optimal split ratio is to check a limited number of randomly selected percentages for the split ratio (Nguyen et al., 2021). Finding the optimal value of a continuous-ranged parameter by checking a limited number of possible options is impossible though (Naseri, Jahanbakhsh, et al., 2020). Hence, developing an efficient method for optimizing the split ratio is essential to maximize the prediction accuracy of machine learning techniques when applying them to travel behavior datasets. Fourth, most of the powerful machine learning techniques are black-box tools, and hard to interpret. Therefore, they cannot easily be applied by policymakers and planners. In this regard, the application of novel approaches to make black-box prediction methods interpretable and apply them to set new transportation policies remains another immense concern.

On the other side, there are limitations associated with the current literature in terms of analyzing individual behavior and travel behavior. As such, although environmental behavior prediction has been investigated in some studies, the prediction of individual Climate Change Stage of Change (CC-SoC) has not received enough attention considering the crisis at hand. This variable (CC-SoC) can explain the environmental motivations in transportation choices, and it is more simple than conventional complex measures (Waygood et al., 2021). However, how a multitude of general environmental behaviors, attitudes, and sociodemographic characteristics related to the CC-SoC is not well known.

Additionally, although research has been conducted on adult Travel Mode Choice (TMC), child TMC has not received enough attention, and this group has been excluded from most transportation studies. That is, most transportation planning is based on adults' needs, and children are more and more reliant on adults for their transportation needs. Children are members of society, and travel mode choice has been found to relate to their overall wellbeing. Children's travel can influence parents' travel behavior. Furthermore, children become adults, and their childhood behaviors can

impact their adult behaviors. Therefore, it is important to examine children's TMC and how different parameters impact healthier and more sustainable mode choices.

Although previous studies investigated the intention to buy EVs or the willingness to pay for EVs, a few prediction models have been generated to determine who prefers EVs over ICEVs. Further, the influence of GHG information framing on individuals' willingness to buy EVs has not been tested in EV buyer prediction models.

7.2 Contributions

Machine learning is an increasingly powerful technique for prediction and analyzing data and it is increasingly applied in the field of transportation studies. However, as discussed in the previous chapters, various problems exist in how to use it and how to interpret outcomes. This Ph.D. project made a number of contributions to improve how machine learning is applied in sustainable travel behavior. The following will discuss the contributions of this study:

7.2.1 Finding optimal variables for CC-SoC prediction

Transportation-based surveys generally include many variables, and it is ambiguous which of the questions (variables) should be applied in the modeling to maximize the performance of machine learning techniques. To this end, a new approach was proposed to predict individuals' environmental attitudes and behaviors (i.e., CC-SoC). Due to the significant effects of transportation on generating harmful emissions, transport-related behavior questions are taken into account as independent variables as well as socio-demographic characteristics, environmental behaviors (GEB) and attitudes (NEP). This large number of variables increases the model's computational complexity and may reduce the prediction accuracy (Ping et al., 2019). Therefore, a feature selection should be applied to optimize the variables used in the modeling. Some of the conventional feature selection techniques can not present the optimal number of features, and they only present the importance weight of each feature. Thus, a new feature selection technique was introduced, capable of finding the optimal number of features and the optimal feature set to maximize the prediction accuracy.

In this regard, an optimization problem is modeled to maximize the prediction accuracy by finding the optimal features; that is, maximizing the prediction accuracy is an optimization problem that should be solved by an optimization algorithm. Since the type of the mentioned problem is Integer

Programming, and the number of decision variables is high, the problem is a non-deterministic polynomial-time (NP-hard). Exact optimization algorithms (e.g., branch and bound) cannot solve NP-hard problems. Moreover, exact optimization cannot be synced with machine learning techniques (Naseri, Fani, et al., 2020). Therefore, a metaheuristic optimization algorithm is employed to solve the mentioned problem and develop the mentioned feature selection technique.

The results suggest that the application of feature selection techniques to the transportation-based survey with many questions increases the prediction accuracy. Therefore, the variables should be analyzed using feature selection techniques, and the optimal feature set should be applied in the machine learning modeling so as to maximize the prediction accuracy. Further, using all the variables does not necessarily increase the prediction accuracy. Ma et al. (2022) stated that the application of optimal variables in machine learning modeling increases prediction accuracy, which is in line with the results of this study.

The developed feature selection technique is applied to determine optimal variables for CC-SoC prediction. The share of GEB, transportation-based, and sociodemographic variables in the optimal variable set is 46%, 20%, and 15%, respectively. Waygood et al. (2021) showed that sociodemographic characteristics and responses to GEB questions significantly influence CC-SoC, which is consistent with the results of the current dissertation. However, this analysis was able to show to what extent each of those different measures relate to the CC-SoC.

7.2.2 Using a multi-objective approach to tune hyperparameters applied to child TMC

Machine learning techniques typically include some parameters, called hyperparameters, that should be tuned by users. Since conventional techniques to tune hyperparameters may not be highly efficient, and they can only optimize a single indicator during the tuning process, a new technique is proposed in this study that can optimize multiple indicators. The proposed technique can be highly effective for imbalanced datasets (e.g., Montreal TMC data) as the F1-score and accuracy can be maximized simultaneously. To develop this model, a powerful machine learning technique is synchronized with a multi-objective metaheuristic optimization algorithm. The developed model is capable of considering more performance indicators. The results show that the proposed approach can increase prediction accuracy by optimizing the hyperparameters' values. Moreover, this technique is faster than comprehensive search methods to tune hyperparameters (i.e., grid

search). Grid search checks all the possible combinations of hyperparameters in the defined set. Hence, MOHT is more efficient than grid search since it applies a heuristic process to find the optimal values of hyperparameters. Liashchynskiy & Liashchynskiy (2019) showed that grid search and random search are not efficient techniques to tune hyperparameters, and the heuristic process can outperform these conventional techniques in terms of maximizing the prediction accuracy. Hence, the outcomes of the current study are in agreement with that finding.

The proposed technique was applied to predict child TMC since child TMC has not received enough attention in the literature, and this group has been excluded from most of the studies. To address this issue, this study developed a model to predict the TMC of children and determine which variables significantly influence child mode choice for all trips (i.e., not only school-related trips) using an ensemble learning approach. The results suggest that trip distance, Walk Score, age, Bike Score, household income, and accessibility to secondary schools are the top-ranked variables. The result that distance is the primary explanatory factor and that age is important is in agreement with previous findings (e.g., Scheiner et al. (2019)). As mentioned, much of the research has focused on trips to school, and this paper also found that Walk Score was an important explanatory factor for children's active travel to school (Kunaratnam et al., 2022). Such results would suggest that more compact, mixed environments are important explanatory factors for how children get around.

7.2.3 Optimizing the split ratio applied to vehicle engine preferences

Although optimizing the split ratio (i.e., the share of testing data) influences the prediction accuracy, its application to transportation studies has been overlooked. Hence, a new method is introduced to determine the optimal share of training and testing datasets. The mentioned method uses a metaheuristic process to maximize prediction accuracy by optimizing the split ratio. The results suggest that optimizing the split ratio can increase the prediction accuracy, and the application of the metaheuristic algorithm to the mentioned problem is efficient. Nguyen et al. (2021) showed that split ratio optimization increases prediction accuracy, which validates the results of this study.

The proposed approach was applied to a vehicle engine preference prediction problem. A robust tree-based ensemble machine learning technique is applied to predict who will buy EVs (with respect to both PHEV and BEV). Using an ensemble machine learning technique, variables are

ranked based on their influence on individuals' intention to buy EVs. Moreover, a new version of XGBoost (i.e., xgbfir) was used to understand the interaction of parameters and the interactions with the highest impact on the likelihood of buying EVs were scrutinized. The results suggest that CC-SoC has the highest relative influence on the intention to select EVs over ICEVs, followed by EV to ICEV purchase price ratio, region (province), age, education, and GHG information framing. Previous studies demonstrated that people with different levels of CC-SoC are statistically different in willingness to buy EVs (Wang et al., 2021; Waygood et al., 2021), though such research did not demonstrate that it was even more consequential than the price ratio for determining who would choose an EV. Monetary incentives have often been shown to significantly influence individuals' choice when selecting between an ICEV and an EV (Bas, Zou, et al., 2021; Wang et al., 2021), but this research would suggest that climate change attitudes is a critical factor. Finally, as an influence that can be better controlled by the government, GHG information framing significantly affected the willingness to pay for EVs as has been previously demonstrated as well (Wang et al., 2021). In contrast to the previous work, this research demonstrated that it was one of the top considerations for choosing an EV.

7.2.4 Interpreting the results of machine learning techniques

Machine learning techniques are black-box, and their results are hard to apply for policy-relevant purposes. To address this, different methods are applied to interpret the results of machine learning techniques. As such, ensemble learning techniques and their important weights are applied in the modeling to make the results of machine learning techniques interpretable. These (i.e., ensemble) techniques can present the relative influence of variables on the response variables. However, these techniques cannot represent in which direction independent variables affect dependent variables. To address this, Accumulated Local Effects and Multiple Logistic Regression are employed to better understand the influence direction of top variables on travel behavior. Application of these interpretation techniques makes the results of machine learning techniques interpretable, and their results can be easily used by policy- and decision-makers.

7.3 Limitations

The limitations of this study are presented in this section. In the first analysis of this dissertation presented in Chapter 4, a model was developed to predict the Climate Change Stage of Change (CC-SoC). The measure, Climate Change Stage of Change, captures individuals' self-assessment

of their climate concerns and behavioral intentions. It does not measure what their actual climate impacts are. It is possible for a person not to be concerned about climate change and lead a low-carbon lifestyle. It should only be considered with respect to how strongly they would likely support or react to climate-related choices. Therefore, this study could not be able to predict individuals' actual climate impacts due to the unavailability of datasets, and it might be one of the limitations of this study. It should be noted that the original survey, applied here to predict CC-SoC, focused on vehicle choice and included only car owners in the US. As such, future research should examine a larger array of transportation behaviors with a general population sample.

To predict CC-SoC, a new feature selection technique was developed (COA-QDA) to maximize the prediction accuracy. Although COA-QDA performed well in CC-SoC prediction, the performance of this technique on different prediction problems should be tested. As another limitation, Coyote Optimization Algorithm was used to propose a COA-QDA. However, the application of other metaheuristic algorithms to develop feature selection algorithms was overlooked. One of the other limitations of this study was to consider testing data accuracy as the performance indicator for optimizing the feature set using COA-QDA. However, the application of other performance indicators (such as F1-score) in the feature selection process was not examined. Moreover, the developed feature selection technique only applied a single performance indicator (accuracy) to optimize the feature set. Although many machine learning techniques were used to predict CC-SoC, the application of powerful ensemble learning algorithms, such as XGBoost, LGBM, and Categorical Boosting, was overlooked.

Although MOHT can consider multiple performance indicators, only two performance indicators (i.e., accuracy and F1-score) were considered as the problem's objective function. Further, MOHT was only applied to predict child TMC and its application on other imbalanced data sets was not tested. Further, MOHT was only applied to tune the hyperparameters of an ensemble learning technique, and its performance on other prediction techniques was not evaluated.

Previous research has demonstrated that BEV-oriented and PHEV-oriented are different groups in Canada in terms of attitudes, socio-demographic variables, and level of concern about the environment (Ferguson et al., 2018). However, to predict individual EV preferences, both battery electric vehicles and plug-in hybrid electric vehicles were considered in the same category (electric vehicles) in this study. For GHG information presentation, tailpipe emissions are taken into account

to design vehicle labels. Due to the significant emissions associated with the life-cycle of EVs, the impact of using life-cycle GHG emissions on the probability of choosing EVs over ICEVs needs to be tested in future research. Although Accumulated Local Effect and Multiple Logistic Regression are applied to interpret machine learning techniques, the application of other interpretation techniques, such as Partial Dependence Plot and SHapley Additive exPlanations, are overlooked in this study.

CHAPTER 8 Conclusions and Recommendations for Future Studies

This chapter presents the conclusions of this Ph.D. project and recommendations for future studies. First, the conclusions are presented. Then, a few suggestions for future studies are presented. Finally, a summary of this Ph.D. project is presented.

8.1 Conclusions

This Ph.D. project aimed to better predict individual travel behavior through improved machine learning approaches. It succeeded by: 1) developing a feature selection technique to identify the optimal feature set when dealing with large numbers of potential explanatory variables; 2) developing a new multi-objective approach to tune machine learning hyperparameters; 3) developing a new hybrid method (COA-XGBoost) to maximize accuracy. These developments will help improve future analyses of sustainable transport.

The three methodological contributions helped to improve understanding on three sustainable transportation questions in this PhD: what explains an individual's climate-change stage-of-change; children's travel mode choice; and individual preferences between internal combustion engine vehicles and electric vehicles. In each case, greater understanding of the importance of different variables in predicting outcomes was achieved and the improved methodologies further increased accuracy compared to previously existing machine learning approaches.

The proposed feature selection technique (COA-QDA) can appropriately detect the optimal feature set to maximize the prediction accuracy of CC-SoC prediction. To predict CC-SoC, 32.6% of optimal features are related to transportation behavior. This suggests that the application of transportation behavior to predict CC-SoC is vital, and transportation is an important distinguishing behavior when it comes to how one views themselves with respect to climate change concern and actions. The proposed feature selection technique (COA-QDA) can increase the average testing data accuracy more than conventional feature selection techniques. Although COA-QDA was the best feature selection technique in terms of average and maximum prediction accuracy, its running time was higher than Lasso and Elastic Net. Furthermore, using all features in the prediction models can result in an accuracy reduction. To predict CC-SoC, the highest testing data accuracy was obtained by combining COA-QDA and Logistic Regression. Logistic Regression and Random Forest outperform other classifiers based on the average prediction accuracy to predict CC-SoC.

A new multi-objective approach (MOHT) was proposed to tune machine learning hyperparameters. MOHT was demonstrated to be an appropriate technique for tuning hyperparameters of imbalanced datasets (such as child travel mode choice) since it can consider multiple machine learning performance indicators in the tuning process. MOHT outperformed other hyperparameter tuning techniques in terms of machine learning performance indicators (e.g., prediction accuracy, F1-score, and AUC). Moreover, this technique significantly reduced the computational cost compared to grid search and Hyperopt. However, the running time of MOHT was considerably higher than the random search, though it can present more accurate solutions.

MOHT was used to tune an ensemble learning technique to better understand the influence of different variables on children's TMC. The outcomes showed that trip distance, Walk Score, age, Bike Score, household income, and accessibility to secondary schools were the top-ranked variables. Multiple Logistic Regression was applied to better understand the influence of these variables on children's TMC. In reference to trips by car, the results suggest that as trip distance decreases, active modes are more likely. The built environment, as measured by Walk Score, was positively associated with all sustainable and independent modes as was Bike Score to a lesser degree. As age increases, children used more sustainable and independent modes. Finally, the highest household income was associated with more passenger trips in cars, but the relationship with active modes was less strong.

Finally, a model was developed to predict who prefers EVs and detect the likely determinants of EV adoption. To address the question of data splitting (optimizing the share of training and testing datasets), a new hybrid method (COA-XGBoost) was introduced to maximize the model's accuracy. COA-XGBoost enhanced the prediction power significantly by finding the optimal data split. The outcomes of this technique indicated that the climate change stage of change (CC-SoC) was the most important parameter on individuals' decisions. The interaction of CC-SoC and EV to ICEV price ratio was the most important interaction based on the relative influence on the likelihood of buying an EV. ALE was applied to interpret the results and determine in which direction variables and their interactions influence EV preference. ALE results showed that individuals who self-identified as being at the top of the climate change stage of change (CC-SoC4 & 5), French Quebecers, those older than 58 years, and those who completed a doctoral degree (e.g., Ph.D.), or a degree in medicine, dentistry, veterinary medicine, or optometry had the highest intention to buy EVs. As well, individuals who do not own a vehicle, part-time workers, females,

people of Oceania origins, and individuals with household incomes of over \$200,000 are more likely to prefer EVs to ICEVs. In terms of GHG framing, frame 4 (integrating color, injunctive norms, and air quality) is the most effective treatment to increase the likelihood of buying EVs.

8.2 Recommendations for future studies

Since COA-QDA was only applied to one prediction problem, it is recommended that assessing the performance of COA-QDA on different prediction problems with different complexities should be considered in future studies. As COA-QDA optimized the features based on the prediction accuracy, it is recommended that F1-score be used to optimize the feature set in future studies. Moreover, it is recommended to consider multiple performance indicators simultaneously to optimize the feature set. In this case, the application of multi-objective optimization algorithms is recommended. As Coyote Optimization Algorithm was used to develop COA-QDA, it is suggested to employ various robust metaheuristic algorithms to generate new feature selection methods using the proposed approach.

Although many machine learning techniques were used to predict CC-SoC, the application of powerful ensemble learning algorithms, such as XGBoost, LGBM, and Categorical Boosting, was overlooked. Hence, it is suggested that powerful ensemble learning algorithms, such as XGBoost, LGBM, and Categorical Boosting, are considered when predicting CC-SoC in future studies.

MOHT was used to tune hyperparameters of LGBM in order to accurately predict child TMC considering accuracy and F1-score. It is suggested that more performance indicators are added to MOHT and to compare the model's accuracy with the one presented in this dissertation. Moreover, it is recommended to apply MOHT to other imbalanced data sets.

It is suggested to consider battery electric vehicles and plug-in hybrid electric vehicles separately when predicting who is more likely to prefer EVs to ICEVs. That is, two different prediction models should be generated: one to predict individual decisions when options are plug-in hybrid electric vehicles and internal combustion engine vehicles, and the other to predict individual decisions when alternatives are battery electric vehicles and internal combustion engine vehicles. Further, it is suggested that the influence of life cycle GHG emissions on the probability of choosing EVs over ICEVs is tested in future research.

8.3 Final summary

This study applied machine learning techniques to predict sustainable travel behavior. A few problems are associated with machine learning techniques, such as being sensitive to the variables applied in the modeling, being affected by the hyperparameter values, being influenced by the split ratio, and being black-box tools. Hence, this study developed new techniques to consider these issues. The research gaps and the contributions of this dissertation are summarized in Table 8.1.

Table 8.1 A summary of this dissertation

Research gaps	Contributions	Chapter
Detecting the optimal feature set may be a complicated task	A new feature selection technique (COA-QDA) is developed to address this issue because it can detect the optimal number of features	4
Many of the conventional feature selection techniques can not detect the optimal number of features	The developed feature selection technique (COA-QDA) can directly detect the optimal number of features	4
Many of the conventional feature selection techniques are linear while travel behavior does not generally follow a linear trend	COA-QDA is a non-linear model allowing for improved accuracy for such relationships	4
Conventional methods to tune hyperparameters are not efficient techniques	A new multi-objective approach (MOHT) is proposed to tune machine learning hyperparameters	5
Conventional methods to tune hyperparameters use a single performance indicator	MOHT can apply multiple performance indicators simultaneously and improve prediction (testing data) accuracy	5

Table 8.1 A summary of this dissertation (continued)

Research gaps	Contributions	Chapter
Determining the optimal split for training and testing datasets has been of immense concern	A new hybrid method (COA-XGBoost) is developed to optimize the split ratio with respect to accuracy	6
Finding the optimal value of a continuous-ranged parameter by checking a limited number of possible options is impossible	COA-XGBoost considers a wide range for splits within a reasonable time giving researchers confidence that a near-best split was identified	6
Most of the powerful machine learning techniques are black-box tools	Accumulated Local Effects, importance weight of ensemble learning techniques, and Logistic Regression were used to interpret the results of machine learning techniques	4,5,6
The prediction of individual Climate Change Stage of Change (CC-SoC) has not received enough attention	A prediction model is designed to predict CC-SoC, and the predictors of CC-SoC are determined demonstrating the importance of travel behavior	4
Child travel mode choice has not received enough attention, and this group has been excluded from most of the studies	A prediction model is developed to predict child travel mode choice with distance, Walk Score, and age being the key considerations	5
Few prediction models have been generated to determine who prefers electric vehicles over internal combustion engine vehicles	A prediction model is developed to predict individual's engine choices, and the strongest variables on EV purchase likelihood are determined with CC-SoC and price ratio being the top two	6
The influence of GHG information framing on individuals' willingness to buy electric vehicles has not been tested in electric vehicle buyer prediction models	Framing was one of the top variables that governments can control and framing the emissions with respect to the societal reduction target along with an emoticon that presents approval/disapproval was the most effective.	6

REFERENCES

- Aboutaleb, Y. M., Danaf, M., Xie, Y., & Ben-Akiva, M. (2021). Discrete Choice Analysis with Machine Learning Capabilities. Retrieved 5 April 2022 from <https://doi.org/10.48550/arxiv.2101.10261>
- Ahmad, J., Fiaz, M., Kwon, S., Sodanil, M., Vo, B., & Baik, S. W. (2016). Gender Identification using MFCC for Telephone Applications - A Comparative Study. Retrieved 9 December 2021 from <https://arxiv.org/abs/1601.01577v1>
- Amasyali, K., & El-Gohary, N. (2021). Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. *Renewable and Sustainable Energy Reviews*. Retrieved from <https://doi.org/10.1016/j.rser.2021.110714>
- Anable, J. (2005). ‘Complacent Car Addicts’; or ‘Aspiring Environmentalists’? Identifying travel behaviour segments using attitude theory. *Transport Policy*. Retrieved from <https://doi.org/10.1016/j.tranpol.2004.11.004>
- Andani, I. G. A., La Paix Puello, L., & Geurs, K. (2021). Modelling effects of changes in travel time and costs of toll road usage on choices for residential location, route and travel mode across population segments in the Jakarta-Bandung region, Indonesia. *Transportation Research Part A: Policy and Practice*, 145, 81–102. Retrieved from <https://doi.org/10.1016/j.tra.2020.12.012>
- Andrade, K., Uchida, K., & Kagaya, S. (2006). Development of Transport Mode Choice Model by Using Adaptive Neuro-Fuzzy Inference System. *Transportation Research Record: Journal of the Transportation Research Board*, 1977(1), 8–16. Retrieved 29 October 2021 from <https://doi.org/10.1177/0361198106197700102>
- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(4), 1059–1086. Retrieved 28 March 2022 from <https://doi.org/10.1111/rssb.12377>
- Aschwanden, G. D. P. A., Wijnands, J. S., Thompson, J., Nice, K. A., Zhao, H., & Stevenson, M. (2021). Learning to walk: Modeling transportation mode choice distribution through neural networks. *Environment and Planning B: Urban Analytics and City Science*, 48(1), 186–199.

Retrieved 5 May 2022 from <https://doi.org/10.1177/2399808319862571>

- Assi, K. J., Shafiullah, M., Nahiduzzaman, K. M., & Mansoor, U. (2019). Travel-to-school mode choice modelling employing artificial intelligence techniques: A comparative study. *Sustainability (Switzerland)*, 11(16), 4484. Retrieved 4 May 2022 from <https://doi.org/10.3390/su11164484>
- Augeri, M. G., Greco, S., & Nicolosi, V. (2019). Planning urban pavement maintenance by a new interactive multiobjective optimization approach. *European Transport Research Review*, 11(1). Retrieved from <https://doi.org/10.1186/s12544-019-0353-9>
- Austmann, L. M. (2021). Drivers of the electric vehicle market: A systematic literature review of empirical studies. *Finance Research Letters*, 41, 101846. Retrieved from <https://doi.org/10.1016/j.frl.2020.101846>
- Barros, V., Cruz, C. O., Júdice, T., & Sarmento, J. M. (2021). Is taxation being effectively used to promote public transport in Europe? *Transport Policy*, 114, 215–224. Retrieved from <https://doi.org/10.1016/j.tranpol.2021.10.003>
- Bas, J., Cirillo, C., & Cherchi, E. (2021). Classification of potential electric vehicle purchasers: A machine learning approach. *Technological Forecasting and Social Change*, 168, 120759. Retrieved from <https://doi.org/10.1016/j.techfore.2021.120759>
- Bas, J., Zou, Z., & Cirillo, C. (2021). An interpretable machine learning approach to understanding the impacts of attitudinal and ridesourcing factors on electric vehicle adoption. *Transportation Letters*. Retrieved 14 March 2022 from <https://doi.org/10.1080/19427867.2021.2009098>
- Bedard, S. A. N., & Tolmie, C. R. (2018). Millennials' green consumption behaviour: Exploring the role of social media. *Corporate Social Responsibility and Environmental Management*, 25(6), 1388–1396. Retrieved 7 December 2021 from <https://doi.org/10.1002/csr.1654>
- Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *30th International Conference on Machine Learning, ICML 2013* (pp. 115–123). Retrieved 29 June 2022 from <https://proceedings.mlr.press/v28/bergstra13.html>
- Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation*

- Research Part B: Methodological*, 45(7), 923–939. Retrieved from <https://doi.org/10.1016/j.trb.2011.04.005>
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, 738. Retrieved 17 June 2022 from <https://link.springer.com/book/9780387310732>
- Brazeau, A., Denoncourt, J.-C., Gruber, V., & Lachapelle, E. (2021). *LIMITLESS: CAR ADVERTISING IN CANADA PRACTICES, REGULATORY FRAMEWORK AND RECOMMENDATIONS FINANCIAL SUPPORT THE RISE OF LIGHT-DUTY TRUCK IN CANADA: REVERSING THE TREND METHODOLOGICAL SUPPORT RESEARCH AND WRITING TEAM*.
- Brenner, A., Wu, M., & Amin, S. (2022). Interpretable Machine Learning Models for Modal Split Prediction in Transportation Systems (pp. 901–908). Institute of Electrical and Electronics Engineers (IEEE). Retrieved 4 December 2022 from <https://doi.org/10.1109/itsc55140.2022.9921938>
- Buehler, R. (2011). Determinants of transport mode choice: a comparison of Germany and the USA. *Journal of Transport Geography*. Retrieved from <https://doi.org/10.1016/j.jtrangeo.2010.07.005>
- Buijs, R., Koch, T., & Dugundji, E. (2021). Using neural nets to predict transportation mode choice: Amsterdam network change analysis. *Journal of Ambient Intelligence and Humanized Computing*, 12(1), 121–135. Retrieved 5 May 2022 from <https://doi.org/10.1007/s12652-020-02855-6>
- Canada Statistics. (2021). Population and demography statistics. Retrieved 22 September 2021, from https://www.statcan.gc.ca/eng/subjects-start/population_and_demography
- Chang, X., Wu, J., Liu, H., Yan, X., Sun, H., & Qu, Y. (2019). Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data. *Transportmetrica A: Transport Science*. Retrieved from <https://doi.org/10.1080/23249935.2019.1620380>
- Chao, Y. (2021). Predicting Mode Choice on Urban Work Trips by Non-Private Vehicles. In *CICTP 2021: Advanced Transportation, Enhanced Connection - Proceedings of the 21st COTA International Conference of Transportation Professionals* (pp. 2160–2169). American

- Society of Civil Engineers. Retrieved 5 May 2022 from <https://doi.org/10.1061/9780784483565.208>
- Chapleau, R., Gaudette, P., & Spurr, T. (2019). Application of Machine Learning to Two Large-Sample Household Travel Surveys: A Characterization of Travel Modes. *Transportation Research Record*. Retrieved from <https://doi.org/10.1177/0361198119839339>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 13-17-Aug, pp. 785–794). New York, NY, USA: ACM. Retrieved 1 January 2022 from <https://doi.org/10.1145/2939672.2939785>
- Chen, X., Liu, X., & Li, F. (2013). Comparative study on mode split discrete choice models. *Journal of Modern Transportation*, 21(4), 266–272. Retrieved 8 November 2021 from <https://doi.org/10.1007/s40534-013-0028-5>
- Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*. Retrieved from <https://doi.org/10.1016/j.tbs.2018.09.002>
- Cheung, M. L., Pires, G. D., Rosenberger, P. J., Leung, W. K. S., & Salehuddin Sharipudin, M. N. (2021). The role of consumer-consumer interaction and consumer-brand interaction in driving consumer-brand engagement and behavioral intentions. *Journal of Retailing and Consumer Services*, 61, 102574. Retrieved from <https://doi.org/10.1016/j.jretconser.2021.102574>
- Climate Watch. (2018). Global GHG emissions. [Dataset]-Available at: <https://www.climatewatchdata.org>.
- Cui, L., Bai, L., Wang, Y., Jin, X., & Hancock, E. R. (2021). Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection. *Pattern Recognition*. Retrieved from <https://doi.org/10.1016/j.patcog.2021.107835>
- Daziano, R. A., Waygood, E. O. D., Patterson, Z., & Braun Kohlová, M. (2017). Increasing the influence of CO2 emissions information on car purchase. *Journal of Cleaner Production*. Retrieved from <https://doi.org/10.1016/j.jclepro.2017.07.001>
- Daziano, R., Waygood, E. O. D., Patterson, Z., Feinberg, M., & Wang, B. (2021). Reframing

- greenhouse gas emissions information presentation on the Environmental Protection Agency's new-vehicle labels to increase willingness to pay. *Journal of Cleaner Production*. Retrieved from <https://doi.org/10.1016/j.jclepro.2020.123669>
- Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. In *2018 15th Learning and Technology Conference, L and T 2018*. Retrieved from <https://doi.org/10.1109/LT.2018.8368509>
- Dong, L., Wesseloo, J., Potvin, Y., & Li, X. (2016). Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression. *Rock Mechanics and Rock Engineering*. Retrieved from <https://doi.org/10.1007/s00603-015-0733-y>
- Dong, Y. (2022). Analysis of Consumers' Willingness to Accept of Government Subsidies for Electric Vehicles. *Transportation Research Procedia*, 61, 90–97. Retrieved from <https://doi.org/10.1016/j.trpro.2022.01.016>
- Dong, Y., Sun, Y., Waygood, O., Wang, B., Huang, P., & Naseri, H. (2022). Insight into the Nonlinear Effect of COVID-19 on Well-Being in China: Commuting, a Vital Ingredient. *Journal of Transport & Health*, 25, 101424. Retrieved 5 November 2022 from <https://doi.org/10.1016/j.jth.2022.101424>
- Duca, A. Lo, Bacciu, C., & Marchetti, A. (2017). A K-nearest neighbor classifier for ship route prediction. In *OCEANS 2017 - Aberdeen*. Retrieved from <https://doi.org/10.1109/OCEANSE.2017.8084635>
- Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). Measuring endorsement of the new ecological paradigm: A revised NEP scale. *Journal of Social Issues*, 56(3), 425–442. Retrieved 9 December 2021 from <https://doi.org/10.1111/0022-4537.00176>
- Eluru, N., Chakour, V., & El-Geneidy, A. M. (2012). Travel mode choice and transit route choice behavior in Montreal: Insights from McGill University members commute patterns. *Public Transport*. Retrieved from <https://doi.org/10.1007/s12469-012-0056-2>
- Environment and Climate Change Canada. (2020). *National Inventory Report 1990–2018: Greenhouse Gas Sources and Sinks in Canada: Executive Summary. Canada's Greenhouse Gas Inventory*.

- Fan, J., Wang, X., Zhang, F., Ma, X., & Wu, L. (2020). Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data. *Journal of Cleaner Production*. Retrieved from <https://doi.org/10.1016/j.jclepro.2019.119264>
- Fonti, V. (2017). Feature Selection using LASSO. *VU Amsterdam*.
- Freeman, C., & Tranter, P. (2012). *Children and their urban environment: Changing worlds. Children and their Urban Environment: Changing Worlds*. Taylor and Francis. Retrieved 21 July 2022 from <https://doi.org/10.4324/9781849775359>
- Funke, S. Á., Sprei, F., Gnann, T., & Plötz, P. (2019). How much charging infrastructure do electric vehicles need? A review of the evidence and international comparison. *Transportation Research Part D: Transport and Environment*, 77, 224–242. Retrieved from <https://doi.org/10.1016/j.trd.2019.10.024>
- Gaker, D., & Walker, J. L. (2013). Revealing the value of ‘Green’ and the small group with a big heart in transportation mode choice. *Sustainability (Switzerland)*. Retrieved from <https://doi.org/10.3390/su5072913>
- Gao, K., Yang, Y., Zhang, T., Li, A., & Qu, X. (2021). Extrapolation-enhanced model for travel decision making: An ensemble machine learning approach considering behavioral theory. *Knowledge-Based Systems*, 218, 106882. Retrieved from <https://doi.org/10.1016/j.knosys.2021.106882>
- Golafshani, E. M., & Behnood, A. (2019). Estimating the optimal mix design of silica fume concrete using biogeography-based programming. *Cement and Concrete Composites*, 95–105.
- Golob, T. F. (2003, January 1). Structural equation modeling for travel behavior research. *Transportation Research Part B: Methodological*. Pergamon. Retrieved from [https://doi.org/10.1016/S0191-2615\(01\)00046-7](https://doi.org/10.1016/S0191-2615(01)00046-7)
- Guo, Q., Wu, W., Massart, D. L., Boucon, C., & De Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*. Retrieved from [https://doi.org/10.1016/S0169-7439\(01\)00203-9](https://doi.org/10.1016/S0169-7439(01)00203-9)
- Hajmeer, M., & Basheer, I. (2003). Comparison of logistic regression and neural network-based

- classifiers for bacterial growth. *Food Microbiology*. Retrieved from [https://doi.org/10.1016/S0740-0020\(02\)00104-1](https://doi.org/10.1016/S0740-0020(02)00104-1)
- Handy, S. (1996). Methodologies for exploring the link between urban form and travel behavior. *Transportation Research Part D: Transport and Environment*, 1(2), 151–165. Retrieved from [https://doi.org/10.1016/S1361-9209\(96\)00010-7](https://doi.org/10.1016/S1361-9209(96)00010-7)
- Hardman, S. (2019). Understanding the impact of reoccurring and non-financial incentives on plug-in electric vehicle adoption – A review. *Transportation Research Part A: Policy and Practice*, 119, 1–14. Retrieved from <https://doi.org/10.1016/j.tra.2018.11.002>
- Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*. Retrieved from <https://doi.org/10.1109/ACCESS.2020.2989857>
- Hensher, D. A., & Rose, J. M. (2007). Development of commuter and non-commuter mode choice models for the assessment of new public transport infrastructure projects: A case study. *Transportation Research Part A: Policy and Practice*, 41(5), 428–443. Retrieved from <https://doi.org/10.1016/j.tra.2006.09.006>
- Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, 12, 100591. Retrieved from <https://doi.org/10.1016/j.softx.2020.100591>
- Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). Weak teachers: Assisted specification of discrete choice models using ensemble learning. *HEART 2019, 8th Symposium of the European Association for Research in Transportation*, (Ml), 1–12.
- Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., & Jin, Y. (2021, March 1). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*. Elsevier. Retrieved from <https://doi.org/10.1016/j.jocm.2020.100221>
- Hu, G., Yin, C., Wan, M., Zhang, Y., & Fang, Y. (2020). Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. *Biosystems Engineering*. Retrieved from <https://doi.org/10.1016/j.biosystemseng.2020.03.021>
- Jahanbakhsh, H., Karimi, M. M., Naseri, H., & Nejad, F. M. (2020). Sustainable asphalt concrete containing high reclaimed asphalt pavements and recycling agents: Performance assessment,

- cost analysis, and environmental impact. *Journal of Cleaner Production*, 244, 118837. Retrieved 1 April 2020 from <https://doi.org/10.1016/j.jclepro.2019.118837>
- Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*. Retrieved from <https://doi.org/10.1109/WCCCT.2016.25>
- Jalilian, N., Najafpour, G. D., & Khajouei, M. (2020, February 1). Macro and Micro Algae in Pollution Control and Biofuel Production – A Review. *ChemBioEng Reviews*. John Wiley & Sons, Ltd. Retrieved 16 March 2023 from <https://doi.org/10.1002/cben.201900014>
- Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study. *International Journal of Injury Control and Safety Promotion*, 28(4), 408–427. Retrieved 12 November 2022 from <https://doi.org/10.1080/17457300.2021.1928233>
- Jeon, H., Seo, W., Park, E., & Choi, S. (2020). Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services. *Technological Forecasting and Social Change*, 161, 120303. Retrieved from <https://doi.org/10.1016/j.techfore.2020.120303>
- Kaiser, F. G., & Wilson, M. (2004). Goal-directed conservation behavior: The specific composition of a general performance. *Personality and Individual Differences*, 36(7), 1531–1544. Retrieved from <https://doi.org/10.1016/j.paid.2003.06.003>
- Kim, E. J. (2021a). Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach. *Journal of Advanced Transportation*. Retrieved from <https://doi.org/10.1155/2021/6685004>
- Kim, E. J. (2021b). Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach. *Journal of Advanced Transportation*, 2021. Retrieved from <https://doi.org/10.1155/2021/6685004>
- Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram

- signals based on the wrist-motion directions. *Current Applied Physics*. Retrieved from <https://doi.org/10.1016/j.cap.2010.11.051>
- Kim, K., Yang, H., Yi, J., Son, H. E., Ryu, J. Y., Kim, Y. C., ... Kim, S. (2021). Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: External validation and model interpretation. *Journal of Medical Internet Research*, 23(4), e24120. Retrieved 28 March 2022 from <https://doi.org/10.2196/24120>
- Klinger, T., & Lanzendorf, M. (2016). Moving between mobility cultures: what affects the travel behavior of new residents? *Transportation*, 43(2), 243–271. Retrieved 10 June 2022 from <https://doi.org/10.1007/s11116-014-9574-x>
- Koushik, A. N. P., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behaviour research: a review. *Transport Reviews*, 40(3), 288–311. Retrieved 28 February 2022 from <https://doi.org/10.1080/01441647.2019.1704307>
- Kucukoglu, I., Dewil, R., & Cattrysse, D. (2021). The electric vehicle routing problem and its variations: A literature review. *Computers and Industrial Engineering*, 161, 107650. Retrieved from <https://doi.org/10.1016/j.cie.2021.107650>
- Kunaratnam, V., Schwartz, N., Howard, A., Mitra, R., Saunders, N., Cloutier, M. S., ... Rothman, L. (2022). Equity, walkability, and active school transportation in Toronto, Canada: A cross-sectional study. *Transportation Research Part D: Transport and Environment*, 108, 103336. Retrieved from <https://doi.org/10.1016/j.trd.2022.103336>
- Lee, D., Kang, S., & Shin, J. (2017). Using deep learning techniques to forecast environmental consumption level. *Sustainability (Switzerland)*. Retrieved from <https://doi.org/10.3390/su9101894>
- Lee, D., Kim, M., & Lee, J. (2016). Adoption of green electricity policies: Investigating the role of environmental attitudes via big data-driven search-queries. *Energy Policy*. Retrieved from <https://doi.org/10.1016/j.enpol.2015.12.021>
- Li, J., Pei, X., Wang, X., Yao, D., Zhang, Y., & Yue, Y. (2021). Transportation mode identification with GPS trajectory data and GIS information. *Tsinghua Science and Technology*, 26(4), 403–416. Retrieved from <https://doi.org/10.26599/TST.2020.9010014>
- Li, L. L., Zhao, X., Tseng, M. L., & Tan, R. R. (2020). Short-term wind power forecasting based

- on support vector machine with improved dragonfly algorithm. *Journal of Cleaner Production*. Retrieved from <https://doi.org/10.1016/j.jclepro.2019.118447>
- Liang, L., Xu, M., Grant-Muller, S., & Mussone, L. (2020). Household travel mode choice estimation with large-scale data—an empirical analysis based on mobility data in Milan. *International Journal of Sustainable Transportation*, 15(1), 70–85. Retrieved 5 May 2022 from <https://doi.org/10.1080/15568318.2019.1686782>
- Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. *ArXiv Preprint*. Retrieved 26 January 2022 from <https://doi.org/doi.org/10.48550/arXiv.1912.06059>
- Liu, J., Wang, B., & Xiao, L. (2021). Non-linear associations between built environment and active travel for working and shopping: An extreme gradient boosting approach. *Journal of Transport Geography*, 92, 103034. Retrieved from <https://doi.org/10.1016/j.jtrangeo.2021.103034>
- Liu, Yan, & Cirillo, C. (2018). A generalized dynamic discrete choice model for green vehicle adoption. *Transportation Research Part A: Policy and Practice*. Retrieved from <https://doi.org/10.1016/j.tra.2018.01.034>
- Liu, Yang, Lyu, C., Khadka, A., Zhang, W., & Liu, Z. (2020). Spatio-Temporal Ensemble Method for Car-Hailing Demand Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(12), 5328–5333. Retrieved from <https://doi.org/10.1109/TITS.2019.2948790>
- Louviere, J. J., Hensher, D. A., Swait, J. D., & Adamowicz, W. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University Press. Cambridge University Press. Retrieved 27 May 2022 from <https://doi.org/10.1017/cbo9780511753831>
- Lu, J., Meng, Y., Timmermans, H., & Zhang, A. (2021). Modeling hesitancy in airport choice: A comparison of discrete choice and machine learning methods. *Transportation Research Part A: Policy and Practice*, 147, 230–250. Retrieved from <https://doi.org/10.1016/j.tra.2021.03.006>
- Ma, J., Wang, Y., Niu, X., Jiang, S., & Liu, Z. (2022). A comparative study of mutual information-based input variable selection strategies for the displacement prediction of seepage-driven landslides using optimized support vector regression. *Stochastic Environmental Research and*

- Risk Assessment*, 36(10), 3109–3129. Retrieved 11 July 2023 from <https://doi.org/10.1007/s00477-022-02183-5>
- Mahmoud, N., Abdel-Aty, M., Cai, Q., & Yuan, J. (2021). Predicting cycle-level traffic movements at signalized intersections using machine learning models. *Transportation Research Part C: Emerging Technologies*, 124, 102930. Retrieved from <https://doi.org/10.1016/j.trc.2020.102930>
- Majidifard, H., Jahangiri, B., Rath, P., Urrea Contreras, L., Buttlar, W. G., & Alavi, A. H. (2021). Developing a prediction model for rutting depth of asphalt mixtures using gene expression programming. *Construction and Building Materials*. Retrieved from <https://doi.org/10.1016/j.conbuildmat.2020.120543>
- Manifesty, O. R., & Park, J. Y. (2022). A Case Study of a 15-Minute City Concept in Singapore's 2040 Land Transport Master Plan: 20-Minute Towns and a 45-Minute City. *International Journal of Sustainable Transportation Technology*, 5(1), 1–11. Retrieved from <https://doi.org/10.31427/ijstt.2022.5.1.1>
- Martín-Baos, J. Á., García-Ródenas, R., & Rodríguez-Benitez, L. (2021). Revisiting kernel logistic regression under the random utility models perspective. An interpretable machine-learning approach. *Transportation Letters*, 13(3), 151–162. Retrieved 5 May 2022 from <https://doi.org/10.1080/19427867.2020.1861504>
- McCright, A. M., Marquart-Pyatt, S. T., Shwom, R. L., Brechin, S. R., & Allen, S. (2016). Ideology, capitalism, and climate: Explaining public views about climate change in the United States. *Energy Research and Social Science*. Retrieved from <https://doi.org/10.1016/j.erss.2016.08.003>
- Meti, N., Saednia, K., Lagree, A., Tabbarah, S., Mohebpour, M., Kiss, A., ... Tran, W. T. (2021). Machine Learning Frameworks to Predict Neoadjuvant Chemotherapy Response in Breast Cancer Using Clinical and Pathological Features. *JCO Clinical Cancer Informatics*, (5), 66–80. Retrieved from <https://doi.org/10.1200/cci.20.00078>
- Mi, X., Wang, S., Shao, C., Zhang, P., & Chen, M. (2021). Resident travel mode prediction model in Beijing metropolitan area. *PLoS ONE*, 16(November), e0259793. Retrieved 27 February 2022 from <https://doi.org/10.1371/journal.pone.0259793>

- Minal, S., Sekhar, C. R., & Madhu, E. (2019). Development of neuro-fuzzy-based multimodal mode choice model for commuter in Delhi. *IET Intelligent Transport Systems*, 13(2), 406–416. Retrieved 3 May 2022 from <https://doi.org/10.1049/iet-its.2018.5112>
- Mohd Ali, N. F., Mohd Sadullah, A. F., Abdul Majeed, A. P. P., Mohd Razman, M. A., & Musa, R. M. (2022). The identification of significant features towards travel mode choice and its prediction via optimised random forest classifier: An evaluation for active commuting behavior. *Journal of Transport and Health*, 25, 101362. Retrieved from <https://doi.org/10.1016/j.jth.2022.101362>
- Mohd Ali, N. F., Mohd Sadullah, A. F., P.P. Abdul Majeed, A., Mohd Razman, M. A., & Muazu Musa, R. (2020). Mode Choice Prediction using Machine Learning Technique for A Door-to-Door Journey in Kuantan City. *MEKATRONIKA*, 2(1), 73–78. Retrieved 27 February 2022 from <https://doi.org/10.15282/mekatronika.v2i1.6745>
- Momenzadeh, A., Shamsa, A., & Meyer, J. G. (2022). Bias or biology? Importance of model interpretation in machine learning studies from electronic health records. *JAMIA Open*, 5(3), 1–13. Retrieved 20 November 2022 from <https://doi.org/10.1093/jamiaopen/ooac063>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 109–139). Springer, Cham. Retrieved 15 June 2023 from https://doi.org/10.1007/978-3-030-89010-0_4
- Morikawa, T., Ben-Akiva, M., & McFadden, D. (2002). Discrete choice models incorporating revealed preferences and psychometric data. *Advances in Econometrics*, 16, 29–55. Retrieved from [https://doi.org/10.1016/S0731-9053\(02\)16003-8](https://doi.org/10.1016/S0731-9053(02)16003-8)
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44), 22071–22080. Retrieved 20 November 2022 from <https://doi.org/10.1073/pnas.1900654116>
- Muthukrishnan, R., & Rohini, R. (2017). LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*. Retrieved from

<https://doi.org/10.1109/ICACA.2016.7887916>

- Nam, D., & Cho, J. (2020). Deep neural network design for modeling individual-level travel mode choice behavior. *Sustainability (Switzerland)*, 12(18), 7481. Retrieved 27 February 2022 from <https://doi.org/10.3390/SU12187481>
- Naseri, H. (2019). Cost Optimization of No-Slump Concrete Using Genetic Algorithm and Particle Swarm Optimization. *Int J Innov Manag Technol*, 10(1). Retrieved from <https://doi.org/10.18178/ijimt.2019.10.1.832>
- Naseri, H., Ehsani, M., Golroo, A., & Moghadas Nejad, F. (2021). Sustainable pavement maintenance and rehabilitation planning using differential evolutionary programming and coyote optimisation algorithm. *International Journal of Pavement Engineering*, 0(0), 1–18. Retrieved from <https://doi.org/10.1080/10298436.2021.1873331>
- Naseri, H., Fani, A., & Golroo, A. (2020). Toward equity in large-scale network-level pavement maintenance and rehabilitation scheduling using water cycle and genetic algorithms. *International Journal of Pavement Engineering*. Retrieved from <https://doi.org/10.1080/10298436.2020.1790558>
- Naseri, H., Golroo, A., Shokoohi, M., & Gandomi, A. H. (2022). Sustainable pavement maintenance and rehabilitation planning using the marine predator optimization algorithm. *Https://Doi.Org/10.1080/15732479.2022.2095407*, 1–13. Retrieved 15 July 2022 from <https://doi.org/10.1080/15732479.2022.2095407>
- Naseri, H., Hosseini, P., Jahanbakhsh, H., Hosseini, P., & Gandomi, A. H. (2022). A novel evolutionary learning to prepare sustainable concrete mixtures with supplementary cementitious materials. *Environment, Development and Sustainability 2022*, 1–35. Retrieved 9 April 2022 from <https://doi.org/10.1007/S10668-022-02283-W>
- Naseri, H., Jahanbakhsh, H., Foomajd, A., Galustanian, N., Karimi, M. M., & Waygood, E. O. D. (2022). A newly developed hybrid method on pavement maintenance and rehabilitation optimization applying Whale Optimization Algorithm and random forest regression. *Https://Doi.Org/10.1080/10298436.2022.2147672*, 1–13. Retrieved 5 December 2022 from <https://doi.org/10.1080/10298436.2022.2147672>
- Naseri, H., Jahanbakhsh, H., Hosseini, P., & Moghadas Nejad, F. (2020). Designing sustainable

- concrete mixture by developing a new machine learning technique. *Journal of Cleaner Production*, 258, 120578. Retrieved 19 March 2020 from <https://doi.org/10.1016/j.jclepro.2020.120578>
- Naseri, H., Jahanbakhsh, H., Khezri, K., & Shirzadi Javid, A. A. (2021). *Toward sustainability in optimizing the fly ash concrete mixture ingredients by introducing a new prediction algorithm. Environment, Development and Sustainability*. Springer Netherlands. Retrieved from <https://doi.org/10.1007/s10668-021-01554-2>
- Naseri, H., Jahanbakhsh, H., & Moghadas Nejad, F. (2019). Developing a novel machine learning method to predict the compressive strength of fly ash concrete in different ages. *AUT Journal of Civil Engineering*, 0. Retrieved 12 August 2019 from <https://doi.org/10.22060/AJCE.2019.16124.5569>
- Naseri, H., Shokoohi, M., Jahanbakhsh, H., Golroo, A., & Gandomi, A. H. (2021). Evolutionary and swarm intelligence algorithms on pavement maintenance and rehabilitation planning. *International Journal of Pavement Engineering*, 1–15. Retrieved 26 September 2021 from <https://doi.org/10.1080/10298436.2021.1969019>
- Naseri, H., Shokoohi, M., Jahanbakhsh, H., Karimi, M. M., & Waygood, E. O. D. (2023). Novel Soft-Computing Approach to Better Predict Flexible Pavement Roughness. *Transportation Research Record: Journal of the Transportation Research Board*, 036119812311610. Retrieved 14 June 2023 from <https://doi.org/10.1177/03611981231161051>
- Naseri, H., Waygood, E. O. D., Wang, B., & Patterson, Z. (2022). Application of Machine Learning to Child Mode Choice with a Novel Technique to Optimize Hyperparameters. *International Journal of Environmental Research and Public Health*, 19(24), 16844. Retrieved 7 March 2023 from <https://doi.org/10.3390/ijerph192416844>
- Naseri, H., Waygood, E. O. D., Wang, B., Patterson, Z., & Daziano, R. A. (2021). A Novel Feature Selection Technique to Better Predict Climate Change Stage of Change. *Sustainability*, 14(1), 40. Retrieved 26 December 2021 from <https://doi.org/10.3390/su14010040>
- Nguyen-Sy, T., Wakim, J., To, Q. D., Vu, M. N., Nguyen, T. D., & Nguyen, T. T. (2020). Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method. *Construction and Building Materials*, 260, 119757.

Retrieved from <https://doi.org/10.1016/j.conbuildmat.2020.119757>

- Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., ... Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021. Retrieved from <https://doi.org/10.1155/2021/4832864>
- OECD. (2016). *Energy and air pollution: world energy outlook special report. Comprehensive Energy Systems*. Retrieved 10 July 2023 from www.iea.org/t&c/
- Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., & Valdes-Sosa, M. (2017). Fast Gaussian Naïve Bayes for searchlight classification analysis. *NeuroImage*. Retrieved from <https://doi.org/10.1016/j.neuroimage.2017.09.001>
- Panahi, H. K. S., Dehghani, M., Kinder, J. E., & Ezeji, T. C. (2019, September 1). A review on green liquid fuels for the transportation sector: A prospect of microbial solutions to climate change. *Biofuel Research Journal*. Green Wave Publishing of Canada. Retrieved 20 October 2021 from <https://doi.org/10.18331/brj2019.6.3.2>
- Pham, H., Jiang, X., & Zhang, C. (2022). Causality and Advanced Models in Trip Mode Prediction: Interest in Choosing Swissmetro. *OSF Preprints*. Retrieved 28 February 2022 from <https://doi.org/10.31219/OSF.IO/M4W38>
- Pierezan, J., & Coelho, L. dos S. (2018). Coyote Optimization Algorithm : A new metaheuristic for global optimization problems. *2018 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. Retrieved from 978-1-5090-6017-7/18/\$31.00 ©2018 IEEE
- Pierezan, J., Coelho, S., Mariani, V. C., & Lebensztajn, L. (2019). Electromagnetic Optimization. *2019 22nd International Conference on the Computation of Electromagnetic Fields (COMPUMAG)*, 1–4.
- Pierezan, J., Maidl, G., Massashi Yamao, E., dos Santos Coelho, L., & Cocco Mariani, V. (2019). Cultural coyote optimization algorithm applied to a heavy duty gas turbine operation. *Energy Conversion and Management*, 199(August), 111932. Retrieved from <https://doi.org/10.1016/j.enconman.2019.111932>
- Pineda-Jaramillo, J., & Arbeláez-Arenas, Ó. (2022). Assessing the Performance of Gradient-Boosting Models for Predicting the Travel Mode Choice Using Household Survey Data.

- Journal of Urban Planning and Development*, 148(2). Retrieved 14 February 2022 from [https://doi.org/10.1061/\(asce\)up.1943-5444.0000830](https://doi.org/10.1061/(asce)up.1943-5444.0000830)
- Ping, P., Qin, W., Xu, Y., Miyajima, C., & Takeda, K. (2019). Impact of driver behavior on fuel consumption: Classification, evaluation and prediction using machine learning. *IEEE Access*. Retrieved from <https://doi.org/10.1109/ACCESS.2019.2920489>
- Pirra, M., & Diana, M. (2019). A study of tour-based mode choice based on a Support Vector Machine classifier. *Transportation Planning and Technology*, 42(1), 23–36. Retrieved 3 May 2022 from <https://doi.org/10.1080/03081060.2018.1541280>
- Posada-Quintero, H. F., Molano-Vergara, P. N., Parra-Hernández, R. M., & Posada-Quintero, J. I. (2020). Analysis of risk factors and symptoms of burnout syndrome in colombian school teachers under statutes 2277 and 1278 using machine. *Social Sciences*, 9(3), 30. Retrieved 20 November 2022 from <https://doi.org/10.3390/socsci9030030>
- Příbýl, O., & Goulias, K. G. (2003). Application of Adaptive Neuro-Fuzzy Inference System to Analysis of Travel Behavior. In *Transportation Research Record* (pp. 180–188). SAGE PublicationsSage CA: Los Angeles, CA. Retrieved 10 June 2022 from <https://doi.org/10.3141/1854-20>
- Prochaska, J., Colleen, O., Redding, A., & Evers, K. E. (2015). *The transtheoretical model and stages of change." Health behavior: Theory, research, and practice.*
- Prochaska, J. O., & Colleen, A. (2015). Health Behavior: Theory, Research, and Practice. *Health Behavior: Theory, Research, and Practice*. Retrieved 19 October 2021 from https://books.google.ca/books?hl=en&lr=&id=PhUWCgAAQBAJ&oi=fnd&pg=PA125&dq=The+Transtheoretical+Model+and+Stages+of+Change&ots=-eoRjSG9GC&sig=mnXbWBzi-ugpRquLmtY0M-4ndgg&redir_esc=y#v=onepage&q=The+Transtheoretical+Model+and+Stages+of+Change&f=false
- Prochaska, J. O., & Velicer, W. F. (1997). The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12(1), 38–48. Retrieved 19 October 2021 from <https://doi.org/10.4278/0890-1171-12.1.38>
- Qian, Y., Aghaabbasi, M., Ali, M., Alqurashi, M., Salah, B., Zainol, R., ... Hussein, E. E. (2021). Classification of imbalanced travel mode choice to work data using adjustable svm model.

- Applied Sciences (Switzerland)*, 11(24), 11916. Retrieved 27 February 2022 from <https://doi.org/10.3390/app112411916>
- Ramachandran, A., Anupama, K. R., Adarsh, R., & Pahwa, P. (2018). Machine Learning-Based Techniques for Fall Detection in Geriatric Healthcare Systems. In *Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018* (pp. 232–237). Institute of Electrical and Electronics Engineers Inc. Retrieved from <https://doi.org/10.1109/ITME.2018.00059>
- Rau, C. S., Wu, S. C., Chien, P. C., Kuo, P. J., Chen, Y. C., Hsieh, H. Y., & Hsieh, C. H. (2017). Prediction of mortality in patients with isolated traumatic subarachnoid hemorrhage using a decision tree classifier: A retrospective analysis based on a trauma registry system. *International Journal of Environmental Research and Public Health*. Retrieved from <https://doi.org/10.3390/ijerph14111420>
- Richards, M. J., & Zill, J. C. (2019). Modelling mode choice with machine learning algorithms. In *Australasian Transport Research Forum, ATRF 2019 - Proceedings*.
- Riggs, W. (2017). Painting the fence: Social norms as economic incentives to non-automotive travel behavior. *Travel Behaviour and Society*, 7, 26–33. Retrieved from <https://doi.org/10.1016/j.tbs.2016.11.004>
- Salas, P., De la Fuente, R., Astroza, S., & Carrasco, J. A. (2022). A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. *Expert Systems with Applications*, 193, 116253. Retrieved from <https://doi.org/10.1016/j.eswa.2021.116253>
- Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*. Retrieved from <https://doi.org/10.1016/j.ijmedinf.2018.05.006>
- Sanchez, O. R., Repello, M., Carrega, A., & Bolla, R. (2021). Evaluating ML-based DDoS Detection with Grid Search Hyperparameter Optimization. In *Proceedings of the 2021 IEEE Conference on Network Softwarization: Accelerating Network Softwarization in the Cognitive Age, NetSoft 2021* (pp. 402–408). Institute of Electrical and Electronics Engineers Inc. Retrieved from <https://doi.org/10.1109/NetSoft51509.2021.9492633>

- Scheiner, J., Huber, O., & Lohmüller, S. (2019). Children's mode choice for trips to primary school: a case study in German suburbia. *Travel Behaviour and Society*, 15, 15–27. Retrieved from <https://doi.org/10.1016/j.tbs.2018.09.006>
- Sharma, J., Giri, C., Granmo, O. C., & Goodwin, M. (2019). Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *Eurasip Journal on Information Security*. Retrieved from <https://doi.org/10.1186/s13635-019-0098-y>
- Shen, J. (2009). Latent class model or mixed logit model? A comparison by transport mode choice data. *Applied Economics*, 41(22), 2915–2924. Retrieved 8 November 2021 from <https://doi.org/10.1080/00036840801964633>
- Solaymani, S. (2019). CO2 emissions patterns in 7 top carbon emitter economies: The case of transport sector. *Energy*, 168, 989–1001. Retrieved from <https://doi.org/10.1016/j.energy.2018.11.145>
- Song, F., Guo, Z., & Mei, D. (2010). Feature selection using principal component analysis. In *Proceedings - 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, ICSEM 2010*. Retrieved from <https://doi.org/10.1109/ICSEM.2010.14>
- Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*.
- Statistics Canada. (2020). Proximity measures database - early release. [Data Set]. <https://www150.statcan.gc.ca/N1/Pub/17-26-0002/172600022020001-Eng.Htm>.
- Statistics Canada. (2021a). Dissemination Block (DB), Dictionary, Census of Population, 2021. Retrieved from <https://www12-2021.statcan.gc.ca/census-recensement/2021/ref/dict/az/definition-eng.cfm?ID=geo014>
- Statistics Canada. (2021b). *Table 23-10-0067-01 Vehicle registrations, by type of vehicle*. Retrieved from <https://doi.org/10.25318/2310006701-eng>
- Sun, X., & Wandelt, S. (2021). Transportation mode choice behavior with recommender systems: A case study on Beijing. *Transportation Research Interdisciplinary Perspectives*, 11, 100408. Retrieved from <https://doi.org/10.1016/j.trip.2021.100408>

- Suresh, A., Udendhran, R., & Balamurgan, M. (2020). Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. *Soft Computing*. Retrieved from <https://doi.org/10.1007/s00500-019-04066-4>
- Susilo, Y. O., Williams, K., Lindsay, M., & Dair, C. (2012). The influence of individuals' environmental attitudes and urban design features on their travel patterns in sustainable neighborhoods in the UK. *Transportation Research Part D: Transport and Environment*. Retrieved from <https://doi.org/10.1016/j.trd.2011.11.007>
- Tamim Kashifi, M., Jamal, A., Samim Kashefi, M., Almoshaogeh, M., & Masiur Rahman, S. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, 29, 279–296. Retrieved from <https://doi.org/10.1016/j.tbs.2022.07.003>
- Tariq, F., & Shakeel, N. (2021). Modeling and Predicting Commuters' Travel Mode Choice in Lahore, Pakistan. *Global Social Sciences Review*, VI(III), 106–118. Retrieved from [https://doi.org/10.31703/gssr.2021\(vi-iii\).12](https://doi.org/10.31703/gssr.2021(vi-iii).12)
- Tariq, M. T., Hadi, M., & Saha, R. (2021). Using high-resolution signal controller data in the calibration of signalized arterial simulation models. In *Transportation Research Record* (Vol. 2675, pp. 1043–1055). SAGE PublicationsSage CA: Los Angeles, CA. Retrieved 29 June 2022 from <https://doi.org/10.1177/03611981211031882>
- Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland)*. Retrieved from <https://doi.org/10.3390/s18010018>
- Thanh, T. T. M., Ly, H. B., & Pham, B. T. (2020). A possibility of AI application on mode-choice prediction of transport users in Hanoi. In *Lecture Notes in Civil Engineering* (Vol. 54, pp. 1179–1184). Springer, Singapore. Retrieved 27 February 2022 from https://doi.org/10.1007/978-981-15-0802-8_189
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. Retrieved from <https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In *Explanatory Item Response*

- Models* (pp. 75–109). Springer, New York, NY. Retrieved 17 March 2022 from https://doi.org/10.1007/978-1-4757-3990-9_3
- Ullah, I., Liu, K., Yamamoto, T., Shafiullah, M., & Jamal, A. (2022). Grey wolf optimizer-based machine learning algorithm to predict electric vehicle charging duration time. *Transportation Letters*. Retrieved 5 December 2022 from <https://doi.org/10.1080/19427867.2022.2111902>
- Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., & Koppen, M. (2017). Detecting malicious URLs using machine learning techniques. In *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*. Institute of Electrical and Electronics Engineers Inc. Retrieved from <https://doi.org/10.1109/SSCI.2016.7850079>
- Viegas de Lima, I., Danaf, M., Akkinapally, A., De Azevedo, C. L., & Ben-Akiva, M. (2018). Modeling Framework and Implementation of Activity- and Agent-Based Simulation: An Application to the Greater Boston Area. *Transportation Research Record*, 2672(49), 146–157. Retrieved 5 April 2022 from <https://doi.org/10.1177/0361198118798970>
- Wade, B. S. C., Joshi, S. H., Gutman, B. A., & Thompson, P. M. (2017). Machine learning on high dimensional shape data from subcortical brain surfaces: A comparison of feature selection and classification methods. *Pattern Recognition*. Retrieved from <https://doi.org/10.1016/j.patcog.2016.09.034>
- Walk Score. (2021). [dataset]. Available at: <https://www.walkscore.com/CA-QC/Montr%C3%A9al>.
- Wang, B., Waygood, E. O. D., Daziano, R. A., Patterson, Z., & Feinberg, M. (2021). Does hedonic framing improve people's willingness-to-pay for vehicle greenhouse gas emissions? *Transportation Research Part D: Transport and Environment*, 98, 102973. Retrieved from <https://doi.org/10.1016/j.trd.2021.102973>
- Wang, Fangru, & Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record*, 2672(47), 35–45. Retrieved 5 December 2022 from <https://doi.org/10.1177/0361198118773556>
- Wang, Fei, Yu, Y., Wang, X., Ren, H., Shafie-Khah, M., & Catalão, J. P. S. (2018). Residential electricity consumption level impact factor analysis based on wrapper feature selection and

- multinomial logistic regression. *Energies*, 11(5), 1180. Retrieved 18 June 2022 from <https://doi.org/10.3390/en11051180>
- Wang, S., Mo, B., Hess, S., & Zhao, J. (2021). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. *ArXiv Preprint ArXiv:2102.01130*.
- Wang, Yan, & Sherry Ni, X. (2019). A XGBOOST RISK MODEL VIA FEATURE SELECTION AND BAYESIAN HYPER-PARAMETER OPTIMIZATION. *International Journal of Database Management Systems*, 11(01), 01–17. Retrieved 12 November 2022 from <https://doi.org/10.5121/ijdms.2019.11101>
- Wang, Yong, & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. *Proceedings of the 9th European Conference on Machine Learning Poster Papers*. Retrieved 27 May 2022 from <https://researchcommons.waikato.ac.nz/handle/10289/1183>
- Waygood, E., & Avineri, E. (2011). Does ‘500g of CO₂ for a mile trip’ mean anything? Towards More Effective Presentation of CO₂ Information. CHIME (CHildren’s Independent Mobility Environments) View project The role of traffic, road design and parent’s perceptions on children active transp. *Proceedings of the Transportation Research Board 90th Annual Meeting*. Retrieved 9 November 2021 from <https://www.researchgate.net/publication/263256511>
- Waygood, E. O. ., Wang, B., Daziano, R. A., Patterson, Z., & Kohlová, M. B. (2020). Vehicle Choice And Co₂ Emissions Information: Framing Effects And Individual Climate Change Stage Of Change. *Annual Meeting of the Transportation Research Board*. Washington, D.C.
- Waygood, E. O. D., & Avineri, E. (2016). Communicating transportation carbon dioxide emissions information: Does gender impact behavioral response? *Transportation Research Part D: Transport and Environment*. Retrieved from <https://doi.org/10.1016/j.trd.2016.08.026>
- Waygood, E. O. D., Friman, M., Olsson, L. E., & Taniguchi, A. (2017). Transport and child well-being: An integrative review. *Travel Behaviour and Society*, 9, 32–49. Retrieved from <https://doi.org/10.1016/j.tbs.2017.04.005>
- Waygood, E. O. D., Wang, B., Daziano, R. A., Patterson, Z., & Kohlová, M. B. (2021). The Climate Change Stage of Change Measure Compared to GEB and NEP: Vehicle Choice Experiment.

- Journal of Environmental Planning and Management*. Retrieved from <https://doi.org/10.1080/09640568.2021.1913107>
- Wynes, S., & Nicholas, K. A. (2017). The climate mitigation gap: Education and government recommendations miss the most effective individual actions. *Environmental Research Letters*. Retrieved from <https://doi.org/10.1088/1748-9326/aa7541>
- Xun, J., Waygood, E. O. D., Wang, B., Naseri, H., Loiselle, L. A., & Patterson, Zachary Feinberg, M. (2022). Exploring the Effects of New Framing Techniques for Greenhouse Gas Emissions. In *Transportation Research Board 101st Annual Meeting*. Washington DC, United States. Retrieved 22 May 2022 from <https://annualmeeting.mytrb.org/OnlineProgram/Details/17423>
- Yamauchi, T. (2013). Mouse trajectories and state anxiety: Feature selection with random forest. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*. Retrieved from <https://doi.org/10.1109/ACII.2013.72>
- Yang, J., & Ma, J. (2019). Compressive sensing-enhanced feature selection and its application in travel mode choice prediction. *Applied Soft Computing Journal*. Retrieved from <https://doi.org/10.1016/j.asoc.2018.10.048>
- Yang, M. X., Tang, X., Cheung, M. L., & Zhang, Y. (2021). An institutional perspective on consumers' environmental awareness and pro-environmental behavioral intention: Evidence from 39 countries. *Business Strategy and the Environment*, 30(1), 566–575. Retrieved from <https://doi.org/10.1002/bse.2638>
- Zha, D., Yang, G., Wang, W., Wang, Q., & Zhou, D. (2020). Appliance energy labels and consumer heterogeneity: A latent class approach based on a discrete choice experiment in China. *Energy Economics*. Retrieved from <https://doi.org/10.1016/j.eneco.2020.104839>
- Zhang, H. (2021). Research on solvency measurement and prediction-based on the analysis of financial indicators of insurance companies. *Academic Journal of Business & Management*, 3(9), 20–23. Retrieved 30 May 2022 from <https://doi.org/10.25236/AJBM.2021.030905>
- Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 20, 22–35. Retrieved from <https://doi.org/10.1016/j.tbs.2020.02.003>
- Zhou, X., Wang, M., & Li, D. (2019). Bike-sharing or taxi? Modeling the choices of travel mode

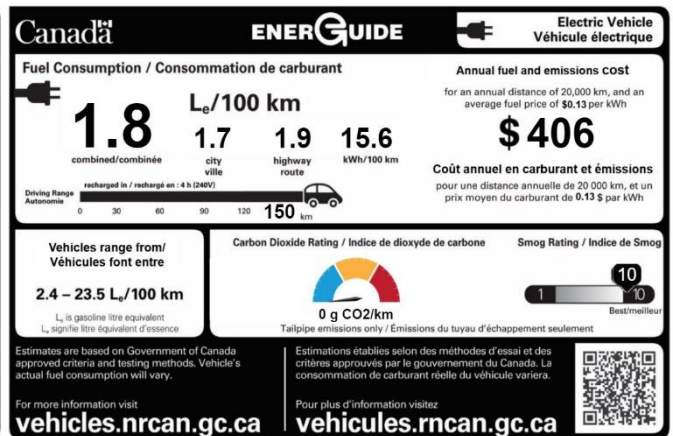
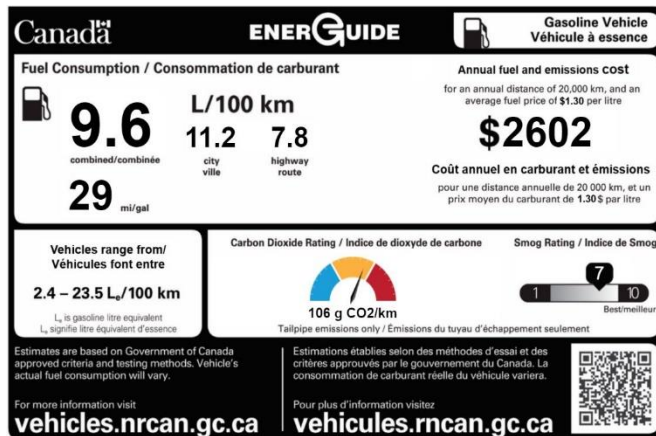
- in Chicago using machine learning. *Journal of Transport Geography*, 79, 102479. Retrieved from <https://doi.org/10.1016/j.jtrangeo.2019.102479>
- Zhou, Z., Wang, M., Huang, J., Lin, S., & Lv, Z. (2021). Blockchain in Big Data Security for Intelligent Transportation With 6G. *IEEE Transactions on Intelligent Transportation Systems*. Retrieved from <https://doi.org/10.1109/TITS.2021.3107011>
- Zhu, S., & Zhu, F. (2019). Cycling comfort evaluation with instrumented probe bicycle. *Transportation Research Part A: Policy and Practice*, 129, 217–231. Retrieved from <https://doi.org/10.1016/j.tra.2019.08.009>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. Retrieved from <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- Zwerts, E., Janssens, D., & Wets, G. (2007). *How the presence of children affects parents' travel behavior*.

APPENDIX A

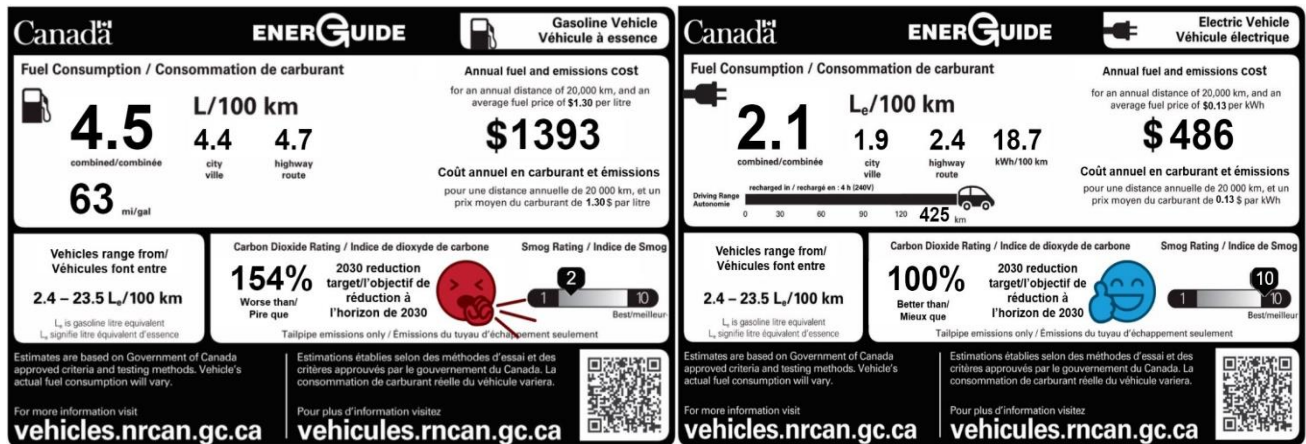
Table A1 Characteristics of respondents.

Variable	Frequency	Variable	Frequency
Gender		CC-SoC	
Male	970	CC-SoC1	222
Female	1042	CC-SoC2	259
Other	3	CC-SoC3	387
Education attainment		CC-SoC4	728
No formal education	1	CC-SoC5	419
Elementary school education	4	Province*language	
Less than high school equivalent	26	BC	257
High school diploma or equivalent	382	AB	258
Registered Apprenticeship or other trades certificate or diploma	165	SK	123
College, CEGEP or other non-university certificate or diploma	542	MB	142
University bachelor's degree	656	ON	258
Degree in medicine, dentistry, veterinary medicine or optometry (M.D., D.D.S., D.M.D., D.V.M., O.D.)	11	QC-English	227
Master's degree (e.g., M.A., M.Sc., M.Ed., M.B.A.)	203	QC-French	750
Doctoral degree (e.g., Ph.D.)	25	Employment	
Household income		Full-time (more than 30 hours per week)	1094
Less than \$5,000	11	Part-time/casual job	195
\$5,000-\$7,499	6	Homemaker	75
\$7,500-\$9,999	12	Full-time student	98
\$10,000-\$12,499	18	Retired	383
\$12,500-\$14,999	13	Not currently employed, but looking for work	105
\$15,000-\$19,999	33	Other	65
\$20,000-\$24,999	60	Urbanization	
\$25,000-\$29,999	60	Urban	1786
\$30,000-\$34,999	69	Rural	226
\$35,000-\$39,999	68	Ethnicity*	
\$40,000-\$49,999	132	North American	62
		Aboriginal	

\$50,000-\$59,999	156	Other North American	506
\$60,000-\$74,999	217	European	1025
\$75,000-\$84,999	129	Latin, Central and	19
\$85,000-\$99,999	203	South American	28
\$100,000-\$124,999	213	African	30
\$125,000-\$149,999	153	Asian	225
\$150,000-\$174,999	85	Oceania	4
\$175,000-\$199,999	54	Prefer not to answer	116
\$200,000 or more	105	Age	
Prefer not to answer	218	Min	19
Own a vehicle currently		Max	87
Yes	1821	Median	48
No	194	Mean	47.14



(a) frame3: Pressure gauge label and vehicle choice



(b) frame4: -30% societal goal with new emoticon label and vehicle choice



(c) frame5: Patriotic goal label and vehicle choice



(d) frame6: Dirty air label and vehicle choice



(e) frame7: Tree scale label and vehicle choice



(f) frame8: -30% societal goal with thumbs up/down label and vehicle choice

Figure A1 Other new-vehicle labels with different framing techniques designed to analyze the impacts of different GHG framings on individual's vehicle choices (framings three to eight)