Case Studies on Transport Policy 21 (2025) 101557

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

# Key factors affecting transportation choices in school commuting in Lisbon – A machine learning approach

Vivek Bhosale [a], Miguel San Payo [a], Gabriel Cipriano [b], António R. Andrade [a,*]

[a] IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon 1049-001, Portugal
[b] Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Iscte – Instituto Universitário de Lisboa, Lisbon, Portugal

ARTICLE INFO

*Keywords:*
Active Commuting to School
School transportation mode choice
Hands up survey
Machine learning classification

ABSTRACT

Understanding the behaviour of students in choosing a transportation mode to school is crucial to promote Active Commuting to School (ACS) and the adoption of healthier lifestyles. Therefore, analysing all types of transportation modes with multiple factors/features is essential, though it can be a challenge in statistical modelling. The main objective of the present study was to determine the factors that contribute to the choice of a particular mode in school transportation, by using Machine Learning (ML) algorithms: Extreme Gradient Boosting (XGB), Random Forest (RF), Decision Tree (DT) and Multinomial Logistic Regression (MNL). Data from the 'Hands Up' Survey in Lisbon, Portugal, between 2018 and 2021, with 10 different modes of transportation were analysed. A range of factors including safety around school, socioeconomic status of schools' parishes, school regime, school grades and the proximity of schools to the different public transportation modes were considered. The algorithms have been compared in terms of accuracy scores. The XGB algorithm shows the best performance (64 % accuracy and 0.33 Macro F1) for multi-class classification, while RT, DT and MNL provide accuracy of 40 %, 37 % and 47 % respectively. Weighted Average Feature Importance (WAFI) have been determined for all variables. For the best-performing algorithm, the XGB, the combination factor of school regime and school grade is the most relevant factor, contributing to around 21.2 % for multi-class classification. WAFI scores for each variable suggest that the proximity of schools to various public transports is an important factor contributing more than 50 % for the predominance of private car in school transportation.

## 1. Introduction

Active Commuting to School (ACS) refers to the practice of students traveling to and from school by physically active means, such as walking or cycling. ACS can encourage a healthier lifestyle in students by increasing their levels of Physical Activity (PA) (Chillon et al., 2010; Jesus et al., 2021) and helping to reduce the risk of obesity and other cardiovascular diseases (Saris et al., 2003; Schoeppe et al., 2013). Additionally, ACS can contribute to more sustainable and healthier cities by decreasing traffic congestion (Black et al., 2001) and mitigating the harmful effects of noise and air pollution on students' health (Mattioli et al., 2020; Pantelaki et al., 2024). Since it is estimated that by 2050, 70 % of people will live in urban areas (United Nations, 2022), understanding and promoting ACS in urban areas seems to be crucial for addressing various issues related to multiple Sustainable Development Goals (SDGs), such as SDG 3 on "Good Health and Well-being," SDG 11 on "Sustainable Cities and Communities," and SDG 13 on "Climate Action."

In our case study of Lisbon municipality (the capital of Portugal), passive modes of transportation, which do not involve physical activity (e.g. car, bus), account for around 75 % of students' school commuting choices, with private car use representing more than 45 % of students' transportation modes. ACS represents only about 25 % of students' choices, with a clear prevalence of walking over cycling (CML, 2021). With our study, a technical exercise on the use of ML techniques applied to school commuting in a highly urbanized area, we wanted to understand which are the key factors that determine Lisbon students' choices when commuting to school. While more traditional statistical approaches have various restrictions in handling highly skewed data with multiple variables, this study shows how ML algorithms can be integrated to study school commuting phenomena, helping stakeholders and policymakers to be informed on how to promote and improve ACS

* Corresponding author.
*E-mail addresses:* vivek.bhosale@tecnico.ulisboa.pt (V. Bhosale), miguel.san.payo@tecnico.ulisboa.pt (M. San Payo), gabriel.cipriano@iscte-iul.pt (G. Cipriano), antonio.ramos.andrade@tecnico.ulisboa.pt (A.R. Andrade).

https://doi.org/10.1016/j.cstp.2025.101557
Received 21 January 2025; Received in revised form 27 July 2025; Accepted 27 July 2025
Available online 28 July 2025

policies in highly urbanized areas.

## 2. Literature review

### 2.1. Factors affecting school commuting choices

Understanding human behaviour in school commuting is complex, as it depends on multiple factors and features. Furthermore, different case studies and different contexts have led to the identification of different determinants. For example, in Yogyakarta, Indonesia, Irawan et al. (2022) found that motorcycling becomes a transport choice among students once they turn 16, as part of gaining independence and pursuing status as motorcyclists. However, this appears to be a context-specific phenomenon that does not occur in many other parts of the world. Even so, some other factors that emerge in various studies appear to be fairly consistent across different countries and contexts.

When analysing factors affecting school commuting choices in Sri Lanka, Dias et al. (2022) found that the type of school students attend (i.e., private or public) plays a significant role in the choice between using a school bus or a private vehicle, with students from private schools being less likely to use public transport. Socio-economic factors and inequalities were also explored by Xiao et al. (2021), who revealed that socio-economic characteristics are also crucial when considering transportation mode choices. To encourage the use of public transportation for school commuting, policies should enhance the reliability and safety of public transportation, as parents prioritize safety, convenience, and reliability when choosing a mode of transport for their children (Nanthawong et al., 2024).

Other studies, such as Lopes et al. (2014) and Huertas-Delgado et al. (2017), focused on the built environment, showing a decrease in ACS in highly urbanized areas due to parental fears of traffic and accidents, with barriers to ACS being related to the built environment. Palma et al. (2020) also analysed sociodemographic characteristics to understand perceived parental barriers to active transport among children and adolescents. The authors note that, among children, the parental barrier was the distance to school, while among adolescents, traffic was the main barrier. Moreover, Palm and Farber (2020) identified weak but significant association between transit accessibility and the participation in after-school activities among high-school students in the Greater Toronto Area. Multiple sources of data were also fused and analysed into a multi-method study on school commuting in Dunedin, New Zealand, by Mindell at al. (2021), who have identified different factors that are major barriers to the use of public transport to school, such as distance, cost, parental trip chaining, built environment features, the weather, convenience, and also safety perceptions.

Some other recent studies have also focused on the intersection of transportation and students' well-being. For example, Waygood et al. (2019) explored the relations between children's life satisfaction, travel satisfaction and the transport modes chosen in school commuting, suggesting that children do not enjoy frequent travel. In addition, Humberto et al. (2022) concluded that the perceptions of children regarding urban mobility, along with their views and needs, are commonly disregarded in mobility-related studies, suggesting that more inclusive approaches are needed in school commuting. Therefore, policies should consider the mutual dependencies between parents' and children's commuting needs, knowing that parents are the ultimate decision-makers for their children's mode of transport (Ehteshamrad et al., 2022).

When focusing on ACS specifically, factors affecting ACS choices among adolescents in Oporto (also in Portugal), a city with a very hilly and rugged landscape, were analysed by Pizzaro et al. (2016), who observed that walking is the most frequent mode of ACS, while cycling is quite uncommon. For students in Oporto, a distance of 2.0 km seems to be the maximum reasonable distance for walking between home and school, and increasing distances are associated with higher levels of passive transportation to school (Pizzaro et al., 2016). Walking time and walking distance were also identified in other studies as the most influential factors in the decision to walk-to-school (Benita et al., 2023; Corral-Abós et al., 2021), and Helbich (2017) also concluded that urban policies should target walking and cycling separately rather than combining them into a single category, as they have different characteristics. In line with these studies, regarding trip length affecting ACS choices, Jesus et al. (2021) noted that special attention should be given to cycling on trips over 2.0 km, as it is associated with greater daily physical activity and a way to maintain stable daily physical activity levels. Significant shifts in ACS were also observed by Gálvez-Fernández et al. (2023) when students changed from pre-primary to primary and then to secondary school.

All these conditional factors affecting school commuting and ACS seem to result in children gaining autonomy and independence at a later age. Accordingly, data from Cordovil et al. (2015) shows that in Portugal only 21 % of primary school students and 45 % of secondary school students travel home from school actively and independently. Therefore, based on the literature, we believe that the most prominent factors that might affect the school commuting choices of students in Lisbon municipality are the safety levels of the built environment, socio-economic factors, school regime (public or private), students' age (grade level), and the distance from home to school. Additionally, it is believed that features as the proximity and accessibility of schools to public transport systems will also play a significant role in determining transport choices.

### 2.2. Methodological approaches to study commuting phenomena

Modelling travel mode choices has been thoroughly explored in the past five decades. The pioneering method for determining travel mode has been the Multinomial Logit (MNL) (McFadden, 1973). Since then, this model has been widely used in transportation mode choice problems, as it works on the utility maximization principle and easily estimates parameters. Thus, it is based on a comparison between a pair of alternatives at a single time, and for each comparison it neglects the dependence characteristics from all other alternatives. Discrete choice models have also been commonly used to model travel mode choices (see, e.g., Ben-Akiva and Lerman, 1985; Ben-Akiva and Bierlaire, 1999).

More recently, various Machine Learning (ML) models have been used in studies to model travel mode choice. Wang and Ross (2018) and Hagenauer and Helbich (2017) have provided an effective comparison between ML classification algorithms and the MNL model using household travel data. In these studies, the RF showed a good advantage over other algorithms (Hagenauer and Helbich, 2017), while for some cases the Extreme Gradient Boosting (XGB) has an advantage (Wang and Ross, 2018). In both studies, it has been highlighted that in the case of imbalanced data sets, ML algorithms outperformed the traditional MNL model.

Wang et al. (2021) have also performed an enriching study using hundreds of ML classifiers to predict travel behaviour. The study involved the use of various parameters under a ML algorithm and a comparison of it to get to the model with the best parameters. The authors have concluded that ML algorithms performed with lesser computational effort than MNL on larger datasets even though the accuracy for ML algorithms is better by 4 or 5 percentage points. Gao et al. (2021) have used a novel method of mode choice prediction that is a combination of ML and decision-making models. The model used by the authors provided extrapolation ability over normal classification algorithms.

Numerous more studies have been performed around ML compared to MNL, which include Cheng et al. (2019), Ermagun et al. (2015), Hillel et al. (2018), Jahangiri and Rakha (2015), Sekhar et al. (2016), Shukla et al. (2015), Wang et al., 2020b, Wang et al. 2020a, Xian-Yu (2011) and Zhou et al. (2019). In these studies, Wang et al. (2020c, 2020a) have mainly focused their research on the use of deep neural networks for choice analysis using utility functions and deriving the best possible economic interpretation. Cheng et al. (2019) and Sekhar et al. (2016)

have considered RF as first choice classifier for determining choice and have been able to produce better results compared to MNL model. Xian-Yu (2011) has explored the support vector machine classifier for performing the choice analysis. Shukla et al. (2015) have considered the use of fuzzy logic and data mining, while Hillel et al. (2018) have performed a case study on generating passenger choice sets for transportation in London. Ermagun et al. (2015) have explored ML and nested logit models for school trips. The study involves dealing with each alternative separately to determine how each mode is affected by independent parameters. The study also highlighted that ML classifiers have better accuracy compared to the nested logit model.

Nevertheless, none of the previous studies have been performed for 10 modes of transportation with very large sample sizes. Hence, as far as we know, previous studies on school transportation and ACS have not included classification machine learning techniques. Therefore, the present study aims to explore the use of ML classification techniques for school commuting, offering an alternative approach to inform education, transportation, and urban planning policies. Table 1 provides a summary of all the previous studies performed using ML, along with sample size, number of modes studied, and models used.

**Table 1**
Summary of previous related studies.

| Sr No. | Authors – Year | Sample Size | Number of Modes studied | Models Used |
|---|---|---|---|---|
| 1 | (Xian-Yu, 2011) | 4,725 | 5 | MNL, BM, BOOSTING, DT |
| 2 | (Ermagun et al., 2015) | 4,700 | 6 | NL, RF |
| 3 | (Jahangiri and Rakha, 2015) | N.A. | 5 | KNN, SVM, DT, BAGGING, RF, MNL |
| 4 | (Shukla et al., 2015) | 100,000 | 5 | DNN, DT |
| 5 | (Sekhar et al., 2016) | 4,976 | 8 | RF, DT, MNL |
| 6 | (Hagenauer and Helbich, 2017) | 230,608 | 4 | MNL, DNN, NB, SVM, BOOSTING, BAGGING, RF |
| 7 | (Hillel et al., 2018) | N.A. | 4 | DNN, BAGGING, BOOSTING, KNN, GLM, BM, RF, SVM |
| 8 | (F. Wang and Ross, 2018) | 51,910 | 4 | BOOSTING, MNL |
| 9 | (Cheng et al., 2019) | 7,276 | 5 | RF, SVM, BOOSTING, MNL |
| 10 | (Zhou et al., 2019) | 30,000 | 2 | MNL, KNN, DT, SVM, BM, BOOSTING, BAGGING, RF |
| 11 | (Wang et al., 2020b) | 8,418 | 5 | NL, MNL, DNN |
| 12 | (Wang, et al., 2020a) | 80,000 | 5 | MNL, DNN |
| 13 | (Kim, 2021) | 172,889 | 4 | ANN, RF, XGB |
| 14 | (Gao et al., 2021) | 2,316 | 4 | RF |
| 15 | (Chen and Cheng, 2023) | 81,086 | 4 | MNL, XGB, DNN |
| 16 | (Gálvez-Fernández et al., 2023) | 42,074 | 2 | MNL |
| 17 | (Etaati et al. 2024) | 1,387 | 2 | MNL, RF, SVM |
| 18 | (Kolidakis et al., 2024) | 496 | 2 | MNL, SVM, DT, KNN, RF, XGB |
| 19 | Present Study | 140,702 | 10 | XGB, RF, DT, MNL |

Notes: DNN – Deep Neural Network, RF- Random Forest, DT – Decision Trees, NL – Nested logit, MNL – Multinomial logit, BOOSTING – Boosting methods, BAGGING – Bagging methods, SVM – Support Vector Machine, KNN – K nearest neighbours, BM – Bayesian Models, GLM – Generalized linear Models, ANN – Artificial neural network, N.A. – Not Available.

## 3. Research opportunities and objectives

Although many studies, such as Zhang and Xie (2008), have examined the use of classification algorithms for travel mode prediction, the present study, to the best of our knowledge, makes the following contributions to the existing literature:

- Examines ten different commuting modes, including public transport, private vehicles, and active modes, an approach not previously undertaken.
- Explores the large set of ten transport modes simultaneously, along with ten independent variables or features, using multiclass classification algorithms.
- Identifies the factor with the highest influence on students' decision-making by analysing classifier feature importance.

While other studies have addressed the use of classification algorithms for travel mode prediction, only recently has specific attention been given to school commuting, with studies like Assi et al. (2019), Etaati et al. (2024), and Kolidakis et al. (2024). Therefore, the objective of this study is to deepen an understanding of the use of ML models in school commuting and ACS, focusing on the commuting behaviour of students in Lisbon municipality, by analysing real data regarding all modes of school commuting and incorporating various influencing factors. Even though some transport modes have a proportion of less than 1 %, with the help of ML algorithms and resampling techniques, we will analyse all modes regardless of their proportion. This will be achieved by determining the parameters that influence the choice of a particular mode of transportation using ML algorithms such as XGB, RF, DT, and the MNL model.

In fact, in high-dimensional multi-class classification problems, the ML algorithms are particularly suited, enhancing predictive accuracy compared with traditional modelling approaches. However, it may not necessarily provide better interpretability and insights to support policy design. Therefore, the present paper explores both the ML algorithms and more traditional modelling approaches. This will hopefully provide interesting common findings/insights, while discussing how researchers should be cautious and avoid using a single model to draw conclusions.

## 4. Material and methods

### 4.1. Dataset

The present study uses the database obtained from the 'Hands Up' survey conducted by the Lisbon Municipality, Portugal (CML, 2021). The 'Hands Up' survey began in 2018 and is currently conducted every year in October. The survey targets students in the compulsory school system. Table 2 presents the structure of the compulsory school system in Portugal, specifying the cycles and grade levels (Law No. 46/86), where students aged 6 to 18 are required to attend school (Law No. 85/2009).

Students from military schools, professional schools, and schools with specialized artistic programs are excluded from the 'Hands Up' survey due to school characteristics. The survey is then distributed to all other public and private schools in Lisbon within the compulsory school system, and data is collected and aggregated by grade level, school

**Table 2**
Compulsory school system with cycles and grades.

| Cycle | Grades |
|---|---|
| 1st cycle (basic education) | 1–4 |
| 2nd cycle (basic education) | 5–6 |
| 3rd cycle (basic education) | 7–9 |
| High school (secondary education) | 10–12 |

Sources: Law n. 46/86; Law n. 85/2009.

regime, and school name. The 'Hands Up' survey aims to understand how students typically travel to school. To achieve this, students in the compulsory school system are asked, 'How do you normally travel to school?' and are then sequentially presented with different transport modes (e.g., 'By car?'). The teacher counts the number of hands raised and records it on the survey sheet (CML, 2021). Due to the data collection characteristics of the 'Hands Up' survey, concerns about data validity and reliability usually arise. Note that in the past, Mammen et al. (2014) and Vitale et al. (2019) have also analysed data from the 'Hands up' survey; while Wit et al. (2012) confirmed the validity and reliability of the 'Hands up' survey, cross-checking the answers that children gave with their parents, obtaining consistent results (with Cronbach's alphas ranging from 0.81 to 1.00).

Note that in the public school system, in order to assign a student to a school close to their home and, indirectly, to promote ACS, Legislative Order n. 6/2018 states that a student should be prioritized for enrolment in a specific school when guardians demonstrably live in the area of influence of the intended educational establishment. For students attending compulsory school in the private school system, assignments to schools are based on each private school's standards and rules, regardless of home to school distance. However, despite Legislative Order n. 6/2018 aimed at promoting walkability and reducing distances from home to school, Table 3 shows a dependence on private vehicles (automobile mode) for school commuting in Lisbon. Approximately 45 % of students travel by car to school, while ACS is chosen by 24–27 % (walking + cycling) and public transport by 19–25 % (CML, 2021).

### 4.2. Participants

Table 4 shows the number of schools that participated in the 'Hands Up' surveys, as well as the total number of schools in Lisbon. It also shows the number of answers and the total number of students in Lisbon.

In the current study, data as well as the main statistics collected between 2018 and 2021 by the Lisbon Municipality (CML) have been used. It is important to note that the CML 'Hands Up' survey reports mainly contain descriptive statistics, and our study has not been included or analysed in the current CML studies.

### 4.3. Measures

Along with the transportation mode indicated by students in the CML 'Hands Up' survey, we aimed to integrate factors and features known to affect students' school commuting choices. However, it is important to note that the information collected by the CML 'Hands Up' survey pertains to each school, making it impossible to determine students' residences, the public transportation options available near their homes, or the distances from students' homes to school. Therefore, we could only create and consider factors and features related to each school, requiring some creativity in generating these variables based on the previous literature.

**Table 3**
School commuting distribution in Lisbon by year and across modes (2018–2021) (%).

| Mode | 2018 (%) | 2019 (%) | 2020 (%) | 2021 (%) |
|---|---|---|---|---|
| Automobile | 42.8 | 48.6 | 51.4 | 48.2 |
| Foot | 30.1 | 23.4 | 24.2 | 25.8 |
| Bus | 16.6 | 15.5 | 13.1 | 14.1 |
| Metro | 5.6 | 5.1 | 3.5 | 4.9 |
| Train | 1.5 | 2.5 | 2.0 | 2.2 |
| Bicycle | 0.8 | 1.0 | 1.5 | 1.5 |
| School Transport | 0.9 | 1.3 | 1.5 | 1.4 |
| Motorcycle | 0.9 | 1.0 | 0.8 | 0.9 |
| Tram | 0.2 | 0.5 | 0.4 | 0.4 |
| Other | 0.6 | 1.1 | 1.6 | 0.6 |

Source: Lisbon Hands Up Survey report – CML, 2021.

**Table 4**
'Hands Up' survey summary.

| Year | Schools | | Students | |
|---|---|---|---|---|
| | Sample (n) | Population (N) | Sample (n) | Population (N) |
| 2018 | 85 | 228 | 15,689 | 35,068 |
| 2019 | 188 | 228 | 47,141 | 55,392 |
| 2020 | 159 | 229 | 39,883 | 49,669 |
| 2021 | 179 | 226 | 38,078 | 53,616 |

Source: Lisbon Hands Up Survey report – CML, 2021.

To create a socio-economic status variable for each school, they were classified based on the characteristics of the parishes in which they are located. Since Lisbon is divided into 24 parishes, the socio-economic status variable for each school was developed through the characterisation of each parish's socio-economic profile. The implementation of this variable involved conducting an Exploratory Factor Analysis (FA) in RStudio, incorporating five social indicators from the 2015–2016 Social Diagnostic of Lisbon (CML, 2016). These indicators from each parish are statistical measures that quantitatively represent social concepts and provide insights into specific aspects of social reality. The FA considered the following indicators from the Social Diagnostic of Lisbon: i) early school dropout rate, ii) percentage of food aid program beneficiaries, iii) percentage of registered unemployed people, iv) percentage of families benefiting from social inclusion income, and v) percentage of people receiving unemployment benefits. The FA results showed that 14 parishes had a positive score and 10 parishes had a negative score for the 'socio-economic status' variable. Those with the highest scores represent a higher socio-economic status, such as São Domingos de Benfica (1.261) and Belém (1.141), while those with the lowest scores represent a lower socio-economic status, such as Marvila (−1.646) and Santa Clara (−2.646).

To define school surroundings' safety levels in terms of built environment, Google Maps and Google Street View photos were used. The safety around the school were classified into 4 levels, which range from 0 (worst) to 3 (best). Table 5 highlights the criteria for each level of categorization.

To define the school proximity to different transportation modes and bike stations, GIRA Bike Sharing System (BSS) map and Google Maps were used. It was assumed that the effects of different transportation modes and their proximity could vary for shorter versus longer distances/proximities.

### 4.4. Procedures

Data from the 'Hands Up' survey was compiled with data regarding parishes' socio-economic status, school surroundings-built environment' safety and school proximity to different transportation modes. Table 6 shows the detailed description and type of variable used in the present study.

All data were inserted into an Excel.csv file, which is 11.1 MB in size and contains 140,702 rows and 13 columns. The file includes all the previously mentioned factors/variables, along with each student's transportation mode choice. Each row in the Excel file represents an individual student and their choice.

Then, all programming for the present study was performed in Python. Fig. 1 shows the detailed flowchart of the procedure followed for

**Table 5**
Safety levels classification and description next to each school main access.

| Level | Description |
|---|---|
| 0 | It does not have traffic signs, and it does not have crosswalks. |
| 1 | It has either traffic signs or crosswalks, not both. |
| 2 | It has traffic signs and crosswalks, but it either has narrow roads with the potential of traffic or narrow sidewalks with little space |
| 3 | It has traffic signs, crosswalks and larger roads or sidewalks |

**Table 6**
Variable name, description, and type.

| Variable | Description | Type |
|---|---|---|
| Safety level | Indicates safety around school environment. | Ordinal |
| Socio-economic | Socio-economic status of a school area. | Ordinal |
| Proximity to Bicycle sharing station (BSS) – (ProxEG) | Distance (in metres) to nearest BSS. | Continuous |
| Proximity to Subway – (ProxEM) | Distance (in metres) to nearest metro station. | Continuous |
| Proximity to Bus Station (ProxEB) | Distance (in metres) to nearest bus station. | Continuous |
| Proximity to Train Station (ProxEC) | Distance (in metres) to nearest train station. | Continuous |
| Proximity to Tram Station (ProxEE) | Distance (in metres) to nearest tram station. | Continuous |
| School Regime | Highlights – public or private school. | Boolean |
| School Grades | Grade level in which student is studying. | Discrete |
| School Regime X School Grades (SxG) | Interaction terms between School regime and Grade levels. | Discrete |
| School Regime X Socio-economic (SxSe) | Interaction terms between School regime and socio-economic status factor. | Ordinal |

the ML model analysis, highlighting key sections referenced throughout the process. Data pre-processing involved cleaning and balancing the data according to the requirements of the classification algorithms. Afterwards, the data were split into two parts: the training set (80 %) and the testing set (20 %). The testing data were fed into a trained model to make predictions, which were compared with actual values to get accurate results. Following the testing phase, the model was validated using the k-fold cross-validation technique. Finally, to end the modelling procedure, feature importance was derived to reach conclusions regarding the major factor affecting the decision-making by students.

### 4.5. Imbalanced dataset

Table 2 highlights the distribution of various modes across the years, and it can be observed that the dataset is imbalanced, i.e., highly skewed. A class imbalance has been a common issue considering travel choice classification (Hagenauer and Helbich, 2017; Hillel et al., 2021; Kim, 2021; Pirra and Diana, 2019; F. Wang and Ross, 2018; S Wang et al., 2020). Imbalanced data can be usually dealt with by using two approaches: firstly, data must be pre-processed using resampling (Batista et al., 2004); secondly, using cost-sensitive learning (Domingos, 1999). Fernández et al. (2011) have successfully analysed both approaches and have pointed out the key challenges on dealing with imbalanced datasets. Chen and Cheng (2023) have effectively highlighted that limited efforts have been put to solve this problem. Multiple studies on imbalanced datasets using various resampling techniques and classification algorithms have been conducted. The study involves resample methods like Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN), while the classifiers used are MNL, XGB and DNN. Important conclusions show that for each choice mode, different combinations work effectively and overall F1 scores improve for the minority class. Ling et al. (2004) and Sun et al. (2007) have studied cost-sensitive learning for imbalanced data. Cost-sensitive learning approach involves the minimization of misclassification costs.

In our study, the XGB classifier uses a preprocessing approach with SMOTE oversampling technique (Chawla et al., 2002), while RF, DT and MNL model use a cost-sensitive learning approach. SMOTE (Chawla et al., 2002) is a widely used and effective approach for tackling class imbalance in classification datasets, as SMOTE addresses the issue of data scarcity in the minority class by generating synthetic samples that bridge the gap between minority class instances in feature space. Table 7 shows the distribution of classes after resampling. Considering RF, DT and MNL model hyperparameter 'class-weight' has been heuristically set to 'balanced' for considering cost-sensitivity of the algorithm.

### 4.6. Classifiers

ML models the mode choice prediction as a classification problem, which is given values of independent variables, and it determines the most likely variable of all the dependent variables. The present study analyses three ML classifiers including RF, DT, and XGB. The traditionally used MNL has also been used to compare the results with ML classifiers. All the classifiers have been previously used in travel mode choice determination (Biagioni et al., 2008; Cheng et al., 2019; Ermagun et al., 2015; Sekhar et al., 2016). Wang et al. (2021) has broadly represented various studies carried out around ML classifiers. That study highlights that tree-based techniques and neural networks have gained more interest due to higher prediction accuracy and for being able to quantify feature importance.

RF is an ensemble decision tree-based classifier that works using bagging technique (Breiman, 2001). The bagging procedure often has a better performance, since it reduces the variance with changing bias. RF is a specific version of bagging, which performs feature bagging at splitting of each candidate. In our study, the RF algorithm consists of 100 trees and at every tree node each split considers 2 randomly selected variables.

DT is a classifier performing classification using a tree like structure. In this tree structure, nodes represent binary rules of decision-making that split feature space and leaves illustrate the classification choices (Breiman, 2017). DT has been considered a powerful classification method when handling a non-linear feature space. Ensemble DT have been more accurate in terms of overfit reduction and noisy data reduction (Breiman, 1996). In our study, the DT consists of the following hyperparameters: criterion='gini', splitter='best', min_samples_split = 2, min_samples_leaf = 1.

XGB is also a DT-based ensemble classifier that works with a boosting principle (Friedman, 2001). Additive learning forms the basis of this ensemble classifier. In simpler terms, this algorithm builds by iteratively sequencing low-accuracy trees with the aim of minimising a loss function. For every iteration, every misclassified tree is weighted more than a properly classified tree. The final results of the model are weighted results of all the involved decision trees. The hyperparameters used in the present study for the XGB algorithm are: 'multi = softprob' and 'n_estimators' = 100.

Hyperparameters is the term mostly used in ML that represents the parameters that must be defined for any ML algorithm. By tuning these hyperparameters using cross-validation technique, the performance of the model can be improved to greater extent.

MNL models have traditionally been used for discrete choice modelling and most frequently used in travel mode choice classification (Ben-Akiva and Lerman, 1985). This model is considered as a baseline or benchmark classifier in the current study. MNL model is the Logistic Regression model from the scikit-learn python library. It has been used by defining hyperparameters 'multi_class' as 'ovr' and 'solver' as 'liblinear'.

Table 8 highlights the Python library used for the creation of the ML classifier and MNL models. Scikit-Learn is a crucial package for many classifiers across ML.

### 4.7. Classification performance

The measurement of imbalanced classification performance is crucial as it is highly likely that the model predicts the majority class (i.e., the one that occurs the largest number of times in the dataset). Thus, overall accuracy will always be high but the performance of the model on a minority class (i.e., the ones that occur a lower number of times in the dataset) would be poor. Therefore, it is essential to consider various

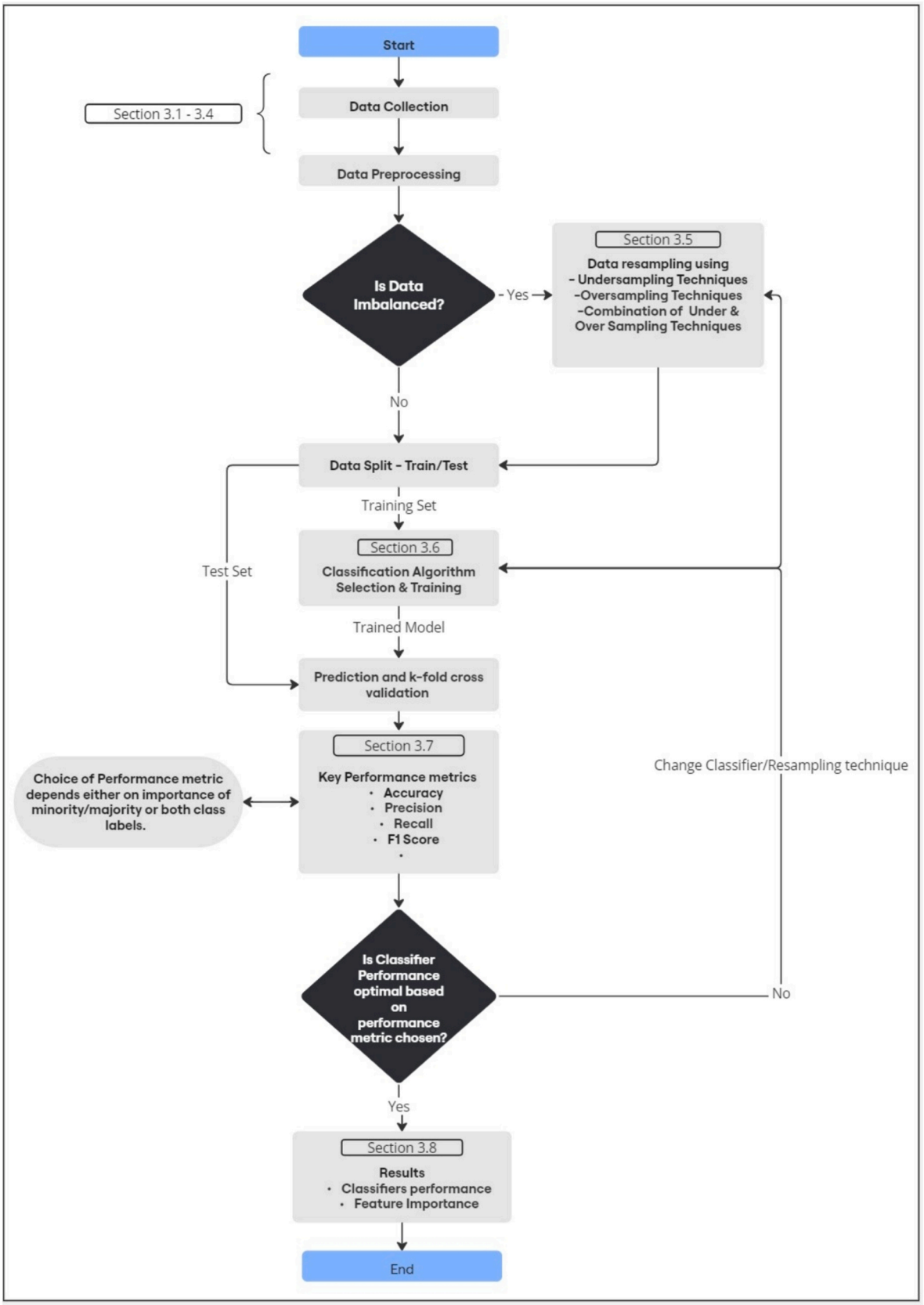


**Fig. 1.** Process Flowchart.

**Table 7**
Mode's distribution after using SMOTE oversampling.

| Modes | Count | Percentage % |
|---|---|---|
| Foot | 28,146 | 19.82 |
| Bicycle | 2,500 | 1.76 |
| Bus | 16,437 | 11.58 |
| School Transport | 5,000 | 3.52 |
| Train | 5,000 | 3.52 |
| Metro | 10,000 | 7.04 |
| Tram | 10,000 | 7.04 |
| Automobile | 54,879 | 38.66 |
| Motorcycle | 5,000 | 3.52 |
| Other | 5,000 | 3.52 |

**Table 8**
Python libraries for modelling.

| Python Package | Version | Use |
|---|---|---|
| scikit-learn | 1.2.1 | Import of RF, DT and MNL model |
| xgboost | 1.7.6 | Import of XGB model |
| imbalanced-learn | 0.10.1 | Resampling – under sampling or over sampling |

metrics which have been previously used (Chen and Cheng, 2023; Hagenauer and Helbich, 2017) in travel mode choice, such as the recall, the precision, and the F1 score.

Fig. 2 represents the confusion matrix which forms the basis for derivation of metrics. Eqs. (1)–(4) represent the formula for Accuracy, Precision, Recall, and F1 score, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

In our study, by using resampling and cost-sensitive techniques, metrics such as Recall, Precision and F1 scores for minority classes have been optimized. The Macro F1 score, which is an unweighted mean of the F1 scores of all classes (Chen and Cheng, 2023) has been derived to compare it with the accuracies.

A k-fold cross-validation is the most used error-estimation method in Machine Learning predictive models. The aim of cross-validation is to test the model before even using it for actual testing data. The procedure of validation involves splitting the dataset into k subsets or folds. Then, training and evaluation is performed k times, every time the validation set has a different fold. At the end, the model's general performance is the average of the performance metrics from each fold. In the case of class imbalances, to use stratified k-fold cross-validation, which ensures that the proportion of positive to negative examples found in the original distribution is respected in all the folds (Japkowicz, 2013). The stratified

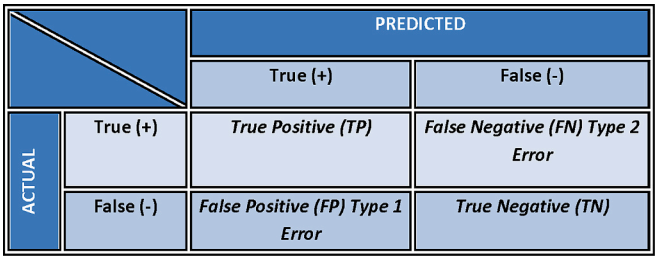


**Fig. 2.** Confusion Matrix.

k-fold cross-validation is a k-fold cross-validation that preserves the imbalanced class distribution in each fold. Thus, considering the imbalanced dataset that is used in the present study, a stratified 5-fold cross-validation has been used. Mean accuracy derived from 5-fold cross validation for various models has been considered during overall comparison.

### 4.8. Feature importance

One of the major objectives of the present study is to understand the factors that influence the choice of a transportation mode in school commuting. To achieve this objective, Feature Importance has been explored, which supports valid explanation. All the ML algorithms used are based out of DTs and are modelled using training data. The trained model produces the Feature Importance for each feature, thus indicating how important the respective feature/factor was in the classification.

Feature Importance derivation is possible in three ways: i) by using built-in attribute present in each classifier (Wang and Ross, 2018); ii) by using SHAP (Shapley Additive exPlanations) package (Lundberg and Lee, 2017); and iii) by using the permutation-based method (Hagenauer and Helbich, 2017). In the current study, Feature Importance for all the classifiers has been calculated using in-built attribute. Built-in attribute in RF and DT is based on 'Gini impurity'. Gini Impurity is being decreased at each node of the tree and the final importance is ranked taking the aggregate of all trees (Hastie et al., 2001). XGB built-in attribute is based on 'gain', 'frequency', and 'cover'. Gain measures the overall value addition made by a feature to the nodes present on it, the frequency measures how many times the feature is used among generated trees, the cover measures the comparative quality of concerned observations by a feature (Wang and Ross, 2018). Feature importance for the MNL model has been derived from the standardised regression coefficients (Kwak and Clayton-Matthews, 2002).

## 5. Results

### 5.1. Preliminary analysis

A descriptive statistics analysis was performed on the dataset used in the current case study. The analysis reports information about the mean, standard deviation, minimum and maximum values. Table 9 summarizes these descriptive statistics of all variables.

### 5.2. Analysis of classification models

Table 10 provides a summary of accuracies calculated for the prediction and validation phase. Training and testing accuracies have been compared to examine whether the model is overfitted or underfitted (Qin, 2020). For every model studied in the present work, the training and testing accuracies have minor variations, highlighting that all the models are neither overfitted nor underfitted.

**Table 9**
Descriptive statistics (mean, standard deviation, minimum and maximum) of variables.

| Variables | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| Safety | 1.774 | 0.6117 | 0 | 3 |
| Socio-economic | −0.212 | 0.928 | −1.261 | 2.646 |
| ProxEG [m] | 1,621.689 | 1,367.016 | 5 | 6,500 |
| ProxEM [m] | 1,544.785 | 1,714.464 | 10 | 7,400 |
| ProxEB [m] | 129.964 | 103.201 | 5 | 500 |
| ProxEC [m] | 2,614.507 | 4,510.517 | 5 | 57,000 |
| ProxEE [m] | 3,480.498 | 2,675.794 | 15 | 10,000 |
| School regime | 0.674 | 0.469 | 0 | 1 |
| Grade | 5.483 | 3.473 | 0 | 12 |
| S x G | 3.414 | 3.988 | 0 | 12 |
| S x Se | 0.398 | 1.945 | −1.26114 | 2.646 |

**Table 10**
Prediction and Validation Accuracies for each ML classifier (RF, DT and XGB) and MNL Model.

| Model | Prediction | | Validation | |
|---|---|---|---|---|
| | Training | Testing | Mean | SD (%) |
| RF | 0.414 | 0.399 | 0.396 | 0.3 |
| DT | 0.381 | 0.367 | 0.366 | 0.5 |
| XGB | 0.623 | 0.613 | 0.643 | 0.2 |
| MNL | 0.475 | 0.468 | 0.476 | 0.6 |

Fig. 3 shows the accuracy box plot for all the models used. These accuracies have been derived from the 5-fold stratified cross-validation. Concerning mean accuracy, XGB performs the best with 64.3 %, followed by the traditional MNL model with 47.6 %. RF and DT classifiers perform poorly with lower accuracies of 39.6 % and 36.6 %, respectively. XGB has the lowest standard deviation (SD) of 0.2 %, whereas MNL has the highest of 0.6 %. RF and DT provide an SD of 0.3 % and 0.5 %, respectively. Box plots for each classifier show that there is a small variation across all the validation accuracies.

Fig. 4 is a representation of F1 scores for all the classifiers used. The automobile mode has the highest, while the bicycle mode has the lowest F1 score, across all the classifiers. F1 scores for automobile range from 0.6 (for DT) to 0.8 (for XGB), whereas for bicycle they range from 0.01 (for XGB) to 0.09 (for DT and RF). Foot, Bus, and Metro modes have an average performance with F1 scores ranging from 0.21 (MNL for Metro) to 0.54 (XGB for Foot). For motorcycle and other modes, XGB has F1 scores of 0.47 and 0.3, respectively, while other models perform poorly with F1 scores below 0.2. School transport and train modes follow a similar trend with each one delivering F1 scores between 0.02 and 0.29, across all classifiers.

Fig. 5 shows a comparison between the Macro F1 and Accuracy values for each model considered in the present study. XGB has the highest Accuracy (61.3 %) and largest Macro F1 (0.335). MNL has an Accuracy of 46.8 %, but the lowest Macro F1 (0.199). Macro F1 scores for RF and DT are 0.248 and 0.237, respectively.

Tables 11 and 12 provide the summary for the overall results (Prediction phase) for the XGB and the MNL model, respectively. Considering XGB, the commuting mode that obtained the highest Precision was Bicycle with a score of 0.91, but it fails in Recall with a score of 0.08, which leads to an F1 score of 0.15. For all other modes, good values for the Precision and the Recall results into a good F1 score.

For the MNL model, it is observed that modes like train and tram fail in Precision with scores below 0.1, whereas modes such as bicycle, school transport and other fail in Precision, as well as Recall, with scores below 0.05.

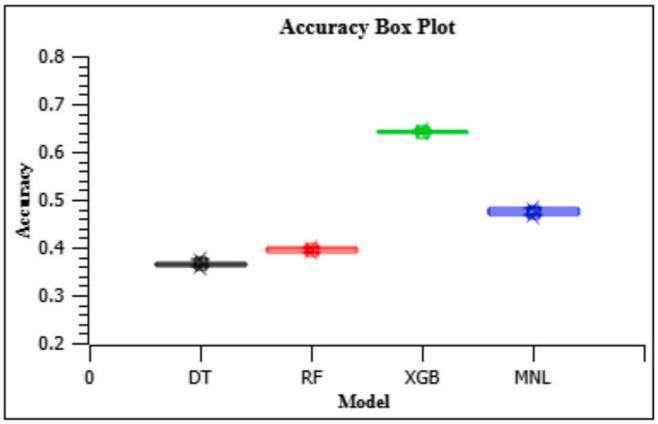


**Fig. 3.** Box Plot for Mean Accuracies – Validation Study.

### 5.3. Feature importance

The Feature Importance (FI) for all the factors and models examined in the present case study is represented in Fig. 6. Considering the MNL model, ProxEC (Proximity to Train Station) and SxSe (interaction term between school regime and socio-economic status) have high feature importance of 33 % and 21 %, respectively; while other factors have importance below 10 %, with ProxEB being the least important with a score of only 2.4 %.

Grade and ProxEC account for 20 % and 17 % importance, respectively, for the RF model, while other factors account for less than 15 %. If the DT model is taken into consideration, ProxEC is a factor with almost 19 % importance, whereas all other factors are below 15 %.

When examining the XGB model, which is the model that has the highest accuracy and therefore the model which can predict the commuting mode the best, one can see that the features with the highest importance are SxG and School Regime provide 21 % and 14 % importance, respectively, followed by SxSe with around 12 % and School Regime with around 11 %. These results go hand in hand with results from previous studies such Dias et al. (2022) which emphasize the importance of the School Regime in the school commuting modal choice, but for the Lisbon case study one can see that this feature gains even more importance when interacting with the age of the student, which is also interesting when comparing with results from studies such as Cordovil et al. (2015) and Barranco-Ruiz et al. (2018), which emphasize how important the age of the student is for the commuting choice. In addition to this, although it is in interaction with the school regime, the fact that the socio-economic variable also appears with high feature importance corroborates findings from studies such as Pinto et al. (2017) which shows that socio-economic status of a family has a major influence on ACS.

### 5.4. Determinants in school transportation choice

In the previous section, we analysed the Feature Importance with respect to each classification model. However, it is also important to find the main determinants irrespective of models. Therefore, a weighted average of feature importance (WAFI) has been also computed using the combination factor of Accuracy and Macro F1 (A x MF1). Eqs. (5) and (6) represent the formulas for A x MF1 and WAFI, respectively. Table 13 and 14 shows the results for A x MF1 and WAFI, respectively.

$$A \times MF1 = Accuracy(\%) \times MacroF1 \tag{5}$$

$$WAFI = \frac{\sum_{AllModels} A \times MF1 \times FeatureImportancescore}{\sum_{AllModels} A \times MF1} \tag{6}$$

Results suggest that the top five determinants for overall choice are ProxEC, SxG, SxSe, ProxEE and Grade with WAFI larger than 9 %. While other determinants also have significant contribution, summing all together a total WAFI of 36 %.

## 6. Discussion

### 6.1. Classifier performance

The ensemble tree based XGB classifier outperforms all the other models. This indicates that ensemble algorithms, by having a combination of multiple classifier trees, provide good results for travel mode choice modelling. The reason behind XGB performing notably better than DT and RT is the way the classifier minimises loss function at each node of the tree. As the number of features and modes are sufficiently high, improving the loss function at every point provides great overall performance. The trend of XGB performing better than MNL (Wang and Ross, 2018) and other classifiers (Hillel et al., 2018) has been previously observed, but not in school commuting data. Thus, future studies must aim to improve data quality and optimise the XGB algorithm

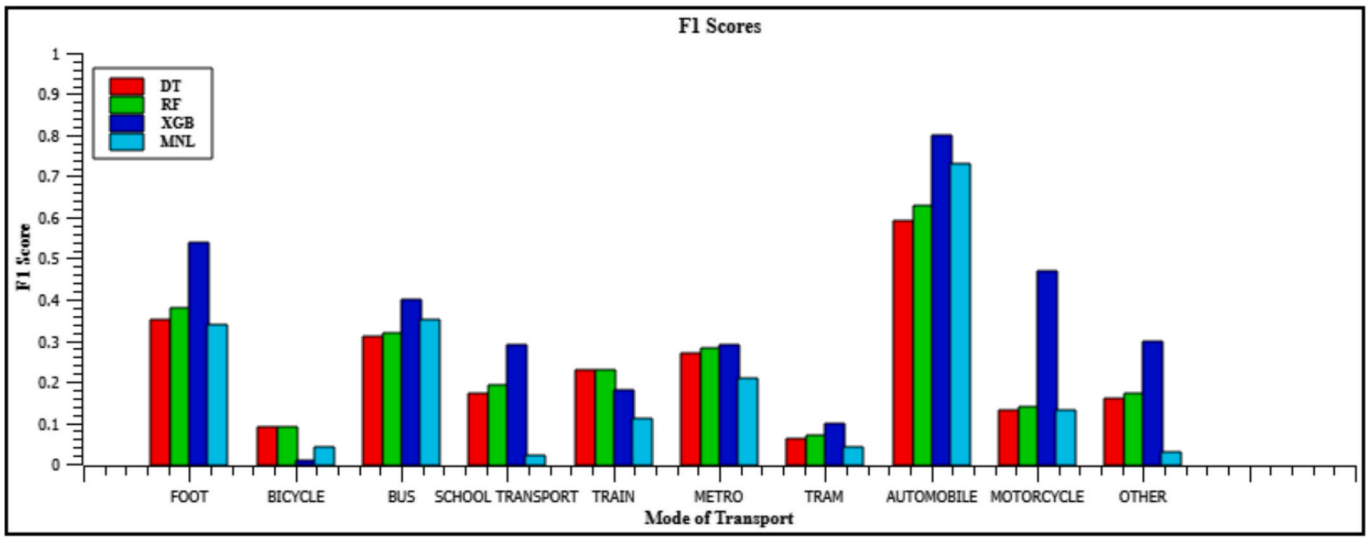


**Fig. 4.** Comparison of the F1 Scores across each mode of transport and for all the models (DT, RF, XGB and MNL).

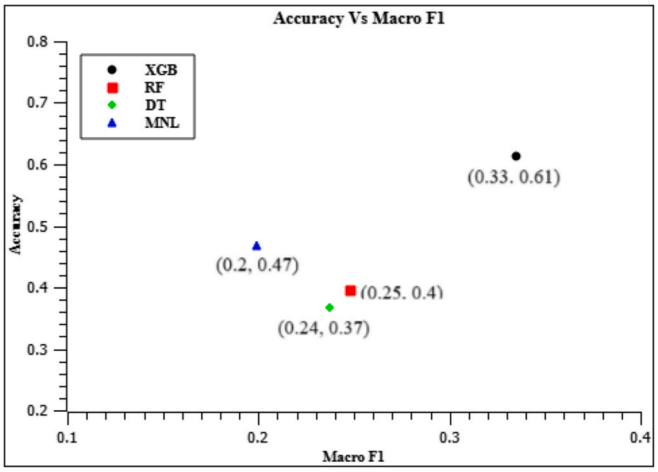


**Fig. 5.** Macro F1 and Accuracy for Predicted data.

**Table 11**
Summary for the Prediction overall result for the XGB Model.

| Mode | Precision | Recall | F1 Score |
|---|---|---|---|
| Automobile | 0.73 | 0.83 | 0.78 |
| Foot | 0.49 | 0.54 | 0.51 |
| Bus | 0.46 | 0.33 | 0.39 |
| Metro | 0.42 | 0.37 | 0.39 |
| Train | 0.52 | 0.16 | 0.25 |
| Bicycle | 0.91 | 0.08 | 0.15 |
| School Transport | 0.56 | 0.55 | 0.56 |
| Motorcycle | 0.84 | 0.65 | 0.73 |
| Tram | 0.68 | 0.86 | 0.76 |
| Other | 0.65 | 0.58 | 0.61 |

parameters, adding further features adapted to the school commuting problem.

Although the MNL model performance is better than DT and RF in terms of Accuracy, it does not produce satisfactory Macro F1 scores. Fig. 4 clearly shows that Macro F1 for MNL has a lower score than for all other models. Since the data is highly skewed, the MNL algorithm fails on F1 scores for each class. Classes like bicycle, school transport and other, which form the lowest percentage of all the classes together, have been wrongly predicted by the MNL algorithm.

**Table 12**
Summary of the Prediction overall result for the MNL Model.

| Mode | Precision | Recall | F1 Score |
|---|---|---|---|
| Automobile | 0.80 | 0.68 | 0.74 |
| Foot | 0.43 | 0.30 | 0.36 |
| Bus | 0.37 | 0.34 | 0.36 |
| Metro | 0.17 | 0.30 | 0.22 |
| Train | 0.09 | 0.15 | 0.11 |
| Bicycle | 0.03 | 0.04 | 0.04 |
| School Transport | 0.01 | 0.02 | 0.02 |
| Motorcycle | 0.11 | 0.18 | 0.13 |
| Tram | 0.01 | 0.42 | 0.04 |
| Other | 0.02 | 0.03 | 0.03 |

DT and RF performances are quite close to each other, with a nominal difference in overall Accuracy, but when Macro F1 scores are compared, RF exceeds DT. As the RF works on the principle of randomized splitting of combined classifier trees, it can comprehend complexity among the various variables, and therefore, RF performs better than DT for F1 scores, which is also true for the present case study.

One of the overall general trends, which can be seen in terms of F1 scores and across all the classifiers/models, is that classes with the lowest percentages are not predicted as accurately as classes with the highest percentage. It has also been observed that XGB performs quite well on all the classes regardless of their frequency in the dataset. Thus, it is very important to study the performances of different classifiers using F1 scores, rather than considering only Accuracy in direct comparison.

### 6.2. *Feature importance*

The findings on Feature Importance, for the overall classification of all the ten modes, do not follow a similar pattern, across all the models studied in the current case study. This clearly indicates a highly complex interaction between independent features and dependent variables/classes. Several modes were studied, and the number of factors analysed, in the present case study, is significantly higher than in previous studies. This suggests that data scientists and modellers must be cautious in the way they attribute importance to specific features/factors, as they highly depend on the predictive model being used.

Considering the best performing XGB model, SxG and School regime contribute the most, while grade and safety contribute the least. In both RF and DT algorithms, a similar trend is followed with Grade and

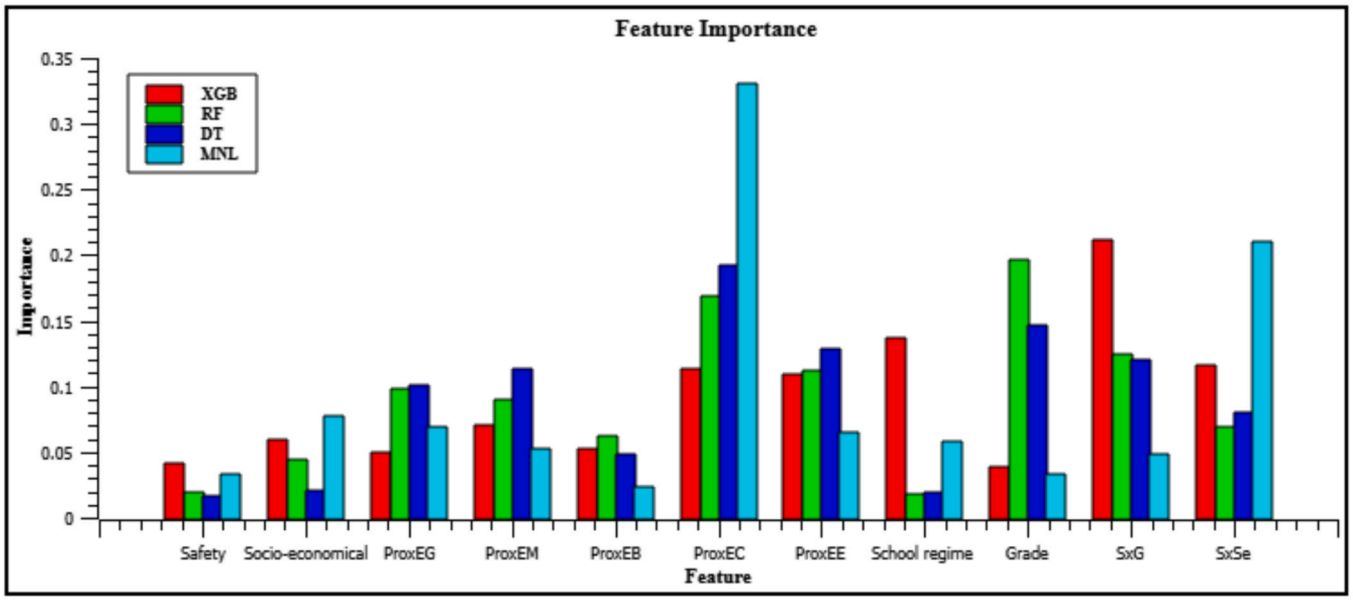


**Fig. 6.** Feature Importance for each classifier (XGB, RF, DT, MNL).

**Table 13**
A x MF1 scores for each model.

| Mode | Accuracy (%) | Macro F1 | A x MF1 (%) |
|---|---|---|---|
| XGB | 61.3 | 0.335 | 20.54 |
| RF | 39.6 | 0.248 | 9.82 |
| DT | 36.6 | 0.237 | 8.67 |
| MNL | 46.8 | 0.199 | 9.31 |

**Table 14**
Weighted Average Feature Importance (WAFI) for each variable.

| Variables | WAFI |
|---|---|
| Safety | 0.032 |
| Socio-economic | 0.054 |
| ProxEG | 0.073 |
| ProxEM | 0.077 |
| ProxEB | 0.048 |
| ProxEC | 0.181 |
| ProxEE | 0.104 |
| School regime | 0.078 |
| Grade | 0.089 |
| SxG | 0.146 |
| SxSe | 0.119 |

ProxEC being the most important, while school regime and safety being the least important. In the case of the traditional MNL model, ProxEC and SxSe have the highest influence on the model performance, whereas ProxEB and safety have the lowest influence.

### 6.3. Determinants for sustainable school transportation

Overall, one of the major findings is that none of the explaining variables/features is contributing more than 25 % individually. Hence, the analysis of the weighted average feature importance score (WAFI) emphasises that each variable has a significant effect on mode choice, since the number of modes being studied is large.

All the proximity variables (ProxEG, ProxEM, ProxEB, ProxEC, and ProxEE) effectively contribute around 50 % to the choice of transportation model, meaning that distances to each of the sustainable transport modes (Bicycle, Bus, Metro, Train and Tram) have a large contribution. Thus, the present study indicates that schools closer to sustainable modes of transport can significantly facilitate the adoption of these transport modes, thereby supporting emission reductions along with a cleaner environment in metropolitan areas. Similar results highlighting that distance to school is a major contributor have been previously observed in other school commuting studies (Ferri-García et al., 2020; Palma et al., 2020). A major contribution of the present study is its examination of nearly all modes of sustainable transportation using multi-class classification. Thus, the interdependence of different modes and factors is analysed in an optimal way, rather than dealing with a single mode at a time. Factors that might affect any single mode of transport are often more predictable and, therefore, the present study used a more innovative approach of multi-class classification for predicting all modes of transport simultaneously.

Another 50 % of contribution is due to other variables (School Regime, Grade, SxG, SxSe and Socio-economic), highlighting how important it is for public or private schools, based on the neighbourhood they are included in. The present study briefly points out that schools need to analyse students based on grade levels, as well as on the socio-economic status of their neighbourhood. Several studies have also previously concluded that the type of school (Gálvez-Fernández et al., 2023) and urban policies (Helbich, 2017) affect active transportation among children. Therefore, a micro-level approach (analysing the specific area around school) is necessary for driving more sustainable and public transportation among students.

These findings are relevant to improve the ACS and the holistic design of urban mobility policies, which should be adapted to school communities, controlling for certain features/variables, such as the schools' grade levels (if primary or secondary school) and/or their socioeconomic status. In a nutshell, proximity variables contribute up to 50 %, whereas socio-economic, school regime, grade and interactions contribute to another 50 %. To leverage the largest impact to shift current ACS must include all dimensions, and design public policies in different domains, such as urban planning, transport infrastructure, educational and social system. For example, in the urban planning domain, even the location of stations/stops and schools as relevant facilities in the urban mobility and the educational systems should be adapted/optimized to these communities, to make ACS a more popular choice. In fact, given the high share of automobile use for school commuting in Lisbon, this study also seems to question the educational policies that assigning students to specific schools, primarily based on

the distance from home to school, without considering other determinants/dimensions (e.g. proximity to public transport). Therefore, these other dimensions appear to be necessary to consider when assigning students to a school, as part of the educational public policy, especially since this study has shown that there are other factors influencing ACS and the use of public transport.

## 7. Conclusions

This research work examines school travel model choice, for ten different modes of transportation, through ML algorithms along with traditionally used MNL, using the 'Hands-Up' survey data for Lisbon schools from 2018 to 2021. All the models are compared for accuracy, as well as the F1 scores, as data is highly imbalanced. Validation has been also performed using a stratified 5-fold cross-validation. The following conclusions have been drawn from the present study regarding the methodological approach and main findings:

1. XGB classifier performs best, in terms of both accuracy as well as F1 scores. This establishes that not only overall prediction, but the XGB is able to predict each mode more accurately than other models, as the XGB classifier minimises loss function at each node of the tree.
2. As the data studied is highly imbalanced, with a large number of features, the ensemble ML classifiers also proved to be more effective than the MNL model.
3. When it comes to active or passive mode choices to commute to school, all models predicted automobile, foot, and bus modes more effectively than other modes, as they form a larger part of the input dataset. Due to a larger number of data points, the model convergence improves.
4. Regarding the determinants affecting these mode choices, all the features significantly affect the choice, and improving each of them will lead to a better adoption of active and public transport. Yet, as the importance of features presented by different models varies widely, the results are not likely to be robust. To verify the reliability of the model results, further research should clarify the influence of the travel model decision through questionnaires or interviews.
5. To achieve greater sustainability in school transportation, it is important to have more schools located closer to sustainable transport modes (Bicycle, Bus, Train, Tram and Metro) or, conversely, to locate sustainable transport modes closer to schools. Additionally, it is crucial to understand the neighbourhood (e.g., socio-economic status) where the school is to be situated.
6. If, on the one hand, current educational policies for allocating students to schools are primarily based on the distance from home to school (Legislative Order n. 6/2018), on the other hand, this study shows that maximizing ACS and the use of public transport in student mobility in urban areas requires a better alignment of this educational policy with urban planning and public transport network design. To this end, appropriate sustainable transport policies should consider the geographical, urban, and infrastructural characteristics of the places in which schools are located and students' travel routes (Pantelaki et al., 2024). This alignment can also be achieved by combining transportation and education policies, with the definition of alternative conditions for assigning students to schools beyond just distance, and considering local contexts such as the built environment, socio-economic status and the transportation network.

## 8. Limitations and future research

Results suggest some future studies to better understand travel mode choice behaviour among students in Lisbon municipality. Although the 'Hands Up' Survey is an excellent starting instrument for studying school commuting modal choice due to its ability to provide a high amount of information regarding the chosen modes for so many students and covering all 24 parishes in Lisbon, the reality is that, when compared with other studies on this topic, the information obtained from this survey is quite limited, if one wishes to explore these behaviours at a deeper level rather than relying solely on descriptive statistics. This study acknowledges these limitations and attempts to include important variables, such as the Socio-economic status and safety level around the schools, in other ways. Moreover, extending the application of ML to this multi-class classification problem may require further analysis, namely on the use of the SMOTE resampling technique through sensitivity analysis.

It would have been beneficial to include individual-level features, or at the family-level or household-level, rather than at the school-level. For instance, it could include information regarding sex, household income, vehicles ownership, distance from home to school, factors leading to the choice of school regime (private or public), parent opinions, the effect of subsidies on public transport for students, and the built environment around the residence/home in their surveys. Furthermore, it would be interesting for them to include additional options, such as mixed modes and carpooling, which would provide more insight into school commuting modes and include trips from school back to home. This approach would make the information more reliable, eliminating the need to perform Factor Analysis using additional data for the socio-economic variable and reducing the necessity of examining each school with Google Street View for the safety level variable; this information could instead be provided by the parents. However, we recognize that a longer and more detailed survey would be more difficult, or even impossible to implement with such a large sample. Therefore, as part of further research, a careful selection of schools has already been made to conduct a more in-depth survey of students' parents, including these additional variables to better understand school commuting in Lisbon. This will allow for the verification of the reliability of the model results in this case study.

Additionally, another future study that is being drawn with a qualitative approach involves personally interviewing students and parents to gain a deeper understanding of the reasons behind their commuting decisions in Lisbon. Exploring the impact of any new variables that might emerge on the performance of our ML algorithms will be valuable, as identifying the feature importance of these variables will be crucial for promoting ACS in Lisbon Municipality. Finally, generalizability of the findings to other case studies or contexts can use the features explored so far, but may require adding other features that may describe the specific contexts within such case studies/cities/regions.

## Legislation

Law n. 46/86 published on the 14th of October 1986.
Law n. 85/2009 published on the 27th of August 2009.
Legislative Order n. 6/2018 published on the 12th of April 2018.

## CRediT authorship contribution statement

**Vivek Bhosale:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Miguel San Payo:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gabriel Cipriano:** Writing – review & editing, Validation, Methodology, Investigation, Data curation, Conceptualization. **António R. Andrade:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

## Funding

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work is supported by the Foundation for Science and Technology (FCT), through IDMEC, under LAETA, project UIDB/5022/2020; and through CIES, project UIDB/03126/2020. The authors express their gratitude to experts from Lisbon Municipality and persons involved in the Lisbon 'Hands Up' survey.

## References

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 11, 4484. https://doi.org/10.3390/su11164484.

Barranco-Ruiz, Y., Guevara-Paz, A.X., Ramírez-Velez, R., Chillón, P., Villa-González, E., 2018. Mode of commuting to school and its association with physical activity and sedentary habits in young ecuadorian students. Int. J. Environ. Res. Public Health 15 (12), 2704. https://doi.org/10.3390/ijerph15122704.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newslett. 6 (1), 20–29.

Ben-Akiva, M.E., Lerman, S.R, 1985. Discrete choice analysis: theory and application to travel demand (Vol. 9). MIT press.

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. In: Handbook of Transportation Science. Springer, US, Boston, MA, pp. 5–33.

Benita, F., Bansal, G., Piliouras, G., Tunçer, B., 2023. Understanding short-distance travel to school in Singapore: A data-driven approach. Travel Behav. Soc. 31, 349–362.

Biagioni, J.P., Szczurek, P.M., Nelson, P.C., Mohammadian, A., 2008. Tour-Based Mode Choice Modeling: Using an Ensemble of (un-) Conditional Data-Mining Classifiers. Transportation Research Board, Washington, DC.

Black, C., Collins, A., Snell, M., 2001. Encouraging walking: The case of journey-to-school trips in compact urban areas. Urban Stud. 38 (7), 1121–1141. https://doi.org/10.1080/00420980124102.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, H., Cheng, Y., 2023. Travel mode choice prediction using imbalanced machine learning. IEEE Trans. Intell. Transp. Syst. 24 (4), 3795–3808. https://doi.org/10.1109/TITS.2023.3237681.

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 Behav. Soc. 14, 1–10. https://doi.org/10.1016/j.tbs.2018.09.002.

Chillon, P., Ortega, F.B., Ruiz, J.R., Veidebaum, T., Oja, L., Mäestu, J., Sjostrom, M., 2010. Active commuting to school in children and adolescents: An opportunity to increase physical activity and fitness. Scand. J. Public Health 38, 873–879. https://doi.org/10.1177/1403494810384427.

CML (2016) II Diagnóstico Social de Lisboa 2015-2016: Sinopse [II Social Diagnosis of Lisbon 2015-2016: Synopsis]. Rede Social de Lisboa last access in June 7, 2023 through www.am-lisboa.pt/documentos/1518709936A8sST5fr2Qg86FJ5.pdf.

Cordovil, R., Lopes, F., Neto, C., 2015. Children's (in)dependent mobility in Portugal. J. Sci. Med. Sport 18 (3), 299–303. https://doi.org/10.1016/j.jsams.2014.04.013.

Corral-Abós, A., Aibar, A., Estrada-Tenorio, S., Julián, J.A., Ibor, E., Zaragoza, J., 2021. Implications of school type for active commuting to school in primary education students. Travel Behav. Soc. 24, 143–151.

Dias, C., Abdullah, M., Lovreglio, R., Sachchithanantham, S., Rekatheeban, M., Sathyaprasad, I.M.S., 2022. Exploring home-to-school trip mode choices in Kandy, Sri Lanka. J. Transp. Geogr. 99. https://doi.org/10.1016/j.jtrangeo.2022.103279.

Domingos, P., 1999. Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155–164.

Ehteshamrad, S., Saffarzadeh, M., Mamdoohi, A., Nordfjærn, T., 2022. Behavior of parents and children in the way they use public transport: A case study of Iranian households. Case Studies Transp. Policy 10 (2), 454–462. https://doi.org/10.1016/j.cstp.2022.01.006.

Ermagun, A., Rashidi, T.H., Lari, Z.A., 2015. Mode choice for school trips. Transp. Res. Record: J. Transp. Res. Board 2513 (1), 97–105. https://doi.org/10.3141/2513-12.

Etaati, B., Jahangiri, A., Fernandez, G., Tsou, M.-H., Ghanipoor Machiani, S., 2024. Understanding active transportation to school behavior in socioeconomically disadvantaged communities: A machine learning and SHAP analysis approach. Sustainability 16, 48. https://doi.org/10.3390/su16010048.

Fernández, A., García, S., Herrera, F., 2011. Addressing the classification with imbalanced data: open problems and new challenges on class distribution. Hybrid Artificial Intelligent Systems: 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011. Proceedings, Part I 6, 1–10.

Ferri-García, R., Fernández-Luna, J.M., Rodríguez-López, C., Chillón, P., 2020. Data mining techniques to analyze the factors influencing active commuting to school. Int. J. Sustain. Transp. 14 (4), 308–323. https://doi.org/10.1080/15568318.2018.1547465.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 1189–1232.

Gálvez-Fernández, P., Chillón, P., Timperio, A., Herrador-Colmenero, M., 2023. The patterns of active commuting to school change along the school-life in spanish youth. Int. J. Sustain. Transp. 17 (9), 982–989. https://doi.org/10.1080/15568318.2022.2133651.

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. Knowl.-Based Syst. 218, 106882. https://doi.org/10.1016/j.knosys.2021.106882.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 78, 273–282. https://doi.org/10.1016/j.eswa.2017.01.057.

CML (2021) Mãos ao ar! Lisboa [Hands up! Lisbon]. Câmara Municipal de Lisboa last access in June 7, 2023 through www.lisboa.pt/cidade/mobilidade/maos-ao-ar-lisboa.

Hastie, T., Friedman, J., Tibshirani, R. (2001). Boosting and Additive Trees (pp. 299–345). https://doi.org/10.1007/978-0-387-21606-5_10.

Helbich, M., 2017. Children's school commuting in the Netherlands: Does it matter how urban form is incorporated in mode choice models? Int. J. Sustain. Transp. 11 (7), 507–517. https://doi.org/10.1080/15568318.2016.1275892.

Hillel, T., Bierlaire, M., Elshafie, M.Z.E.B., Jin, Y., 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. J. Choice Modell. 38, 100221. https://doi.org/10.1016/j.jocm.2020.100221.

Hillel, T., Elshafie, M.Z.E.B., Jin, Y., 2018. Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. Proc. Inst. Civil Eng. - Smart Infrastruct. Constr. 171 (1), 29–42. https://doi.org/10.1680/jsmic.17.00018.

Huertas-Delgado, F.J., Herrador-Colmenero, M., Villa-González, E., Aranda-Balboa, M.J., Cáceres, M.V., Mandic, S., Chillón, P., 2017. Parental perceptions of barriers to active commuting to school in Spanish children and adolescents. Eur. J. Pub. Health 27 (3), 416–421. https://doi.org/10.1093/eurpub/ckw249.

Humberto, M., Moura, F., Giannotti, M., 2022. Incorporating children's views and perceptions about urban mobility: Implementation of the "philosophy with children" inquiry approach with young children. Travel Behav. Soc. 26, 168–177.

Irawan, M.Z., Belgiawan, P.F., Joewono, T.B., 2022. Investigating the effects of individual attitudes and social norms on students' intention to use motorcycles–An integrated choice and latent variable model. Travel Behav. Soc. 28, 50–58.

Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans. Intell. Transp. Syst. 16 (5), 2406–2417. https://doi.org/10.1109/TITS.2015.2405759.

Japkowicz, N., 2013. In: Assessment Metrics for Imbalanced Learning. In Imbalanced Learning. John Wiley & Sons, Inc, pp. 187–206.

Jesus, G.M., de Oliveira Araujo, R.H., Dias, L.A., Barros, A.K.C., dos Santos Araujo, L.D. M., de Assis, M.A.A., 2021. Influence of active commuting to school on daily physical activity among children and adolescents. J. Transp. Health 21, 101071.

Kim, E.-J., 2021. Analysis of travel mode choice in Seoul using an interpretable machine learning approach. J. Adv. Transp. 2021, 1–13. https://doi.org/10.1155/2021/6685004.

Kolidakis, S., Kotoula, K., Botzoris, G., Kamberi, P., Skoutas, D., 2024. Assessing impact factors that affect school mobility utilizing a machine learning approach. Sustainability 16, 588. https://doi.org/10.3390/su16020588.

Kwak, C., Clayton-Matthews, A., 2002. Multinomial_logistic_regression. Nurs. Res. 51 (6) https://journals.lww.com/nursingresearchonline/Fulltext/2002/11000/Multinomial_Logistic_Regression. 9.aspx.

Ling, C.X., Yang, Q., Wang, J., Zhang, S., 2004. Decision trees with minimal costs. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 69.

Lopes, F., Cordovil, R., Neto, C., 2014. Children's independent mobility in Portugal: Effects of urbanization degree and motorized modes of travel. J. Transp. Geogr. https://doi.org/10.1016/j.jtrangeo.2014.10.002.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30.

Mammen, G., Stone, M.R., Faulkner, G., Ramanathan, S., Buliung, R., O'Brien, C., Kennedy, J., 2014. Active school travel: an evaluation of the Canadian school travel planning intervention. Prev. Med. 60, 55–59.

Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: A systems of provision approach. Energy Res. Soc. Sci. 66, 101486. https://doi.org/10.1016/j.erss.2020.101486.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Mindell, J.S., Ergler, C., Hopkins, D., Mandic, S., 2021. Taking the bus? Barriers and facilitators for adolescent use of public buses to school. Travel Behav. Soc. 22, 48–58.

Nanthawong, S., Banyong, C., Janhuaton, T., Wisutwattanasak, P., Champahom, T., Ratanavaraha, V., Jomnonkwao, S., 2024. Exploring parental decision-making in school commutes: A structural equation model of public transport utilization and child safety in Thailand. Case Studies Transp. Policy 18, 101275. https://doi.org/10.1016/j.cstp.2024.101275.

Palm, M., Farber, S., 2020. The role of public transit in school choice and after-school activity participation among Toronto high school students. Travel Behav. Soc. 19, 219–230.

Palma, X., Chillón, P., Rodríguez-Rodríguez, F., Barranco-Ruiz, Y., Huertas-Delgado, F.J., 2020. Perceived parental barriers towards active commuting to school in Chilean children and adolescents of Valparaíso. Int. J. Sustain. Transp. 14 (7), 525–532. https://doi.org/10.1080/15568318.2019.1578840.

Pantelaki, E., Caspani, A.C., Maggi, E., 2024. Impact of home-school commuting mode choice on carbon footprint and sustainable transport policy scenarios. Case Studies Transp. Policy 15, 101110. https://doi.org/10.1016/j.cstp.2023.101110.

Pinto, A.A., Claumann, G.S., Angelo, H.C.C., Menezes, E.C., Dias, D.T., Pelegrini, A., 2017. Active commuting to school and associated factors among adolescents: A systematic review. J. Phys. Educ. 28. https://doi.org/10.4025/jphyseduc.v28i1.2859.

Pirra, M., Diana, M., 2019. A study of tour-based mode choice based on a support Vector Machine classifier. Transp. Plan. Technol. 42 (1), 23–36. https://doi.org/10.1080/03081060.2018.1541280.

Pizarro, A.N., Schipperijn, J., Andersen, H.B., Ribeiro, J.C., Mota, J., Santos, M.P., 2016. Active commuting to school in Portuguese adolescents: Using PALMS to detect trips. J. Transp. Health 3 (3), 297–304.

T. Qin Qin, T. (2020). Machine Learning Basics. In *Dual Learning* (pp. 11–23). Springer Singapore. https://doi.org/10.1007/978-981-15-8884-6_2.

Saris, W.H.M., Blair, S.N., van Baak, M.A., Eaton, S.B., Davies, P.S.W., Di Pietro, L., Fogelholm, M., Rissanen, A., Schoeller, D., Swinburn, B., Tremblay, A., Westerterp, K.R., Wyatt, H., 2003. How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement. Obes. Rev. 4 (2), 101–114. https://doi.org/10.1046/j.1467-789X.2003.00101.x.

Sekhar Minal, C.R., Madhu, E., 2016. Mode choice analysis using random forrest decision trees. Transp. Res. Procedia 17, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119.

Schoeppe, S., Duncan, M.J., Badland, H., Oliver, M., Curtis, C., 2013. Associations of children's independent mobility and active travel with physical activity, sedentary behaviour and weight status: A systematic review. J. Sci. Med. Sport 16 (4), 312–319. https://doi.org/10.1016/j.jsams.2012.11.001.

Shukla, N., Ma, J., Wickramasuriya, R., Huynh, N.N., Perez, P. (2015). Tour-based travel mode choice estimation based on data mining and fuzzy techniques.

Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 40 (12), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009.

United Nations (UN). (2022). The Sustainable Development Goals Report 2022. New York.

Vitale, M., Millward, H., Spinney, J., 2019. School siting and mode choices for school travel: Rural–urban contrasts in Halifax, Nova Scotia Canada. Case Stud. Transp. Policy 7 (1), 64–72.

Wang, F., Ross, C.L., 2018. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transport. Res. Record: J. Transp. Res. Board 2672 (47), 35–45. https://doi.org/10.1177/0361198118773556.

Wang, S., Mo, B., Zhao, J., 2020a. Predicting travel mode choice with 86 machine learning classifiers: an empirical benchmark study. Proc. 99th Annu. Meeting Transp. Res. Board, 279–296.

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. Preprint ArXiv:2102.01130 ArXiv.

Wang, S., Mo, B., Zhao, J., 2020b. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. Transp. Res. Part C Emerging Technol. 112, 234–251. https://doi.org/10.1016/j.trc.2020.01.012.

Wang, S., Wang, Q., Zhao, J., 2020c. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transp. Res. Part C Emerging Technol. 118, 102701. https://doi.org/10.1016/j.trc.2020.102701.

Waygood, E.O.D., Friman, M., Taniguchi, A., Olsson, L.E., 2019. Children's life satisfaction and travel satisfaction: Evidence from Canada, Japan, and Sweden. Travel Behav. Soc. 16, 214–223.

Wit, B., Loman, K., Faithfull, K., Hinckson, E.A., 2012. Reliability and validity of the hands-up survey in assessing commuting to school in New Zealand elementary school children. Health Promot. Pract. 13 (3), 349–354.

Xian-Yu, J.-C., 2011. Travel mode choice analysis using support vector machines. In: In ICCTP 2011: towards Sustainable Transportation Systems, pp. 360–371.

Xiao, Z., Lin, T., Liao, J., Lin, Y., 2021. School travel inequity between students from public and private schools in the city of Shenzhen China. J. Adv. Transp. 2021 (1), 5032726.

Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. Transp. Res. Rec. 2076 (1), 141–150. https://doi.org/10.3141/2076-16.

Zhou, X., Wang, M., Li, D., 2019. Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. J. Transp. Geogr. 79, 102479. https://doi.org/10.1016/j.jtrangeo.2019.102479.