Exploring the enhancement of predictive accuracy for minority classes in travel mode choice models.

Master thesis submitted to Delft University of Technology
in partial fulfillment of the requirements for the degree of
Master of Science in
Engineering and Policy Analysis
Faculty of Technology, Policy and Management

Author: Aspasia Panagiotidou

Academic Year: 2023-2024

Chair & First Supervisor: Dr.ir. Sander van Cranenburgh

Second Supervisor: Dr. Trivik Verma

External Supervisor: Dr.ir. Kingsley Adjenughwure

Advisor: Ir. Gabriel Nova

Acknowledgements

I would like to express my gratitude to everyone who contributed to this project. First of all, I would like to convey my special gratitude to my external supervisor from TNO, Dr.ir. Kingsley Adjenughwure, for all his guidance, the support and the time he devoted throughout this project. His contribution was truly invaluable. Secondly, I would like to express my gratitude to my first supervisor, Dr.ir. Sander van Cranenburgh for entrusting me with this project. I would like to thank him for his guidance as well as for bringing me in contact with TNO and assisting me with my internship. Next, I would like to express my gratitude to my second supervisor Dr. Trivik Verma, for agreeing to be part of my committee. I would also like to thank him for his guidance and for warmly welcoming me to CUSP. Lastly, I would like to express my gratitude to my advisor from TU Delft, Ir. Gabriel Nova for all his help and the time he dedicated. His expertise in the DCM models has been invaluable to me. Furthermore, I want to express my gratitude to all colleagues from SUMS for the inspiring discussions and their assistance on this project. Finally, none of this would have been possible without the unwavering support from the most important people in my life. I feel truly blessed to have them by my side.

Executive summary

Transportation systems are pivotal in shaping the economic and social dynamics of contemporary societies, fostering connectivity and opportunities while reducing geographical distances. Despite these benefits, they also contribute to adverse effects such as emissions, congestion, and traffic fatalities. Effectively developing and maintaining transportation infrastructure and services that cater to evolving population needs and align with environmental goals requires accurate forecasting of travel demand. However, due to inherent uncertainty in individuals' behavior and data limitations, forecasting this demand is a complex task.

A common limitation often encountered in transport datasets is class imbalance, with regard to the utilization of the different modes. Class imbalance in this context refers to the uneven distribution of samples among the various modes. Modes with a higher number of samples are termed majority modes, while those with fewer instances are labeled as minority modes. The existence of class imbalance within the dataset has the potential to compromise the performance of classifiers, especially for the minority modes, leading to inaccurate forecasts. This, in turn, may result in insufficient investments and provisions for these modes, ultimately having adverse consequences for the population segments that rely on them. Existing studies in the literature have either entirely overlooked or only partially addressed the impact of class imbalance. Recognizing the significance of precise demand predictions and acknowledging the identified gaps within the literature, the primary research question of this study was formulated as follows:

"How can the impact of class imbalance in model performance be systematically identified and addressed in transport mode share forecasting?''

To address the main question, a framework was proposed. This framework encompassed various aspects including a) the measurement of class imbalance within a dataset and the assessment of its impact on classification performance, b) the investigation of other challenging factors coexisting in imbalanced datasets, with a specific focus on class overlap, and c) the proper evaluation of classification performance across classes. As an integral part of this framework, the 'Performance Gap Metric' was introduced - a metric employed to evaluate the difference in classification performance between the majority and minority classes. Establishing a threshold of 20%, favorable classifier performance was determined when this metric fell below the threshold, signifying the classifier's equitable treatment of both minority and majority classes. Subsequently, this framework was applied using the ODiN data as a case study to predict mode choices in the Netherlands. Mode choices encompassed car, bike, and transit, with car representing the majority and transit the minority class. Two modeling techniques, namely Random Forest and an MNL model, were employed in conjunction with various sampling techniques, including the SMOTENC, the Neighborhood-based Undersampling, and the Separation scheme.

The key findings of this study can be summarized as follows: Both models were impacted by the presence of class imbalance and class overlap in the dataset, exhibiting a reduced performance on the minority class. Multiclass scenarios proved to be more complex compared to binary scenarios. Among the employed sampling techniques, SMOTENC and Neighborhood-based Undersampling demonstrated superior performance in binary and multiclass scenarios, respectively, when employing the Random Forest model. This resulted in the Performance Gap Metric falling below or nearly reaching the predetermined threshold in each case. Conversely, when employing the MNL model, SMOTENC emerged as the best-performing technique, achieving a Performance Gap Metric below the predetermined threshold in both scenarios. Furthermore, the heightened sensitivity of the minority class after the implementation of sampling techniques was consistently accompanied by a decrease in

its precision, indicating a trade-off between the two metrics. Concurrently, as the sensitivity of the minority class increased, in most scenarios, the sensitivity of the other classes and sometimes the overall accuracy also decreased, suggesting that achieving fairness might necessitate compromises.

Moreover, this study provided practical recommendations concerning all aspects addressed within the proposed framework, while it also suggested considerations beyond the accurate prediction of travel demand that transport planners and policymakers should take into account to ensure an inclusive transportation system. Finally, the primary limitations of the study were also acknowledged. These included the exclusive assessment of the Performance Gap Metric based on the majority (Car) and minority (Transit) classes, overlooking the classifier's performance on the Bike class in multiclass classification tasks, as well as the confined testing of the proposed framework in only one specific case. As recommendations for future research, the consideration of all classes when evaluating the classifiers' performance and an extended validation of the proposed framework were proposed. Furthermore, exploring more advanced techniques, such as the use of generative models to augment the minority classes and the implementation of feature engineering techniques to enhance class separability, were also suggested.

Contents

CHAPTER 1: Introduction 11

Chapter 2: Literature Review & Theoretical Background 15

2.3 Class imbalance in the Transport domain 15

2.1. Class imbalance in classification tasks 17

2.1.2 Techniques to address class imbalance 18

2.2. Class overlap in classification tasks and techniques to address it 20

2.4 Theoretical background 21

2.4.1 Random Forest model – Model description 21

2.4.1.1 Random Forest model – Feature Importance 23

2.4.1.2 Random Forest model – Hyperparameter tuning 23

2.4.1.3 Random Forest model – Testing 23

2.4.2 Multinomial Logit Model (MNL) – Model description & specification 24

2.4.2.1 Multinomial Logit Model – Cross Validation 25

2.4.3 Sampling Techniques 25

2.4.3.1 SMOTENC 26

2.4.3.2 Neighborhood based under-sampling 27

2.4.3.3 Separation between the overlapping and non-overlapping regions 28

CHAPTER 3 : Proposed Framework 30

CHAPTER 4: Data 35

4.1 Literature review - Determinants of travel mode choices 35

4.2 Dataset Description 37

4.2.1 Descriptive Analysis of the dataset 39

4.2.2 Descriptive Analysis of the dataset 41

4.3 Data Preprocessing Steps 42

4.3.1 Feature selection 43

4.3.2 Calculation of trip costs 44

4.3.3 Data Assumptions 45

CHAPTER 5: Application of the proposed framework 46

5.1 Results 49

5.1.1 Random Forest – Binary & Multiclass classification 50

5.1.1.1 Binary classification 50

5.1.1.2 Multiclass classification 54

5.1.2 Multinomial Logit Model – Binary & Multiclass classification 57

5.1.2.1 Binary classification 58

5.1.2.2 Multiclass classification 60

5.1.2.3 MNL - Interpretability 63
Chapter 6: Discussions & Conclusions 65
6.1 Main findings 66
6.2 Practical recommendations 68
6.3 Main contributions 68
6.4 Limitations & Ideas for future research 69
References 71
Appendix 77

Abbreviations

Full name	Abbreviation
Centraal Bureau voor de Statistiek (Statistics Netherlands)	CBS
Cost Benefit Analysis	CBA
Discrete Choice Model	DCM
Imbalance Ratio	IR
Machine Learning	ML
Multinomial Logit Model	MNL
Neighborhood-based Undersampling technique	NBU
Onderweg in Nederland	ODiN
Onderzoek Verplaatsingen in Nederland	OViN
Open Trip Planner	OTP
Random Forest	RF
Random Utility Maximization	RUM
Revealed Preference	RP
Stated Preference	SP
Synthetic Minority Over-Sampling Technique	SMOTE
Synthetic Minority Over-Sampling Technique Nominal Continuous	SMOTENC
t-Distributed Stochastic Neighbor Embedding	t-SNE
Value of Time	VOT

List of Tables

Table 1. Summary of the most recent research studies on mode choice forecasting employing imbalanced datasets. 17
Table 2. Overview of the models and sampling techniques employed in this study. Both the SMOTENC and Neighborhood-based Undersampling techniques are employed to both models, whereas the Separation scheme is exclusively applied to the Random Forest model. It is important to emphasize that no comparative analysis between the two models is conducted in this study. Instead, the classification performance is assessed individually for each model before and after the implementation of the sampling techniques. 21
Table 3. Creation of new instances employing the SMOTE algorithm. 26
Table 4. The Euclidean distance computed within the SMOTENC framework. SMOTENC, a variant of the traditional SMOTE technique, distinguishes itself by altering the calculation of the Euclidean distance between minority samples and their k-nearest neighbors compared to the conventional approach. 26
Table 5. Pseudocode of the Neighborhood-based Undersampling approach in binary classification. .. 27
Table 6. Pseudocode of the separation scheme approach for binary classification. 28
Table 7. Pseudocode of the Separation scheme approach for multiclass classification. 29
Table 8. Confusion matrix for binary classification. TP represents the number of correctly classified positive examples, FN denotes the number of positive examples misclassified as negative, TN signifies the number of correctly classified negative examples, and lastly, FP represents the number of negative examples misclassified as positive. The attribution of the terms "positive" and "negative" to the minority or majority classes is contingent upon researchers' choice. 33
Table 9. Evaluation metrics for classification tasks. Accuracy and Error rate are commonly utilized metrics for evaluating the overall performance. Recall, Precision and F-score are often employed to assess performance independently for each class. 34
Table 10. List of factors influencing mode choices. 35
Table 11. Description of the dataset utilized in the application of this study. The descriptions of the variables correspond to the post-processing stage. 41

Table 12. Descriptive analysis of the dataset utilized in the this study after the pre-processing stage. For numerical variables, mean and standard deviation values are provided for both the datasets utilized in the Random Forest model (left value) and the MNL model (right value). 42
Table 13. Summary of the data assumptions considered in this study. 45
Table 14. Summary of the limitations of the data utilized in this study. 45
Table 15. Imbalance Ratio of the datasets utilized in each model within this study. 47
Table 16. List of libraries and packages employed for classification in the application of this study. 48
Table 17. Summary of the sampling techniques utilized in the application of this study. 49
Table 18. Summary of the sampling techniques utilized in this study. 50
Table 19. Imbalance Ratio and Performance Gap Metric in the context of binary classification using the Random Forest model. The Imbalance Ratio evaluates the difference in the number of samples between the majority and minority classes, whereas the Performance Gap Metric quantifies the discrepancy in their respective classification performance. 51
Table 20. Summary of the results of binary classification using the Random Forest model. The reported values represent the mean performance of the model across five runs, with standard deviation values indicated in parentheses. The values of the Performance Gap Metric are highlighted in the cases where its value surpasses the 20% threshold, indicating the ability of the classifier to predict equally well the majority and the minority classes. 54
Table 21. Imbalance Ratios and Performance Gap Metric in the context of multiclass classification using the Random Forest model. In this case, the Imbalance Ratio was computed for all pairs of classes, signifying the difference in their number of samples, while the Performance Gap Metric was utilized to quantify the disparity in the classification performance between the majority and minority classes. 54
Table 22. Summary of the results of multiclass classification using the Random Forest model. The reported values represent the mean performance of the model across five runs, with standard deviation values indicated in parentheses. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, nearly attaining the 20% threshold. 57
Table 23. Imbalance Ratio and Performance Gap Metric for the case of binary classification employing the MNL model. The Imbalance Ratio evaluates the difference in the number of samples between the majority and minority classes, whereas the Performance Gap Metric quantifies the discrepancy in their respective classification performances. 58
Table 24. Summary of binary classification results obtained with the MNL model. The reported values represent the mean performance of the model following a 3-fold cross-validation, with standard deviation values indicated in parentheses. VOTs are expressed in euros/h. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, falling below the 20% threshold and indicating the ability of the classifier to predict equally well the majority and the minority classes. 60
Table 25. Imbalance Ratios and Performance Gap Metric for the case of multiclass classification with MNL model. In this case, the Imbalance Ratio was computed for all pairs of classes, signifying the difference in their number of samples, while the Performance Gap Metric was utilized to quantify the disparity in the classification performance between the majority and minority classes. 61
Table 26. Summary of multiclass classification results obtained with the MNL model. The reported values represent the mean performance of the model following a 3-fold cross-validation, with standard deviation values indicated in parentheses. VOTs are expressed in euros/h. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, falling below the 20% threshold and indicating the ability of the classifier to predict equally well the majority and the minority classes. 63

Table 27. National averaged Values of Time with uncertainty bandwidths in the Netherlands, in € / hr (Significance, 2023). (Values are reported in euros in price level 2022). 64
Table 28. Value of Time for the Car and Transit alternatives according to the baseline scenarios (before the implementation of sampling techniques) of this study. 64
Table 29. Sensitivity outcomes pertaining to the prediction of transit trips across Dutch provinces are presented. From left to right, the results are displayed for the baseline model (using imbalanced data), the model after implementing SMOTENC to address "between-class" imbalance, the model after implementing SMOTENC to address both "between-class" and "within-class" imbalance, and lastly, the model after implementing Random Oversampling to tackle both types of imbalance. 80
Table 30. Sensitivity outcomes concerning the prediction of transit trips in the Randstad and non-Randstad regions. Results are displayed for the baseline model (using imbalanced data), the model after implementing SMOTENC to address "between-class" imbalance, the model after implementing SMOTENC to address both "between-class" and "within-class" imbalance, and lastly, the model after implementing Random Oversampling to tackle both types of imbalance. 81
Table 31. Top 10 most important features identified in the binary and multiclass classification tasks employing the Random Forest model. The significance of each feature is assessed based on the average reduction in impurity across all splits within the forest where the feature is employed. 82
Table 32. Results from the binary classification task using the Random Forest model. The top row of the table showcases classification results with utilizing all explanatory features, while the second row illustrates classification outcomes considering only the top 10 most important features. 82
Table 33. Results from the multiclass classification task employing the Random Forest model. The first row of the table illustrates outcomes when employing all explanatory features, while the second row showcases results when only the top 10 most important features are considered. 83
Table 34. t values for the binary classification task employing the MNL model. For a degree of freedom (df) equal to 2 and a 95% confidence level, the critical t values are ± 2.92. 83
Table 35. t values for the multiclass classification task employing the MNL model. For a degree of freedom (df) equal to 2 and a 95% confidence level, the critical t values are ± 2.92. 83

List of Figures

Figure 1. Flowchart of the present research study. 14
Figure 2. Overview of strategies utilized for handling imbalanced data. Techniques employed in the existing literature to mitigate the impact of class imbalance are categorized into data-level, algorithmic-level and combination-level approaches. 19
Figure 3. The structure of a decision tree is organized as follows: the topmost node, known as the root node, represents the entire dataset. Moving down the tree, the internal nodes, often called split nodes, make decisions based on specific features, effectively partitioning the dataset into subsets. These nodes are linked by branches, representing the outcomes of the decisions, which can be either True or False. Furthermore, the nodes maintain a hierarchical relationship: the starting point of a branch, referred to as the parent node, is connected to child nodes, symbolizing the subsequent decision paths. Ultimately, the terminal nodes at the end of the branches are known as leaf nodes, each denoting a distinct class label. 23
Figure 4. Illustration of the random forest method. Random Forest is an ensemble learning approach comprising multiple decision trees. Each tree is constructed using the bootstrapping technique and incorporates random feature selection. After training the ensemble, the method consolidates the outcomes from each individual estimator. In classification tasks, this consolidation involves considering the majority vote among the trees 24

Figure 5. Illustration of the 3 fold cross-validation process. In this study, cross-validation is employed for the internal validation of the MNL model. Within each of the three iterations, 2/3 of the dataset is utilized for model estimation, and the remaining 1/3 is reserved for testing. 25
Figure 6. Illustration of the generation of synthetic data through the implementation of SMOTE technique considering the 5 nearest neighbors. 27
Figure 7. Illustration of testing the Separation scheme. 29
Figure 8. Proposed framework comprising several steps that can be integrated in classification tasks with imbalanced datasets. 30
Figure 9. Mode preferences in the Netherlands using ODiN data (2018-2019), highlighting an inherent imbalance in the utilization of the different modes. Note that in the classification task of this study only the Car (driver), Bike and Transit modes are considered. 38
Figure 10. Maps of the Netherlands depicting the distribution of transit and car trips across Dutch provinces. Notably, the majority of transit trips occur in the Randstad area, encompassing the provinces of South-Holland, North-Holland, Utrecht and Flevoland. 38
Figure 11. Map of the Netherlands depicting the distribution of bike trips across Dutch provinces.... 39
Figure 12. Integration of the proposed framework in the context of forecasting travel mode choices in the Netherlands. 46
Figure 13. T-SNE plot for the binary classification between the Car and Transit classes, highlighting both overlapping and non-overlapping regions within the dataset. T-SNE is a dimensionality reduction technique that improves the visualization of multidimensional data by projecting it into a lower-dimensional space. In this case, the data is visualized in a 2D space. In the above plot, two distinct regions are emphasized: the overlapping, encompassing both transit and car samples, and the non-overlapping, comprising exclusively car samples. 51
Figure 14. T-SNE plot for multiclass classification involving the Car, Transit, and Bike classes. T-SNE is a dimensionality reduction technique that improves the visualization of multidimensional data by projecting it into a lower-dimensional space. In this case, the data is visualized in a 2D space. In the context of multiclass classification, overlapping regions are defined as those occupied by the minority class samples and their nearest neighbors from the Car and Bike classes, while non-overlapping regions are considered those occupied solely by samples from the Car and Bike classes. 55
Figure 15. Approach adopted to tackle both ''between-class'' and ''within-class'' imbalance, following the methodology proposed by Japkowicz et al. (2001). The term ''between-class'' imbalance denotes the existence of classes with varying numbers of instances within a dataset. In contrast, “within-class” imbalance indicates the presence of sub-clusters with varying numbers of instances within a specific class. 78
Figure 16. Map of the Netherlands indicating the proportion of transit journeys per province based on the trips' origins. The Netherlands comprises 12 provinces, including South Holland, North Holland, Utrecht, Zeeland, Flevoland, North Brabant, Limburg, Overijssel, Drenthe, Friesland, and Gelderland. Among these provinces, South-Holland stands as the most populous, while Zeeland is recognized as the least populous Dutch province. 79
Figure 17. Map of the Netherlands depicting the percentage of public transport journeys originating in both the Randstad and non-Randstad regions. The Randstad area encompasses four Dutch provinces: South Holland, North Holland, Utrecht, and Flevoland. 81

CHAPTER 1: Introduction

Transport systems hold great significance as they serve as the backbone of economic and social activities, bringing people together, reducing distances, and connecting individuals to a multitude of opportunities, encompassing education, employment and leisure activities. Conversely transportation can also have external effects such as congestion, carbon emissions, noise and traffic fatalities, which negatively impact human well-being (Mackett & Thoreau, 2015).

To create and uphold transportation infrastructure and services that align with both the travel demand of the population and environmental and other sustainability goals, precise assessment of travel demand stands as a paramount necessity for urban planners and transportation authorities. Travel demand has a substantial influence on resource allocation for infrastructure investments and the prioritization of transportation policies (H. Chen & Cheng, 2023). Nevertheless, forecasting this demand is a challenging task. The complexity emerges from the inherent uncertainty in individuals' behavior as they adapt their preferences based on a multitude of factors, limiting modelers from accurately representing the actual decision-making processes individuals undergo when selecting transportation modes. Furthermore, challenges in data quality, such as incomplete data or biases introduced by sampling methods or survey designs focused on collecting pertinent information, contribute to the overall complexity.

A common limitation observed in transportation datasets is the occurrence of class imbalance, especially concerning the distribution of users across different travel modes (Qian et al., 2021; Hillel et al., 2021; H. Chen & Cheng, 2023). Class imbalance, in this context, refers to the uneven distribution of target classes within the dataset (Fernández et al., 2018), where certain modes, termed majority classes, have a significantly larger number of samples compared to others, termed minority classes. Using imbalanced data to predict travel demand can result in inaccurate forecasts, particularly resulting in a diminished number of correctly classified minority samples, as classifiers often demonstrate suboptimal performance on minority classes (Johnson & Khoshgoftaar, 2019). The underestimation of travel demand for minority modes can lead to diminished attention and support for these modes in terms of transportation planning and resource allocation.

In the context of the conventional approach to evaluating transportation projects, an accurate determination of modal shares, especially for less commonly used modes (minority modes), is of paramount importance. Traditionally, the primary method for evaluating transportation projects in most western countries has been the use of Cost-Benefit Analysis (CBA). CBA has its roots in utilitarian theory, which places paramount importance on achieving the greatest good for the largest number of people (Van Wee & Roeser, 2013). Allocating resources to transport infrastructure, driven by higher user demand and potential overall benefits, leads to overlooking the spatial, social, and economic background of various groups within this demand (Pereira et al., 2016; Jafino et al., 2021). This approach can exacerbate disparities among disadvantaged populations, which within the literature, are defined by factors such as income level, gender, age, and health status (Hananel & Berechman, 2016).

For instance, designing transportation systems for sparsely populated rural areas is often seen as economically challenging, leading to a prioritization of investments in urban areas. Due to their larger populations and economic importance, urban regions often occupy a higher position in the political agendas and decision-making processes related to transport, while rural regions frequently experience exclusion (Flipo et al., 2023) Many individuals facing restricted transportation options in outlying areas belong to the low-income demographic, often forced to relocate from city centers due to soaring housing costs. This group encompasses single-parent families, migrant families, newly established

households (e.g., first-time homebuyers), and elderly, some of whom may also lack access to a car or have given up driving (Stanley & Stanley, 2017). Conversely, higher-income residents also inhabit suburban areas, but their residential choices are primarily driven by preference rather than necessity Scott & Horner, 2008; Van Wee and Geurs, 2011).

In the Netherlands, public transport represents a minority mode, serving only 9% of the population (CBS, 2019). Public transport systems, encompassing trains, buses, trams, and metros, play a vital role in reducing carbon emissions, alleviating traffic congestion, and ensuring essential accessibility, especially for population groups such as low-income individuals, the elderly, and students who heavily rely on transit for their daily travel needs. Underestimating the demand for public transport can lead to insufficient investments in these systems and a lack of prioritization in improving service quality, frequencies, and accessibility where it is most needed. Restricting access to essential services and opportunities for specific subgroups may ultimately result in their isolation and social exclusion.

Scientific Relevance

Different levels of transport service provision and transport policies often result in mode choice data exhibiting a significant degree of class imbalance (Hillel et al., 2021). Determining mode choices is primarily a classification task, which in presence of discrepancies in the number of samples among the target classes can become difficult (Qian et al., 2021). Despite being infrequent, minority classes can contain valuable information, often overshadowed by classifiers' focus on classes with larger sample sizes. Nevertheless, the challenge of class imbalance in the field of mode choice modeling has not yet been sufficiently addressed (H. Chen & Cheng, 2023).

Within the domain of travel-behavior research and mode share prediction, discrete choice models, particularly those belonging to the logit family, have traditionally been heavily relied upon. These models, known for their interpretable results, offer insights into the behavioral aspects influencing decision-making processes (Kashifi et al., 2022). The most widely used discrete choice model is the Multinomial Logit (MNL) model (McFadden, 1973), while other statistical models such as the nested logit and mixed logit models have been used as well. Occasionally, Machine Learning (ML) models are also applied due to their enhanced prediction capabilities and efficient handling of large datasets (García-García et al., 2022). Nevertheless, their adoption by choice modelers has been slower, as despite their superior predictive performance, ML models face criticism for operating as 'black boxes' and lacking the theoretical basis for understanding and interpreting human behavior (Brathwaite et al., 2017). Both model categories can be affected by the presence of class imbalance, exhibiting problems in accuracy performance (van Cranenburgh et al., 2022).

Various research studies in the domain of mode share forecasting have essentially neglected the issue of class imbalance (Omrani, 2015; Sekhar et al., 2016; Zhao et al., 2020). On the contrary, there are still some studies that have made efforts to address it by incorporating a range of methods derived from the field of machine learning, which has extensively examined class imbalance over the past two decades (Johnson & Khoshgoftaar, 2019). Hagenauer and Helbich (2017) leveraged oversampling and under-sampling techniques to balance their dataset and improve the prediction accuracy of the minority modes. However, their analysis focused solely on classification results following the implementation of these techniques, without offering a comparison of classifier performance before data balancing. Qian et al. (2021) introduced a novel Support Vector Machine (SVM) model that significantly enhanced the accuracy of minority modes. However, despite its effectiveness, this method is tailored to their specific model and cannot be applied across various classifiers. Kim (2021) took a different approach, increasing the visibility of minority class instances by assigning weights inversely proportional to the frequency

distribution of each class during the training of the models. Nevertheless, their approach did not prove successful, as the performance of the minority class remained poor. Finally, other studies have primarily focused on assessing the overall performance of models without considering metrics that evaluate the performance independently on each class, such as recall, f1-score and others (García-García et al., 2022). Even in cases where class specific evaluation metrics were considered and a thorough comparison of results before and after the implementation of balancing techniques was presented, as observed in the study by H. Chen & Cheng (2023), a comprehensive and generic framework for addressing the impact of class imbalance was absent.

Considering the above, it becomes evident that existing research studies in mode choice forecasting either neglect or only partially address the presence of class imbalance in the data. Given that, the main gap in the current literature is pinpointed in the lack of a systematic and structured approach that researchers in the field could adopt to effectively manage case studies involving imbalanced datasets.

Objective

So far, it has been clear that class imbalance can negatively impact the performance of classifiers, and a diminished predictive performance for minority modes can have adverse effects on future transport provisions aiming to create an inclusive transportation system. Recognizing the importance of attaining consistently precise predictions across various travel modes and taking into account the primary gap in the current body of literature, our principal aim in this research is to introduce a comprehensive framework for systematically identifying and mitigating the implications of class imbalance on classifier performance. Specifically, through this framework, we aim to provide guidelines on: a) quantifying class imbalance within a dataset, b) investigating potential adverse effects on the classification performance of minority modes, c) exploring other challenging factors co-existing in imbalanced datasets, with a particular emphasis on class overlap, and d) evaluating the classification performance across classes both before and after implementing techniques aimed at addressing performance degradation caused by imbalanced datasets. Furthermore, to showcase the practical application of this framework, we plan to implement it in forecasting mode share in the Netherlands, utilizing the 'Onderweg in Nederland' (ODiN) data for the years 2018-2019 as a case study.

The main research question of this study is formulated as follows:

"How can the impact of class imbalance in model performance be systematically identified and addressed in transport mode share forecasting?''

To be able to answer the main research question the following sub-questions will be also addressed:

1. What are the existing techniques for identifying and addressing the impact of class imbalance in model performance in different contexts?

2. How can these techniques be incorporated within a comprehensive framework aiming in systematically addressing class imbalance?

3. How can the application of these techniques be extended across various types of transport mode share forecasting models, encompassing both traditional utility models and machine learning models?

This research study will be conducted within the Sustainable Urban Mobility and Safety department (SUMS) at TNO. Recently, SUMS has shifted its focus towards leveraging the capabilities of machine learning algorithms in the field of travel mode choice modeling. Acknowledging the common occurrence of class imbalance in transportation datasets and its impact on the predictive performance of minority modes, the department is dedicated to thoroughly investigating and addressing the potential implications of class imbalance on the performance of models forecasting mode share.

So far, the problem of this research study has been introduced. The subsequent structure of this document is as follows: Chapter 2 provides an in-depth literature review, highlighting the identified knowledge gaps and establishing the theoretical background. Chapter 3 outlines the proposed framework, constituting the methodology of this study. Moving forward, Chapter 4 offers a comprehensive description of the data used in this study. Chapter 5 presents the application of the proposed methodology as well as the results of our analysis. Finally, Chapter 6 serves as the conclusion, summarizing the main findings, highlighting limitations and proposing ideas for future research (Figure 1).

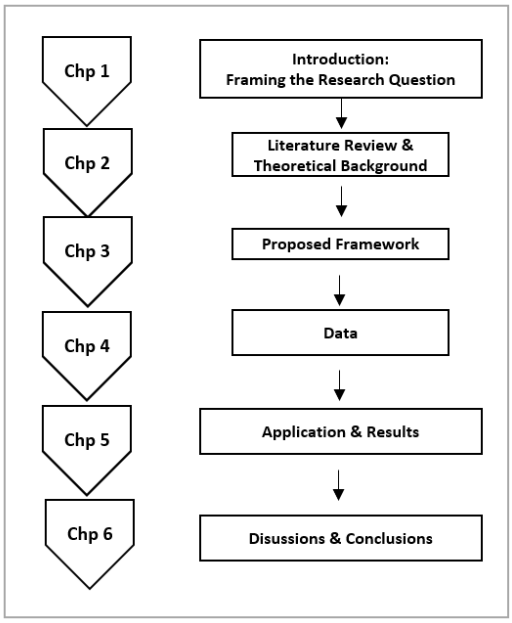

Figure 1. *Flowchart of the present research study.*

Chapter 2: Literature Review & Theoretical Background

Following the introduction of the primary research question and sub-questions in this study, this chapter reveals our findings from a comprehensive review of the existing literature, aimed at identifying gaps related to the treatment of class imbalance within recent mode choice studies. Additionally, it provides readers with the theoretical background necessary for a thorough understanding of the key concepts in this study. Specifically, it delves into defining class imbalance and class overlap, elucidating existing techniques to address them. Subsequently, it provides insights into the models and sampling techniques employed in this study.

2.3 Class imbalance in the Transport domain

In Chapter 1, we briefly outlined the gaps related to addressing class imbalance in studies focusing on mode choices. This section delves deeper, providing a more thorough presentation of these studies. Table 1 outlines key information for each study, including its objectives, the data and models employed and the metrics utilized to assess the classifiers' performance. Additionally, we highlight whether researchers have implemented techniques, specifying which ones, to address class imbalance.

Beyond addressing class imbalance, this study places emphasis on investigating and mitigating the potential existence of class overlap in the data. Class overlap is recognized as one of the factors that can hinder classification performance when learning from imbalanced datasets, often considered the most detrimental (Santos et al., 2023). Consequently, we also provide information on whether the examined studies have explored and/or mitigated the potential presence of class overlap.

Study	Purpose	Data	Models	Performance Evaluation Metrics	Addressing Class Imbalance	Techniques for addressing class imbalance	Investigation of class overlap	Technique for addressing class overlap
Omrani (2015)	Assessment of predictive performance across classifiers	Data from national travel survey in Luxemburg	Multinomial Logit Model, Multilayer Perceptron, Support Vector Machine, Radial Basis Function Network	Root mean square percentage error, Average probability of correct assessment	✗	–	✗	–
Sekhar et al. (2016)	Assessment of predictive performance across classifiers	Data gathered from household interviews in Delhi	Multinomial Logit Model, Random Forest	Overall Accuracy	✗	–	✗	–
Hagenauer and Helbich (2017)	Assessment of predictive performance across classifiers	Dutch travel diary data from the years 2010-2012	Multinomial Logit Model, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Classification Tree, BOOST, BAG, Random Forest	Overall Accuracy, Recall (Sensitivity) per mode	✓	Combination of Random Oversampling and Random Undersampling	✗	–
Wang & Ross (2018)	Assessment of predictive performance across classifiers	Data from household travel survey in Delaware Valley Region (2012)	Multinomial Logit Model, Extreme Gradient Boosting Model	Overall Accuracy, Prediction error per mode	✗	–	✗	–

Zhao et al. (2020)	Assessment of predictive performance among classifiers	Stated preference data from a survey from the University of Michigan	Multinomial Logit Model, Mixed Logit Model, Naïve Bayes, CART, BAG, Random Forest, Support Vector Machine, Neural Network	Overall Accuracy, Accuracy per mode, L1-norm	✗	–	✗	–
Qian et al. (2021)	Assessment of predictive performance among classifiers	Data from national household travel survey from California (2017)	Novel Support Vector Machine algorithm with adjusting kernel scaling technique, other SVM-based models, Artificial Neural Network, XGBoost, Bayesian Network	Overall and per mode Accuracy, Recall per mode, Precision per mode, F1-score per mode	✓	Novel Support Vector Machine algorithm with adjusting kernel scaling technique	✗	–
Kim (2021)	Assessment of predictive performance among classifiers	Data from national household travel survey from Seoul (2016)	Artificial Neural Network, XGBoost, Random Forest	Recall per mode, precision per mode, Balanced accuracy	✓	Training the models by assigning weights inversely proportional to the frequency distribution of each class	✗	–
Rezaei et al. (2021)	Assessment of predictive performance and interpretability of mode choice models under balanced and imbalanced data conditions	Trip data from the city of Nashville	Multinomial Logit Model, Nested Logit Model, Mixed Logit Model	Overall accuracy, Recall per model, Mean absolute percentage error, Signs and magnitude of beta coefficients	✓	Combination of Random Undersampling and Random Oversampling	✗	–
Kashifi et al. (2022)	Assessment of predictive performance across various classifiers	Dutch National travel survey data from the years 2010-2020	Logistic regression, Decision Tree, Random Forest, Multilayer Perceptron, Light Gradient Boosting Decision Tree	Precision per mode, Recall per mode, F1-score per mode, Overall accuracy, Average precision	✓	Random Oversampling, Random Undersampling	✗	–
Chaipanha & Kaewwichian (2022)	Assessment of the predictive performance across classifiers	Trip data from a Thai study and interviews (2015)	k-Nearest Neighbor, Decision Tree, Naïve Bayes	Accuracy, True positive rate, False positive rate, Macro F1-score	✓	SMOTE, Random Undersampling	✗	–
García-García et al. (2022)	Assessment of the predictive performance across classifiers	a) Revealed Preferences dataset from Switzerland (2009-2010) b) Dutch national travel survey data (2010-2012)	Multinomial Logit Model, Multilayer Perceptron, Deep Neural Network, Support Vector Machine, Random Forest	Overall Accuracy	✓	Combination of Random Oversampling and Random Undersampling (balancing techniques were implemented only in the 2nd dataset)	✗	–

H. Chen & Cheng (2023)	Assessment of the predictive performance across classifiers	Data from household travel survey in London (April 2012 – March 2015)	Multinomial Logit Model, XGBoost, Deep Neural Network	F1-score per mode, Macro F1-score, Overall Accuracy, MADMS (Mean Absolute Deviation of Market Share), Economic interpretation through elasticities	✓	SMOTE, ADASYN, One-Sided Selection, Neighborhood Cleansing Rule, Random Under-sampling, Random Oversampling	✗	–
Narayanan & Antoniou (2023)	Estimation of choice model for shared mobility services	Data from household travel survey from Madrid (February 2018-June 2018)	Multinomial Logit Model	–	✓	Combination of SMOTE and Random Undersampling	✗	–

Table 1. Summary of the most recent research studies on mode choice forecasting employing imbalanced datasets.

Identified gaps

Upon examining the research studies outlined in the above table, several gaps in the literature emerge:

- In some studies, the presence of class imbalance in the dataset is entirely neglected.
- Certain studies rely solely on overall accuracy for performance evaluation, neglecting to assess the models' performance on individual classes.
- Certain studies propose alternatives to sampling techniques including the development of novel models explicitly designed to handle imbalanced data or the assignment of higher weights to the minority class during the classifiers' training. While these approaches may have proven successful in enhancing accuracy for the minority class in specific studies, it is crucial to acknowledge that they are tailored to particular models and may lack universal applicability.
- No research study investigates and/or addresses the potential presence of class overlap in the data.
- Notably, there is an absence of studies presenting a comprehensive framework for addressing the impact of class imbalance in classification performance.

2.1. Class imbalance in classification tasks

Classification involves assigning a label (or class) to an observation based on its distinctive features. Classification tasks are commonly divided into two main groups: binary and multiclass. In binary classification, the goal is to differentiate instances between two classes, whereas multiclass classification involves assigning instances to one of three or more classes (Ali et al., 2019). In the realm of transportation planning, a prime illustration of classification involves forecasting travel mode choices. Mode choice prediction constitutes the third stage in the traditional four-step travel demand model, encompassing trip generation, trip distribution, mode choice prediction, and trip assignment. The prediction of transportation mode choices involves estimating the likelihood that a traveler will opt for a specific mode based on diverse factors, such as individual characteristics, preferences, trip attributes,

and built environment features (Omrani et al., 2015; Rezaei et al., 2021). Biases stemming from sampling methods and survey designs aimed at collecting relevant data, as well as variations in the number of users across different transportation modes often result in travel mode choice datasets characterized by ‘’class imbalance”.

Class imbalance is a common issue, observed in diverse domains such as fraud detection, disease diagnosis, and image recognition, which refers to disparities in class representation within a dataset. Classes with fewer samples compared to others are referred to as 'minority classes', while those with more samples are referred to as 'majority classes'. Identifying majority and minority classes in a dataset can be accomplished through various methods. The most commonly used metric for measuring class imbalance is the Imbalance Ratio (IR), defined as the ratio of the number of majority samples to the number of minority samples (Zhu et al., 2020). Class imbalance may manifest to varying degrees, resulting in datasets that are either slightly or highly imbalanced, even though no official thresholds exist for categorizing a dataset into one of these two categories.

Classifiers tend to exhibit superior predictive performance for majority classes compared to minority classes. Despite having fewer samples, minority classes may encompass crucial information that is of great interest to analysts and may be also associated with higher misclassification costs. Consequently, insufficient identification of these classes can result in adverse effects, the nature of which vary across applications (Johnson & Khoshgoftaar, 2019). In the context of predicting travel mode choices, imbalanced datasets can lead to models favoring overrepresented modes. Relying on the outcomes of such models may result in suboptimal and inequitable policies, as insufficient provisions and resource allocation for minority modes fail to address the diverse mobility preferences of the population.

Class imbalance is not a new challenge. In the field of machine learning it has been studied since the last two decades (Johnson & Khoshgoftaar, 2019). According to He & Garcia (2009), class imbalance can be characterized as either ‘intrinsic’ or ‘extrinsic’. Intrinsic imbalance arises from naturally occurring skewed data distributions, while extrinsic imbalance is introduced by external factors, such as data collection processes and privacy issues. While it is often presumed that class imbalance is the main factor contributing to the deterioration in classifier performance, research studies have unveiled that the decline in performance when learning from imbalanced datasets is also affected by other factors. Among the data characteristics that compound the complexity of classification tasks, class overlap is identified as the most detrimental. Both earlier and recent research studies affirm that the performance of learning algorithms diminishes across different levels of class overlap, whereas class imbalance does not consistently have a significant impact (Prati et al., 2004; S V. García et al., 2007 ; Santos et al., 2023).

2.1.2 Techniques to address class imbalance

In this section, we introduce the techniques found in the existing literature that are utilized to address class imbalance. Addressing class imbalance is commonly categorized into solutions at the algorithmic level and data level. Furthermore, hybrid methods result from combining these approaches (Napierala et al., 2010; Johnson & Koshgoftaar, 2019).

At the algorithmic level, addressing class imbalance entails either developing new algorithms specifically designed to handle imbalanced data or modifying the costs assigned by existing algorithms to different classes. Cost-modification entails assigning higher misclassification costs to the minority class during classifiers' training compared to those assigned to the majority class. This elevation of costs for the minority class enhances its importance, reducing the likelihood of the classifier making incorrect classifications for instances belonging to this class (Napierala et al., 2010; Krawczyk, 2016). However, a drawback of this technique is that misclassification costs are not always known (Elrahman & Abraham, 2013).

At the data level, solutions revolve around altering the composition of the dataset itself through the application of sampling techniques. These techniques can be broadly categorized into two types: undersampling techniques and oversampling techniques. Simple methods include Random Under-sampling and Random Oversampling (H. Chen & Cheng, 2023). Random Under-sampling involves balancing classes by randomly removing instances from the majority class. While this method is straightforward to implement, it has the downside of potentially discarding valuable examples, leading to the loss of crucial information. This problem becomes more evident when the class imbalance ratio is high, as it may lead to the removal of a significant amount of data. In contrast, Random Oversampling addresses class imbalance by randomly duplicating instances from the minority class. Like Random Under-Sampling, this technique is also easy to implement; however, the replication of samples might result in overfitting (H. Chen & Cheng, 2023).

In response to the limitations arising from the use of Random Oversampling with replacement and to enhance the classifier's generalization on testing data, Chawla et al. (2002) introduced an advanced approach for augmenting the minority class through the generation of synthetic samples. Their method, known as Synthetic Minority Oversampling Technique (SMOTE), is one of the most widely used sampling algorithms in machine learning (García et al., 2016). Based on SMOTE, new examples are created along the line segments connecting minority class samples to their k-nearest neighbors from the minority class. In specific, synthetic samples are generated by calculating the difference between the sample under consideration and its nearest neighbor. This difference is then multiplied by a random number between 0 and 1 and added to the corresponding feature vector. The selection of neighbors from the k nearest neighbors is done randomly, depending on the desired amount of oversampling. In their research study, Chawla et al. (2002) conducted a comparison between the implementation of random oversampling and oversampling through the creation of synthetic instances in a binary classification task. Their findings indicated that the latter yielded superior results.

Following the development of SMOTE, various adaptations emerged, emphasizing targeted oversampling rather than random oversampling and targeted undersampling rather than random undersampling (Han et al., 2005; He et al., 2008; Johnson & Khoshgoftaar, 2019). Additionally, generative models, from the field of machine learning, have also been leveraged for data augmentation showing promising results (Rezaei et al., 2021; Salas et al., 2023).

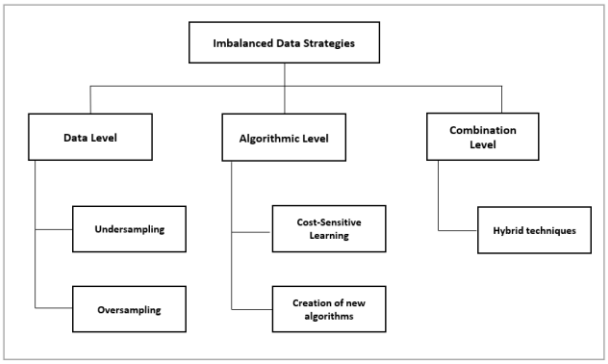

Figure 2. Overview of strategies utilized for handling imbalanced data. Techniques employed in the existing literature to mitigate the impact of class imbalance are categorized into data-level, algorithmic-level and combination-level approaches.

2.2. Class overlap in classification tasks and techniques to address it

As it has been previously mentioned, one of the factors that can impede classification performance when learning from imbalanced datasets is class overlap. In classification tasks, class overlap arises when instances from different classes share similar or identical feature values, leading to ambiguity for the classifier in determining the correct class label. Class overlap is a prevalent challenge encountered in various real-world problems and similar to class imbalance it has been a subject of study in the machine learning community over the last two decades (Trappenberg & Back, 2000; Alogogianni & Virvou et al., 2023).

Prati et al. (2004) were among the first to explore the relationship between class imbalance and class overlap. To determine whether class imbalance singularly contributed to the decline in classifiers' performance or whether other factors were also involved, they conducted a systematic study utilizing binary artificial datasets with varying levels of class imbalance and class overlap. Their findings revealed a strong correlation between class imbalance and class overlap. As the extent of overlap increased the classifier experienced a decline in performance, more pronounced in highly imbalanced datasets. On the contrary, in cases where classes were distinct with minimal overlap, classifier performance showed less dependence on prior probabilities, suggesting that the models' performance was mostly influenced when both class imbalance and class overlap were concurrently present.

Garcia et al. (2007) also explored the relationship between class imbalance and class overlap, and their implications on classification performance. In their study, they applied the Nearest Neighbor algorithm, utilizing binary artificial datasets with a consistent imbalance ratio and varying degrees of class overlap. According to their results, in the absence of overlap, both classes demonstrated comparable performance. However, as overlap increased, their accuracies declined, and the disparity between them widened. In line with observations by Prati et al. (2004), their findings suggested that classifier performance is not affected by class imbalance alone; rather, the influence of the latter becomes more prominent with the escalation of class overlap.

Additionally, other research studies focused on exploring ways to address classification challenges stemming from overlapping data. Trappenberg and Back (2000) proposed a classification scheme, that focused on avoiding predictions in data regions considered highly ambiguous. Their approach comprised two key steps: Initially, a k-nearest neighbor algorithm was applied to reclassify the data samples. If the majority of a sample's k-nearest neighbors belonged to a specific class, that sample was assigned to that class. On the contrary, in the absence of clear majority, the sample was categorized into a new class denoted as "IDK" (I Don't Know). Subsequently, after reclassifying all the samples in the training set, a classifier was trained on the re-labeled data to perform the final classification task. The potential outcomes included the initial classes along with the addition of the "IDK" class. This proposed scheme could be seamlessly integrated with any classifier, and during testing on two real-world datasets, no misclassifications were reported.

Similarly, Xiong et al. (2013) conducted a thorough investigation into addressing class overlap through three distinct strategies: the discarding, the merging, and the separating scheme. Their study incorporated five real-world datasets with varying overlap ratios, and each scheme underwent testing across five classifiers. In the discarding scheme, data within the overlapping region were completely disregarded and the classifiers were exclusively trained on non-overlapping data. The merging scheme, consisting of two models, categorized data from the overlapping region into a new class labeled 'overlapping' using the first model, while the second model was then applied to learn the samples within this new class. Finally, in the separating scheme two models were used, one for the overlapping region and one for the non-overlapping region, respectively. The F1-score for both classes was used to evaluate the performance across the different classifiers. The separating scheme demonstrated superior

performance and was further tested in artificial datasets characterized by both class overlap and class imbalance. Their findings revealed that as the degree of class imbalance increased, the classification performance improved when implementing the separating scheme compared to the baseline scenario where no scheme was applied.

In recent studies, Vuttipittayamongkol and Elyan (2020) introduced an under-sampling framework aiming in identifying and removing majority class instances from the overlapping region. The framework featured four k-NN based methods, each exploring the local surroundings of individual instances and identifying overlapped instances for elimination based on distinct criteria. Extensively tested on both real and artificial datasets, all four methods exhibited superior performance in terms of the sensitivity of the minority class compared to widely used state-of-the-art methods.

2.4 Theoretical background

So far, we have emphasized the identified gaps in the existing literature and provided definitions for the concepts of class imbalance and class overlap. Furthermore, we have introduced various techniques within the literature aimed at addressing their impact. Before delving into our proposed methodology, this section imparts essential information on foundational concepts crucial for readers to comprehend the remainder of this study. Specifically, we delve into the two models -Random Forest and Multinomial Logit model- employed in this study. Furthermore, we provide details about the sampling techniques applied to these models, namely the SMOTENC, the Neighborhood-based Undersampling, and the Separation scheme. Table 2 outlines the techniques utilized in conjunction with each model. It is important to note that this study does not involve a comparative analysis between the two models. Instead, we individually assess the classification performance of each model both before and after implementing the sampling techniques.

	Machine Learning modeling	Discrete choice modeling
	Random Forest	MNL
SMOTENC	✓	✓
Neighborhood-based Undersampling	✓	✓
Separation scheme	✓	✗

Table 2. Overview of the models and sampling techniques employed in this study. Both the SMOTENC and Neighborhood-based Undersampling techniques are employed to both models, whereas the Separation scheme is exclusively applied to the Random Forest model. It is important to emphasize that no comparative analysis between the two models is conducted in this study. Instead, the classification performance is assessed individually for each model before and after the implementation of the sampling techniques.

2.4.1 Random Forest model – Model description

The Random Forest model is an ensemble supervised machine learning algorithm that comprises multiple tree estimators, and is employed for both regression and classification tasks. To promote diversity among the single trees and mitigate their susceptibility to overfitting, Random Forest combines bootstrapping and random feature selection. Bootstrapping is a statistical technique that involves randomly sampling with replacement from a set of observed values. Based on this technique, each tree is constructed using a distinct training set of the same size. Because the sampling is done with replacement, some observations may be duplicated, while others may be omitted. Furthermore, the random feature selection entails considering only a random subset of variables to split each node, rather than using all the explanatory variables (Breiman, 2001). These two layers of randomness contribute to minimizing errors stemming from biased or noisy samples and consequently reducing prediction variance (Cheng et al., 2019).

In the Random Forest model, each tree estimator divides the data into mutually exclusive regions based on the explanatory variables of the dataset, grouping together samples with similar target values. Specifically, as mentioned earlier, at each node, a randomly selected subset of the input features is considered. Using a specified criterion, the best splitting point for each feature is determined, and ultimately, the optimal pair of splitting variable and splitting point is selected to partition the data at each node (Cheng et al., 2019). This process continues until the tree is fully grown or until a constraint is met. Constraints may include reaching the maximum depth of the tree or having fewer than the minimum required samples in one or both branches of a node after a split. Once all trees are constructed, the predicted class for each input sample is determined through majority voting. In other words, the class that gains the highest number of 'votes' from the individual trees is selected as the final prediction. An illustration of the Random Forest model is presented in Figure 4.

While various criteria are available to evaluate the quality of each split and determine the optimal choice, in this research study we employed the Gini Index. This index is used to calculate the impurity of a node after utilizing each splitting feature. Impurity, in this context, indicates the homogeneity of the class labels at a node. A zero value indicates a pure node with samples exclusively from the same class. Conversely, a value of 0.5 (in binary tasks) denotes maximum impurity, signifying an equal distribution among classes. A feature is deemed optimal when a split based on it yields the lowest impurity.

The mathematical calculation of impurity based on the Gini criterion is detailed below, where G represents the impurity at a node m based on split θ, Q_m represents the data at node m, Q_m^{left} and Q_m^{right} are the data in the left and right branches of the node respectively, $n_m$is the total number of samples at node m, $n_m^{left}$and n_m^{right} represent the number of samples at the left and right branches, θ represents each candidate split, and H() denotes the impurity criterion, which in our case is the Gini Index.

In specific, for each candidate split θ = (j, t_m) consisting of a feature j and a threshold t_m the data is partitioned into the $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets, where x is a training vector ($x_i \in R^n, i = 1, \dots, I$) and y a label vector ($y \in R^l$). Next, the quality of the split is calculated based on (2.5).

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\} \quad (2.1)$$

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta) \quad (2.2)$$

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) \quad (2.3)$$

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \quad (2.4)$$

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)) \quad (2.5)$$

The criterion to select a feature for a split is the minimization of impurity. Consequently, the feature which satisfies (6) is finally selected.

$$\theta^* = argmin_\theta G(Q_{m,}\theta) \quad (2.6)$$

2.4.1.1 Random Forest model – Feature Importance

The Gini criterion (equation. 2.4) can be also employed to identify the most influential variables among the considered explanatory features in predicting the target variable. The importance of each feature in a single tree is calculated as the sum of the impurity reduction over all nodes where it was used for splitting. The overall importance of a feature in the forest is then defined as the average of its importance values across all trees (Cheng et al., 2019).

2.4.1.2 Random Forest model – Hyperparameter tuning

Prior to deploying the Random Forest Classifier, we fine-tuned its hyperparameters. This tuning process was essential, as it facilitated the identification of specific model parameters tailored to the unique characteristics of our dataset, with the goal of achieving optimal performance. The parameters subjected to tuning included:

- Number of estimators (the quantity of the trees in the forest).
- Max depth (the maximum depth of individual trees).
- Minimum number of samples in the leaf nodes (the minimum number of samples that should be left in both the left and right branches for a split to be considered).
- Max samples (the number of samples considered for training each tree).
- Max features (the number of features to be drawn for training each tree).

To determine the ideal values for these parameters, we defined a range of potential values for each and systematically trained the model for every possible combination. Subsequently, we evaluated the model's performance on a validation set, which constituted 10% of the dataset. The values of the parameters that yielded the highest accuracy were selected. This iterative procedure was replicated each time a model was trained on a new dataset. Another option could have been to use a k-fold cross-validation method.

2.4.1.3 Random Forest model – Testing

The results obtained with the Random Forest model represent the average performance on the test set across five model runs. This approach was employed to account for the inherent model's stochastic nature, which introduces randomness via the implementation of bootstrapping and random feature selection techniques.

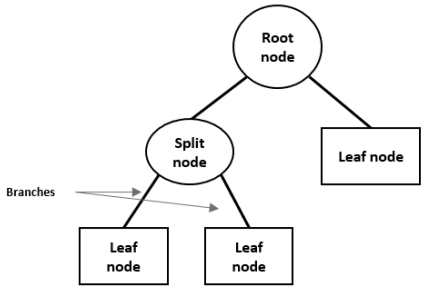

Figure 3. The structure of a decision tree is organized as follows: the topmost node, known as the root node, represents the entire dataset. Moving down the tree, the internal nodes, often called split nodes, make decisions based on specific features, effectively partitioning the dataset into subsets. These nodes are linked by branches, representing the outcomes of the

decisions, which can be either True or False. Furthermore, the nodes maintain a hierarchical relationship: the starting point of a branch, referred to as the parent node, is connected to child nodes, symbolizing the subsequent decision paths. Ultimately, the terminal nodes at the end of the branches are known as leaf nodes, each denoting a distinct class label.

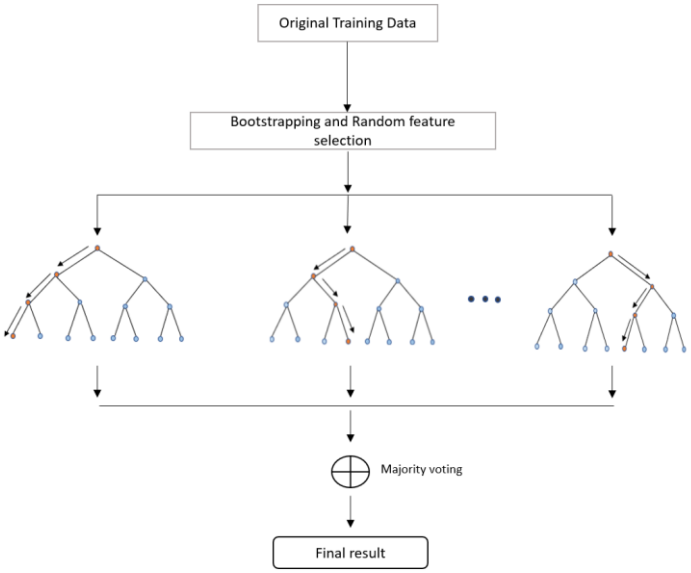

Figure 4. Illustration of the random forest method. Random Forest is an ensemble learning approach comprising multiple decision trees. Each tree is constructed using the bootstrapping technique and incorporates random feature selection. After training the ensemble, the method consolidates the outcomes from each individual estimator. In classification tasks, this consolidation involves considering the majority vote among the trees

2.4.2 Multinomial Logit Model (MNL) – Model description & specification

Discrete choice models constitute a family of models capturing decision makers' choices among a set of alternatives, referred to as the choice set. Specifically, these models are typically formulated on the premise that decision makers select the alternative that maximizes their utility (Train, 2003).

In detail, these models propose that each alternative j provides the decision maker n, with a specific level of utility Unj, j = 1….J. The utility comprises two components: one encompasses the effects of the observed explanatory variables (including attributes of the alternatives, such as travel time and cost, as perceived by the decision maker, and attributes of the decision maker, such as income and age), and the other reflects the influence of the variables that the analyst cannot observe.
As a result the utility of mode j is expressed as follows:

$$U_{nj} = V_{nj} + \varepsilon_{nj} \qquad (2.7)$$

In equation (2.7), V_{nj} is the observed part of the utility, often called ''representative utility'' and ε_{nj} the unobserved part.

A decision maker chooses alternative i if and only if $Uni > Unj \; \forall \; j \neq$ i. The probability that the decision maker n will choose alternative i is defined as the probability that $Uni > Unj \; \forall \; j \neq i$. In order to calculate this probability, the distribution of the unobserved component (random error) εnj must be assumed by the researcher. By specifying distinct types of random errors different logit models arise. In the present study we specifically employed the Multinomial Logit Model (MNL), ,in which the assumption is made

that the unobserved εnj is independently and identically Gumbel-distributed (H. Chen & Cheng, 2023). The strength of this model lies in its explicit closed-form mathematical formulation (Kashifi et al., 2022).

In the MNL model, the probability that a decision maker n chooses mode i is given by the following formula:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \quad (2.8)$$

Model Specification

The utility function employed in this study for each alternative i is defined as follows.

$$V_i = ASC_i + B_{time_i} * time_i + B_{cost_i} * cost_i \quad (2.9)$$

In the above equation, " i " represents the i^{th} mode, while "$time_i$" and "$cost_i$" respectively signify the duration and expenses related to a specific trip undertaken using mode i. B_{time} and B_{cost} serve as the coefficients for the time and cost attributes, while ASC denotes the alternative specific coefficient. This coefficient corresponds to the mean of the error term and captures inherent preferences that are independent of specific attribute values, expressing user inclinations towards the alternatives.

2.4.2.1 Multinomial Logit Model – Cross Validation

For the MNL models utilized in this study, internal validation was carried out following the approach outlined by Parady et al. (2021). Specifically, to assess the models' ability to maintain predictive accuracy across different samples from the same population, a 3-fold cross-validation was implemented. In each iteration, 2/3 of the dataset was used for model estimation, with the remaining 1/3 reserved for testing. Given the dataset's imbalance, a stratified shuffling approach was employed to guarantee that the class distribution was preserved during the splitting process. Ultimately, the predictive performance of the model was evaluated based on its mean performance across all iterations.

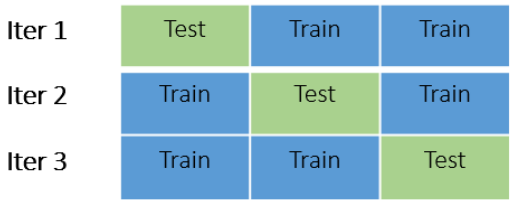

Figure 5. Illustration of the 3 fold cross-validation process. In this study, cross-validation is employed for the internal validation of the MNL model. Within each of the three iterations, 2/3 of the dataset is utilized for model estimation, and the remaining 1/3 is reserved for testing.

2.4.3 Sampling Techniques

In this section, we present a summary of the sampling techniques applied in this study. All three methods belong to the category of data-level approaches, aiming to modify the distribution of the

dataset used for training the classifiers. The key advantage of these techniques is their applicability, as they can be implemented irrespective of the underlying classifier.

2.4.3.1 SMOTENC

The first technique implemented in this study is the Synthetic Minority Over-Sampling Technique Nominal Continuous (SMOTENC). SMOTENC is an adaptation of the popular SMOTE technique, which was initially introduced by Chawla et al. in 2002. Unlike traditional oversampling with replacement, SMOTE addresses class imbalance by generating "synthetic" examples to augment the minority class. Synthetic examples are generated as follows: First, for each minority sample the k-nearest neighbors from the minority class are identified. Then, based on the desired number of new instances, one or more neighbors are randomly selected for each sample. New synthetic examples are generated by adding to each minority sample the product of the difference between itself and its neighbor, and a randomly chosen number ranging from 0 to 1. In case the desired quantity of new instances is smaller than the initial number of the samples in the minority class, random selection is used to choose the minority samples.

Example of creating new instances using SMOTE: Consider a minority sample (a,b) and let (c,b) to be its nearest neighbor. diff_1 = (a – c) diff_2 = (b – d) The new sample will be generated as : (x,y) = (a,b) + rand(0,1) * (diff_1, diff_2) , where rand(0,1) generates a random sample between 0 and 1.

Table 3. Creation of new instances employing the SMOTE algorithm.

SMOTENC is tailored to handle datasets that encompass both continuous and nominal features. Within the SMOTENC framework, the continuous features of the newly generated synthetic minority samples are constructed using the same approach that has been previously described. The key distinction lies in the calculation of the Euclidean distance between the minority samples and their neighbors. Specifically, when a considered minority sample differs in its nominal features from its nearest neighbor, the median of the standard deviations of all continuous features from the minority class is included in the calculation. An illustrative example of this computation is provided in Table 4. Furthermore, nominal features are assigned the value that is most frequently observed among the k-nearest neighbors. Note that in this study, we have utilized the SMOTENC algorithm from the open-source imblearn library. Also, the synthetic data has been created by only using the training data to avoid a potential 'data leakage' between the training and test sets. Finally, similar to the binary classification tasks, in the multiclass classification tasks, synthetic data have been solely created for the minority class.

Example of calculating the Euclidean distance within the SMOTENC framework: Consider a minority sample (continuous_feature_1, continuous_feature_2, nominal_feature_1, nominal_feature_2) and its nearest neighbor (continuous_feature_3, continuous_feature_4, nominal_feature_3, nominal_feature_4). Given that nominal_feature_1 ≠ nominal_feature_3 and nominal_feature_2 ≠ nominal_feature_4, the Euclidean distance is calculated as follows: Euclidean_distance = sqrt [$(\text{continuous\_feature\_3} - \text{continuous\_feature\_1})^2 +$ $(\text{continuous\_feature\_4} - \text{continuous\_feature\_2})^2 + Med^2 + Med^2$] , where Med is the median of the standard deviations of the continuous features of the minority class. The median term is included twice since two nominal features differ among the two samples.

Table 4. The Euclidean distance computed within the SMOTENC framework. SMOTENC, a variant of the traditional SMOTE technique, distinguishes itself by altering the calculation of the Euclidean distance between minority samples and their k-nearest neighbors compared to the conventional approach.

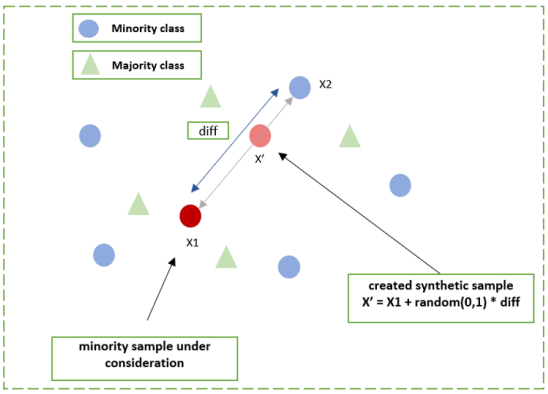

Figure 6. Illustration of the generation of synthetic data through the implementation of SMOTE technique considering the 5 nearest neighbors.

2.4.3.2 Neighborhood based under-sampling

The second technique implemented in this study is the Neighborhood-based Undersampling technique (NBU). The NBU technique was introduced by Vuttipittayamongkol and Elyan (2020). Its primary objective is to cleanse the overlapping area by eliminating any sample from the majority class that possesses at least one neighbor from the minority class. The elimination criterion is established at only one minority neighbor to ensure the visibility of all minority samples. This approach leads to a reduction in the number of majority samples, simultaneously enhancing the visibility of the minority samples. That way it effectively addresses both class imbalance and class overlap.

A brief description of the methodology is provided in the table below. Also, it is important to note that in the case of multiclass classification, samples were eliminated from all classes other than the minority class.

Neighborhood-based under-sampling:

T_{min} → training samples from the minority class
T_{maj} → training samples from the majority class

```
# Step 1: Identify the majority instances to be eliminated
eliminated_instances= []
for sample in T_maj:
   nearest_neighbors = find_nearest_neighbors(sample, T_maj + T_min)
   for neighbor in nearest_neighbors:
      if neighbor.class == minority class:
         add sample to eliminated_instances

# Step 2: Remove the samples that belong to the eliminated_instances from the training samples
T_maj = T_maj – eliminated_instaces
# Step 3: Run the model with the remaining training samples and classify them into the majority and minority classes respectively
model = train_model(T_maj + T_min, labels)
```

Table 5. Pseudocode of the Neighborhood-based Undersampling approach in binary classification.

2.4.3.3 Separation between the overlapping and non-overlapping regions

The third technique implemented in this study is the Separation scheme. The Separation scheme proposed by Xiong et al. (2010) was implemented in this study for both the binary and multiclass classification tasks using the Random Forest model. This scheme involves a neighborhood analysis of the data space, utilizing a geometrical distance (in this case, the Euclidean distance) to partition samples into two regions: the overlapping and non-overlapping regions. After defining these regions, a different model is applied within each region to perform the classification task.

In the binary case, the implementation of the Separation scheme utilized two distinct Random Forest (RF) models. The first RF model classified samples into either the overlapping or non-overlapping region. Samples categorized into the non-overlapping region were identified as belonging to the majority class, as all minority samples were grouped within the overlapping region. The samples categorized as part of the overlapping region were then fed into the second Random Forest model, which further classified them into either the minority or majority class.

A similar approach was employed for multiclass classification. In this scenario, three distinct classes were considered. Unlike binary classification, three RF models were utilized. The first RF classifier categorized samples based on whether they fell within the overlapping or non-overlapping region. Subsequently, within each of these regions, two distinct RF models were applied to precisely determine the class to which each sample belonged. In this case, the overlapping region consisted of samples belonging to minority class and their k-nearest nearest neighbors from the other two classes, while the non-overlapping region consisted solely of samples belonging to classes other than the minority class.

Table 6. Pseudocode of the separation scheme approach for binary classification.

Separation scheme approach for binary classification:

T_{min} → training samples from the minority class
T_{maj} → training samples from the majority class
$T_{overlap}$ → set of training samples belonging to the overlapping region
$T_{nonoverlap}$ → set of training samples belonging to the non-overlapping region

```
# Step 1: Define the overlapping region
overlapping_samples = []

for sample in T_min:
   add sample in T_overlap
   nearest_neighbors = find_nearest_neighbors(sample, T_maj + T_min)
   for neighbor in nearest_neighbors:
      if neighbor.class != minority_class:
         add neighbor to T_overlap

# Step 2: Train a RF model to classify samples in overlapping and non-overlapping regions
RF_model = train_RF_model (T_overlap + T_nonoverlap , labels)
# Step 3: Train a RF model to classify samples in the overlapping region into minority and majority classes
RF_model_overlap = train_RF_model (T_overlap , labels)
```

Separation scheme approach for multiclass classification:

T_{min} → training samples from the minority class
T_{class_1} → training samples from class_1 (class_1 != minority class)
T_{class_2} → training samples from class_2 (class_2 != minority class and class_1 != class_2)
$T_{overlap}$ → set of training samples belonging to the overlapping region
$T_{nonoverlap}$ → set of training samples belonging to the non-overlapping region

```
# Step 1: Define the overlapping region
overlapping_samples = []

for sample in T_min:
  add sample in T_overlap
  nearest_neighbors = find_nearest_neighbors(sample, T_class_1 + T_class_2 + T_min)
  for neighbor in nearest_neighbors:
      if neighbor.class != minority_class:
        add neighbor to T_overlap

# Step 2: Train a RF model to classify samples in overlapping and non-overlapping regions

RF_model = train_RF_model (T_overlap + T_nonoverlap , labels)

# Step 3: Train a RF model to classify samples in the non-overlapping region into class_1 and class_2
RF_model_overlap = train_RF_model (T_nonoverlap , labels)

# Step 4: Train a RF model to classify samples in the overlapping region into class_1, class_2 and minority class
RF_model_overlap = train_RF_model (T_overlap , labels)
```

Table 7. Pseudocode of the Separation scheme approach for multiclass classification.

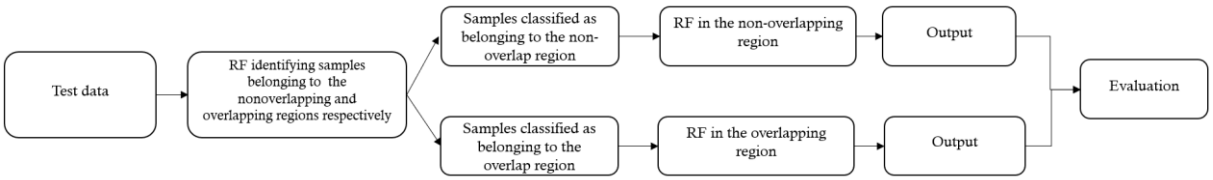

Figure 7. Illustration of testing the Separation scheme.

CHAPTER 3 : Proposed Framework

Building upon the literature review presented in the previous chapter, it becomes evident, based on the identified gaps, that class imbalance is a prevalent issue in classification tasks. Despite being extensively discussed in recent years, a comprehensive, unified framework to address this challenge remains absent (Erlahman & Abraham, 2013). Within the transportation domain, which is the focus of this research study, the problem of class imbalance is often either entirely overlooked or addressed only to a limited extent (H. Chen & Cheng, 2023). Furthermore, the presence of additional complexities in imbalanced datasets, such as class overlap—which is deemed particularly detrimental—is not receiving the necessary attention.

In light of these gaps, we move forward by introducing a framework -depicted in Figure 8-, which comprises several steps that could be integrated in classification tasks involving imbalanced datasets. The framework is model agnostic, meaning that it can be implemented irrespective of the underlying classifier. Our main objective with this approach is to bring researchers' attention to factors that may contribute to the deterioration of classification performance, with a particular focus on class imbalance and class overlap. Additionally, we aim to provide helpful guidelines that researchers can adopt when dealing with imbalanced datasets. These guidelines concern:

- The identification of the presence of class imbalance and/or other ''difficulty factors'' such class overlap within a dataset.
- The selection of right techniques to mitigate their effects.
- The selection of suitable evaluation metrics for model assessment.

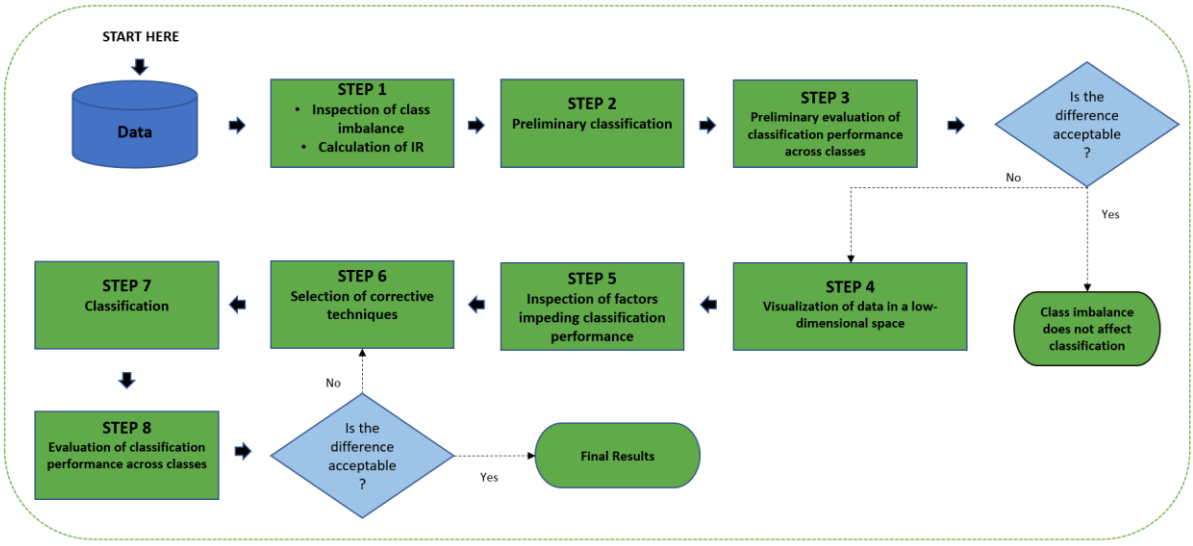

Figure 8. Proposed framework comprising several steps that can be integrated in classification tasks with imbalanced datasets.

In the rest of this section we delve into the details of the steps integrated in our proposed framework.

STEP 1 - Inspection of class imbalance & Calculation of the Imbalance ratio: As illustrated in Figure 8, the 1st step in addressing imbalanced datasets is to assess the degree of class imbalance. In specific, once the minority and majority classes are identified, class imbalance can be quantified. While there exist various methods to measure class imbalance, in the current framework we propose employing the

Imbalance Ratio (IR), as it is the most commonly used metric for characterizing the degree of imbalance in a dataset. Nevertheless, researchers can opt for an alternative approach should they find it more suitable for their specific application.

As depicted in equation (3.1), IR is calculated as the ratio of the number of majority samples to minority samples (Zhu et al., 2020). When IR = 1, then the dataset is perfectly balanced. Conversely, when IR is greater than 1, the higher its value, the greater the imbalance in the dataset.

$$\text{Imbalance Ratio (IR)} = \frac{N_{majority}}{N_{minority}} \qquad (3.1)$$

In the context of classification tasks involving more than two classes, $N_{majority}$ denotes the sample size of the largest class, while $N_{minority}$ signifies the sample size of the smallest class within the dataset (Zhu et al., 2020).

STEP 2 – Preliminary Classification: The 2nd step of the framework involves the preliminary execution of the classification task. Upon reviewing the existing literature, it has become apparent that the presence of class imbalance within a dataset may not always pose a problem (Prati et al., 2004). For instance, the presence of class imbalance seems to have little effect on linearly separable classification tasks (JapkowiczNathalie & StephenShaju, 2002). Therefore, we recommend conducting the classification task initially without applying any techniques to assess whether class imbalance truly hinders it or not.

One of the merits of our proposed framework lies in its adaptability, allowing its integration with any classification model. Instead of mandating expertise in specific modeling techniques, the choice of the classifier is left to the discretion of the researcher. A prime illustration of this is the application of the framework across two different model categories, as described later in this report, showcasing its effectiveness irrespective of the specific model employed.

STEP 3 – Preliminary evaluation of the classification performance across classes: The 3rd step of the framework involves the initial evaluation of the classifier's performance across classes. Comparing the evaluation metrics across different classes can provide insights into potential factors hindering classification performance. Although, defining precise thresholds for the difference among metric values is not an easy task, a notable performance gap should trigger a more thorough analysis of the dataset. Conversely, consistently high performance across all classes, implies that class imbalance does not impact the model's performance, and in that case no further step is required to address it.

To evaluate the performance gap among classes, drawing inspiration from the fairness metric introduced by Zheng et al. (2023), we propose the use of the metric outlined in equation 3.2. This metric quantifies the absolute percentage difference between the sensitivity of the minority and the majority classes. The metric's minimum value is zero, indicating that the model is equally adept at identifying instances of both classes. Conversely, the maximum value it can attain is 1. This situation occurs when the sensitivity for one class is at its maximum value (100%), while for the other class it is at its minimum value (0 %). This suggests that the model is highly effective in identifying instances of one class but falls short in recognizing instances from the other class.

Given the above, to ensure a fair model that identifies instances of both classes equally well, we want the value of the metric to be as close to zero as possible. As noted earlier, establishing precise thresholds is challenging; nevertheless, in assessing performance across classes in this study, a 20% threshold is applied to the metric. If the performance difference among classes falls below this threshold, the model's performance is considered favorable. Should the difference exceed this value, a set of corrective techniques is applied.

It is crucial to acknowledge that the predetermined threshold may vary in different applications. Its value is contingent on the importance researchers place on achieving equally accurate predictions across all classes. Therefore, the selection of an appropriate threshold is at the discretion of the researcher, and alternative values, including more or less strict, may be considered more suitable in different applications.

In equation 3.2, TP represents true positive, TN represents true negative, FN represents false negative, and FP represents false positive. Additional clarification for these terms is provided in the confusion matrix outlined in Table 8 presented later in this section. Furthermore, when we use the term "positive," we are indicating the minority class, whereas "negative" corresponds to the majority class. Finally, in the context of multiclass classification, the metric can be applied to class pairs, where "positive" and "negative" represent the minority and majority classes within each pair.

$$Performance\ Gap\ Metric\ = |Sensitivity_{minority\ class} * 100\% - Sensitivity_{majority\ class} * 100\%|$$

$$= |\frac{TP}{TP+FN} * 100\% - \frac{TN}{TN+FP} * 100\%| \quad (3.2)$$

STEP 4 – Visualization of data samples in a low-dimensional space: The 4th step of the framework involves visualizing the dataset utilized in the classification task. Given the potential difficulty of identifying patterns and relationships in a higher-dimensional space, we propose visualizing the data in a lower-dimensional space for enhanced clarity.

This step is particularly valuable when there's a observed discrepancy in measuring the performance gap across classes, enhancing our understanding of the data structure. Graphic inspection, in general, serves as a method to comprehend the internal characteristics of a dataset (Santos et al., 2023). In the context of imbalanced datasets, it can facilitate the identification of factors beyond class imbalance that might impede classification performance. For high-dimensional data, visualization can be facilitated by applying transformation techniques that allow data representation in two or three dimensions. Note that although we discuss this step at this stage in the framework, researchers can implement it at any point they deem necessary.

STEP 5 - Inspection of factors impeding classification: The 5th step of the framework involves the identification of factors that hinder classification performance based on the graphic inspection of the previous step. Imbalanced datasets often encompass additional factors that contribute to decreased performance in classification tasks. These factors may include the presence of rare sub-concepts within the minority class, known as "within-class" imbalance, or the existence of overlapping regions between classes. Specifically, the combination of class overlap, along with class imbalance, is widely acknowledged as one of the most challenging issues within the machine learning community (Lango & Stefanowski, 2022; Santos et al., 2023). Given the intricacies they introduce to the classifier's learning process, these factors are commonly referred to as "difficulty" factors. Identifying difficulty factors through visualization techniques can guide the selection of appropriate methods to enhance classification performance.

STEP 6 - Selection of corrective techniques: The 6th step of the framework involves the selection of appropriate techniques to mitigate the impact of difficulty factors and improve the classifier's performance. In the existing literature, numerous techniques have proved successful in addressing challenges posed by these factors, with each technique suited to address the problems associated with one or more of them. A comprehensive analysis of these techniques is presented in the literature review

chapter (Chapter 2). In our suggested framework, we advocate for the adoption of data-level techniques due to their independence from specific modeling approaches (Napierala et al., 2010; Elrahman & Abraham, 2013), thereby enhancing the adaptability of the framework. Since the suitability of different techniques may vary depending on the dataset or the modeling technique, researchers are granted discretion in choosing the most appropriate ones for their specific cases.

STEP 7 - Classification: The 7th step of the framework involves executing the classification task. This step aligns with STEP 2 of the framework; however, the classification task is now performed after the implementation of the corrective techniques selected in STEP 6.

STEP 8 – Evaluation of classification performance across classes: Similar to STEP 3, the 8th step of the framework involves evaluating the classifier's performance across classes. In addition to the proposed metric outlined in equation 3.2, there are alternative metrics available, based on which the effectiveness of the applied techniques can be assessed.

Due to the presence of class imbalance, we specifically, advocate for the use of metrics that independently evaluate the performance of each class. These metrics, including recall (sensitivity), precision (specificity), F1-score, balanced accuracy, area under the ROC curve (AUC), and others, are not influenced by disparities in class sizes, offering a more meaningful evaluation.

In general, the proper selection of evaluation metrics in classification tasks with imbalance datasets is a critical consideration, as highlighted in the literature review (Chapter 2). Commonly employed evaluation metrics, such as overall accuracy or the error rate, are mostly influenced by majority classes due to the their significantly larger sample sizes (Prati et al., 2004; García et al., 2007). Furthermore, these metrics assign equal misclassification costs to all classes, whereas highly imbalanced problems often entail non-uniform error costs that prioritize the minority classes, which are typically of greater interest. One of the key objectives in introducing our framework is to underscore the significance of choosing appropriate metrics when working with imbalanced datasets and to encourage fellow researchers to make thoughtful choices in this regard. This is critical to prevent overly optimistic yet potentially misleading outcomes.

Information regarding the evaluation metrics mentioned above is outlined in the subsequent tables. While these details are provided in the context of binary classification, metrics can be extended to apply to multiclass classification scenarios as well.

	Predicted Negative Class	Predicted Positive Class
Actual Positive Class	False Negative (fn)	True Positive (tp)
Actual Negative Class	True Negative (tn)	False Positive (fp)

Table 8. Confusion matrix for binary classification. TP represents the number of correctly classified positive examples, FN denotes the number of positive examples misclassified as negative, TN signifies the number of correctly classified negative examples, and lastly, FP represents the number of negative examples misclassified as positive. The attribution of the terms "positive" and "negative" to the minority or majority classes is contingent upon researchers' choice.

Metric	Formula	Evaluation focus
Accuracy	$\frac{tp + tn}{tp + fp + tn + fn}$	Accuracy measures the proportion of correct predictions over the total number of instances assessed.
Balanced accuracy	$\frac{1}{N}\sum_{i=1}^{N} Recall_i$	Balanced accuracy measures the average recall across all classes.
Error rate	$\frac{fp + fn}{tp + fp + tn + fn}$	Misclassification error measures the proportion of incorrect predictions over the total number of instances assessed.

Recall	$\frac{tp}{tp+fn}, \frac{tn}{tn+fp}$	Recall measures the number of correctly classified positive (negative) samples.
Precision	$\frac{tp}{tp+fp}, \frac{tn}{tn+fn}$	Precision measures the proportion of positive (negative) patterns correctly classified among the total patterns predicted as positive (negative).
F-score	$2 * \frac{Precision * Recall}{Precision + Recall}$	F-score represents the harmonic mean between the recall and precision values.

Table 9. Evaluation metrics for classification tasks. Accuracy and Error rate are commonly utilized metrics for evaluating the overall performance. Recall, Precision and F-score are often employed to assess performance independently for each class.

If the chosen metrics indicate an improvement in classification performance, and the researcher deems this improvement satisfactory, then this step concludes the framework. Alternatively, if further enhancement is desired, STEPS 6,7 and 8 can be reiterated, as illustrated in Figure 8.

CHAPTER 4: Data

Following the introduction of our proposed framework, this chapter offers insights into the data utilized in this study before delving into the framework's application. Initially, we provide a brief literature review on the factors influencing travel mode choices. Subsequently, we present a succinct description of the dataset, along with a descriptive analysis and an overview of the pre-processing steps. Finally, we briefly discuss the assumptions and limitations associated with the data.

4.1 Literature review - Determinants of travel mode choices

A brief literature review on the determinants of mode choices based on studies conducted using data from the Netherlands is summarized in this section. The variables encompass socio-economic attributes at both the household and individual level, along with trip-related factors, built environment characteristics, weather conditions and attitudes-perceptions.

	Variables	Sources
Trip characteristics	Trip distance, Trip purpose, Day of the week, Trip cost	Böcker & Thorsson (2016); Hagenauer and Helbich (2017); Kashifi et al. (2022); Versteijlen et al. (2021)
Socio-demographic characteristics (encompasses both individual and household related variables)	Age, Education, Car ownership, Work status, Gender, Possession of valid driving license, Ethnicity, Health condition (e.g. disability), Household size, Public Transport Card Ownership, Income	Schwanen et al. (2001); Limtanakool et al. (2006); Rassouli & Timermans (2014); Böcker et al. (2016); Hagenauer & Helbich (2017); Kashifi et al. (2022)
Built environment	Degree of urbanization, Green space	Limtanakool et al. (2006); Kemperman & Timmermans (2014); Hagenauer & Helbich (2017); Kashifi et al. (2022)
Weather characteristics	Temperature, Precipitation, Season	Böcker et al.(2013); Böcker & Thorsson. (2016); Kashifi et al. (2017); Ton et al.(2019)
Attitudes-Perceptions	Perception of infrastructure's quality, Travel convenience, Environmental concerns, Perceived benefit, Awareness, Safety concerns	Heinen et al. (2011); La Paix Puello et al. (2020); Versteijlen et al. (2021)

Table 10. List of factors influencing mode choices.

4.1.1 Trip characteristics

Among trip characteristics, distance emerges as a crucial factor in determining mode choices (Hagenauer & Helbich , 2017; Kashifi et al., 2022). Across all age groups, shorter distances are notably more likely to be covered by bicycle and especially on foot, whereas longer distances are significantly more likely to be covered by car and public transport. It is worth noting that the impact of distance on transportation choices appears to be more pronounced among the elderly, which could be attributed to biological limitations or challenges in walking and cycling longer distances as age increases (Böcker et al., 2016). According to Böcker et al. (2016), and regarding trip purposes, leisure trips are predominantly undertaken by car, whereas walking, cycling, and public transport are preferred for work/study, errands, and social visits. Moreover, according to the same study, the day of the week does not exhibit any noticeable influence on the transport mode choices of the elderly, while public transport usage by younger age cohorts is less in the weekends. Finally, travel expenses emerged as a significant determinant in students' transportation choices for commuting to the university. Notably, high parking costs was the primary motivation to transition from using their cars to opting for public transport (Versteijlen et al., 2021).

4.1.2 Weather conditions

With respect to the effect of weather conditions on mode choices, literature presents mixed findings. In a study conducted in Rotterdam by Böcker et al. (2016), higher maximum air temperatures were observed to have a favorable influence on choosing cycling over using a car, for both elderly and non-elderly individuals, while they did not significantly impact the choice of public transport or walking. Similarly, precipitation had a negative effect on cycling among the non-elderly but did not show a significant impact on the elderly population, while wind speed showed no notable effect on transport mode choices for all age groups. The influence of seasonality on mode choices was also observed, especially for cyclists and pedestrians who are more exposed. Summer and autumn were identified as the most favorable seasons for selecting these modes (Böcker et al., 2013). Additionally, temperature was recognized as a pivotal factor, especially in forecasting bicycle and public transport trips, in Hagenauer & Helbich's research (2017), while Kashifi et al. (2022) affirmed the influence of temperature in forecasting all modes. Conversely, a study conducted by Ton et al. (2019), found no impact between weather conditions and the choice of active transportation modes.

4.1.3 Built-Environment attributes

Regarding built-environment attributes, address density, which serves as a metric of the urbanization level, displayed a greater relevance in forecasting public transport trips compared to other modes (Hagenauer & Helbich, 2017; Kashifi et al., 2022). An elevated urbanity level was found to have a positive correlation with choosing public transport over private cars, particularly among the elderly population (Limtanakool et al., 2006; Böcker et al., 2016). In their study, Schwanen et al. (2001) discovered that elderly individuals residing in urban areas are less inclined to cycle and more inclined to use public transport when compared to their counterparts in rural areas. This difference could be attributed to better access to public transport in densely populated areas and the fact that urban areas have heavier traffic, which may make cycling seem less appealing as a choice. Additionally, Kempermann & Timmermans (2014) confirmed that elderly individuals in urban areas are less likely to cycle, while they also concluded that green residential environments are more likely to encourage them to cycle and walk.

4.1.4 Socio-demographic attributes

Concerning the socio-demographic variables, ethnicity appears to be an influential factor, with research indicating that individuals with a non-Western migration background tend to cycle less and rely more on public transport than native Dutch individuals (Böcker et al. , 2016). According to Limtanakool et al. (2006), socio-demographic factors in general and car availability, significantly influence the choice of transportation modes for medium and longer distance trips, regardless of the trips' purpose. Conversely, in the study by Rassouli & Timmermans (2014), car availability was determined to have minimal significance in travel mode choice decisions, while Hagenauer & Helbich (2017) and Kashifi et al., (2022) demonstrated that the number of cars and bicycles per household held importance, along with other significant variables such as age, education, and household income. Furthermore, in line with the study of Böcker et al. (2016), household size plays a significant role, particularly for the elderly. The study's results indicated that elderly individuals living in larger households are more likely to partake in walking, cycling, and the use of public transport compared to those in smaller households. Additionally, car ownership, bicycle ownership, and the possession of a driving license and a public transport card influence transportation mode choices regardless of age, as evidenced by studies such as Schwanen et al. (2001) , Böcker et al. (2016) and Kashifi et al. (2022). Finally, concerning health characteristics, disability is linked to a decrease in trip frequencies for both the elderly and non-elderly populations, while among the non-elderly, obesity is negatively correlated with the use of active transportation modes (Böcker et al., 2016).

4.1.5 Perceptions & Attitudes

In the context of perceptions and attitudes, La Paix Puello et al. (2020) found that an enhanced perception of the quality of cycling infrastructure is strongly associated with a greater likelihood of choosing the bicycle as the preferred mode of transport for reaching the train station. Travel convenience was also discovered to be a crucial factor influencing students to favor public transport over cars, when commuting to the university (Versteijlen et al., 2021). This preference was attributed to the fact that public transport provided them with the opportunity to engage in other activities during the commute. In the same study, environmental considerations were also mentioned as a motive for selecting public transport, albeit rarely. In their research regarding bicycle commuting, Heinen et al. (2011) discovered that the decision to use a bicycle as a means of commuting to work is shaped by the perceived advantages of time savings, comfort, and flexibility, as well as individuals' appraisal of these benefits. Moreover, consideration of the impacts of cycling on personal health, the environment, and safety concerns significantly contributes to influencing the decision to commute by bicycle.

4.2 Dataset Description

The dataset employed in the present study combines information from two main sources: the 'Centraal Bureau voor de Statistiek' (CBS) and the open-source software OpenTripPlanner (OTP). The CBS data used in the current research study concerns the years 2018-2019 and is derived from the ODiN survey (formerly known as OViN); a national revealed-preference travel survey, which is conducted via an online questionnaire and is designed to gather statistical insights into the daily mobility patterns of the Dutch population. The survey's target population includes Dutch residents aged 6 years and older residing in private households, with the exclusion of individuals living in institutions or other types of communal settings. The data collected through the survey encompasses sociodemographic information about the respondents (e.g. age, income, car ownership, education) at both individual and household levels, as well as details about their daily trips (e.g. origin-destination postcode, start time, mode choices). In the present study, specific features from the ODiN survey were selected based on their significance in influencing travel mode choice, as indicated by relevant literature (Table 10). Next, the OTP software was used to compute the distance and duration for all available travel options for the individual trips recorded in the survey. This computation was carried out by using the centroids of the origin and destination postcodes associated with these particular trips. Finally, data regarding the expenses associated with the car and public transport journeys was obtained from a dataset provided by TNO. The final dataset includes both numerical and categorical attributes. The categorical features are encoded with numerical values, with each category assigned a specific numerical representation. Detailed information about each feature is presented in Table 11.

Figure 9 illustrates the overall distribution of mode preferences in the Netherlands using ODiN data (2018-2019). It is important to note that, in the this study, classification tasks exclusively focused on the car (driver), bike and transit modes. Figure 10 and Figure 11 showcase the distributions of trips involving these three modes across Dutch provinces.

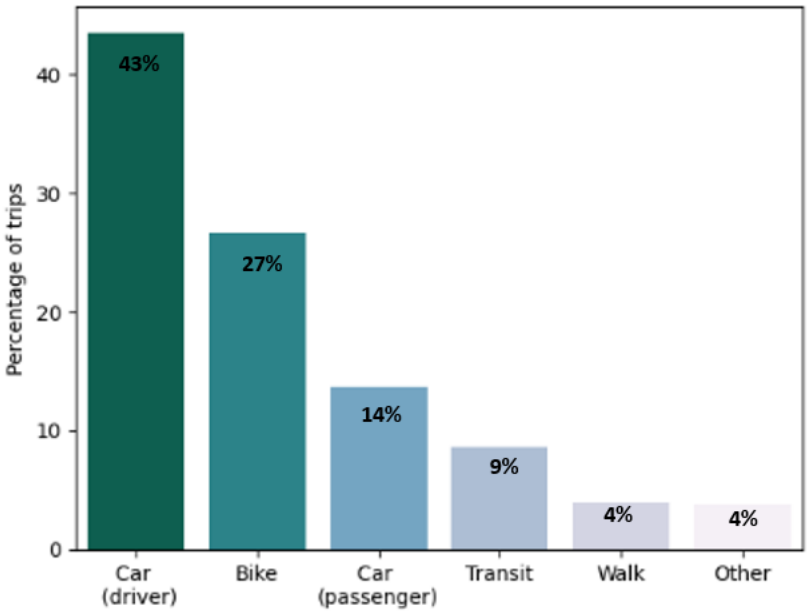

Figure 9. Mode preferences in the Netherlands using ODiN data (2018-2019), highlighting an inherent imbalance in the utilization of the different modes. Note that in the classification tasks of this study only the Car (driver), Bike and Transit modes are considered.

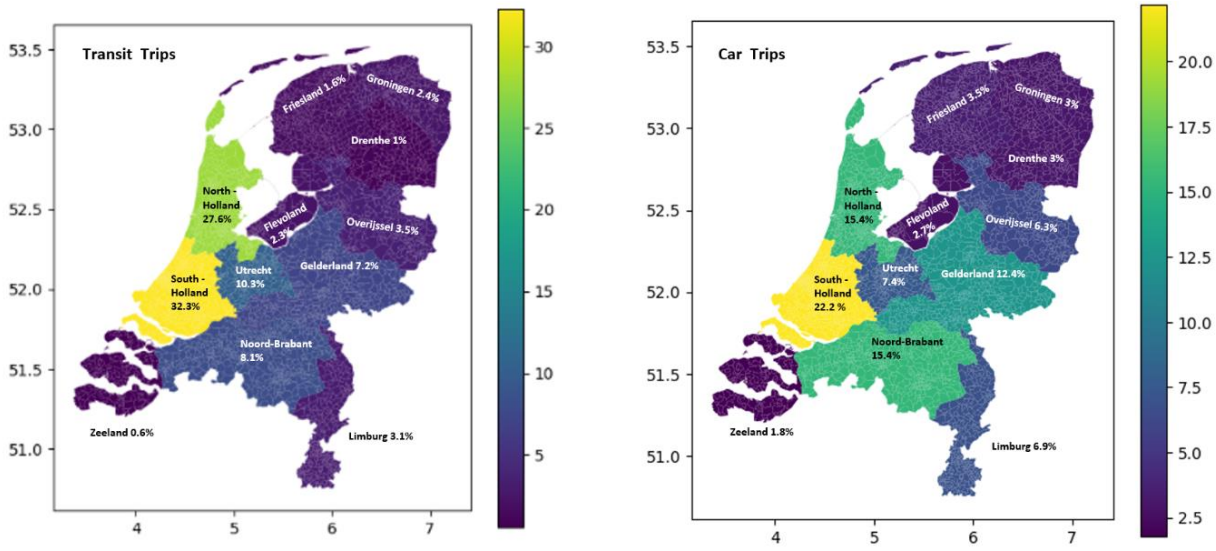

Figure 10. Maps of the Netherlands depicting the distribution of transit and car trips across Dutch provinces. Notably, the majority of transit trips occur in the Randstad area, encompassing the provinces of South-Holland, North-Holland, Utrecht and Flevoland.

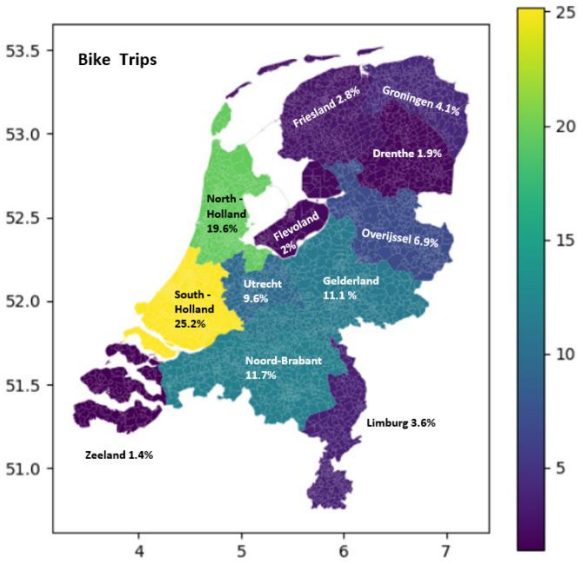

Figure 11. Map of the Netherlands depicting the distribution of bike trips across Dutch provinces.

4.2.1 Descriptive Analysis of the dataset

	Description	Type	Source
Socio-demographic attributes			
Individual level			
Age	1: age < 18 , 2: 18 ≤ age ≤ 54, 3: age ≥ 55	Categorical	CBS data (2018-2019)
Sex	male – 1 ; else – 0	Categorical	CBS data (2018-2019)
Education	higher education - 1 ; else – 0	Categorical	CBS data (2018-2019)
Driving license	driving license – 1 ; else – 0	Categorical	CBS data (2018-2019)
Paid occupation	yes – 1 ; else – 0	Categorical	CBS data (2018-2019)
Background	1: native Dutch, 2: Western migration background, 3: non-Western migration background	Categorical	CBS data (2018-2019)
Possession of student public transport card	yes – 1 ; else - 0	Categorical	CBS data (2018-2019)
Household level			
Income	1: low, 2: high	Categorical	CBS data (2018-2019)
Car ownership	car ownership - 1; else - 0	Categorical	CBS data (2018-2019)
E-bike ownership	e-bike ownership - 1 ; else - 0	Categorical	CBS data (2018-2019)

Children	yes - 1 ; else - 0	Categorical	CBS data (2018-2019)
Number of people in the household	single-person household – 1 ; else - 0	Categorical	CBS data (2018-2019)
Trip attributes			
Origin postcode	origin postcode (PC4)	Categorical	CBS data (2018-2019)
Destination postcode	destination postcode (PC4)	Categorical	CBS data (2018-2019)
Origin province	1: Groningen, 2: Friesland, 3: Drenthe, 4: Overijssel, 5:Flevoland, 6: Gelderland, 7:Utrecht, 8: Noord-Holland, 9: Zuid-Holland, 10: Zeeland, 11:Noord-Brabant,12: Limburg	Categorical	CBS data (2018-2019)
Purpose	purpose of the trip, 1: commute trip, 2: business trip, 3: other	Categorical	CBS data (2018-2019)
Weekday	weekday – 1; else - 0	Categorical	CBS data (2018-2019)
Holiday	public holiday -1 ; else - 0	Categorical	CBS data (2018-2019)
Departure time	1: morning, 2: afternoon, 3: evening	Categorical	CBS data (2018-2019)
Season	1: winter, 2: spring, 3: summer, 4: autumn	Categorical	CBS data (2018-2019)
Mode choice	1: car as driver, 2: car as passenger, 3: transit*, 4: bike, 5: walk *(train, bus, tram, metro)	Categorical	CBS data (2018-2019)
Car duration	travel time by car in sec	Numerical	OTP
Transit duration	travel time by public transportation in sec	Numerical	OTP
Cycling duration	cycling travel time in sec	Numerical	OTP
Walking duration	walking travel time in sec	Numerical	OTP
Car cost	cost in euros for driving journeys	Numerical	TNO
Transit cost	cost in euros for public transportation journeys	Numerical	TNO
Number of mode transitions during a transit journey	number of mode transitions	Numerical	OTP
Activity duration	duration of the activity the respondent undertook after reaching the destination of their trip in min	Numerical	CBS data (2018-2019)
Peak-hour	Peak-hour – 1; else - 0	Categorical	CBS data (2018-2019)

Built environment attributes			
Urbanity level of the origin postcode	1: very high, 2 : high, 3: moderate, 4: low, 5: rural area	Categorical	CBS data (2018-2019)
Urbanity level of the destination postcode	1: very high, 2:high, 3: moderate, 4: low, 5: rural area	Categorical	CBS data (2018-2019)
Population class * *of the respondent's residential municipality	1: inhabitants ≤ 50.000, 2: 50.000 <inhabitants ≤ 150.000, 3: inhabitants > 150.000	Categorical	CBS data (2018-2019)

Table 11. Description of the dataset utilized in the application of this study. The descriptions of the variables correspond to the post-processing stage.

4.2.2 Descriptive Analysis of the dataset

Variable		Values/ Percentages		Variable		Values/ Percentages
Mode choice	Car Bike Transit	55% 34% 11%		*Possession of driving license*	Yes No	82% 18%
				Car ownership	Yes No	87% 13%
Sex	Male Female	53% 47%		*E-bike ownership*	Yes No	23% 77%
Trip purpose	Commute Business Other	32%, 4% 64%		*Possession of student public transport card*	Yes No	7% 93%
Single household	Yes No	17% 83%		*Public holiday*	Yes No	1% 99%
Children in the household	Yes No	53% 47%		*Peak- hour*	Yes No	39% 61%
*Population class ** **of the respondent's residential municipality*	1 2 3	39% 31% 30%		*Weekday*	Yes No	79% 21%
				Car duration (sec)	mean sd	1313,842 1004,374
Age	1 2 3	11% 75% 14%		*Income** **High income category includes both medium and high income households*	High Low	76% 24%
Paid occupation	Yes No	64% 36%		*Departure time*	1 2 3	37% 45% 18%
Higher Education	Yes No	45% 55%		*Migration Background*	1 2 3	82% 9% 9%
Urbanity level of the origin postcode	1 2 3 4 5	29% 28% 17% 14% 11%		*Urbanity level of the destination postcode*	1 2 3 4 5	29% 29% 17% 14% 11%

Season	1 2 3 4	24% 25% 24% 27%		*Transit duration (sec)*	mean sd	3036,2119 2802,1498
				Bike duration (sec)	mean sd	2911,1169 3991,797
Car cost (euros)	mean sd	7, 3 9, 4		*Number of mode transitions during a transit journey*	mean sd	3, 2 2, 1
Transit cost (euros)	mean sd	3, 2 4, 1		*Origin province*	1 2 3 4 5 6 7 8 9 10 11 12	6% 3% 2% 6% 2% 11% 8% 18% 24% 2% 13% 5%

Table 12. Descriptive analysis of the dataset utilized in the this study after the pre-processing stage. For numerical variables, mean and standard deviation values are provided for both the datasets utilized in the Random Forest model (left value) and the MNL model (right value).

4.3 Data Preprocessing Steps

The preprocessing of the dataset utilized in the present study can be summarized in the following steps:

- Trips from the ODiN data were filtered to encompass only those conducted using one of the following modes: car (as a driver), bike, and transit.
- Intermediate legs for trips consisting of multiple segments were eliminated.
- Trips with zero postcode coordinates were removed.
- Trips with identical origin and destination postcodes were removed from the dataset since the OTP software cannot generate travel alternatives for such trips. It is worth noting that this decision resulted in a notable reduction in the number of cycling and walking trips.
- Trips in which a driving license was not present, and car was reported as the chosen mode were excluded.
- The exact start time of the trips were replaced with one of the following categories : 'morning', 'afternoon', 'evening'. Additionally, based on the start times, trips were categorized as occurring during peak hours or not.
- The weekday feature was encoded by assigning a value of 1 for weekdays and 0 for weekends, instead of using distinct numerical representations for each day of the week.
- For the Random Forest model, missing values were substituted with the median value of their corresponding features. This process was executed after the dataset was partitioned into

training and test sets to avoid any 'data leakage' from the test set to the training set. To clarify the procedure, the missing values were initially replaced in the training set, and afterward, the median values from the training set were used to fill in the missing values in the test set.

- For the MNL model , the treatment of NaN values differed. Specifically, for each alternative mode, a corresponding variable indicating its availability was created. When the OTP software produced NaN values for the duration and distance of a particular mode, indicating the unavailability of that option for a particular trip, the availability of that mode for the trip in question was set to zero. Conversely, if the OTP software provided valid values, the availability was defined as 1. Modes with zero availability were excluded from the user's choice set during the computation of the utility functions.

- Samples for which the OTP software generated no information for the duration and distance features of the respondent's chosen mode were excluded.

- Samples with zero transit costs were excluded in the MNL model.

- The availability of the car alternative in the MNL model was dependent on the possession of a valid driving license.

- Data binning was performed to address the presence of categories within the features with very few samples. This step was taken to mitigate potential issues when dividing the dataset into the training and test sets.

- Features without any inherent category order, such as trip purpose, background of the respondent, departure time etc were subjected to one-hot encoding.

- Missing values from the activity duration feature were imputed with the feature's median value derived from trips with the same trip purpose.

4.3.1 Feature selection

A crucial step taken after the data pre-processing was feature selection. This step is of paramount importance, as irrelevant features can hamper the performance of classification algorithms. Furthermore, in our case it could contribute to addressing the presence of class overlap in the dataset. The selection of the appropriate techniques to assess correlation between variables should be contingent on the variables' type. Considering the presence of both numerical and categorical data in our data, we evaluated correlation using three different metrics.

Specifically, to identify potential correlation among the numerical explanatory variables, we employed the Pearson coefficient, which is the most widely used measure for assessing linear correlation between numerical variables. The values of Pearson coefficient range from -1 to 1. A correlation coefficient of zero signifies no linear relationship between the variables; a value of -1.00 indicates a perfect negative linear relationship, while a value +1.00 indicates a perfect positive linear relationship (Prematunga, 2012).

By setting a threshold of 0.95, the distance feature for the car and bike modes was excluded due to its high correlation with the trip duration of these alternatives. Similarly, the distance feature for the public transport mode was omitted as it exhibited a high correlation with the cost feature of this alternative.

To assess variable correlation among categorical variables, we employed Cramer's V association (Khamis, 2008). For examining the correlation between continuous and categorical variables, we used the point-biserial correlation test, a specific case of the Pearson coefficient tailored for exploring correlation between continuous and dichotomous (binary) variables. In both cases, no features displayed high correlation.

Mutual Information was utilized to detect weak dependencies between the features and the target, with the goal of potentially excluding certain features. However, a challenge arose as mutual information values, spanning from 0 to +∞, consistently remained close to zero for all features in our case, rendering the selection process difficult. Moreover, literature evidence indicated that all considered variables play a role in determining mode choices, and consequently, the decision was made to retain all variables for further analysis.

4.3.2 Calculation of trip costs

In this sub-section we outline the methodology employed to compute the costs for both the car and transit trips in our dataset.

To compute the car costs, we considered two factors: the cost per kilometer and parking expenses. The cost per kilometer was set at 0.34 euros/km, relying on data gathered by TNO. Parking costs were computed by multiplying the parking tariff by the duration of the respondent's activity after reaching the trip destination. Activity durations were sourced from the ODiN data, and parking tariffs were extracted from a dataset developed by TNO. This dataset associates each postcode in the Netherlands with a parking tariff, representing the weekly average tariff for the respective postcode. The detailed calculation of car costs is outlined in the following equation:

$$car\ cost\ (euros) = 0.34\ * distance_{car} + parking\ tariff_{\substack{destination \\ postcode}} * activity\ duration \quad (4.1)$$

Similarly, to compute transit costs, we considered two factors: a base tariff of 0.96 euros and the cost per kilometer traveled .The same travel fare (euros/km) was applied to all means of public transport and all regions. The value of the travel fare corresponds to the average train fare from the Hague and Rotterdam areas. The calculation of transit costs is expressed by the following equation:

$$transit\ cost\ (euros) = 0.96 + \frac{(0.147+0.166)}{2} *\ distance_{transit} \quad (4.2)$$

4.3.3 Data Assumptions

In this subsection, we outline the assumptions made in the current study with regard to the data.

Data assumptions
o In this study, we focused only on the following modes: car, transit and cycling, while the remaining mode choice alternatives were not considered. o In the ODiN survey, the primary mode for each trip is determined by considering the mode that covered the greatest distance. o To generate the public transport alternatives through the OTP software, a maximum walking distance of 2 km was considered. Modifying this walking distance limit, could possibly lead to the generation of alternative options for the public transport mode. o Intermediate legs for trips consisting of multiple segments were excluded in order to simplify the process and reduce complexity. o Trips intended for going home were regarded as having no associated parking costs. o While the ODiN survey originally classified means of public transport into two distinct categories—namely, train, and tram/bus/metro—we consolidated them into a single category referred to as "public transport." o When calculating the cost of transit trips, the same fares were considered for all modes within the public transport category, including the train, tram, bus, and metro. Additionally, the costs associated with bike trips were assumed to be zero. o Although certain respondents of the survey have reported multiple trips, we treated all samples in the dataset as independent observations.

Table 13. Summary of the data assumptions considered in this study.

4.3.4 Data Limitations

Data Limitations
The primary limitations of our dataset pertain to the travel alternatives derived from the OpenTripPlanner software, outlined as follows: • The OTP software generates travel alternatives using separate origin and destination points. In our methodology, we supplied the software with the centroids of the origin and destination postcodes for each trip. As a result, travel alternatives for trips within the same postcode could not be generated. This predominantly impacted walking trips, given that these trips typically cover short distances compared to other modes that are more frequently used for longer journeys. Consequently, a substantial number of walking trips had to be excluded. Coupled with a substantial amount of missing values for the remaining trips (approximately 30%), we opted to exclude the walking class entirely from our analysis. • Due to the fact that postcode centroids may land within inaccessible areas, such as lakes or fields distant from the road network, the software may deem a requested trip impossible. As a result, no travel alternatives will be generated for such trip, even though, in reality, it may have been undertaken. • The data derived from the OTP software may be subject to both overestimation and underestimation. For example, a very short trip from one side of a postcode's border to the other might be overestimated, as the software will provide information related to a longer route between the centroids of the two neighboring postcodes. Conversely, an underestimation may occur for trips conducted between the furthest edges of two postcodes.

Table 14. Summary of the limitations of the data utilized in this study.

CHAPTER 5: Application of the proposed framework

Having introduced our proposed framework (Chapter 3) and discussed the data utilized in this study (Chapter 4), this section presents the practical application of our framework, specifically in the context of predicting travel mode choices in the Netherlands. In contrast to numerous prior studies, that delve into the issues of class imbalance and class overlap using artificial datasets, our approach employs Revealed-Preference (RP) data. In specific, we employed the CBS ODiN dataset (for the years 2018-2019), which provides information on the daily mobility patterns of the Dutch population, data from the OpenTripPlanner (OTP) software, and a parking cost dataset developed by TNO. Comprehensive details about each of these datasets are thoroughly outlined in Chapter 4. Furthermore, unlike the majority of studies that primarily focus on binary scenarios, resulting in a relatively limited exploration of multiclass classification tasks with imbalanced datasets (Lango & Stefanowski, 2022), our framework was employed for both binary and multiclass classification tasks.

Figure 12, illustrates the sequential steps we undertook, showcasing how our proposed framework can be integrated in a classification task involving imbalanced datasets. Further explanation of these steps is provided in the remainder of this section.

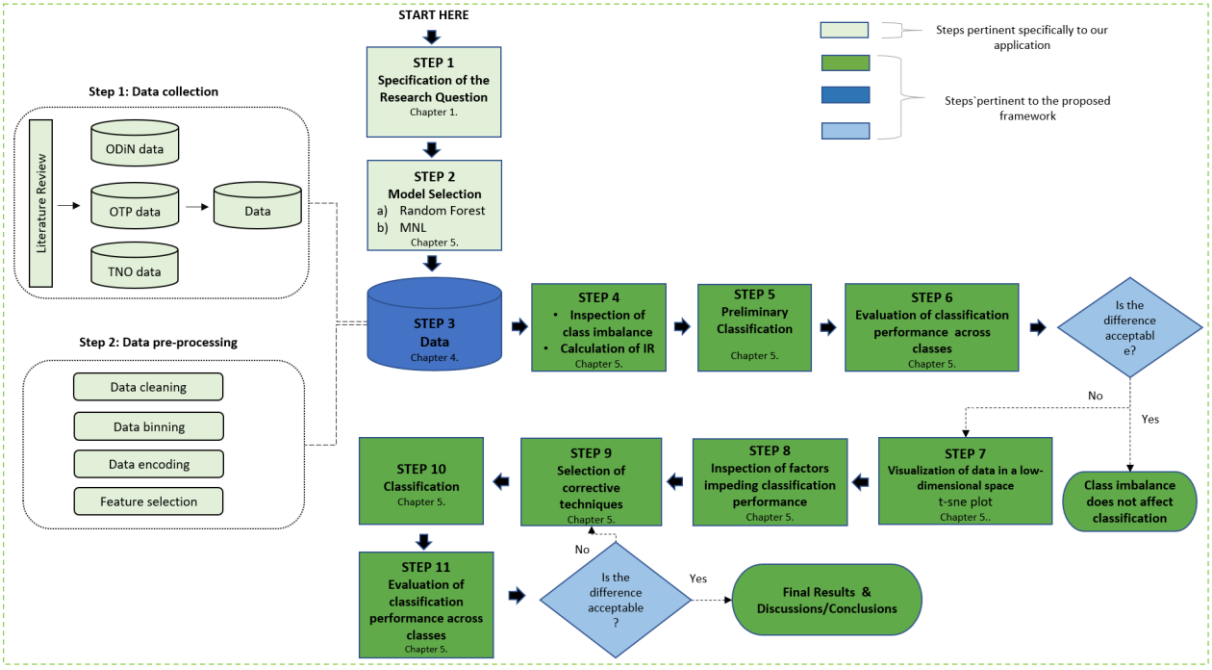

Figure 12. Integration of the proposed framework in the context of forecasting travel mode choices in the Netherlands.

STEP 1 - Specification of the Research Question: The 1st step in our methodology involved articulating our research question. As previously mentioned, the central focus of our study revolves around identifying and addressing the impact of class imbalance within the context of forecasting mode choices in the Netherlands.

STEP 2 - Model selection: The 2nd step encompassed the selection of the modeling techniques employed for the classification task. As mentioned earlier, we applied our proposed framework, using two distinct model categories: machine learning and Random Utility Maximization (RUM) theory – based models. Within the array of machine learning models we opted for the Random Forest Classifier. Our decision was primarily influenced by its superior performance in recent studies focused on forecasting travel mode choices, compared to alternative models (Hagenauer & Helblich, 2017; García-García et al.,

2022). Within the array of RUM theory-based models, we opted for the Multinomial Logit Model (MNL), recognized as the most widely used discrete choice model (Salas et al., 2022).

Typically, this step involves an iterative process wherein researchers experiment with various models to identify those most suitable to their specific cases. In our current application, we have chosen the aforementioned models, recognizing that a more comprehensive exploration could potentially lead to finding more optimal alternatives. Nevertheless, our primary goal in this study is to demonstrate the applicability of our proposed framework across different models rather than determining the most optimal classifier.

STEP 3 - Data: The 3rd stage in our application involved both data collection and pre-processing. This step followed model selection, recognizing that the chosen models can profoundly influence the pre-processing of the data. To identify the factors influencing travel mode choices in the Netherlands, we conducted a comprehensive literature review, elaborated upon in Chapter 4. As mentioned earlier, our dataset comprises information from three distinct sources: ODiN data for the years 2018-2019, data from the OTP software, and data developed by TNO.

In specific, when employing the Random Forest classifier, we incorporated most of the mode choice determinants identified in the existing literature, based on their availability in the ODiN travel survey. In contrast, the Multinomial Logit (MNL) model includes only the cost and duration features. While we acknowledge the potential improvement in predictive performance with an increased number of relevant features, as mentioned earlier, our study's primary objective is not to find the most optimal classifiers. Instead, we aim to demonstrate the adaptability of our proposed framework to any classifier, leaving the choice of models to the discretion of researchers. Additionally, even with a simplified specification of the MNL model, we can effectively capture the Value of Time (VoT), a crucial parameter in the appraisal of transport projects, and explore whether and to what extent its values are affected when addressing the impact of class imbalance. In-depth details regarding all variables and pre-processing steps are included in Chapter 4.

Finally, it is noteworthy that the two models employed in this study differ not only in terms of explanatory features but also in the pre-processing steps applied to their samples, making a direct comparison between them impractical. Nevertheless, it is essential to emphasize that our study does not intend to directly compare these models. Rather, our goal is to illustrate the versatility of the proposed framework when implemented employing different models.

STEP 4 – Inspection of class imbalance & Calculation of the Imbalance ratio: In the 4th step of our application, we identified the minority and majority classes and calculated the Imbalance Ratio (IR) for each dataset. This was necessary because the number of samples in the datasets used in each model differed , given the latter's distinct preprocessing requirements. In all cases, the majority and minority classes were the Car and Transit classes, respectively. In the context of classification tasks using Random Forest, the Imbalance Ratio (IR) was found to be 5, whereas the dataset employed in the MNL model had an Imbalance Ratio equal to 6.8.

	Imbalance Ratio
Random Forest	5
MNL	6.8

Table 15. Imbalance Ratio of the datasets utilized in each model within this study.

STEP 5 & STEP 6 - Preliminary Classification - Preliminary evaluation across classes: The 5th and 6th steps in our application involved the preliminary execution of the classification task and the subsequent evaluation of the performance across classes, by employing the Performance Gap Metric outlined in equation 3.2 of STEP 3 in Chapter 3.

The packages and libraries used for classification are outlined in the following table.

Libraries & packages	
Scikit-learn	Random Forest Classifier
Biogeme	MNL

Table 16. List of libraries and packages employed for classification in the application of this study.

STEP 7 & STEP 8 - Visualization of data in a low-dimensional space – Inspection of difficulty factors: The 7th and 8th steps in our application involved the data visualization and inspection. Given the noted disparities in the performance among classes, in both the binary and multiclass scenarios, our goal in this step was to gain a more comprehensive insight into the data structure, with a particular emphasis on exploring the closeness between samples.

Among the various visualization techniques found in the literature, in this study we opted for a 2D data visualization using the t-distributed Stochastic Neighbor Embedding (t-SNE) plot (Van Der Maaten & Hinton, 2008). t-SNE is as a non-linear dimensionality reduction technique, transforming multidimensional data into a two or three-dimensional space. Its primary objective is to maintain local distances between data points, grouping similar samples together. The initial step involves converting high-dimensional Euclidean distances among data points into conditional probabilities, which capture the similarities between them. Proximity results in higher conditional probabilities, while distant points yield infinitesimal probabilities. These probabilities are then calculated for the lower-dimensional space. Through an optimization process, the mismatch between distributions in the higher and lower dimensional spaces is minimized, ensuring that the positions of data points in the lower-dimensional space faithfully represent their relationships observed in the higher-dimensional space. For an in-depth description of t-SNE's mathematical foundations, readers are referred to Van Der Maaten and Hinton (2008).

As depicted in Figure 13 and Figure 14, in both binary and multiclass scenarios, in addition to the challenge posed by class imbalance our dataset exhibits significant overlap, which serves to validate the observed performance deviations.

Also, it is important to note that while the application of this study concentrates on the class imbalance between different classes (''between-class'' imbalance) as well as the class overlap, a brief analysis related to the ''within-class'' imbalance mentioned in STEP 5 of Chapter 3 is also provided in the Appendix of this study.

STEP 9 - Selection of corrective techniques: The 9th step of our application involves implementing techniques to address both class imbalance and class overlap. Drawing from existing methodologies, the solution space encompasses sampling approaches, data decomposition, modification of existing algorithms, and the creation of new learning algorithms specifically tailored for imbalanced data. As

mentioned earlier, our proposed framework advocates for the use of data-level techniques, due to their independence from the underlying classifier. In particular, the techniques we employed either focus on augmenting the number of minority class samples to enhance their visibility to the classifier or on treating overlapping and non-overlapping regions of the data space separately, following a separation scheme similar to that proposed by Xiong et al. (2010). In this context, "overlapping" pertains to regions in the data space where samples from the minority class and their nearest neighbors from other classes coexist. On the other hand, "non-overlapping" signifies regions exclusively occupied by samples from classes other than the minority class. A thorough descriptions of each method is provided in Chapter 2.

Table 17 provides an overview of the techniques employed in this study. Concerning data augmentation, we chose to utilize the Synthetic Minority Over-Sampling Technique Nominal Continuous (SMOTENC) as it is considered the state-of-the-art approach for augmenting imbalanced tabular datasets. The selection of the other two techniques was motivated by identifying significant overlap in the dataset and consulting relevant literature for its effective resolution. Similar to our approach in selecting modeling techniques, we recognize that there may exist more suitable techniques than those implemented. Additionally, different techniques may be more appropriate for specific cases. Consequently, researchers should select techniques that align better with the specifics of their individual cases.

Selected techniques
SMOTENC
Neighborhood-based Undersampling
Separation scheme

Table 17. Summary of the sampling techniques utilized in the application of this study.

STEP 10 & STEP 11 – Classification & Evaluation of performance across classes: The 10th and 11th steps encompassed the execution of the classification task and the evaluation of classification performance across classes. Detailed results from these steps are presented later in this chapter. With the completion of these steps our application was concluded.

5.1 Results

In this section the final results of our application are presented. Initially we provide a brief overview of the applied techniques and subsequently we proceed to showcase the results achieved through their implementation.

Techniques utilized in the this study	
SMOTENC	Creation of synthetic data for the minority class along the line segments that connect minority samples and their nearest neighbors, which belong also to the minority class. Note that SMOTENC is a version of the traditional SMOTE technique, able to handle both numerical and categorical variables. Parameters to be determined: • k : number of nearest-neighbors to be considered. • N : number of samples to be created.

Neighborhood-based Undersampling	Elimination of samples from the majority class that have at least one nearest neighbor belonging to the minority class. Note that in the context of the multiclass classification, samples were eliminated from both the car and bike classes. Parameters to be determined: • k : number of nearest neighbors to be considered
Separation scheme	Classification is conducted in two (binary case) or three (multiclass case) stages. In the initial stage, a classifier categorizes samples by determining whether they belong to the overlapping or non-overlapping regions. Following this classification, two distinct classifiers are employed in each of the regions to predict the actual classes of the samples. Specifically, in the binary case, all samples within the non-overlapping region are assigned to the car class. Consequently, the second stage of classification exclusively takes place within the overlapping region. Parameters to be determined: • k: number of nearest neighbors to be considered when defining the overlapping region

Table 18. Summary of the sampling techniques utilized in this study.

5.1.1 Random Forest – Binary & Multiclass classification

In this section we present the results obtained with the Random Forest model. Table 20 presents the results for binary classification, encompassing the Car and Transit classes, whereas Table 22 presents the results for multiclass classification, involving the Car, Transit and Bike classes. In both cases, Car constitutes the majority class, while Transit the minority class. The reported values represent the mean performance of the model across five runs, with standard deviation values indicated in parentheses. As mentioned earlier, this method was utilized to accommodate the inherent stochastic nature of the model, which introduces randomness through the implementation of bootstrapping and random feature selection techniques.

In both scenarios, Random Forest models were trained using 70% of the dataset. For hyperparameter tuning, 10% of the data served as a validation set, with the remaining 20% designated for testing. Given the imbalanced nature of our dataset, we adopted a stratified splitting approach to ensure the preservation of the target class ratio across all sets. Finally the models' performance was assessed based on both aggregate (overall & balanced accuracy) and mode-specific (precision, recall, f1-score) evaluation metrics as presented in the tables that follow.

5.1.1.1 Binary classification

As mentioned earlier, the Imbalance Ratio (IR) in the dataset used for classification tasks employing the Random Forest model was determined to be 5. To assess whether class imbalance indeed impacted the classifier's performance and following the steps outlined in the proposed framework, we ran the model and calculated the Performance Gap Metric (eq. 3.2). The metric yielded an approximate value of 31%, surpassing the established threshold and prompting us to proceed with a more thorough analysis of the dataset's structure.

Imbalance ratio	5
Performance Gap Metric (baseline model)	≈ 31 %

Table 19. Imbalance Ratio and Performance Gap Metric in the context of binary classification using the Random Forest model. The Imbalance Ratio evaluates the difference in the number of samples between the majority and minority classes, whereas the Performance Gap Metric quantifies the discrepancy in their respective classification performances.

Figure 13 depicts the data projection into a 2D space. The t-SNE plot reveals the existence of both clean and noisy regions within the data space. Clean regions in this study are characterized by the presence of samples belonging to classes other than the minority class. In this specific case, clean regions are occupied by samples from the Car class. Noisy regions, on the other hand, contain samples from both classes. In this study, regions where no minority samples overlap with samples from other classes are labeled as "non-overlapping", whereas areas with sample overlap between the minority class and the rest of the classes are designated as "overlapping".

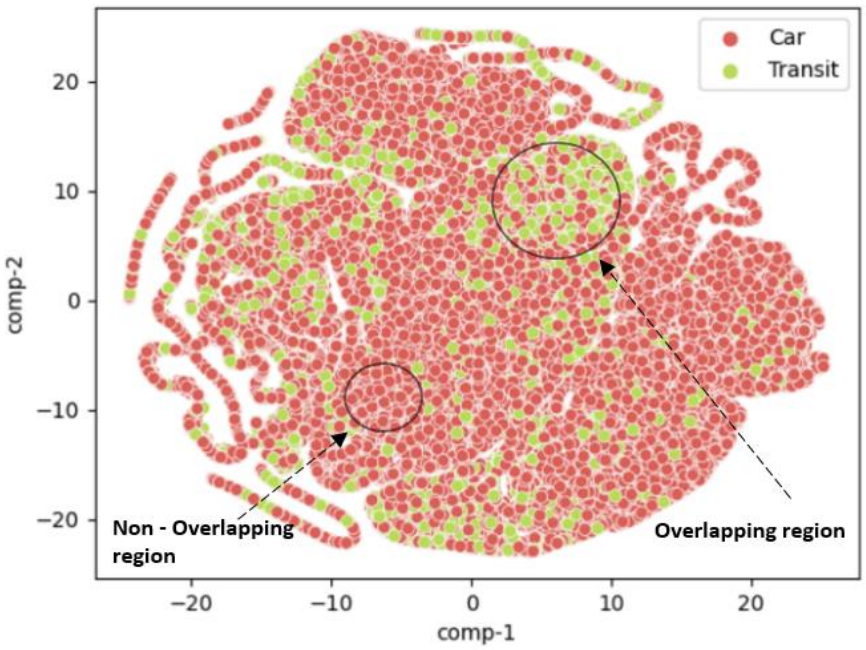

Figure 13. T-SNE plot for the binary classification between the Car and Transit classes, highlighting both overlapping and non-overlapping regions within the dataset. T-SNE is a dimensionality reduction technique that improves the visualization of multidimensional data by projecting it into a lower-dimensional space. In this case, the data is visualized in a 2D space. In the above plot, two distinct regions are emphasized: the overlapping, encompassing both transit and car samples, and the non-overlapping, comprising exclusively car samples.

To address the characteristics of our dataset, marked by both class imbalance and class overlap, we employed all three techniques briefly outlined in Table 18. The results of these techniques are discussed as follows.

With regard to the SMOTENC technique and contrary to other studies (Kashifi et al., 2022; H.Chen & Cheng, 2023) that oversampled the minority class to make it equal with the majority class, without

performing any exploration with regard to the number of created samples, in this study we conducted three experiments. Each experiment involved the generation of a different number of minority instances. Specifically, in the first experiment, we aimed for a number of minority samples equal to 30% of the number of majority samples, while in the second and third experiments, this percentage was set to 50% and 100%, respectively.

Upon comparing the outcomes of applying SMOTENC to the baseline model (prior to any technique implementation) the most notable enhancement in the sensitivity (recall) of the minority class (approximately 11%) is observed when achieving an equal class distribution. In this case, the Performance Gap Metric also exhibited improvement, falling below its established threshold and reaching a value of 16.6%. Conversely, in the first and second experiments, where the number of generated samples equaled 30% and 50% of the majority samples, sensitivity exhibited a more modest increase of approximately 4% and 8%, respectively. Additionally, the Performance Gap Metric did not fall below the 20% threshold, although in the second experiment it nearly attained it.

Furthermore, we observe that in all three experiments, the heightened sensitivity was coupled with a reduction in precision. The trade-off between these two metrics is reflected in the F1-score of the minority class, which remained relatively stable, exhibiting only a minimal decrease in the third experiment.

Simultaneously, with the increase in the sensitivity of the minority class, the sensitivity of the majority class exhibited a slight decrease, which became more pronounced (approximately 4%) when the two classes became equal. In contrast, the total accuracy remained relatively stable, while the balanced accuracy exhibited an increase, which is attributed to heightened sensitivity of the minority class.

In the implementation of the Neighborhood-based Undersampling technique, two experiments were conducted, each utilizing a distinct value for the k parameter. In the initial experiment, k was set to 3, while in the subsequent one, it was set to 5. In both experiments, the sensitivity of the minority class increased, with the improvement being more pronounced, reaching approximately 12% compared to the baseline model, when k was equal to 5. Comparing the two experiments with different k values reveals a greater increase in the sensitivity of the minority class with the higher value of k. This can be attributed to the more substantial elimination of majority samples as k increases. Specifically, with k = 3, the Imbalance Ratio between the two classes was 3.9, while in the case of k = 5, the Imbalance Ratio was reduced to 3.1.

The heightened sensitivity of the minority class in the second experiment led to the Performance Gap Metric falling below its predetermined threshold, reaching a value of approximately 15%, signifying the model's effective prediction with regard to both classes. Similar to the SMOTENC technique, however, the improved sensitivity of the minority class was counterbalanced by a decrease in its precision (approximately 16%, when k was equal to 5), resulting in a relatively unchanged F1-score. Simultaneously, with the increase in the sensitivity of the minority class, the sensitivity of the majority class exhibited a slight decrease, more pronounced in the second experiment. The total accuracy remained relatively stable, while the balanced accuracy slightly increased, following the trend of the sensitivity of the minority class.

In the implementation of the Separation scheme, three experiments were conducted, each utilizing a distinct value for the k parameter. In the first experiment, k was equal to 2, while in the second and third experiments, k was equal to 3 and 5, respectively. By implementing the separation scheme, the sensitivity of the minority class increased in all three experiments. Nevertheless, as the value of k increased from 3 to 5, the incremental improvement in sensitivity compared to the baseline model diminished. This phenomenon can be attributed to the increasing class imbalance within the

overlapping area with the rising value of k, involving more nearest neighbors from the majority class. Specifically, when k was equal to 3, the Imbalance Ratio was 1.2, while for k = 5, the Imbalance ratio was 2.

In the scenario with k equal to 3, the Performance Gap Metric exhibited improvement, falling below the predetermined threshold; however, once again precision had to be compromised. Concurrently, a minimal decrease was observed in the sensitivity of the majority class, while the total accuracy remained unchanged and the balanced accuracy slightly increased.

In summary, the application of all three techniques led to an enhancement in the sensitivity of the minority class, with the Neighborhood-based Undersampling proving to be the most effective, resulting in the minimum value of the Performance Gap Metric. Furthermore, in all cases, total accuracy remained relatively stable, indicating that fairness in terms of accurately predicting both classes could be achieved without compromising the overall accuracy of the classifier. Lastly, when comparing the optimal results obtained by the SMOTENC and the Separation scheme techniques, it is evident that while both methods achieve a relatively similar increase in the sensitivity of the minority class, the latter's precision is further diminished during the implementation of the SMOTENC. This may be attributed to the 'blindness' of this technique in generating synthetic samples, disregarding the presence of majority samples. Consequently, in the presence of substantial overlap in the dataset, an increase in the number of synthetic samples may elevate noise, resulting in a higher misclassification of majority samples. On the contrary, when employing the Separation scheme, the classifier focuses exclusively on the overlapping area, potentially acquiring a better ability to differentiate between samples belonging to the two classes.

	Precision		Recall		F1-score		Acc.	B. Acc.	Perf. Gap Metric
	Car	PT	Car	PT	Car	PT	Total	Total	Car-PT
Baseline model	94 (0)	90 (0)	98.4 (0.55)	67 (0)	96.1 (0.26)	77 (0)	93 (0)	83 (0)	31.4%
SMOTE-NC (k = 5, N = 30% of majority samples)	94 (0)	85.2 (0.45)	97.6 (0.55)	71 (0)	95.8 (0.26)	77.2 (0.45)	93 (0)	84 (0)	26.6%
SMOTE-NC (k = 5, N = 50% of majority samples)	95 (0)	80.2 (0.45)	96 (0)	74.8 (0.45)	95.5 (0)	77.2 (0.45)	93 (0)	85.4 (0.55)	21.2%
SMOTE-NC (k = 5, N = 100% of majority samples)	95 (0)	73.8 (0)	94.2 (0.45)	77.6 (0.55)	94.6 (0.22)	75.4 (0.55)	92 (0)	86 (0)	**16.6%**
NBU (k = 3)	95 (0)	82 (0)	97 (0)	73.8 (0.45)	96 (0)	77.8 (0.45)	93 (0	85 (0)	23.2%
NBU (k = 5)	96 (0)	74 (0)	94 (0)	79.2 (0.45)	95 (0)	76.2 (0.45)	92 (0)	87 (0)	**14.8%**
Separation scheme (k = 2)	95 (0)	81 (0)	96 (0)	74.8 (0.45)	95.5 (0)	77.8 (0.24)	93 (0)	85.4 (0.22)	21.2%
	95	78.8	96	76.4	95.5	77.6	93	86.2	**19.6%**

Separation scheme (k = 3)	(0)	(0.45)	(0)	(0.55)	(0)	(0.23)	(0)	(0.27)	
Separation scheme (k = 5)	95 (0)	82.6 (0.55)	97 (0)	73 (0)	96 (0)	77.5 (0.24)	93 (0)	85 (0)	24%

Table 20. Summary of the results of binary classification using the Random Forest model. The reported values represent the mean performance of the model across five runs, with standard deviation values indicated in parentheses. The values of the Performance Gap Metric are highlighted in the cases where its value surpasses the 20% threshold, indicating the ability of the classifier to predict equally well the majority and the minority classes.

5.1.1.2 Multiclass classification

In the multiclass classification, the Bike class was introduced alongside the existing Car and Transit classes. The ratio between Car and Bike samples was 1.6, while for Bike and Transit samples, it was 3.

The Performance Gap Metric (eq. 3.2), focusing on the Car and Transit classes, reached 40%, surpassing the established threshold of 20%, as well as exceeding the metric's value in the case of binary classification (≈ 31%). This further increase in the Performance Gap Metric compared to the binary scenario indicated a reduced capability of the classifier to identify samples from the minority class in the presence of additional classes.

Imbalance ratio Car-Transit	5
Imbalance ratio Car - Bike	1.6
Imbalance ratio Bike - Transit	3
Performance Gap Metric (baseline model)* considering the majority and minority classes	≈ 40 %

Table 21. Imbalance Ratios and Performance Gap Metric in the context of multiclass classification using the Random Forest model. In this case, the Imbalance Ratio was computed for all pairs of classes, signifying the difference in their number of samples, while the Performance Gap Metric was utilized to quantify the disparity in the classification performance between the majority and minority classes.

The results of the data projection for the multiclass scenario are displayed in Figure 14. The t-SNE plot indicates substantial overlap among samples from all classes. While certain regions appear to be primarily occupied by car and transit samples, we chose to simplify the analysis and refrain from decomposing the data space into multiple regions. Instead, we adopted the same separation scheme introduced in binary classification. In this context, the overlapping region encompassed all samples from the minority class and their k-nearest neighbors from the other two classes. Conversely, the non-overlapping regions included only samples from the Car and Bike classes.

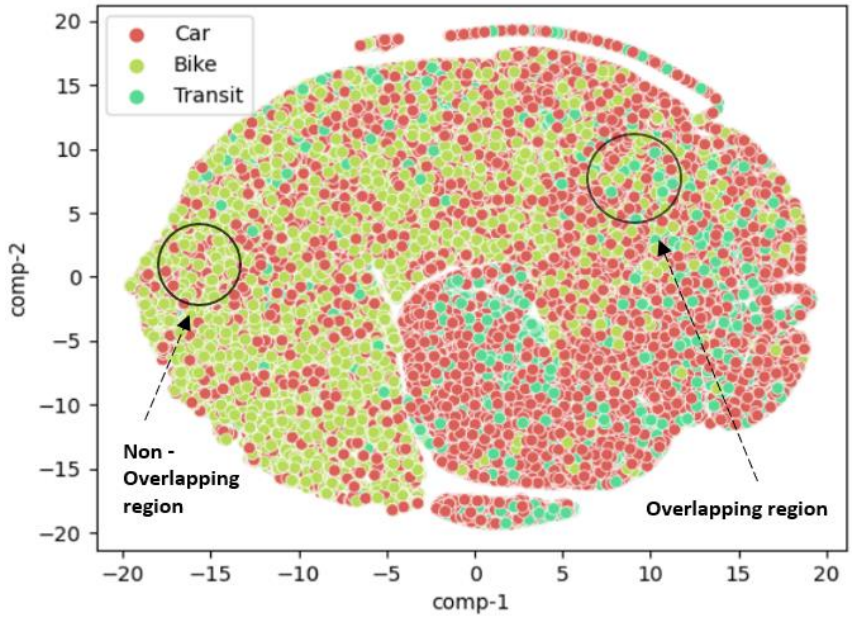

Figure 14. T-SNE plot for multiclass classification involving the Car, Transit, and Bike classes. T-SNE is a dimensionality reduction technique that improves the visualization of multidimensional data by projecting it into a lower-dimensional space. In this case, the data is visualized in a 2D space. In the context of multiclass classification, overlapping regions are defined as those occupied by the minority class samples and their nearest neighbors from the Car and Bike classes, while non-overlapping regions are considered those occupied solely by samples from the Car and Bike classes.

Regarding the SMOTENC technique, similar to the binary case, three experiments were also performed in the multiclass scenario, entailing the generation of varying numbers of synthetic samples. Minority samples were synthesized until their total number reached 30%, 50%, and 100% of the number of majority samples (Car samples). Across all three experiments, the sensitivity of the minority class improved, with the maximum increase (15%) observed when the sizes of the minority and majority classes became equal. In that case, the Performance Gap Metric demonstrated its maximum improvement, nearly attaining the 20% threshold when the classes became equal.

In all three experiments, the increased sensitivity was offset by a reduction in precision, which reached its maximum decline (approximately 21%) when the minority and majority classes achieved equal representation. This trade-off between sensitivity and precision metrics resulted in the F1-score of the minority class remaining relatively consistent.

Contrary to the increased sensitivity of the minority class, the sensitivity of the other two classes declined, with a more pronounced decrease as the number of generated samples increased. Meanwhile, the total accuracy remained stable, while there was a slight rise in the balanced accuracy, attributed to the heightened sensitivity of the minority class.

With regard to the Neighborhood-based Undersampling technique, as previously mentioned, in the context of multiclass classification, apart from the car class, overlapped samples were also eliminated from the bike class. This step was taken due to considerable overlap observed among samples from all classes, aiming to ensure the clear visibility of minority samples to the model. Similar to the binary case two experiments were conducted with k =2 and k = 3, respectively. In both cases, there was an enhancement in the sensitivity of the minority class, with the improvement being more pronounced, reaching approximately 14% compared to the baseline model, when k was equal to 5. Comparing the two experiments with different k values reveals a greater increase in the sensitivity of the minority class with the higher value of k. As previously stated, this can be ascribed to the more substantial elimination

of majority samples as k increases. In specific, with k = 3, the Imbalance Ratio between the Car and Transit classes was 4, while in the case of k = 5, the Imbalance Ratio between the two classes decreased to 3.4. Correspondingly, between the Bike and Transit classes, these values were 2.8 and 2.6, respectively.

The heightened sensitivity of the minority class also resulted in an improvement in the Performance Gap Metric, nearly attaining the predetermined threshold in the second experiment (k =5). Similar to the implementation of the SMOTENC technique, however, the improved sensitivity of the minority class was counterbalanced by a decrease in its precision. The maximum decrease in precision was evident in the scenario where k = 5, reaching approximately 18%. This trade-off between the two metrics resulted in the F-score for the minority class relatively consistent compared to the baseline scenario.

Contrary to the increased sensitivity of the minority class, the sensitivity of the two other classes declined. Concurrently, the total accuracy remained unchanged, while balanced accuracy exhibited a slight increase in line with the trend observed in the sensitivity of the minority class.

Concerning the Separation scheme, similar to the binary scenario, three experiments were performed, each employing a distinct value for the parameter k. In the initial experiment, k was set to 2, while in the subsequent two experiments, k took on values of 3 and 5, respectively.

In all three experiments, the sensitivity of the minority class increased, with a more pronounced enhancement observed at higher values of k (approximately 9%). However, despite this enhancement, in none of these cases the Performance Gap Metric attained or fell below the predefined threshold. Concurrently precision experienced a decline, reaching its maximum drop of around 12% at when k was equal to 5. This trade-off between the two metrics resulted in a relatively stable F1-score for the minority class.

Contrary to the sensitivity of the minority class, the performance of the remaining two classes exhibited a decline, while the total accuracy remained stable and the balanced accuracy exhibited a minimal increase compared to the baseline model, aligning with the trend observed in the sensitivity of the minority class.

In summary, the application of all three techniques led to enhanced sensitivity for the minority class. However, in none of the scenarios the Performance Gap Metric drop below the predefined threshold. Notably, SMOTENC emerged as the most effective technique, with the Performance Gap Metric nearly attaining the 20% threshold, closely followed by Neighborhood-based Undersampling. Throughout all cases, the overall accuracy remained relatively consistent, suggesting that achieving fairness in terms of accurately predicting both majority and minority classes was feasible without undermining the classifier's overall accuracy.

	Precision			Recall			F1-score		
	Car	Bike	Transit	Car	Bike	Transit	Car	Bike	Transit
Baseline model	81 (0)	83 (0)	81.6 (0.55)	93 (0)	72.2 (0.45)	52.2 (0.45)	86.6 (0)	77.2 (0.25)	63.6 (0.55)
SMOTE-NC (k = 5, N = 30% of majority samples)	81 (0)	83.4 (0.55)	75 (0)	92 (0)	71.4 (0.55)	57 (0)	86.2 (0)	76.9 (0.55)	65 (0)
SMOTE-NC (k = 5, N = 50% of majority samples)	82 (0)	84 (0)	68 (0)	91 (0)	70 (0)	63 (0)	86.7 (0)	76.4 (0)	65 (0)
SMOTE-NC (k = 5, N = 100% of majority samples)	82 (0)	84.2 (0.45)	60.4 (0.55)	89 (0)	68.8 (0.45)	67.2 (0.45)	85.4 (0)	75.7 (0.36)	64.2 (0.45)

NBU (k=3)	81 (0)	83 (0)	72.4 (0.55)	91 (0)	71 (0)	60.4 (0.55)	85.7 (0)	76.5 (0)	65.6 (0.55)
NBU (k=5)	81 (0)	83 (0)	63.4 (0.55)	89 (0)	69.8 (0.45)	66.6 (0.55)	84.8 (0)	75.8 (0.26)	64.8 (0.45)
Separation scheme (k = 2)	81 (0)	82.8 (0.45)	73.2 (0.45)	91.4 (0.55)	71.2 (0.45)	58 (0.71)	86 (0)	76.8 (0.45)	(64.7 (0.47)
Separation scheme (k = 3)	81 (0)	83 (0)	70.3 (0.58)	91 (0)	70 (0)	61 (0)	86 (0)	76 (0)	65.3 (0.25)
Separation scheme (k = 5)	81 (0)	83.3 (0.58)	70 (0)	91.7 (0.58)	69.3 (0.58)	61 (0)	86 (0)	75.7 (0.58)	65.2 (0)

	Accuracy	Balanced accuracy	Perf. Gap Metric
	Total	Total	Car-PT
Baseline model	81.6 (0.55)	72.8 (0.45)	40.8 %
SMOTE-NC (k = 5, N = 30% of majority samples)	81 (0)	74 (0)	35%
SMOTE-NC (k = 5, N = 50% of majority samples)	81 (0)	75 (0)	28%
SMOTE-NC (k = 5, N = 100% of majority samples)	80 (0)	75.4 (0.55)	**21.8%**
NBU (k=3)	81 (0)	74 (0)	30.6%
NBU (k=5)	80 (0)	75 (0)	22.4%
Separation scheme (k = 2)	81 (0)	73.5 (0.3)	33.4%
Separation scheme (k = 3)	81 (0)	74 (0)	30%
Separation scheme (k = 5)	81 (0)	74 (0)	30.7%

Table 22. Summary of the results of multiclass classification using the Random Forest model. The reported values represent the mean performance of the model across five runs, with standard deviation values indicated in parentheses. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, nearly attaining the 20% threshold.

5.1.2 Multinomial Logit Model – Binary & Multiclass classification

In this section we present the results obtained with the MNL model. Table 24 presents the results for binary classification, encompassing the Car and Transit classes, whereas Table 26 presents the results for multiclass classification, involving the Car, Transit and Bike classes. In both cases Car constituted the majority class, while Transit the minority class. The reported values represent the mean performance across 3 fold cross-validation, with standard deviation values indicated in parentheses.

For every iteration in the cross-validation process, 2/3 of the dataset was used for model estimation, and 1/3 for validation. To prevent "data leakage" between the estimation and test sets, all techniques were implemented only in the estimation set, while the test set remained untouched. Given the imbalanced nature of our dataset, we adopted a stratified splitting approach to ensure the preservation of the target class ratio across all sets. Finally, the models' performance was assessed based on both aggregate metrics (overall & balanced accuracy) and mode-specific metrics (precision, recall, f1-score), as presented in the tables that follow.

Unike Random Forest model and given the nature of the MNL model , the separation scheme could not be applied in this case. Therefore results are only presented for the SMOTE and the Neighborhood-based Undersampling techniques. Furthermore, concerning the samples utilized for synthetic data generation, only those with all alternatives valid were taken into consideration. When employing the MNL model, in contrast to the Random Forest, NaN values are not imputed. As a result, samples containing such values become unsuitable for the calculations involved in the SMOTENC algorithm. The same restriction applies to samples utilized in the Neighborhood-based Undersampling technique.

5.1.2.1 Binary classification

As previously mentioned, in the datasets employed in the classification tasks with the MNL model, the Imbalance Ratio (IR) was equal to 6.8. The difference in the Imbalance Ratio between the data used for the MNL model and that used for the Random Forest model arises from the distinct data requirements of the two models in terms of pre-processing steps, which result in a different number of samples for each class in each model. Also, the Performance Gap Metric in the baseline model was approximately 50%. In comparison with the binary Random Forest model, the MNL model demonstrates significantly lower performance. Although the two models are not directly comparable, a lower performance is expected for the MNL model, due to its inherent high bias and its simplistic specification, making it challenging to capture complex patterns in the data.

Imbalance ratio	6.8
Performance Gap Metric (baseline model)	≈ 50 %

Table 23. Imbalance Ratio and Performance Gap Metric for the case of binary classification employing the MNL model. The Imbalance Ratio evaluates the difference in the number of samples between the majority and minority classes, whereas the Performance Gap Metric quantifies the discrepancy in their respective classification performances.

Similar to the Random Forest model, in the application of the SMOTENC technique, three experiments were conducted, generating minority samples equivalent to 30%, 50%, and 100% of the majority samples, respectively. Across all three experiments, the sensitivity of the minority class improved, with the most notable enhancement (approximately 27%) observed when achieving an equal class distribution. In this last experiment, the Performance Gap Metric experienced its greatest improvement, falling below the predetermined threshold by reaching a value of approximately 12%.

In all three experiments, the heightened sensitivity of the minority class, was accompanied by a decrease in its precision -revealing the usual trade off between the two metrics-, reaching its maximum value of approximately 43%, when the two classes became equal. Additionally, a declining trend was also observed for the F1-score metric.

Simultaneously, with the increase in the sensitivity of the minority class, both the sensitivity of the majority class and the total accuracy decreased, with the decline becoming more pronounced as more synthetic samples were generated. Conversely, the balanced accuracy increased following the trend of the sensitivity of the minority class.

In the application of the Neighborhood-Undersampling technique, two experiments were conducted, each employing a distinct value for the k parameter. In the initial experiment, k was set to 3, while in the subsequent one, it was set to 5. In both scenarios, there was an improvement in the sensitivity of the minority class, with the enhancement being more notable, reaching approximately 12% compared to the baseline model when k was equal to 5. A comparison between the two experiments with different k values highlights a greater increase in the sensitivity of the minority class with the higher k value. As mentioned earlier, this can be attributed to the more substantial elimination of majority samples as k increases, further reducing the imbalance as well as the overlap between the minority and majority classes. Despite the increase in sensitivity of the minority class, however, the Performance Gap Metric did not manage to reach or fall below the predetermined threshold, improving only up to 36% in the second experiment (k = 5).

Additionally, similar to the implementation of the SMOTENC technique, the improved sensitivity of the minority class was counterbalanced by a decrease in precision. The maximum decline in precision was evident in the scenario where k was equal to 5, reaching 23%. This trade-off between the two metrics resulted in the F-score for the minority class remaining stable compared to the baseline model. Concurrently, as the sensitivity of the minority class increased, the sensitivity of the majority class declined. Additionally, a minimal decrease was also observed in the total accuracy, while balanced accuracy increased, aligning with the trend of the sensitivity of the minority class.

In summary, both techniques led to an enhancement of the sensitivity of the minority class. However, the SMOTENC proved to be the most effective, resulting in the Performance Gap Metric falling below the 20% threshold. Unlike the case of the Random Forest model, though, in this case total accuracy decreased indicating that fairness, in terms of accurately predicting both classes could be achieved at the cost of the classifier's accuracy.

In summary, both approaches improved the sensitivity of the minority class. However, SMOTENC demonstrated superior effectiveness, with the Performance Gap Metric falling below the 20% threshold. Unlike the scenario with the Random Forest model, though, total accuracy decreased in this case, suggesting that achieving fairness, in terms of accurately predicting both the majority and minority classes, came at the expense of the classifier's overall accuracy.

	Precision		Recall		F1-score	
	Car	PT	Car	PT	Car	PT
Baseline scenario	93 (0)	89.7 (0.577)	99 (0)	48.3 (0.577)	95.9 (0)	62.7 (0.577)
SMOTE (k = 5, 30%)	94 (0)	74.3 (0.577)	97 (0)	55.3 (0.577)	95.5 (0)	63.3 (0.577)
SMOTE (k = 5, 50%)	94.4 (0.577)	63 (0)	95 (0)	62 (1)	94.7 (0.288)	62.3 (0.577)
SMOTE (k = 5, 100%)	96 (0)	46.3 (0.577)	87.3 (0.577)	75 (1)	91.5 (0.32)	57 (0)
NBU (k = 3)	94 (0)	74.7 (0.577)	97 (0)	55.3 (0.577)	95.5 (0)	63.3 (0.577)
NBU (k = 5)	94 (0)	66.3 (0.57)	95.7 (0.577)	60 (1)	94.8 (0.283)	63 (0)

	VOT Car	VOT PT	TT Car	TT PT	TC car	TC PT	ASC Car
Baseline scenario	44.4 (0.78)	5.45 (0.27)	-0.076 (0.001)	-0.024 (0.001)	-0.102 (0.002)	-0.263 (0.001)	2.77 (0.023)
SMOTE(k=5, 30%)	44.1 (1.022)	4.63 (0.45)	-0.08 (0.001)	-0.022 (0.001)	-0.109 (0.002)	-0.289 (0.01)	2.143 (0.02)
SMOTE (k = 5, 50%)	42.8 (0.381)	3.59 (0.365)	-0.083 (0.001)	-0.02 (0.001)	-0.117 (0.001)	-0.327 (0.014)	1.743 (0.032)
SMOTE(k = 5, 100%)	41.6 (2.075)	2.08 (0.59)	-0.0871 (0.002)	-0.033 (0.002)	-0.163 (0.003)	-0.37 (0.005)	1.228 (0.05)
NBU(k =3)	35.1 (1.452)	5.36 (0.346)	-0.096 (0.007)	-0.033 (0.002)	-0.160 (0.007)	-0.365 (0.018)	2.62 (0.02)
NBU (k = 5)	32.1 (1.347)	5.204 (0.346)	-0.112 (0.004)	-0.039 (0.002)	-0.209 (0.004)	-0.451 (0.004)	2.53 (0.022)

	Total Accuracy	Balanced Accuracy	Final Loglikelihood	Performance Gap Metric
				Car - PT
Baseline scenario	92.6 (0.081)	73.7 (0.289)	-14631 (68.5)	50%
SMOTE(k=5, 30%)	91.8 (0.084)	76.2 (0.289)	-24640 (176.4)	41.7%
SMOTE (k = 5, 50%)	90.4 (0.04)	78.5 (0.5)	-34225 (188.9)	33%
SMOTE(k = 5, 100%)	85.6 (0.15)	81.2 (0.29)	-50715 (435.9)	**12.3%**
NBU(k =3)	91.9 (0.094)	76.2 (0.289)	-12705 (70.7)	41.7%
NBU (k = 5)	91 (0.115)	77.8 (0.29)	-11399 (87.2)	35,7%

Table 24. Summary of binary classification results obtained with the MNL model. The reported values represent the mean performance of the model following a 3-fold cross-validation, with standard deviation values indicated in parentheses. VOTs are expressed in euros/h. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, falling below the 20% threshold and indicating the ability of the classifier to predict equally well the majority and the minority classes.

5.1.2.2 Multiclass classification

In the multiclass classification, the Bike class was introduced alongside the existing Car and Transit classes. The ratio between Car and Bike samples was 1.5, while for the Bike and Transit samples, it was 4.4.

In the multiclass scenario, the Performance Gap Metric (eq. 3.2) reached a value of 70%, surpassing the predefined threshold and demonstrating the most substantial difference in predictive accuracy between the Car and Transit classes among all models in this study. Additionally, when comparing this value with that obtained in the binary scenario, it becomes clear that the introduction of additional travel alternatives increases the complexity of the classification task.

Imbalance ratio Car-Transit	6.8
Imbalance ratio Car - Bike	1.5
Imbalance ratio Bike - Transit	4.4
Performance Gap Metric (baseline model)* considering the majority and minority classes	≈ 70 %

Table 25. Imbalance Ratios and Performance Gap Metric for the case of multiclass classification with MNL model. In this case, the Imbalance Ratio was computed for all pairs of classes, signifying the difference in their number of samples, while the Performance Gap Metric was utilized to quantify the disparity in the classification performance between the majority and minority classes.

In the multiclass scenario, similar to all scenarios utilizing the SMOTENC technique, three experiments were conducted, generating minority samples equivalent to 30%, 50%, and 100% of the majority samples, respectively. Across all three experiments, the sensitivity of the minority class improved, with the maximum increase, equal to 39%, observed when the sizes of the minority and majority classes became equal. In this case, the Performance Gap Metric exhibited its greatest improvement, falling below the established threshold by reaching a value of approximately 19%.

However, in all three experiments, the heightened sensitivity was counterbalanced by a decrease in precision, with its maximum drop (≈46%) occurring when the minority and majority classes achieved equal representation. Notably, as the number of generated samples increased, the decline in precision became more pronounced.

Contrary to the sensitivity of the minority class, the sensitivity of the other two classes exhibited a decrease as more synthetic samples were generated, followed by a subsequent decrease in their F1 score metrics. Total accuracy mirrored their trend, while balanced accuracy exhibited an increase, attributed to the heightened sensitivity of the minority class.

In the application of the Neighborhood-based Undersampling technique, the elimination of overlapped samples extended beyond the car class to include also the bike class. Similar to the binary scenario, two experiments were conducted with different k values, specifically k=2 and k=3. In both scenarios, there was an enhancement in the sensitivity of the minority class. This improvement was more prominent, reaching approximately 8% compared to the baseline model, when k was set to 5. A comparative analysis of the two experiments with varying k values reveals a more significant increase in the sensitivity of the minority class with the higher k value. This outcome can be attributed to the more substantial elimination of majority samples as k increases, further diminishing the imbalance and overlap between the minority and majority classes.

In contrast to the SMOTENC technique, the Neighborhood-based Undersampling approach resulted in a smaller increase in the sensitivity of the minority class, preventing the Performance Gap Metric from attaining or falling below the 20% threshold. Concurrently, the rise in sensitivity of the minority class was counterbalanced by a decrease in precision ranging between 10% and 15% in the two experiments.

Concurrently, with the rise in sensitivity of the minority class, there was a slight decrease in the sensitivity of the car class, whereas the sensitivity of the bike class exhibited a contrasting trend. The total accuracy remained consistent, while a slight increase was observed in the balanced accuracy.

In summary, both techniques improved the sensitivity of the minority class. However, only SMOTENC achieved a Performance Gap Metric falling below the 20% threshold, proving to be the most effective between the two techniques. Similar to the binary scenario presented above, in this case, total accuracy

decreased, indicating that achieving fairness, in terms of accurately predicting both the minority and majority classes, came at the expense of the classifier's overall accuracy.

	Precision			Recall			F1-score		
	Car	Bike	PT	Car	Bike	PT	Car	Bike	PT
Baseline scenario	72.7 (0.577)	80.7 (0.577)	80.7 (0.577)	95 (0)	55.7 (0.577)	24.7 (1.528)	82.3 (0.37)	65.8 (0.355)	37.7 (1.528)
SMOTE (k = 5, 30%)	73 (0)	80.3 (0.577)	61 (1)	93 (0)	55.6 (0.577)	33.7 (1.155)	81.8 (0)	65.8 (0.214)	43.7 (1.155)
SMOTE(k = 5, 50%)	73.7 (0.577)	80.6 (1.155)	50 (0)	90.3 (0.577)	55 (1)	43.7 (1.15)	81.2 (0.121)	65.4 (0.42)	46.3 (0.577)
SMOTE(k = 5, 100 %)	73.7 (0.064)	81.7 (0.577)	34.3 (0.577)	82.3 (0.58)	50.3 (0.577)	63.7 (1.155)	77.8 (0.063)	62.3 (0.271)	44.7 (0.577)
NBU (k = 3)	73 (0)	79.7 (0.577)	70 (1.732)	93.3 (0.577)	56 (1)	29 (1)	81.9 (0.219)	65.8 (0.717)	41 (1)
NBU (k =5)	73.3 (0.577)	79 (0)	65.7 (2.31)	92 (0)	57.3 (0.577)	32.6 (2.31)	81.6 (0.352)	66.4 (0.387)	43.7 (2.3)

	VOT Car	VOT PT	TT Car	TT PT	TT bike	TC car	TC PT	ASC CAR	ASC PT
Baseline scenario	28.8 (1.267)	7.38 (0.832)	-0.048 (0.002)	-0.021 (0.002)	-0.0819 (0.001)	-0.098 (0.002)	-0.169 (0.005)	-0.074 (0.012)	-2.594 (0.03)
SMOTE(k = 5, 30%)	33.9 (1.608)	5.53 (0.29)	-0.059 (0.002)	-0.02 (0.001)	-0.09 (0.001)	-0.105 (0.002)	-0.223 (0.002)	-0.027 (0.008)	-1.908 (0.01)
SMOTE(k = 5, 50%)	35.9 (1.186)	4.5 (0.427)	-0.07 (0.001)	-0.02 (0.001)	-0.0935 (0.001)	-0.111 (0.002)	-0.266 (0.008)	-0.001 (0.085)	-1.46 (0.03)
SMOTE(k = 5, 100%)	37.3 (1.98)	3.27 (0.82)	-0.075 (0.004)	-0.018 (0.004)	-0.1 (0.003)	-0.121 (0.001)	-0.325 (0.014)	0.0444 (0.013)	-0.885 (0.061)
NBU (k = 3)	26.51 (0.266)	7.13 (0.581)	-0.055 (0.001)	-0.024 (0.001)	-0.091 (0.0002)	-0.124 (0.001)	-0.21 (0.001)	-0.059 (0.018)	-2.429 (0.018)
NBU (k = 5)	26.31 (1.33)	5.79 (0.478)	-0.0613 (0.0022)	-0.0244 (0.001)	-0.098 (0.001)	-0.139 (0.002)	-0.2538 (0.012)	-0.061 (0.0136)	-2.384 (0.017)

	Total Accuracy	Balanced Accuracy	Final Loglikelihood	Performance Gap Metric
				Car- PT
Baseline scenario	74.9 (0.117)	58.4 (0.694)	-61466 (110.8)	70.3%
SMOTE (k = 5, 30%)	74.52 (0.17)	60.8 (0.51)	-76183 (162.6)	59.3%
SMOTE(k=5, 50%)	73.6 (0.12)	63 (0.33)	-90364 (272.1)	46.6%
SMOTE(k = 5, 100%)	69.4 (0.054)	65.4 (0.193)	-115408 (646.5)	**18.6%**

NBU(k =3)	74.7 (0.13)	59.4 (0.509)	-40173 (51.24)	64.4%
NBU (k = 5)	74.7 (0.26)	60.7 (0.882)	-37748 (124.7)	59.4%

Table 26. Summary of multiclass classification results obtained with the MNL model. The reported values represent the mean performance of the model following a 3-fold cross-validation, with standard deviation values indicated in parentheses. VOTs are expressed in euros/h. The highlighted value of the Performance Gap Metric corresponds to the scenario that produced the optimal result, falling below the 20% threshold and indicating the ability of the classifier to predict equally well the majority and the minority classes.

5.1.2.3 MNL - Interpretability

So far , we have discussed the results with regard to the predictive accuracy of the MNL model in both the binary and multiclass scenarios. However, when evaluating discrete choice models, it is crucial to consider not only the predictive accuracy but also the interpretability of the models (Rezaei et al., 2021). To assess the interpretability of the MNL model in this study, we primarily focused on the sign and statistical significance of the coefficients, as well as on the Value of Time (VOT) estimated for each mode. It is worth mentioning that, in the case of the Bike alternative, no Value of Time was estimated due to the absence of associated costs with this mode.

In the binary model, the coefficients consistently exhibited statistical significance, while they also had negative signs, aligning with the expectations for the time and cost features (Rezaei et al., 2021). Similarly, in the multiclass scenario, the coefficients were also statistically significant and negative. An exception arose in the multiclass scenario utilizing SMOTENC (N = 50%), in which the Alternative Specific Constant (ASC) for the car alternative, was found insignificant, indicating that no inherent preference for the car mode existed in that specific scenario. The results of the t-tests conducted to infer the significance of the coefficients are provided in the Appendix of this study. Additionally, in the multiclass scenarios, the Bike alternative was found to be the most appealing, while between the Car and Transit alternatives, the Car alternative was more attractive. Conversely, in all binary scenarios the Car emerged as the most attractive mode.

Regarding the Values of Time (VOTs), a comparison between the estimates from the baseline models of this study and the most recent national VOTs for the Netherlands (Significance, 2023), as presented in Table 27, reveals a discrepancy. In specific, our estimated VOT for the Car alternative appears significantly higher than its corresponding national values, while the VOT for the transit alternative is comparatively lower. Variations in Value of Time (VOT) observed in this study may be attributed to several factors. Firstly, it is important to highlight that national VOTs are derived from Stated Preference (SP) choice experiments, utilizing a questionnaire distributed among the target audience (Significance, 2023). Conversely, our study relies on Revealed Preference (RP) data obtained from actual trips. Therefore, the disparity in the values between the two sources could be partly attributed to the hypothetical bias frequently observed in SP data (Krčál et al., 2019).

Regarding the Value of Travel Time (VTT) associated with the transit alternative, as shown in Table 27, national VOTs distinguish between train and local public transport (including bus, tram, and metro). In contrast, our study consolidates these modes into a single category labeled "transit," resulting in a unified VOT. Additionally, while national VOTs are reported separately for each travel purpose, our study does not make such distinctions. Furthermore, our study does not differentiate between in-vehicle and out-of-vehicle travel time, encompassing all stages of the transit trip, including walking, waiting, and transfer time. These differences may account for the relatively lower VOT observed for the transit alternative in our study.

Concerning the car alternative, as mentioned earlier, its VOT in our study appears notably higher than the values reported in Significance (2023). This difference can be attributed to the comprehensive calculation of car costs in our study using data from TNO, covering both fixed and variable components such as fuel costs, motor vehicle taxes, depreciation costs, maintenance costs, insurance costs, interest expenses, and parking costs. In contrast, in SP surveys, individuals often overlook several of these factors. For example, in an Israeli survey conducted by Shiftan & Bekhor (2002), most participants considered only gas costs in their assessment of car expenses. The perceived car costs by respondents in the survey were significantly lower than the costs calculated through a vehicle cost survey, which is considered the primary source of auto costs in the country. The authors concluded that it is more challenging for people to perceive the cost of a single car trip, as the driver never directly pays for a single trip out of their pocket. Consequently, these disparities between actual and perceived costs could possibly be reflected in the VOTs calculated from the two distinct sources.

Finally, the implementation of the sampling techniques influences the VOTs for both the Car and the Transit alternatives, reflecting the alterations introduced in the dataset's composition during the model estimation process. Specifically, for the Transit alternative, the implementation of sampling techniques leads to a reduction in its VOT, with the most significant decrease (62% and 56% in the binary and multiclass scenarios, respectively) observed after applying the SMOTENC (N =100%) technique, which induces the most substantial alteration in the composition of the Transit class. In the case of the Car class, in the binary scenario, the VOT also experiences a decrease, with the most substantial reduction (28%) observed after the implementation of the NBU (k = 5) technique. Conversely, in the multiclass scenario, the most notable change occurs after implementing the SMOTENC technique, resulting in an approximate 30% increase in the VOT.

Mode	Value of Travel Time		
	Commute	Business	Other
Car	10.78 ± 0.63	21.20 ± 3.06	9.60 ± 0.40
Train	12.05 ± 0.26	17.96 ± 1.75	8.64 ± 0.17
Local public transport (bus/tram/metro)	7.62 ± 0.20	14.39 ± 2.59	6.66 ± 0.20

Table 27. National averaged Values of Time with uncertainty bandwidths in the Netherlands, in € / hr (Significance, 2023). (Values are reported in euros in price level 2022).

Mode	Value of Travel Time	
	Binary case	Multiclass case
Car	44.4 (0.78)	28.8 (1.27)
Transit (train/bus/metro/tram)	5.45 (0.27)	7.38 (0.83)

Table 28. Value of Time for the Car and Transit alternatives according to the baseline scenarios (before the implementation of sampling techniques) of this study.

Chapter 6: Discussions & Conclusions

This section marks the conclusion of this study, providing a summary of its contents and key findings. Furthermore, it emphasizes the study's contribution and evaluates its alignment, or lack thereof, with other research studies in the literature. Lastly, it outlines the study's limitations and presents practical recommendations, along with ideas for future research.

To begin with, this study aimed to investigate and address the potential adverse effects of imbalanced data on the classification performance of minority classes in mode choice models, providing an answer to the main research question formulated as follows:

"How can the impact of class imbalance in model performance be systematically identified and addressed in transport mode share forecasting?''

In response to identified gaps in the existing literature and to answer the main research question of this study, we introduced a comprehensive framework that is applicable regardless of the underlying classifier. This framework places emphasis on critical aspects, including the measurement of class imbalance within the data, the examination of its impact on classification performance, -particularly for the minority modes-, the exploration of additional challenging factors such as class overlap, and the proper evaluation of classification performance across classes. Concurrently, as part of the framework, we introduced the ''Performance Gap Metric'', a metric to assess the difference in classification performance between the majority and minority classes. A threshold of 20% was set for this metric, recognizing that different threshold values may be more suitable in different applications. If the metric's value fell below the predetermined threshold, resulting in a reduction in the performance gap between the two classes (<20%) after conducting a classification task, the classifier's performance was deemed favorable, as prediction outcomes demonstrated fairness. In this context, fairness refers to the classifier's equitable treatment of both majority and minority classes, ensuring accurate predictions for both. Following that, our framework was applied using the ODiN data as a case study to predict mode choices among car, transit, and cycling in the Netherlands. Specifically, we utilized two modeling techniques – a Random Forest and a Multinomial Logit model – in combination with various sampling techniques, including the SMOTENC, the Neighborhood-based Undersampling, and the Separation scheme.

In the application of this study, the minority class corresponded to the transit mode. This aligns with the recognition of the transit mode as the minority mode in other mode choice studies (Rezaei et al., 2021 ; Kashifi et al. , 2022). In different studies, both soft modes (comprising cycling and walking) (Omrani et al., 2015; Wang & Ross, 2018; H. Chen & Cheng, 2023), and shared mobility services (encompassing bike-sharing, car-sharing, and ride-hailing) (Narayanan & Antoniou, 2023) have also been observed as minority modes. Accurate predictions for minority modes are crucial for population groups that heavily rely on them. Concerning public transport systems, systematically underestimating travel demand can result in insufficient transit services. This reduction in services may lead to a decline in ridership, potentially setting off a negative loop (Zheng et al., 2023). The consequences of inaccurate transit demand predictions can be significant, particularly in rural and peripheral areas where accessing essential activities requires covering greater distances. Transit services in these regions are inherently limited compared to urban areas, mainly due to diminished demand stemming from declining populations in these areas. The combination of increased distances and limited travel options makes

rural and peripheral areas more car dependent, as well as more susceptible to "transport poverty", meaning that individuals residing in these areas may face greater constraints in reaching essential destinations (Pot et al., 2020).

In the Netherlands, particularly in the Randstad region, encompassing the provinces of South-Holland, North Holland, Utrecht, and Flevoland, where nearly half of the population resides, public transport systems are well-developed (Kasraian et al., 2016). The remaining peripheral and rural provinces outside the Randstad region are still recognized for having higher accessibility levels compared to other European rural regions. However, in the upcoming years, the outskirts of the country are anticipated to undergo a population decline coupled with population aging (PBL, 2019), which may heighten the risk of experiencing transport poverty. Consequently, it is evident that accurate predictions of travel demand will play a pivotal role in mitigating this risk in the future, through transport planning and provisions tailored to the actual needs of the population.

6.1 Main findings

Having provided a brief summary of this study, its main findings can be summarized as follows. To begin with, apart from class imbalance our dataset exhibited also substantial overlap. Classification results revealed that both the Random Forest and MNL models were affected by those factors, exhibiting reduced predictive performance for the minority class compared to the majority class. Concurrently, performance metrics in the multiclass classification tasks were lower than those in the binary classification tasks, highlighting the heightened complexity associated with additional classes. While many studies typically compare the classification performance among classifiers, highlighting the superior performance of the Random Forest model over the MNL model (Hagenauer & Helbich, 2017; García-García et al., 2022), this study diverges in not directly comparing the two models. This decision stems from two main considerations: firstly, as mentioned earlier, our primary goal in this study is not to determine the most optimal classifier, but rather to demonstrate the applicability of our proposed framework across different models. Secondly, the datasets used in each model include distinct explanatory features and samples, due to the models' different data requirements, rendering direct comparisons impractical.

Furthermore, the selection of the most effective techniques characterized by the smallest achievable value for the Performance Gap Metric, varied depending on the specific model employed. For the Random Forest model, in the binary classification task, the NB-Undersampling technique demonstrated the smallest difference in the predictive performance between the minority and majority classes. Conversely, in the multiclass classification task, the best-performing technique was SMOTENC, closely followed by the Neighborhood-based Undersampling, with only a marginal difference. Also, while in the binary scenario the Performance Gap Metric exceeded the 20% threshold reaching a value of 14.8%, in the multiclass scenario it nearly attained it, with a value of 21.8%. In the case of the MNL model, both in the binary and multiclass classification tasks, the best-performing technique was SMOTENC, with the Performance Gap Metric reaching the values of 12.3% and 18.6%, respectively. Drawing from the results obtained by the two models, our proposed framework successfully enhanced the classification performance of the minority mode. As a result, the classifiers in this study could be characterized by an improved sense of fairness, ensuring more equitable treatment of both minority and majority classes. The effectiveness of the SMOTENC technique in achieving robust classification performance for the minority class was also highlighted in the study by H. Chen and Cheng (2023), ranking among the top three techniques that exhibited the most substantial improvement out of the six sampling techniques employed in their research. However, in the study conducted by Chaipanha and Kaewwichian (2022), SMOTE demonstrated lower performance compared to the Random Undersampling technique. These findings across the different studies emphasize the significance of acknowledging that diverse techniques may yield more favorable outcomes when coupled with different models and datasets.

Therefore, researchers should select techniques based on the specific models and requirements applicable to their study.

Additionally, another finding of this study, is that the increase in the sensitivity of the minority class in both models and classification tasks, was consistently accompanied by a decrease in its precision, revealing a trade-off between the two metrics. While the ideal scenario involves classifiers exhibiting high precision and high recall simultaneously, in this case the compromise in precision is justifiable as the primary goal is to improve the models' ability to accurately predict the minority class. In any case, the extent to which a reduction in precision is acceptable is contingent on the researchers' discretion, considering the importance they attribute to enhancing classifier fairness (with fairness in this context referring to the classifier's ability to accurately predict both the minority and majority classes) as well as the potential cost associated with misclassifying the majority class.

Regarding the Values of Time (VOTs) estimated from the MNL models, our findings diverge from the national VOTs for the Netherlands as reported in Significance (2023). As previously highlighted in this study, several factors could contribute to this discrepancy, including: a) variations in data sources (SP vs RP data), b) the consolidation of all means of public transport in this study into a single category labeled ''transit,'' resulting in the calculation of a unified VOT, c) the computation of car costs in this study considering various fixed as well as variable factors, d) the absence of differentiation based on trip purpose, and e) the inclusion of both in-vehicle and out-of-vehicle travel times in the calculation of the VOT for the transit mode. After implementing the sampling techniques, and since the composition of the dataset is altered, the values of the VOTs are also changing. Unfortunately, there is a scarcity of literature addressing the implementation of sampling techniques in discrete choice models, providing limited opportunities for comparing our results. One of the limited studies available for comparing the travel time coefficients is the study by Rezaei et al. (2021). In this study, transit represents the minority class, while single occupancy vehicles constitute the majority class. After balancing the classes through a combination of the Random Undersampling and Random Oversampling techniques, the coefficients of travel time for both classes demonstrated a reduction of 50%. Regarding the best-performing scenarios within our study, the travel time coefficient for the car class also experienced a decrease, showing a reduction of 14.6% in the binary scenario and 56.25% in the multiclass scenario. Meanwhile, the travel time coefficient for the transit class decreased by 37.5% in the binary scenario, while it increased by approximately 14.3% in the multiclass scenario. Considering the substantial significance of the Value of Time (VOT) as a vital factor in evaluating transport projects (Significance, 2021), we recommend that researchers carefully evaluate the results of their specific cases to determine the acceptability of changes resulting from the implementation of such techniques. Additionally, we argue that in cases where fairness, with respect to accurate predictions for both minority and majority classes, is a primary concern, a slight decrease in the models' interpretability may be justified. In any case, researchers must approach these techniques with caution, ensuring meticulous and accurate implementation to prevent any adverse effects on the models' behavioral outputs.

Finally, in both the Random Forest and MNL models, experiments demonstrating the minimum value of the Performance Gap Metric showed, on average, a reduction in the sensitivity of classes other than the minority class. In some scenarios, a decline in the total accuracy was also observed, particularly pronounced in the case of the MNL model. These findings suggest that achieving fairness in this context might necessitate a compromise in accuracy. Simultaneously, they underscore the importance of not solely relying on metrics like overall accuracy, as the most accurate models may not always be the most fair. Faced with the accuracy-fairness dilemma, the responsibility falls on the researchers to prioritize the more crucial aspect for their specific case and determine when and to what extent compromises are acceptable.

6.2 Practical recommendations

In light of the main findings, this section offers practical guidelines. As discussed in preceding chapters, the key motivation for introducing our proposed framework was to raise awareness among fellow researchers about the impact of class imbalance and shed-light into the factors they should consider when confronted with it.

The classification results obtained with imbalanced data in this study revealed a decline in classifiers' performance concerning minority classes, underscoring that, unlike the approach followed by many existing studies, the presence of class imbalance in the data should not be overlooked. This is especially crucial in applications where fairness, measured by accurately predicting both minority and majority classes, outweighs simply achieving the highest overall accuracy, as our findings suggest that the most accurate classifier is not always the fairest. Additionally, alongside class imbalance, the analysis of data structure also unveiled the presence of high overlap, identified in the literature as the most detrimental factor coexisting in imbalanced data. Consequently, we emphasize the importance of thorough data exploration, shedding light on the internal characteristics of the data and aiding in the informed selection of techniques to address the impact of challenging factors within the data. Furthermore, as demonstrated, different techniques yield varied results when applied to diverse cases. Therefore, researchers are advised to select techniques that are more suitable for their specific cases and ensure their correct implementation to avoid compromising the accuracy and behavioral outputs of their models. Additional, a crucial aspect is the proper evaluation of classification performance. Here, we want to emphasize the importance of avoiding the use of misleading metrics and focusing on metrics that provide insights into the realistic performance of the models. Such metrics, as mentioned earlier is this study can be the ones evaluating performance individually on each class.

Thus far it has become evident, that this study has exclusively focused on improving the predictive accuracy of the minority class, corresponding to transit. Accurately forecasting transit demand is crucial for ensuring sufficient transport availability, playing a paramount role in establishing an inclusive transport system. However, broader considerations are essential in transport planning. A recent investigation by Pot et al. (2020) shed light on the perspectives of individuals in specific regions of the Zeeland province regarding public transport accessibility. Their findings revealed a gap between perceived and actual accessibility, emphasizing the impact of individual experiences and local social norms. Consequently, we stress the importance of transport planners and policymakers taking into account a diverse array of factors when striving for designing a transport system that caters to all. Particularly in the realm of public transport systems, a profound understanding of the demographic groups utilizing them is vital for tailoring transit services to meet their specific needs. Therefore, it is crucial to closely consider the individuals constituting public transport demand and prioritize factors such as their perceptions of safety and competence in utilizing transit modes, the accessibility of information regarding the latter’s availability and operation, and travel costs. That way we can ensure comprehensive service for all population segments, leaving no one overlooked.

6.3 Main contributions

Having outlined our primary findings as well as provided guidelines for transport practitioners, this section provides a summary of the key contributions made by this study.

To begin with, this study introduces a comprehensive framework designed to identify and alleviate the impacts of class imbalance, similar to which - to the best of our knowledge- has not been presented by any other study within the existing literature. While H. Chen and Cheng (2023) have provided a detailed framework in their study, focusing on the evaluation of the classification performance following the

implementation of sampling techniques, their work lacks a methodology encompassing all the essential steps preceding the assessment of classifiers. On the contrary, our framework offers a structured methodology, including various aspects researchers should consider when working with imbalanced datasets, while it also provides the advantage of being implementable across various classifiers, datasets, and domains.

Furthermore, beyond addressing class imbalance, this study investigates and tackles the impact of class overlap. While numerous studies in the field of mode choice forecasting have addressed the impact of class imbalance (Hagenauer and Helbich, 2017; Qian et al., 2021; Rezai et al., 2021; Kashifi et al., 2022; Chaipanha & Kaewwichian, 2022; García-García et al., 2022; Narayanan & Antoniou, 2023; H. Chen & Cheng, 2023), none of these studies explores the potential existence of class overlap or other factors that might hinder classification performance when learning from imbalanced datasets. Conversely, several studies within other domains have simultaneously addressed both class imbalance and class overlap (L. Chen et al., 2016, Devi et al., 2019, Vuttipittayamongkol & Elyan, 2020). Therefore, this study contributes by drawing the attention of researchers within the transport field to the presence of various challenging factors within imbalanced data, which might be even more detrimental than imbalance itself (Garcia et al. 2007), and proposes techniques to address them.

Among the sampling techniques employed in this study, the SMOTE technique has consistently found application in various research studies across different domains, including the transport field. This research expands its contribution by demonstrating the effectiveness of the Neighborhood-based Undersampling and the Separation scheme techniques, adopted from the field of machine learning and not previously explored in transport studies. Furthermore, although these techniques have traditionally been used in binary classification tasks, this study extends their application to multiclass classification scenarios.

Finally, a notable contribution of this study is the integration of sampling techniques with the MNL model, filling a gap in the literature concerning the scarcity of research addressing class imbalance in discrete choice models. Exceptions include the mode choice studies by Rezaei et al. (2021), who tackled class imbalance through a combination of the Random Oversampling and Random Undersampling techniques, as well as the work of H. Chen and Cheng (2023), who employed various sampling techniques alongside an MNL model. Another relevant study is by Salas et al. (2023), where the oversampling of the minority class, when employing an MNL model, was accomplished through leveraging a Variational Autoencoder. Our study aligns with these works, particularly with regard to the enhancement of the classification performance of the minority class, emphasizing the significance of addressing class imbalance for improved model estimation. Moreover, distinguishing itself from the majority of studies employing discrete choice models that rely on Stated Preference (SP) data, this study stands out by utilizing Revealed Preferences (RP) data. This choice enables us to estimate and assess model performance using actual trip data, eliminating the potential hypothetical bias associated with SP data.

6.4 Limitations & Ideas for future research

Finally, while recognizing the contributions of this study, we also acknowledge its limitations, which are summarized in this section. Concurrently, we conclude this study by providing suggestions for future research.

Firstly, a limitation of this study arises from the reliance on the Performance Gap Metric, computed exclusively between the majority (Car) and minority (Transit) classes, to select the best-performing techniques across all examined scenarios. This approach confines the evaluation of classifiers' equitable

performance to only these two classes, overlooking the assessment of other classes, such as the Bike class in this study. Adopting this approach in multiclass scenarios may result in a classifier showing enhanced performance for the minority class but potentially exhibiting less accuracy for the remaining classes. To overcome this limitation, we suggest as a further enhancement to consider the classification performance of all classes and compare the results with those of this study to identify any notable discrepancies.

Moreover, an additional limitation in this study is the confined testing of the proposed framework to one specific case. To enhance the validity of our framework and thoroughly assess its effectiveness, we propose, as a potential avenue for future research, its application in diverse cases, extending even across other domains. In this study, the datasets for each model exhibited imbalance ratios of 5 and 6.8, respectively. However, literature reports studies with imbalance ratios ranging from nearly equal to 1, and even up to 10000 (Johnson & Khoshgoftaar, 2019). Therefore, a future direction could involve the implementation of the proposed framework in datasets with varying Imbalance Ratios to examine how the effectiveness of the sampling techniques is possibly altered. Additionally, another idea could be to validate our models using data from subsequent years to evaluate their temporal robustness and generalization capabilities. This form of validation, referred to as external validation (Parady et al., 2021), assesses the temporal transferability of a model, indicating how well it generalizes using data from a different period. Concerning the ODiN data, it could be worthwhile to consider validating our models using data from the year 2022, while avoiding the years 2019-2021 during which the data distribution might have undergone changes due to the consequences of the pandemic with regard to mobility patterns.

Lastly, another idea could be to incorporate more advanced techniques into our specific case. For instance, delving into the utilization of generative models for synthetic data generation could be an avenue worth exploring. Although Salas et al. (2023) previously attempted this, in their study model evaluation was restricted to predictive accuracy, and therefore our approach could extend to comprehensively assess behavioral outputs as well. Additionally, delving into feature engineering methods may prove beneficial in addressing class overlap by improving class separability.

References

Ali, H., Salleh, M. N. M., Saedudin, R. R., Hussain, K., & Mushtaq, M. F. (2019a). Imbalance class problems in data mining: a review. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1552. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Alogogianni, E., & Virvou, M. (2023). Handling class imbalance and class overlap in machine learning applications for undeclared work prediction. Electronics, 12(4), 913. https://doi.org/10.3390/electronics12040913

Böcker, L., Prillwitz, J., & Dijst, M. (2013). Climate change impacts on mode choices and travelled distances: a comparison of present with 2050 weather conditions for the Randstad Holland. Journal of Transport Geography, 28, 176–185.

Böcker, L., Van Amen, P., & Helbich, M. (2016a). Elderly travel frequencies and transport mode choices in Greater Rotterdam, the Netherlands. Transportation, 44(4), 831–852. https://doi.org/10.1007/s11116-016-9680-z

Brathwaite, T., Vij, A., & Walker, J. L. (2017). Machine Learning meets microeconomics: the case of decision trees and discrete choice. arXiv (Cornell University). https://arxiv.org/pdf/1711.04826.pdf

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

CBS. (2019). Indicator risico op vervoersarmoede. Retrieved from https://www.cbs.nl/nl-nl/achtergrond/2019/42/indicator-risico-op-vervoersarmoede

Chaipanha, W., & Kaewwichian, P. (2022). Smote vs. Random Undersampling for Imbalanced Data - Car Ownership Demand Model. Komunikácie, 24(3), D105–D115. https://doi.org/10.26552/com.c.2022.3.d105-d115

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, L., Fang, B., Shang, Z., & Tang, Y. (2016). Tackling class overlap and imbalance problems in software defect prediction. Software Quality Journal, 26(1), 97–125. https://doi.org/10.1007/s11219-016-9342-6 (In text: L. Chen et al., 2016)

Chen, H., & Cheng, Y. (2023). Travel mode choice prediction using imbalanced machine learning. IEEE Transactions on Intelligent Transportation Systems, 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 Behaviour and Society, 14, 1–10. https://doi.org/10.1016/j.tbs.2018.09.002

Devi, D., Biswas, S. K., & Purkayastha, B. (2019). Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. Connection Science, 31(2), 105–142. https://doi.org/10.1080/09540091.2018.1560394 (Devi et al., 2019)

Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. Journal of Network and Innovative Computing, 1(2013), 332-340.

Fernández A, García S, Galar M, Prati R, Krawczyk B, Herrera F (2018) Data Intrinsic Characteristics. Springer, Cham, pp 253–277

Flipo A., Ortar N., Sallustio M. (2023). Can the transition to sustainable mobility be fair in rural areas? A stakeholder approach to mobility justice. Transport Policy, vol. 139, pp. 136-143, https://doi.org/10.1016/j.tranpol.2023.06.006.

García, V., Mollineda, R. A., & Sánchez, J. S. (2007). On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 11(3–4), 269–280. https://doi.org/10.1007/s10044-007-0087-5

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge Based Systems, 98, 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

García-García, J. C., García-Ródenas, R., López-Gómez, J. A., & Martín-Baos, J. Á. (2022). A comparative study of machine learning, deep neural networks and random utility maximization models for travel mode choice modelling. Transportation Research Procedia, 62, 374–382. https://doi.org/10.1016/j.trpro.2022.02.047

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

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: a new Over-Sampling method in imbalanced Data sets learning. In Lecture Notes in Computer Science (pp. 878–887). https://doi.org/10.1007/11538059_91

Hananel, R., & Berechman, J. (2016). Justice and transportation decision-making: The capabilities approach. Transport Policy, 49, 78–85. https://doi.org/10.1016/j.tranpol.2016.04.005

He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE. https://doi.org/10.1109/ijcnn.2008.4633969

Heinen, E., Maat, K., & Van Wee, B. (2011). The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances. Transportation Research Part D: Transport and Environment, 16(2), 102–109. https://doi.org/10.1016/j.trd.2010.08.010

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. Journal of Choice Modelling, 38, 100221. https://doi.org/10.1016/j.jocm.2020.100221

Jafino, B. A. (2021). An equity-based transport network criticality analysis. Transportation Research Part A: Policy and Practice, 144, 204–221. https://doi.org/10.1016/j.tra.2020.12.013

Japkowicz, N. (2001). Concept-Learning in the presence of Between-Class and Within-Class imbalances. In Lecture Notes in Computer Science (pp. 67–77). https://doi.org/10.1007/3-540-45153-6_7 (In text: Japkowicz (2001))

Japkowicz Nathalie, & StephenShaju. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis. https://doi.org/10.5555/1293951.1293954

Johnson, J., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0192-5

Kashifi, M. T., Jamal, A., Kashefi, M. S., Almoshaogeh, M., & Rahman, S. M. (2022c). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. Travel Behaviour and Society, 29, 279–296. https://doi.org/10.1016/j.tbs.2022.07.003

Kasraian, D., Maat, K., & van Wee, B. (2016). Development of rail infrastructure and its impact on urbanization in the Randstad, the Netherlands. Journal of Transport and Land Use, 9(1), 151- 170

Kemperman, A., & Timmermans, H. H. (2012). Environmental correlates of active travel behavior of children. Environment and Behavior, 46(5), 583–608. https://doi.org/10.1177/0013916512466662

Khamis, H. J. (2008). Measures of Association: how to choose? Journal of Diagnostic Medical Sonography, 24(3), 155–162. https://doi.org/10.1177/8756479308317006 (In text: Khamis (2008))

Kim, E. (2021). Analysis of travel mode choice in Seoul using an interpretable machine learning approach. Journal of Advanced Transportation, 2021, 1–13. https://doi.org/10.1155/2021/6685004

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krčál, O., Peer, S., Staněk, R., & Karlínová, B. (2019). Real consequences matter: Why hypothetical biases in the valuation of time persist even in controlled lab experiments. Economics of Transportation, 20, 100138. https://doi.org/10.1016/j.ecotra.2019.100138

Lango, M., & Stefanowski, J. (2022). What makes multi-class imbalanced problems difficult? An experimental study. Expert Systems With Applications, 199, 116962. https://doi.org/10.1016/j.eswa.2022.116962

La Paix Puello, L. C., Cherchi, E., & Geurs, K. (2020). Role of perception of bicycle infrastructure on the choice of the bicycle as a train feeder mode. International Journal of Sustainable Transportation, 15(6), 486–499. https://doi.org/10.1080/15568318.2020.1765223

Li, X., Wang, Y., Wu, Y., Chen, J., & Zhou, J. (2021). Modeling Intercity Travel Mode Choice with Data Balance Changes: A Comparative Analysis of Bayesian Logit Model and Artificial Neural Networks. Journal of Advanced Transportation, 2021, 1–22. https://doi.org/10.1155/2021/9219176

Limtanakool, N., Dijst, M., & Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. Journal of Transport Geography, 14(5), 327–341. https://doi.org/10.1016/j.jtrangeo.2005.06.004

Lucas, K. (2012). Transport and social exclusion: Where are we now? Transport Policy, 20(1), 105-133

Mackett, R., & Thoreau, R. (2015). Transport, social exclusion and health. Journal of transport and health, 2(4), 610–617. https://doi.org/10.1016/j.jth.2015.07.006

Makarewicz, C., Dantzler, P. A., & Adkins, A. (2020). Another look at location affordability: understanding the detailed effects of income and urban form on housing and transportation expenditures. Housing Policy Debate, 30(6), 1033–1055. https://doi.org/10.1080/10511482.2020.1792528

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in econometrics . New York, NY: Academic Press.

Napierała, K., Stefanowski, J., & Wilk, S. (2010). Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In Lecture Notes in Computer Science (pp. 158–167). https://doi.org/10.1007/978-3-642-13529-3_18

Narayanan, S., & Antoniou, C. (2023). Shared mobility services towards Mobility as a Service (MaaS): What, who and when? Transportation Research Part A: Policy and Practice, 168, 103581. https://doi.org/10.1016/j.tra.2023.103581

Omrani, H. (2015). Predicting travel mode of individuals by machine learning. Transportation Research Procedia, 10, 840–849. https://doi.org/10.1016/j.trpro.2015.09.037

Parady, G. T., Ory, D. T., & Walker, J. L. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. Journal of Choice Modelling, 38, 100257. https://doi.org/10.1016/j.jocm.2020.100257

PBL (2019). Feiten en cijfers over krimp. https://www.pbl.nl/onderwerpen/krimp/feiten-en-cijfers (Accessed 2 May 2019)

Pereira, R. H. M., Schwanen, T., & Banister, D. (2016). Distributive justice and equity in transportation. Transport Reviews, 37(2), 170–191. https://doi.org/10.1080/01441647.2016.1257660

Pot, F. J., Koster, S., Tillema, T., & Jorritsma, P. (2020). Linking experienced barriers during daily travel and transport poverty in peripheral rural areas: the case of Zeeland, the Netherlands. European Journal of Transport and Infrastructure Research, 20(3). https://doi.org/10.18757/ejtir.2020.20.3.4076 (In text: Pot et al. (2020))

Prati, R. C., Batista, G. E. a. P. A., & Monard, M. C. (2004). Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior. In Lecture Notes in Computer Science (pp. 312–321). https://doi.org/10.1007/978-3-540-24694-7_32

Prematunga, R. (2012). Correlational analysis. Australian Critical Care, 25(3), 195–199. https://doi.org/10.1016/j.aucc.2012.02.003

Qian, Y., Aghaabbasi, M., Ali, M., Alqurashi, M., Salah, B., Zainol, R., Moeinaddini, M., & Hussein, E. E. (2021). Classification of imbalanced travel mode choice to work data using adjustable SVM model. Applied Sciences, 11(24), 11916. https://doi.org/10.3390/app112411916

Rasouli, S., & Timmermans, H. H. (2014). Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates. European Journal of Transport and Infrastructure Research, 14(4), 412–424. https://doi.org/10.18757/ejtir.2014.14.4.3045

Rezaei, S., Khojandi, A., Haque, A. M., Brakewood, C., Jin, M., & Cherry, C. (2021). Performance evaluation of mode choice models under balanced and imbalanced data assumptions. Transportation Letters: The International Journal of Transportation Research, 14(8), 920–932. https://doi.org/10.1080/19427867.2021.1955567

Salas, P., De La Fuente, R., Astroza, S., & Carrasco, J. M. C. (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. https://doi.org/10.1016/j.eswa.2021.116253

Salas, P., De La Fuente, R., Astroza, S., & Carrasco, J. A. (2023). Improving the Predictive Performance of the Multinomial Logit Model in Travel Mode Choice Application Using Conditional Variational Autoencoder. Elsevier.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Domingues, I. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228–253. https://doi.org/10.1016/j.inffus.2022.08.017

Schwanen, T., Dijst, M., & Dieleman, F. M. (2001b). Leisure trips of senior citizens: determinants of modal choice. Tijdschrift Voor Economische En Sociale Geografie, 92(3), 347–360. https://doi.org/10.1111/1467-9663.00161

Scott, D. M., & Horner, M. W. (2008). The role of urban form in shaping access to opportunities: An exploratory spatial data analysis. Journal of Transport and Land Use, 1(2), 89–119. http://www.jstor.org/stable/26201615

Sekhar, C. R., Minal, & Errampalli, M. (2016). Mode choice analysis using random forest decision trees. Transportation Research Procedia, 17, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119

Shiftan, Y. & Bekhor, S. (2002). Investigating individual's perception of auto travel cost. 29. 151-166. (Shiftan & Bekhor, 2002)

Significance. (2021). Growth Model 4: The new Dutch passenger transport model. European transport conference 2021. (Significance, 2021)

Significance. (2023). Values of Time, Reliability and Comfort in the Netherlands 2022: New values for passenger travel and freight transport. https://www.kimnet.nl (Significance, 2023)

Stanley, J., & Stanley, J. (2017). The importance of transport for social inclusion. Social Inclusion, 5(4), 108–115. https://doi.org/10.17645/si.v5i4.1289

Ton, D., Duives, D. C., Cats, O., Hoogendoorn-Lanser, S., & Hoogendoorn, S. P. (2019). Cycling or walking? Determinants of mode choice in the Netherlands. Transportation Research Part A: Policy and Practice, 123, 7–23. https://doi.org/10.1016/j.tra.2018.08.023

Train, K., (2003). Discrete Choice Methods with Simulation. Cambridge: University Press

Trappenberg, T., & Back, A. (2000). A classification scheme for applications with ambiguous data. IEEE. https://doi.org/10.1109/ijcnn.2000.859412

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., and Walker, J. (2022). Choice modelling in the age of machine learning-discussion paper. Journal of Choice Modelling, 42:100340.

Van Der Maaten, L., & Hinton, G. E. (2008). Visualizing data using T-SNE. Journal of Machine Learning Research, 9(86), 2579–2605. http://isplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf

Van Wee, B., & Geurs, K. (2011). Discussing equity and social exclusion in accessibility evaluations. European Journal of Transport and Infrastructure Research, 11(4), 350–367. https://doi.org/10.18757/ejtir.2011.11.4.2940

Van Wee, B., & Roeser, S. (2013). Ethical Theories and the Cost–Benefit Analysis-BasedEx AnteEvaluation of transport Policies and Plans. Transport Reviews, 33(6), 743–760. https://doi.org/10.1080/01441647.2013.854281

Versteijlen, M., Van Wee, B., & Wals, A. (2021). Exploring sustainable student travel behaviour in The Netherlands: balancing online and on-campus learning. International Journal of Sustainability in Higher Education, 22(8), 146–166. https://doi.org/10.1108/ijshe-10-2020-0400

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Information Sciences, 509, 47–70. https://doi.org/10.1016/j.ins.2019.08.062

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. Transportation Research Record, 2672(47), 35–45. https://doi.org/10.1177/0361198118773556

Xiong, H., Li, M., Jiang, T., & Zhao, S. (2013). Classification Algorithm based on NB for Class Overlapping Problem. Applied Mathematics & Information Sciences, 7(2L), 409–415. https://doi.org/10.12785/amis/072l05

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. https://doi.org/10.1016/j.tbs.2020.02.003

Zheng, Y., Wang, Q., Zhuang, D., Wang, S., & Zhao, J. (2023). Fairness-Enhancing deep learning for Ride-Hailing demand prediction. IEEE Open Journal of Intelligent Transportation Systems, 4, 551–569. https://doi.org/10.1109/ojits.2023.3297517

Zhu, R., Guo, Y., & Xue, J. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognition Letters, 133, 217–223. https://doi.org/10.1016/j.patrec.2020.03.004

Appendix

1. Spatial Analysis – Addressing the ''within-class'' imbalance

As highlighted in STEP 5 of the proposed framework (Chapter 3), the presence of "within-class" imbalance is acknowledged as one of the difficulty factors in imbalanced datasets, potentially leading to reduced performance in classification tasks. Unlike "between-class" imbalance, which involves the disparity in the number of instances between different classes, "within-class" imbalance, according to Japkowicz (2001), refers to the presence of sub-clusters with varying number of instances within a class. Encountering imbalance, particularly within the minority class and in conjunction with between-class imbalance, can contribute to increased misclassification rates in classification algorithms. While this study predominantly addresses ''between-class" imbalance in combination with class overlap, this section briefly integrates an analysis considering "within-class" imbalance – an aspect that has not received the same level of attention in the literature compared to "between class" imbalance.

Specifically, we focused on the observed imbalance within the transit class concerning the quantity of trips conducted per Dutch province, assessed according to the origin of the trips. As illustrated in Figure 16, it is evident that the majority of public transport trips take place within the Randstad area, particularly in the provinces of South-Holland (32.3%) and North-Holland (27.6%). In contrast, only a minimal percentage of trips is observed in the peripheral and more rural areas, with the province of Zeeland recording the lowest percentage, accounting for only 0.6 % of the total transit trips. Upon evaluating the prediction results of the Random Forest model (Table 29), we noted that the province of Zeeland exhibited the least favorable prediction performance, with only 50% of its trips correctly classified. In contrast, other peripheral provinces demonstrated relatively good accuracy, which in certain cases surpassed the accuracy observed for larger urban centers.

Accurate mode choice predictions are in general of paramount importance given evolving factors such as population growth and aging. Specifically, for the Netherlands, population is expected to grow from 17.3 to 18.5 million inhabitants by 2050, with the growth mostly concentrated in the Randstad area, while a population decline is expected in the outskirts of the Netherlands. Simultaneously, in the upcoming years, all regions will face the challenge of an aging population, a phenomenon expected to accelerate more rapidly in the outskirts than in the Randstad (PBL, 2019). Due to their aging and decreasing populations, coupled with a decline in services and less developed transport systems due to lower demand, peripheral and rural areas are typically more prone to experiencing transport poverty (Lucas et al., 2012). This vulnerability arises due to increased distances to reach essential services and activities, coupled with diminished travel opportunities beyond car usage, potentially leading to the social exclusion of individuals due to physical or financial reasons (Pot et al., 2020). Therefore, accurate predictions of travel demand are particularly crucial for these regions.

Accessibility levels in Dutch peripheral areas may exceed those in other European rural regions due to higher population densities and dense road networks with close links to urban centers. Nonetheless, specific peripheral Dutch regions, such as those within the province of Zeeland, albeit on a smaller scale, still demonstrate patterns of population decline and service reduction similar to other peripheral regions in Europe (Pot et al., 2020).

Recognizing the significance of precise travel forecasts for less urban regions and taking inspiration from the research of Pot et al. (2020), who delved into the mechanisms behind experienced transport poverty in regions within the province of Zeeland by assessing the population's perception of accessibility, we

sought to investigate whether addressing both "within-class" and "between-class" imbalance could enhance predictive accuracy for the transit trips of the province of Zeeland, aiming to better capture its travel demand.

To conduct our analysis, we followed the methodology outlined by Japkowicz et al. (2001). In her study, Japkowicz et al. (2001) explored the impact of concurrently addressing ''between-class'' and ''within-class'' imbalance on the error rate of each sub-cluster within the minority class. In specific, she balanced the classes by oversampling the minority class, ensuring that all of its sub-clusters had an equal size. To increase the number of minority samples, Random Oversampling was employed. This approach resulted in a lower error rate for each sub-cluster compared to the scenario where balancing the majority and minority classes occurred without addressing the imbalance within the minority class.

In our analysis, we focused on the binary classification task involving the Car and Transit classes. Within the transit class, 12 sub-clusters were defined corresponding to the 12 Dutch provinces. Both Random Oversampling and SMOTENC techniques were utilized to augment the number of transit trips for each province, ensuring that the total number of transit samples equaled the number of car samples. The specific steps undertaken are depicted in Figure 15.

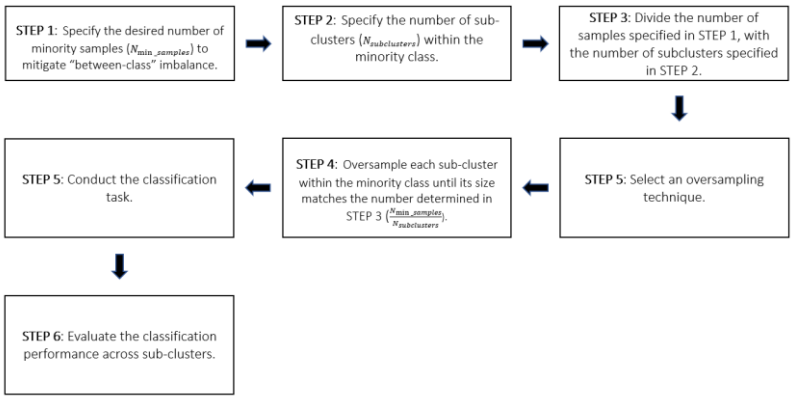

Figure 15. Approach adopted to tackle both ''between-class'' and ''within-class'' imbalance, following the methodology proposed by Japkowicz et al. (2001). The term ''between-class'' imbalance denotes the existence of classes with varying numbers of instances within a dataset. In contrast, "within-class" imbalance indicates the presence of sub-clusters with varying numbers of instances within a specific class.

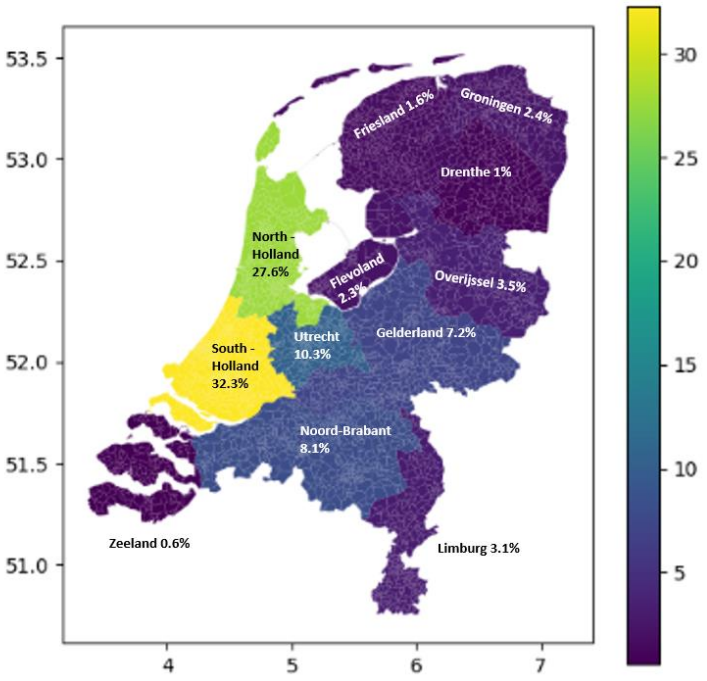

Figure 16. Map of the Netherlands indicating the proportion of transit journeys per province based on the trips' origins. The Netherlands comprises 12 provinces, including South Holland, North Holland, Utrecht, Zeeland, Flevoland, North Brabant, Limburg, Overijssel, Drenthe, Friesland, and Gelderland. Among these provinces, South-Holland stands as the most populous, while Zeeland is recognized as the least populous Dutch province.

Table 29 displays the outcomes of our analysis, focusing on the sensitivity of individual subclusters, corresponding to Dutch provinces, within the transit class. The results are showcased for various models, including the baseline model (using imbalanced data), the model after implementing the SMOTENC technique to targe "between-class" imbalance, the model after implementing the SMOTENC technique to address both "between-class" and "within-class" imbalances, and the model after implementing the Random Oversampling technique also aimed at mitigating both "between-class" and "within-class" imbalances.

Regarding the Zeeland province, which is of particular interest for the present analysis, we observe that employing the SMOTENC technique to address both "within-class" and "between-class" imbalance does not result in a substantial improvement in sensitivity compared to using SMOTENC solely for mitigating the "between-class" imbalance, as previously executed in this study. Conversely, when the Random Oversampling technique is applied to tackle both types of imbalance, sensitivity exhibits a more significant enhancement. The minimal improvement observed after implementing the SMOTENC technique for both types of imbalance could be attributed to the province's very small sample size and particularly to the position of its samples in the feature space, which might make it more challenging to generate meaningful and truly informative synthetic samples.

Province	Representation in the training set (baseline model)	Baseline model (imbalanced data)	Representation in the training set (SMOTENC, between class imbalance)	SMOTENC (between-class imbalance)	Representation in the training set (ROS/SMOTENC, between class imbalance & within-class imbalance)	SMOTENC (both between-class and within-class imbalance	Random Oversampling (both between-class and within-class imbalance)
1.Groningen	2.4%	79 % (0.016)	1.5%	82% (0.001)	8.3%	84% (0.013)	83% (0.011)
2.Friesland	1.6%	72 % (0.014)	0.8%	75% (0.012)	8.3%	77% (0.012)	75% (0.009)
3.Drenthe	1 %	70 % (0.019)	0.4%	70% (0.019)	8.3%	75% (0.015)	71% (0.012)
4.Overijssel	3.6%	61 % (0.009)	2.3%	73% (0.014)	8.3%	74% (0.013)	69% (0.007)
5.Flevoland	2.3%	59 % (0.011)	1%	67% (0.015)	8.3%	73% (0.004)	68% (0.024)
6.Gelderland	7.2%	68 % (0.007)	5.4%	73% (0.009)	8.3%	73% (0.004)	73% (0.011)
7.Utrecht	10%	61 % (0.009)	8.1%	75% (0.005)	8.3%	71% (0.004)	67% (0.005)
8.Noord-Holland	27.5%	67 % (0.006)	31.6%	80% (0.001)	8.3%	73% (0.004)	68% (0.002)
9.Zuid-Holland	32.6%	68 % (0.002)	40.4%	81% (0.002)	8.3%	71% (0.004)	69% (0.006)
10.Zeeland	**0.6%**	**50 % (0)**	**0.3%**	**53% (0.027)**	**8.3%**	**54% (0.022)**	**56% (0.022)**
11.Noord-Brabant	8%	72 % (0.01)	6.4%	78% (0.006)	8.3%	79% (0.01)	76% (0.008)
12.Limburg	3.1%	64 % (0.008)	1.7%	67% (0.013)	8.3%	71% (0.016)	68% (0.011)
Recall transit class	-	67 % (0)	-	78% (0)	-	73.2% (0.447)	69.8% (0.447)
Recall car class	-	98.4 % (0.548)	-	94% (0)	-	97% (0)	98% (0)

Table 29. Sensitivity outcomes pertaining to the prediction of transit trips across Dutch provinces are presented. From left to right, the results are displayed for the baseline model (using imbalanced data), the model after implementing SMOTENC to address "between-class" imbalance, the model after implementing SMOTENC to address both "between-class" and "within-class" imbalance, and lastly, the model after implementing Random Oversampling to tackle both types of imbalance.

In the final step of our analysis, we expanded our investigation by consolidating the sub-clusters within the minority class into two broader groups—namely, the Randstad and non-Randstad sub-clusters. This decision was driven by the observation that the majority of transit trips occur within the Randstad area (72.5%), where transit systems are already well developed, while only a small number of trips take place in the remaining provinces. Given the lower connectivity in these provinces compared to the Randstad region, precise predictions for them are of heightened importance (Kasraian et al., 2016).

Table 30 presents the sensitivity results for the Randstad and non-Randstad regions. When addressing the "within-class" imbalance, the application of the SMOTENC technique results in a modest increase in sensitivity for the non-Randstad provinces compared to exclusively addressing the "between-class" imbalance. Although not substantial, this increase might still remain noteworthy, particularly considering the lower development of transit systems in these provinces.

In conclusion, it is important to note that in this analysis, we exclusively experimented with the Random Oversampling and SMOTENC techniques. However, we acknowledge that the implementation of alternative techniques might have resulted in more favorable outcomes.

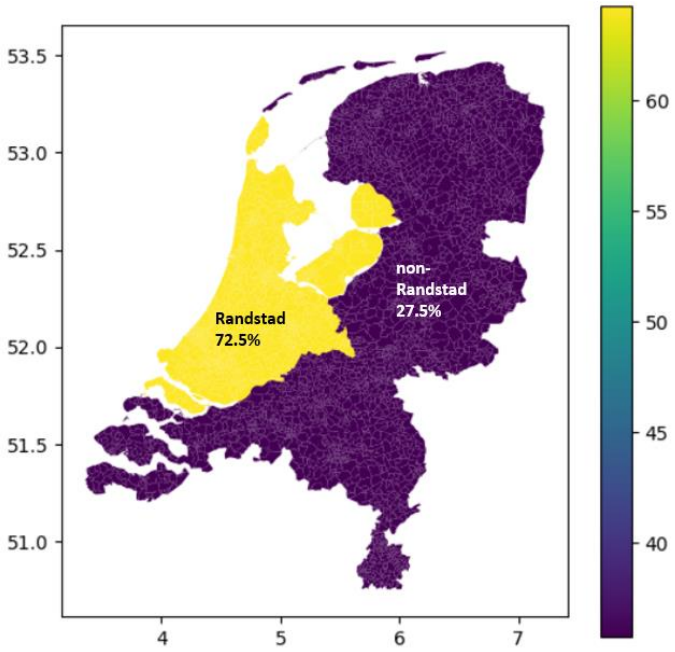

Figure 17. Map of the Netherlands depicting the percentage of public transport journeys originating in both the Randstad and non-Randstad regions. The Randstad area encompasses four Dutch provinces: South Holland, North Holland, Utrecht, and Flevoland.

	Recall	
	Randstad region	**Non-Randstad region**
Baseline model (imbalanced data)	66.3%	68.7%
SMOTENC, between class imbalance)	79.3%	74%
SMOTENC (between-class imbalance & within-class imbalance)	71.8%	75.5%
Random Oversampling (between-class imbalance & within-class imbalance)	68.3%	73.3%

Table 30. Sensitivity outcomes concerning the prediction of transit trips in the Randstad and non-Randstad regions. Results are displayed for the baseline model (using imbalanced data), the model after implementing SMOTENC to address "between-class" imbalance, the model after implementing SMOTENC to address both "between-class" and "within-class" imbalance, and lastly, the model after implementing Random Oversampling to tackle both types of imbalance.

2. Random Forest-Feature Importance

In this section we present the top 10 most important features (Table 31) identified in the binary and multiclass classification tasks using the Random Forest model. Feature importance serves as a metric, uncovering the significance of individual features in improving the accuracy of the model's predictions. As detailed earlier in this study, each feature's importance is assessed by calculating the average reduction in impurity caused by splitting on that feature across all trees in the forest. Features that exhibit a notable decrease in impurity during tree splits are considered more crucial and can be selected during feature selection, contributing to enhanced predictive accuracy.

Upon reviewing the results presented in Table 32 and Table 33, it is evident that, in both scenarios and particularly in binary classification, using only the top 10 most important features yields results very close to the ones obtained when employing the baseline model, incorporating all explanatory features. When working with a high-dimensional dataset, opting for including only the most important features in a classification task can prove beneficial, as while this choice might result in a slight decrease in the predictive accuracy, it can save computational time. Nevertheless, in our specific scenario, we are not dealing with such a dataset; therefore, we have opted to retain all features for our analysis.

	Top 10 Most Important Features
Binary case	Driving license, Car ownership, Transit cost, Car cost, Car trip duration, Transit trip duration, Urbanity level(trip origin), Possession of OV card, Urbanity level(trip destination), Age
Multiclass case Car-Transit-Cycle	Bike trip duration, Driving License, Transit Cost, Car trip duration, Car Cost, Transit Trip Duration, Age, Car ownership, Urbanity level (trip origin), Urbanity level (trip destination)

Table 31. Top 10 most important features identified in the binary and multiclass classification tasks employing the Random Forest model. The significance of each feature is assessed based on the average reduction in impurity across all splits within the forest where the feature is employed.

	Precision		Recall		F1-score		Accuracy	Balanced accuracy
	Car	PT	Car	PT	Car	PT	Total	Total
Baseline model	94 (0)	90 (0)	98.4 (0.55)	67 (0)	96.1 (0.26)	77 (0)	93 (0)	83 (0)
Baseline model-Top 10 most important features	94 (0)	83.8 (0.45)	97 (0)	67.4 (0.55)	95.5 (0)	74.8 (0.45)	92 (0)	82 (0)

Table 32. Results from the binary classification task using the Random Forest model. The top row of the table showcases classification results with utilizing all explanatory features, while the second row illustrates classification outcomes considering only the top 10 most important features.

	Precision			Recall			F1-score		
	Car	AM	Transit	Car	AM	Transit	Car	AM	Transit
Baseline model	81 (0)	83 (0)	81.6 (0.55)	93 (0)	72.2 (0.45)	52.2 (0.45)	86.6 (0)	77.2 (0.25)	63.6 (0.55)
Baseline model-Top 10 most important features	78.4 (0.55)	75 (0)	73.2 (0.45)	88 (0)	68 (0)	51 (0)	82.9 (0.31)	71.3 (0)	60 (0)

	Accuracy	Balanced accuracy
	Total	Total
Baseline model	81.6 (0.55)	72.8 (0.45)
Baseline model-Top 10 most important features	77 (0)	69 (0)

Table 33. Results from the multiclass classification task employing the Random Forest model. The first row of the table illustrates outcomes when employing all explanatory features, while the second row showcases results when only the top 10 most important features are considered.

3. t-tests

In the subsequent two tables, we display the computed values derived from the t-test, assessing the statistical significance of the coefficients utilized in the MNL model.

	TT Car	TT PT	TC Car	TC PT	ASC Car
Baseline model	-131.6	-41.6	-88.3	-455.5	208.6
SMOTENC (k=5, N = 30%)	-138.6	-38.1	-94.4	-50.1	185,6
SMOTENC (k =5, N = 50%)	-143.7	-34.6	-202.6	-40.45	94.3
SMOTENC (k=5, N = 100%)	-75.43	-28.8	-94.1	-128.17	42.54
NBU (k = 3)	-23.75	-28.58	-39.6	-35.12	90.76
NBU (k = 5)	-48.49	-33.7	-90.5	-195.3	199.19

Table 34. t values for the binary classification task employing the MNL model. For a degree of freedom (df) equal to 2 and a 95% confidence level, the critical t values are ± 4.3 (two-tailed test).

	TT Car	TT PT	TT Bike	TC Car	TC PT	ASC Car	ASC PT
Baseline model	-41.6	-18.18	-141.85	-84.87	-58.54	-10.68	-149.76
SMOTENC (k=5, N = 30%)	-51.1	-34.64	-155.88	-90.93	-193.12	-5.85	-330.5
SMOTENC (k =5, N = 50%)	-121.24	-34.64	-161.95	-96.13	-57.59	-0.02	-84.89
SMOTENC (k=5, N = 100%)	-32.47	-7.79	-57.74	-209.6	-40.2	5.92	-25.13
NBU (k = 3)	-95.26	-41.57	-788.08	-214.77	-363.75	-5.677	-233.73
NBU (k = 5)	-42.47	-42.26	-167.74	-120.37	-36.63	-7.768	-242.89

Table 35. t values for the multiclass classification task employing the MNL model. For a degree of freedom (df) equal to 2 and a 95% confidence level, the critical t values are ± 4.3 (two-tailed test).