Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084555

ARTICLE

Hyperparameter Optimisation and Comparative Analysis of Machine Learning Models for Travel Mode Choice Prediction

Mujahid Ali*

Department of Transport Systems, Traffic Engineering and Logistics, Faculty of Transport and Aviation Engineering, Silesian University of Technology, Krasinskiego 8 Street, Katowice, Poland

*Corresponding Author: Mujahid Ali. Email: mali@polsl.pl

Received: 24 April 2026; Accepted: 10 June 2026; Published: 23 July 2026

ABSTRACT: Understanding the determinants of travel mode choice (TMC) in urban contexts is essential for effective transport planning and policy development. Past studies predominantly employed traditional discrete choice models because of their simplicity, diversity, and high interpretability; however, they rely on restrictive assumptions. Although machine learning (ML) techniques have shown promising predictive capabilities, comparative assessments of traditional and ML approaches, particularly considering hyperparameter optimisation, remain limited. This study addresses this gap by comparing a traditional model with four ML algorithms: decision tree (DT), random forest (RF), support vector machine (SVM), and k-nearest neighbour (KNN). In addition, systematic hyperparameter optimisation is performed to evaluate its impact on predictive performance relative to default model settings. Feature importance analysis is also conducted to identify the key determinants of TMC. The analysis is based on a multi-dimensional, three-week household time-use and activity diary dataset comprising 508 individuals from 191 households in the Bandung Metropolitan Area, Indonesia. The results demonstrate that ML models outperform traditional methods, while hyperparameter optimisation substantially improves model performance across all considered algorithms against default models. Notably, the KNN model exhibits a 16.67% increase in accuracy, followed by the SVM model with an 11.15% improvement. Among the optimised evaluated models, SVM achieves the best overall performance, with a macro-averaged accuracy of 0.588 and a precision of 0.591. Feature importance analysis reveals that total travel time is the most influential determinant of TMC. These findings highlight the importance of model tuning and hyperparameter optimisation in ML-based TMC prediction and provide insights into the factors shaping travel behaviour. The outcomes can support more informed decision-making in urban transport planning and policy formulation.

KEYWORDS: Travel mode choice; hyperparameter optimisation; random forest; travel behaviour; sustainable transportation

1 Introduction

There have been major environmental and societal issues like traffic congestion, air pollution, and energy consumption as a result of the fast urbanisation trend and the rising need for transportation systems. The transport sector is the second-largest CO_2 emission sector after energy, which contributes 23% of CO_2, whereas land transport contributes over 75% of GHG in the transport sector [1]. Consequently, promoting sustainable transportation systems has become a major priority for policymakers and researchers worldwide, where replacing motorized transport (MT) with active transport (AT) can achieve 14.52%–17%

GHG emission reduction. Therefore, accurately predicting travel mode choice (TMC) is essential for developing effective transportation policies and intelligent transportation systems (ITS) that support sustainable urban mobility.

Traditionally, discrete choice models (DCMs), particularly the Multinomial Logit (MNL) model, have been extensively used to analyse TMC behaviour and examine the relationship between travel outcomes and socio-demographic, economic, built-environment, and trip-related variables [2,3]. Previous studies identified several important determinants of travel behaviour, including age, trip purpose, health conditions, land-use patterns, road infrastructure, accessibility to PT, household vehicle ownership, and transportation costs [4]. Ali showed that parking prices significantly influence TMC [5]. Although conventional models provide strong theoretical interpretability, they often rely on restrictive assumptions regarding linearity, independence, and predefined utility structures, which may limit their predictive capability in complex travel behaviour analyses.

With advancements in computational techniques and data-driven analytics, machine learning (ML) algorithms have increasingly been adopted for TMC prediction because of their ability to capture nonlinear relationships and complex interactions among variables with fewer statistical assumptions. Recent studies reported that ML approaches frequently outperform traditional DCMs in terms of predictive accuracy and classification performance [6]. Among the commonly used ML algorithms, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) have demonstrated strong predictive capability and adaptability across transportation applications [7–9]. Comparative studies between DCMs and ML models further revealed the superior performance of ML techniques in predicting travel behaviour and mode choice outcomes [10–12]. Le and Teng performed a comparative analysis of DCM and ML algorithms and professed that ML algorithms give high accuracy and precision, where passengers are more sensitive to the last-mile trip [13].

Despite these advantages, several methodological challenges remain in ML-based TMC prediction. One major issue is class imbalance, where certain travel modes, particularly MT, dominate the dataset and bias model performance toward majority classes. Imbalanced datasets can significantly reduce the predictive reliability of both traditional and ML models, especially for minority travel modes such as public transport (PT) and AT [14]. To address this issue, recent studies employed the Synthetic Minority Oversampling Technique (SMOTE) to balance transportation datasets and improve model generalisation [15]. Another important challenge is hyperparameter selection, as the predictive performance of ML algorithms is highly sensitive to parameter configurations [16]. Previous studies demonstrated that hyperparameter optimisation techniques, including Grid Search and Bayesian Optimisation (BO), can substantially improve model accuracy, precision, and overall classification performance [17,18].

In addition to predictive performance, model interpretability has become increasingly important in transportation research. Understanding the relative importance of explanatory variables can help policymakers identify the key determinants influencing TMC decisions and formulate targeted sustainable transportation strategies. Recent studies, therefore, integrated interpretable ML techniques, such as feature importance analysis, permutation importance, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME), to improve the transparency and interpretability of ML-based transportation models.

Although previous studies compared traditional and ML-based approaches for TMC prediction, limited research has simultaneously addressed dataset imbalance, baseline vs. optimised ML models, and feature importance analysis within a unified framework, particularly in the context of the Bandung Metropolitan Area (BMA), Indonesia. Moreover, few studies have systematically evaluated the influence of hyperparameter optimisation on multiple ML algorithms for TMC prediction using balanced datasets.

Therefore, this study performs a comparative analysis of several ML algorithms, including RF, DT, SVM, and KNN, for TMC prediction. First, the Synthetic Minority Oversampling Technique (SMOTE) is applied to address class imbalance in the dataset. Subsequently, baseline ML models using default parameters are developed and evaluated using several classification metrics, including macro-averaged measures. Hyperparameter optimisation is then conducted to improve model performance and identify the optimal parameter configurations for each algorithm. Finally, global feature importance and permutation importance analyses are performed to determine the most influential factors affecting TMC decisions.

The main contributions of this study are threefold. First, it evaluates the effectiveness of hyperparameter optimisation in enhancing the predictive performance of multiple ML algorithms for TMC prediction. Second, it provides a comprehensive comparative analysis between default and optimised ML models using balanced transportation data. Third, it identifies the key factors influencing TMC decisions through interpretable feature importance analysis, thereby supporting sustainable transportation planning and policy development in BMA Indonesia.

The current study will be able to answer the following questions.

- What are the most significant factors influencing TMC?
- Does hyperparameter optimisation of ML algorithms enhance the model performance compared to the baseline ML models?
- Among several ML algorithms, which model is the best predictive model, and what is the most significant factor for the TMC?

The structure of the paper is as follows: Section 2 describes the methods and materials, such as survey design, data collection, and a brief description of ML algorithms that are used for the analysis. Results and discussion are discussed in Section 3, which is followed by the conclusion in Section 4.

2 Methodology

2.1 Survey Design

The current study used a multidimensional three-week household and individual survey that was gathered in September 2013 in the BMA, Indonesia. The survey consists of three main categories: (1) Individual and household characteristics, (2) time-use and activity travel participation, and (3) travel behaviour, daily activities, and health-related quality of life (QoL) to study the correlation among different activities, TMC, and its influence on health parameters such as physical health, social health, and mental health. However, based on the aim and scope of the study, the current research only used the first two parts of the survey to investigate the correlation between socio-demographic and economic variables and TMC to work on daily activities using hyperparameter optimisation of ML algorithms.

The individual and household characteristics survey contained the basic information of individuals, such as age, gender, occupation, marital status, education level, and household information, such as household income, household composition, dependent children in a household, disabled persons in a household, and perceived accessibility to basic amenities. The perceived accessibility gathers information to know the availability and accessibility of households to basic amenities such as workplaces, offices, schools, shopping malls, restaurants, hospitals, grocery shops, etc. Based on the latest studies, socio-demographic factors and economic variables are the most significant factors for the prediction of TMC. Besides, past studies professed that household members, household vehicle ownership, and household income are the most significant features for the TMC prediction; therefore, the current study considered socio-demographic and economic variables of individuals and households to investigate the work TMC in BMA and its significant features.

In the second part of the survey, travel behaviour and transport modes information was gathered, which contained the information of the number of vehicles in a household, main transport modes, transport modes used for daily leisure, maintenance, and mandatory, and intensity, such as frequency and duration of using diverse transport modes. In addition, for consecutive 21 days, the time-use and activity travel participation survey was gathered, which contained all types of activities and motorised modes used for those activities. The transport modes were categorised into 15 different modes, such as cars, taxis, railways, buses, angkot, minibus, walking, cycling, subways, etc. However, the daily activities were categorised into in-home and out-of-home leisure, mandatory, and maintenance activities. The data was gathered at fifteen-minute intervals, which makes a total of 96 slices. For instance, 24 h a day multiplied by 60 min in an hour, which makes a total of 1440 min in a day, is divided by 15 min to make 96 slices. ($24 * 60 = 1440/15 = 96$). Several individuals used to multitask activities and use more than one transport mode, such as walking to public transit stations and using the bus for daily commuting. However, the current study did not consider multitasking activities with multi-transport options.

2.2 Boundary Selection

Since several factors, including weather, availability and accessibility, and income, influence travel behaviour, it is imperative to collect data from a particular area or nation and study travel behaviours. Consequently, it is critical to establish precise boundary conditions for research on travel behaviour. The target population of the current study is the residents of BMA Indonesia.

The Southeast Asian archipelago of Indonesia is large and varied, with thousands of islands dispersed around the equator. Indonesia is the fourth most populous country in the world, home to more than 270 million people. Indonesia is a well-liked travel destination because of its varied landscapes, rich cultural heritage, and stunning natural surroundings. Though other locations also provide distinctive experiences, Bali is a popular worldwide destination due to its beaches and lively culture. Due to its strategic location, economic potential, educational hub, tourism, and urban development initiatives, Bandung, the capital of West Java province, is one of Indonesia's most important urban hubs. The reasons behind Bandung's choice as a development and investment destination are its strategic location, economic potential, educational prominence, and continuous urban growth. The city is a prospective site for a range of activities, from economic enterprises to urban planning projects, due to its well-developed transit infrastructure, diverse urban layout, and dynamic demography. To meet the demands of its citizens and visitors, Bandung, Indonesia, developed a transportation system that included a variety of means of transportation. To reduce traffic and encourage more environmentally friendly forms of transportation, the city has been enhancing its transportation infrastructure. Due to Bandung's vast road network, many inhabitants rely mostly on public buses and Angkot, which are ubiquitous and reasonably priced modes of transportation despite their tendency to be crowded. Based on the population, diverse transport modes, and low-to-middle-income households, the current study chose BMA Indonesia as a target site.

2.3 Survey Implementation

Considering the low educational level of the inhabitants of BMA Indonesia and poor resident registration, the survey was prepared in the local language, Bahasa Indonesia, and was not sent via post or email; therefore, the recruitment was done with direct interaction and a paper version between the surveyors and respondents, which was mediated by the community leader, which may introduce selection bias and limit the representativeness of the dataset. Although the sample size was determined using standard sample-size calculation procedures to ensure sufficient statistical observations, the sampling strategy does not guarantee full representation of the BMA population. Households with greater accessibility and willingness to participate may have been overrepresented in the survey. Before starting the data collection, several meetings among surveyors and respondents were conducted to ensure the accuracy of the survey. Personal and good interaction should be developed between the surveyor and respondents to ensure the success of the survey. A convenience sampling technique was used to collect the data from the respondents via direct interaction. A commitment letter was signed by the respondent stating that during the 21 days of the survey, they will not pull back from the survey and will fill out the time-use and activity travel participation survey daily.

During the data collection stage, the geographical condition was considered, where the total area was divided into ten zones, and an equal number of samples was gathered from each zone. Furthermore, due to data limitations, zone-based sampling weights and detailed comparisons with census or population benchmarks were not available in the current study. Therefore, the findings should be interpreted cautiously, particularly regarding their generalizability to the broader BMA population. At the end of August, the participants and the surveyors were recruited, whereas a pilot study was conducted in the first week of September, where approximately 40 respondents were collected to check the reliability and accuracy of the questionnaire survey. The travel behaviour survey was gathered in the second week of September, whereas the time-use and activity travel participation survey were collected in the preceding week for three consecutive weeks. Development of infrastructure and ICT significantly grew in a decade and has had a significant impact on travel behaviours; however, according to the author's view, there were minor alterations in BMA infrastructure and transport modes, which can't affect the collected data. Therefore, the collected data is still valid for the correlation and predictions.

2.4 Data Collection

Enough samples need to be gathered from a given population to perform the statistical analysis and prediction. An entire population is represented by a sample, which is a subset of the population. An appropriate sample size allows for reliable statistical analysis and provides a fair representation of the population; therefore, the current study used a formula that was introduced by Daryle et al. for the sample size as shown in Eq. (1). According to the Daryle equation, circa 400 samples need to be collected considering the BMA population, whereas the current study gathered 732 samples and utilised 508 respondents after cleaning and removing the missing and those under 15 years old. The responses for dependent children were filled in by their household member, especially parents. Besides, due to the aim and scope of the current study, those under 15 years old were excluded from the current study. These are excluded due to limited decision-making autonomy, methodological challenges in collecting accurate data from children, and reliance on adult-centric models. Nonetheless, the missing value further reduced the final dataset. Therefore, a total of 508 samples are utilised for the statistical analysis. The description of the samples is depicted in Table 1.

$$\boldsymbol{n} = \frac{\boldsymbol{N}}{1 + \boldsymbol{N}\boldsymbol{e}^2} \tag{1}$$

Table 1: Description of the respondents (N = 508).

Factors/Study Variables	Proportion or Mean
Socio-demographic characteristics	
Dependent Child	16.54%
In between 15–22	21.46%
In between 23–44	49.39%

(Continued)

Table 1 (continued)

Factors/Study Variables	Proportion or Mean
In between 45–55	17.81%
Greater than 55 years old	12.80%
Male	57.39%
Female	43.61%
Low income (<3 million IDR)	51.39%
Medium income (3–6 IDR)	40.19%
High income (>6 million)	8.42%
Worker	53.19%
Non-worker	32.29%
Student	14.52%
Married	56.42%
Cohabiting	1.22%
Single	38.12%
Divorced	4.32%
Elementary	14.01%
Junior	16.06%
Senior_High	42.10%
Undergraduate	10.01%
Postgraduate	12.22%
No_education	6.10%
Household features	
Average household members	4.58
Average reliant kids per family	0.86
Motorised transport/household	1.073
Percentage of use of private vehicles	48.59%
Percentage of use of non-motorised vehicles	29.01%
Percentage of use of public transport	22.41%
Average total travel time/day (min)	81.01

Small "n" denotes sample size when population size is represented by N, and "e" indicates the necessary degree of precision (3%, 5%, or 7%). For instance, the present study used a 5% desired degree of precision.

In terms of age, those who are 45 years and below show the highest percentage among all ages. Besides, the contribution of males was a bit higher than that of females in the dataset, where males are 57%, while females are 43%. In addition, most of the individuals are from low and middle-income households, whereas most of the individuals are workers, followed by non-workers and students. Among marital statuses, married people show the highest percentage, followed by singles. The highest education level was postgraduate studies, where senior high school students showed the highest percentage among all educational levels. As per the descriptive statistics, on average, at least every household has one MT (either car or motorcycle) used for daily commuting, whereas the percentage of private vehicles (mostly motorcycles) is circa 50%, showing the highest percentage, followed by non-motorised transport (NMT) such as walking and cycling. Due to the low-density area, farther away from basic amenities, and a poor PT network with low accessibility, PT shows the lowest percentage; however, due to low-income households, individuals still walk and cycle to the public transit station to use PT for their daily commuting.

2.5 Machine Learning Algorithms

2.5.1 Random Forest

Like other ML algorithms, RF is an ensemble ML algorithm that was first presented by Adele Cutler and Leo Breiman. RF is primarily employed for problems involving regression and classification. During training, it builds several decision trees and outputs the mean prediction (for regression) or mode (for classification) of each tree. An RF is a strong, adaptable algorithm that combines the output of numerous decision trees to increase prediction accuracy and decrease overfitting. In large datasets, RF performs well due to being less prone to overfitting, which gives high accuracy and predictions; however, complexity is one of the disadvantages of RF, which can be difficult to interpret.

2.5.2 Decision Tree

Like RF, DT is a non-parametric supervised learning ensemble ML algorithm that could potentially be applied to both regression and classification problems. The feature for the split is chosen using a criterion like mean squared error (for regression) or Gini impurity (for classification). It has a hierarchical tree structure with a root node, branches, internal nodes, and leaf nodes. The root node, which represents the complete dataset, is the highest node in the tree. The root node's outer branches provide information to the interior nodes, also known as decision nodes. Both kinds of nodes generate homogeneous subsets, which are represented by leaf nodes or terminal nodes, by conducting evaluations based on available attributes. The leaf nodes in the dataset represent every possible outcome. The interpretability of DT makes them popular, but to increase accuracy and resilience, they are frequently employed in conjunction with other techniques.

2.5.3 Support Vector Machine

Like DT and RF, SVM is an ensemble-based ML algorithm that has a low risk of overfitting, works in high-dimensional space, and can be mainly used for classification tasks, but can be modified for regression analysis. SVM mainly contains an optimal hyperplane, a maximal margin, and support vectors. Finding the optimal hyperplane to split the dataset into two or more separate groups is one of SVM's primary goals. This is critical when selecting a hyperplane. SVM employs a method known as the kernel trick to transform data into a higher-dimensional space where classes can be divided using a hyperplane when the data is not linearly separable. TMC is usually imbalanced.

2.5.4 K-Nearest Neighbours (KNN)

As implied by the term k-nearest neighbours, the model is referred to as a lazy learner since it selects neighbours (k) to forecast the feature space based on the excess of k-nearest data points. Smaller k-neighbours will make the model more sensitive to even modest changes in the data, whereas larger k-neighbours may smooth out local patterns while increasing model robustness. KNN is a simple ML technique that is frequently used for variable classification and regression analysis. Despite being computationally straightforward, KNN struggles with high-dimensional data and is readily influenced by anomalies. It is frequently employed as a baseline model or in situations where interpretability is more important than forecast accuracy.

2.6 Hyperparameter Optimisation

Hyperparameter optimisation plays a crucial role in improving the model predictive performance, robustness, and generalisation ability of ML algorithms. Unlike model parameters learned during the training process, hyperparameters are predefined settings that control the learning behaviour of an algorithm and directly influence classification accuracy and macro-averaged evaluation metrics, including Precision, Recall, and F1-score. In this study, the performance of RF, DT, SVM, and KNN was first evaluated using their default parameter settings and subsequently improved through hyperparameter optimisation. The optimisation process aimed to minimise overfitting and underfitting while enhancing the predictive capability of the models. Furthermore, a 5-fold cross-validation approach was employed to validate the stability and reliability of the optimised models.

Table 2 presents the hyperparameter ranges, default settings, and best parameter configurations obtained for each ML algorithm. For the RF model, the optimal performance was achieved with 300 estimators, a maximum depth of 20, and a minimum sample split of 5 while retaining the "gini" criterion and "sqrt" feature selection. The DT model showed improved classification performance using a maximum depth of 10, minimum sample leaf of 2, maximum leaf nodes of 30, and "sqrt" as the maximum feature selection method. In the case of SVM, the radial basis function (RBF) kernel provided the best predictive performance with $C = 10$ and $\gamma = 1$, indicating improved nonlinear classification capability compared with the default settings. Similarly, the KNN model achieved optimal results with 5 neighbours and distance-based weighting using the Minkowski distance metric. The comparative analysis between baseline and optimised models demonstrated that hyperparameter tuning substantially enhanced the classification performance and macro-averaged metrics of all four ML algorithms for travel mode choice prediction.

Table 2: Hyperparameter optimisation of ML algorithms.

Models	Hyperparameters	Ranges	Default	Best
RF	bootstrap	[True, False]	True	True
	criterion	['gini', 'entropy']	gini	gini
	max_depth	[None, 10, 20, 30, 40, 50]	None	20
	min_samples_split	[2, 5, 10]	2	5
	min_samples_leaf	[1, 2, 4]	1	1
	n_estimators	[100, 200, 300, 500]	100	300
	max_features	['sqrt', 'log2', None]	sqrt	sqrt
	n_jobs	[−1]	−1	−1
	random_state	[42]	42	42
DT	criterion	['gini', 'entropy']	gini	gini
	max_depth	[None, 5, 10, 15, 20, 30]	None	10
	min_samples_leaf	[1, 2, 4, 6]	1	2
	min_samples_split	[2, 5, 10]	2	5
	max_features	['sqrt', 'log2', None]	None	sqrt
	max_leaf_nodes	[None, 10, 20, 30, 50]	None	30
	splitter	['best', 'random']	best	best
	random_state	[42]	42	42

(Continued)

Table 2 (continued)

Models	Hyperparameters	Ranges	Default	Best
SVM	kernel	[‘linear’, ‘rbf’, ‘poly’]	rbf	rbf
	gamma	[0.001, 0.01, 0.1, 1]	Scale	1
	C	[0.1, 1, 10, 100]	1	10
	degree	[2, 3, 4]	3	3
	probability	[True]	True	True
	random_state	[42]	42	42
KNN	n_neighbors	[3, 5, 7, 9, 11, 15]	7	5
	weights	[‘uniform’, ‘distance’]	uniform	distance
	metric	[‘euclidean’, ‘manhattan’, ‘minkowski’]	minkowski	minkowski
	p	[1, 2]	2	2
	algorithm	[‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’]	auto	auto

Fig. 1 depicts the detailed methodology flowchart that is used to conduct this research. The data were collected using convenience sampling and analysed using ML algorithms. Besides, Pearson correlation was applied to study the correlation between input and outcome variables. The dataset was imbalanced; therefore, before applying any ML algorithms, the data was balanced using the SMOTE technique. Moreover, the dataset was split into 80:20 for training and testing. In addition, the models were trained using the training dataset and default parameters and tested using the test dataset. The models were assessed based on the classification metrics, including per-class and macro-averaged metrics. Moreover, the correlation metrics, Precision-Recall-curves (PR-curves), Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), and feature importance were used to assess the models. Similarly, the models were trained using the best parameters, and the comparative analysis between the defaults and optimised models’ outcomes was performed to suggest the best predictive model.

In the current study, the ML models were developed using individual and household-level observations derived from the survey dataset rather than the full 21-day activity-travel diary records at 15-min intervals. Therefore, the unit of analysis in the modelling dataset is the individual respondent and their associated household characteristics. The dependent variable represents each respondent’s primary TMC, while the explanatory variables include socio-demographic, household, and travel-related attributes. Multiple trips, multimodal trips, and repeated temporal observations from the diary records were not explicitly modelled in the current analysis, and clustering effects at the individual or household level were not considered.

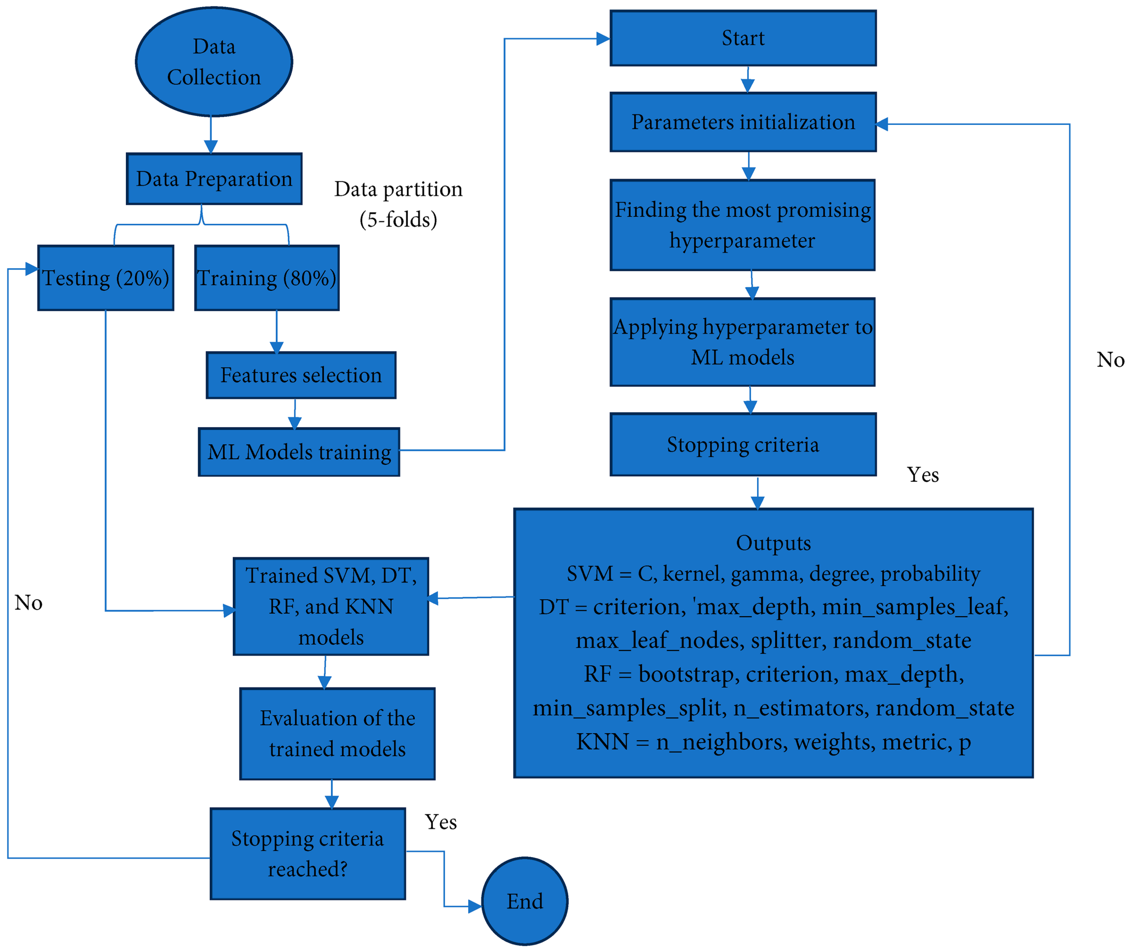
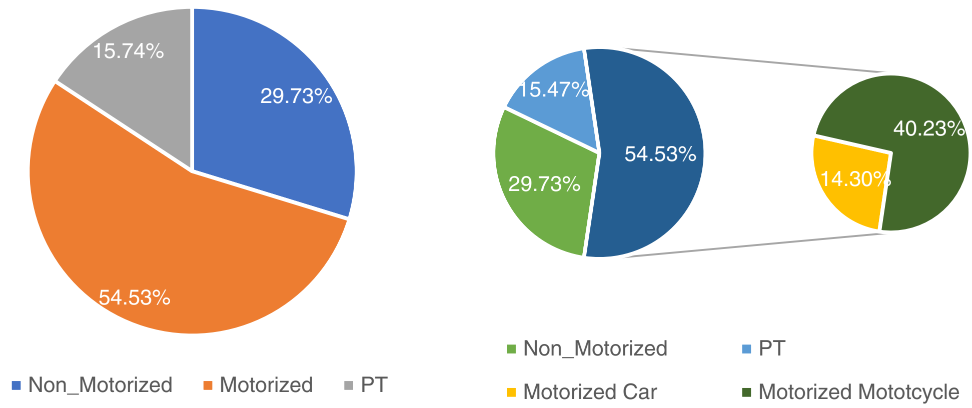


Figure 1: Methodology flowchart.

3 Results and Discussion

The main transport modes for daily commuting were categorised as NMT, MT, and PT. Fig. 2a depicts the percentage of each mode of transport, where MT shares over half of the total transport used for daily commuting, followed by NMT, which contributes circa 30%. However, PT shows only 15.74% of the total transportation, which represents a lack of availability, accessibility, and connectivity. The MT was further categorised as a car and a motorcycle to have a clear picture of private vehicle dependency; however, it was limited to descriptive statistics. Due to low-income households, most of the individuals used motorcycles as their private vehicles for longer distances and AT for shorter distances. Therefore, among 54.53% of private vehicle ownership, motorcycles contribute 40.23%, which is nearly three times that of cars, as depicted in Fig. 2b.

Besides, NMT is categorised as walking and cycling, where among circa 30% of NMT share, walking shows around 20% of NMT, as depicted in Fig. 3, representing the lack of availability of cycling infrastructure, fear of safety, and perceived accessibility. Latest studies claimed that lack of availability of cycling infrastructure, accessibility, fear of safety, and connectivity encourage individuals to use MT, while low income and accessibility to basic amenities encourage them to use NMT [19].



(**a**) (**b**)

Figure 2: Main transport modes categorisation and percentages (**a**) Percentage of using MT, NMT, and PT, (**b**) Classification of MT (cars and motorcycles) and their percentages.

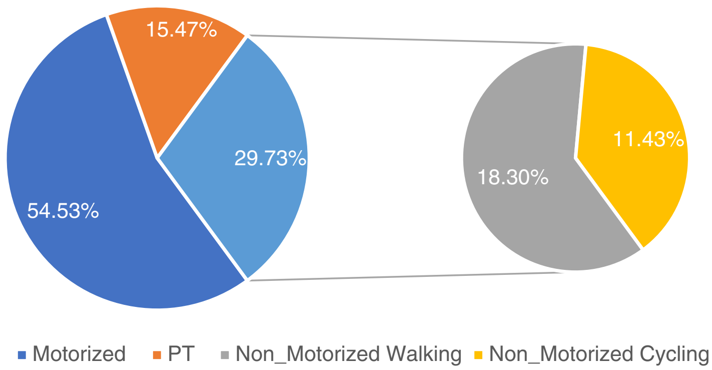


Figure 3: Categorisation of AT into walking and cycling.

Fig. 4 depicts the TMC categorisation based on the socio-demographic and economic variables. Married females with non-working status are more likely to use NMT and PT, while single males with working status are more likely to use MT. The comparative analysis of green mobility (PT and NMT) shows that low-income females and non-workers are more likely to use NMT, whereas medium-income males, with single and student status, are more likely to use PT for daily commuting. Nonetheless, those married workers and those from low-income families are highly dependent on all transport modes. Recent studies claimed that males participate in out-of-home mandatory, leisure, and maintenance activities and use MT, while females mostly engage in in-home activities and use NMT or PT for their daily commuting [20]. Regarding education level, they have almost similar percentages for all modes of transport. Zhang et al. concluded that car ownership, poor walking/cycling environment, and longer distances significantly stimulate the use of cars in school commuting [21].

The TMC dataset used in this study exhibited class imbalance, where MT (54.53%) is substantially overrepresented compared with PT (15.47%). Such an imbalance can bias ML algorithms toward the majority class and reduce the predictive performance for minority travel modes. To address this issue, SMOTE was applied before model development. SMOTE is a data augmentation technique that balances the dataset by generating synthetic samples for minority classes through interpolation between existing neighbouring observations rather than simple duplication. This approach improves class representation, reduces model bias, and enhances the generalisation capability of ML algorithms. By balancing the distribution of travel modes, the SMOTE technique contributed to improving the classification performance and macro-averaged evaluation metrics, including Precision, Recall, and F1-score, thereby enabling more reliable and unbiased travel mode choice prediction. Fig. 5 depicts the comparative analysis of the original and the synthetic dataset, where MT was reduced, and the NMT and PT were substantially increased to balance the dataset.

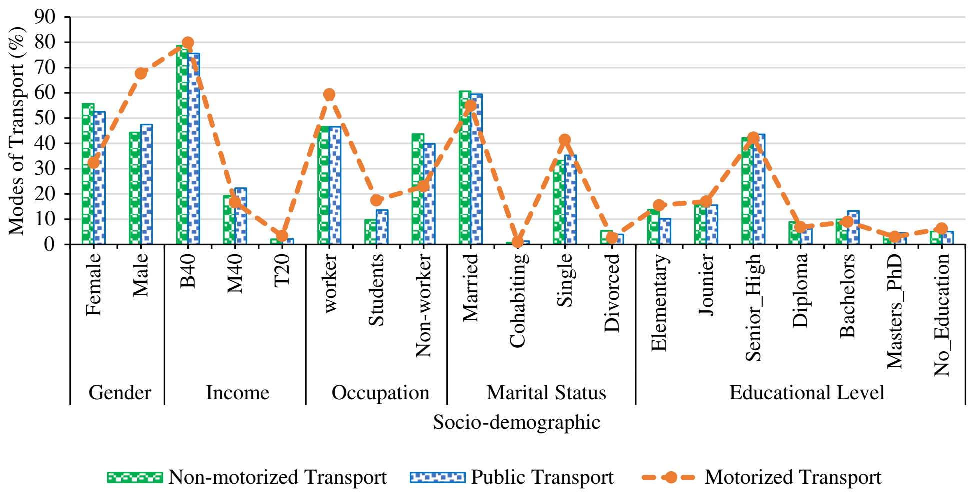


Figure 4: The percentage of TMC based on socio-demographic and economic variables.

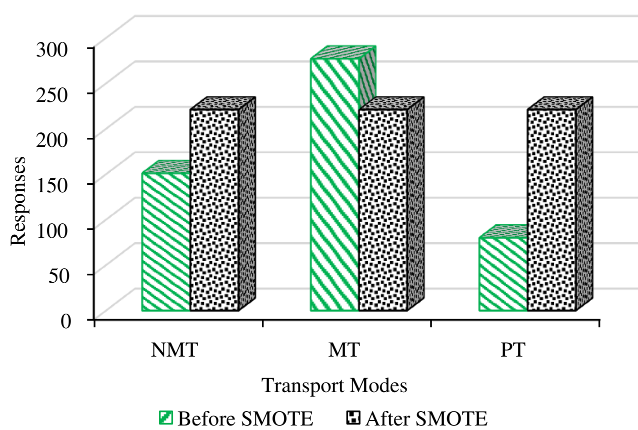


Figure 5: SMOTE comparison.

Fig. 6 depicts a heatmap that was developed from the correlation between the variables using the Pearson correlation. A heatmap is a type of data visualisation where a colour code represents each value in a matrix style. It is frequently used in data science to show how variables are correlated. Each cell's colour intensity indicates how strongly two variables are correlated. A heatmap's primary characteristic is that the colour intensity corresponds to the magnitude of the values. Higher values are usually represented by darker or more intense colours, whereas lower values are represented by lighter hues.

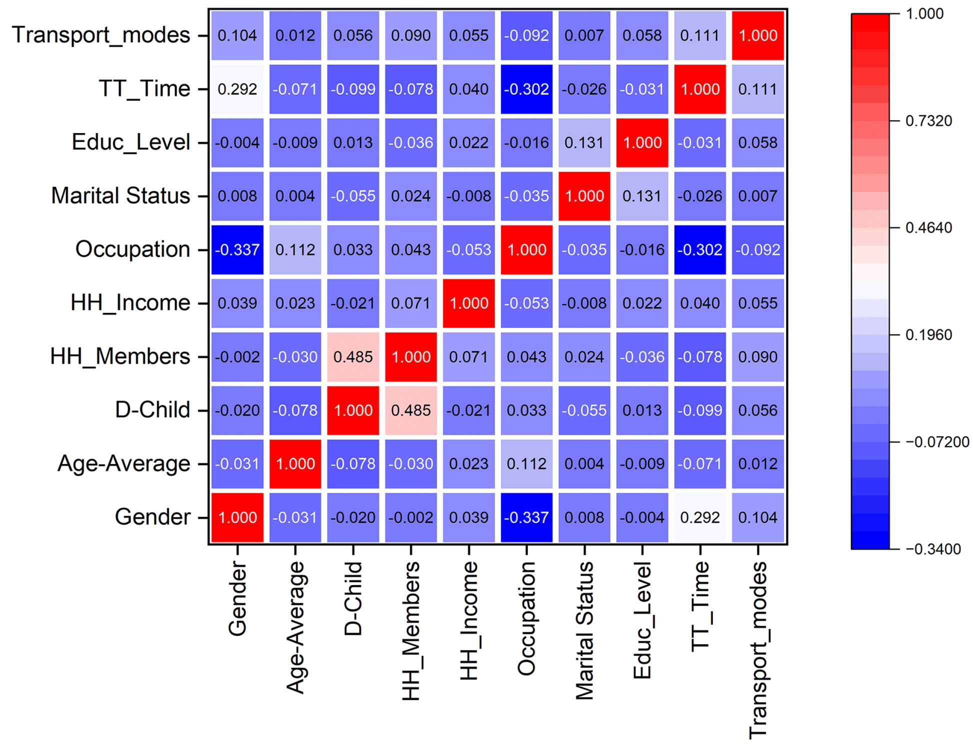


Figure 6: Pearson correlation between the input and outcome variables.

In the current study, household members and dependent children show the strongest correlation, which is 0.485 (circa 50%), followed by gender with total travel time, which is 0.292 (circa 30%). Besides, occupation and age show a strong correlation of 0.112. The most significant factor for TMC was found to be the total travel time, which is 0.111. The higher the distance, the more individuals need to choose diverse transport modes, where for long distances, they prefer to use private vehicles. Besides, the total travel time from origin to destination significantly influences TMC, where the number of stops and waiting times significantly affects individual TMC, and prefers to use private vehicles. Gender was found to be the second most significant factor after total travel time, which is 0.104, positively correlated with TMC. However, occupation shows a negative correlation with TMC, whereas marital status shows a weak correlation with TMC.

Fig. 7 presents the comparative classification accuracy of the baseline (default parameter) and optimised (best parameter) versions of the RF, DT, SVM, and KNN models. The results indicate that hyperparameter optimisation improved the predictive performance of all four ML algorithms. Among the evaluated models, the optimised KNN achieved the highest accuracy, followed by SVM, whereas DT and RF showed comparatively lower predictive performance. The improvement in accuracy after optimisation was particularly noticeable for KNN and SVM, demonstrating the strong influence of parameter tuning on model generalisation and classification capability. Similarly, RF and DT also exhibited moderate improvements compared with their default configurations. Overall, the findings demonstrate that hyperparameter optimisation significantly enhances model accuracy and contributes to more reliable travel mode choice prediction compared with baseline ML models using default settings.

Table 3 depicts the hyperparameter optimisation of several ML algorithms with cross-validation to predict the model performance. A comparative analysis of default and best parameter optimisation for per-class and macro-average metrics is performed. Among all ML models, RF and SVM show the best accuracy and precision in per-class and macro-averaged metrics. In terms of per-class, RF shows the highest precision of 0.667 (66.7%) for MT, whereas SVM shows the highest precision of 0.450 (45%) for NMT and 0.412 (41.2%) for PT. In terms of macro-averaged metrics, RF shows the highest accuracy of 0.539 (53.9%), whereas SVM shows the highest precision of 0.502 (50.2%). The current study is in line with the latest studies and shows that RF outperforms other ML in terms of classification metrics for TMC to work [17].

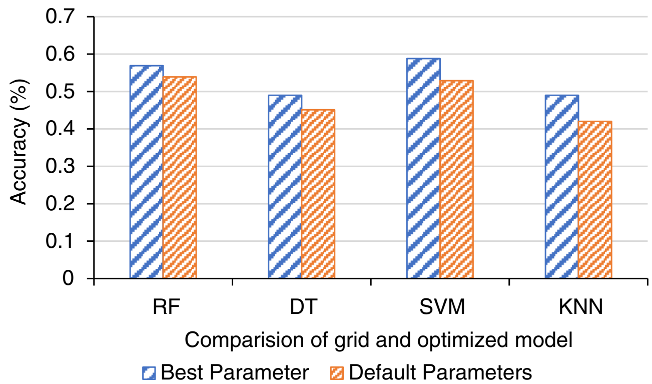


Figure 7: Comparison of grid and optimised model accuracy.

Table 3: Default and best hyperparameter optimisation of ML models' prediction outcomes (best values are marked in bold).

Models	Transport	Precision	Recall	F1-Score	Accuracy	Macro Averaged		
						Precision	Recall	F1-Score
Default Parameters								
RF	NMT	0.421	0.533	0.471	**0.539**	0.491	0.484	0.484
	MT	**0.667**	0.607	0.636				
	PT	0.385	0.312	0.345				
DT	NMT	0.348	0.533	0.421	0.451	0.424	0.416	0.409
	MT	0.591	0.464	0.520				
	PT	0.333	0.250	0.286				
SVM	NMT	**0.450**	0.600	0.514	0.529	**0.502**	0.518	0.504
	MT	0.644	0.518	0.574				
	PT	**0.412**	0.438	0.424				
KNN	NMT	0.42	0.43	0.43	0.42	0.40	0.38	0.38
	MT	0.63	0.46	0.54				
	PT	0.13	0.25	0.17				
Best Parameters								
RF	NMT	0.472	0.567	0.515	0.569	0.519	0.522	0.519
	MT	**0.686**	0.625	0.654				
	PT	0.400	0.375	0.387				

(Continued)

Table 3 (continued)

Models	Transport	Precision	Recall	F1-Score	Accuracy	Macro Averaged		
						Precision	Recall	F1-Score
DT	NMT	0.386	0.567	0.459	0.490	0.450	0.445	0.438
	MT	0.630	0.518	0.569				
	PT	0.333	0.250	0.286				
SVM	NMT	**0.583**	0.233	0.333	**0.588**	**0.591**	0.438	0.443
	MT	0.588	0.893	0.709				
	PT	**0.600**	0.188	0.286				
KNN	NMT	0.46	0.57	0.51	0.49	0.45	0.46	0.45
	MT	0.65	0.50	0.57				
	PT	0.23	0.31	0.26				

The hyperparameter optimisation significantly enhances the classification metrics of all ML models. KNN shows the highest improvement, followed by SVM, DT, and RF in both per-class and macro-averaged metrics. Like the default parameters, RF shows the highest precision of 0.686 for MT, whereas SVM has the highest precision of 0.600 for PT and 0.583 for NMT. Moreover, in terms of macro-averaged metrics, SVM shows the highest accuracy and precision and robustly outperforms all other ML models.

Table 4 depicts the comparative analysis of grid search and optimised models' accuracy and their percentage differences. The outcomes demonstrate that hyperparameter optimisation improved the predictive accuracy of all ML algorithms, although the magnitude of improvement varied across models. Among the optimised models, SVM achieved the highest accuracy (58.8%), followed by RF (56.9%), whereas DT and KNN both attained an accuracy of 49.0%. In terms of relative improvement, KNN exhibited the largest increase in accuracy (16.67%), improving from 42.0% to 49.0%. This substantial improvement suggests that KNN is highly sensitive to hyperparameter selection, particularly the number of neighbours and the weighting strategy, as these parameters directly influence local decision boundaries and classification stability. Similarly, SVM demonstrated an 11.15% increase in accuracy, indicating that optimisation of the kernel parameters (C and γ) significantly enhanced its ability to capture complex nonlinear relationships within the TMC dataset. DT showed a moderate improvement of 8.65%, primarily because controlling tree depth and node splitting criteria reduced overfitting and improved model generalisation. In contrast, RF exhibited the smallest increase in accuracy (5.57%), likely because the ensemble structure of RF already provides relatively stable and robust performance under default settings, making it less sensitive to parameter tuning compared with other algorithms.

Table 4: Comparison of optimised and grid search mode accuracy and percentage changes.

Models	Best Parameters Accuracy (%)	Default Parameters Accuracy (%)	Changes (%)
RF	0.569	0.539	5.57
DT	0.490	0.451	8.65
SVM	0.588	0.529	11.15
KNN	0.490	0.420	16.67

A normalised confusion matrix is used to evaluate the performance by comparing the true labels with the predicted labels of the classification model, as shown in Fig. 8. KNN effectively captures complex patterns in all transport modes, followed by RF. Therefore, these models show the highest precision and accuracy in per-class and macro-averaged metrics. The variation in algorithms is due to feature interaction, class imbalance, and nonlinear relationships. For TMC prediction and planning, policymakers should rely on high-performing models, such as KNN and RF, for comprehensive insights, while recognising that model choice can influence the accuracy of mode-specific predictions. This underscores the importance of using robust, interpretable models to guide transport policy and resource allocation effectively across all travel modes.

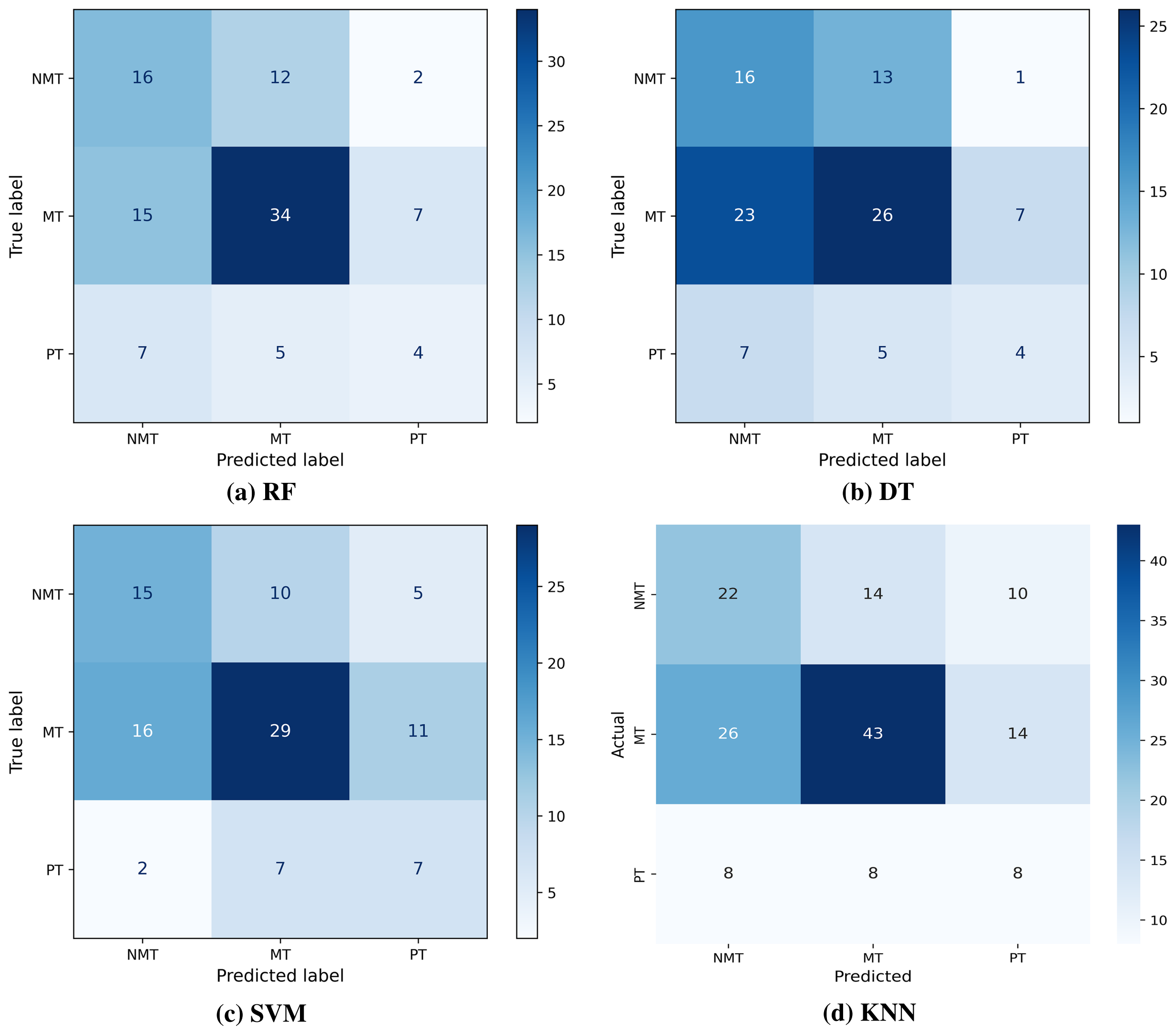
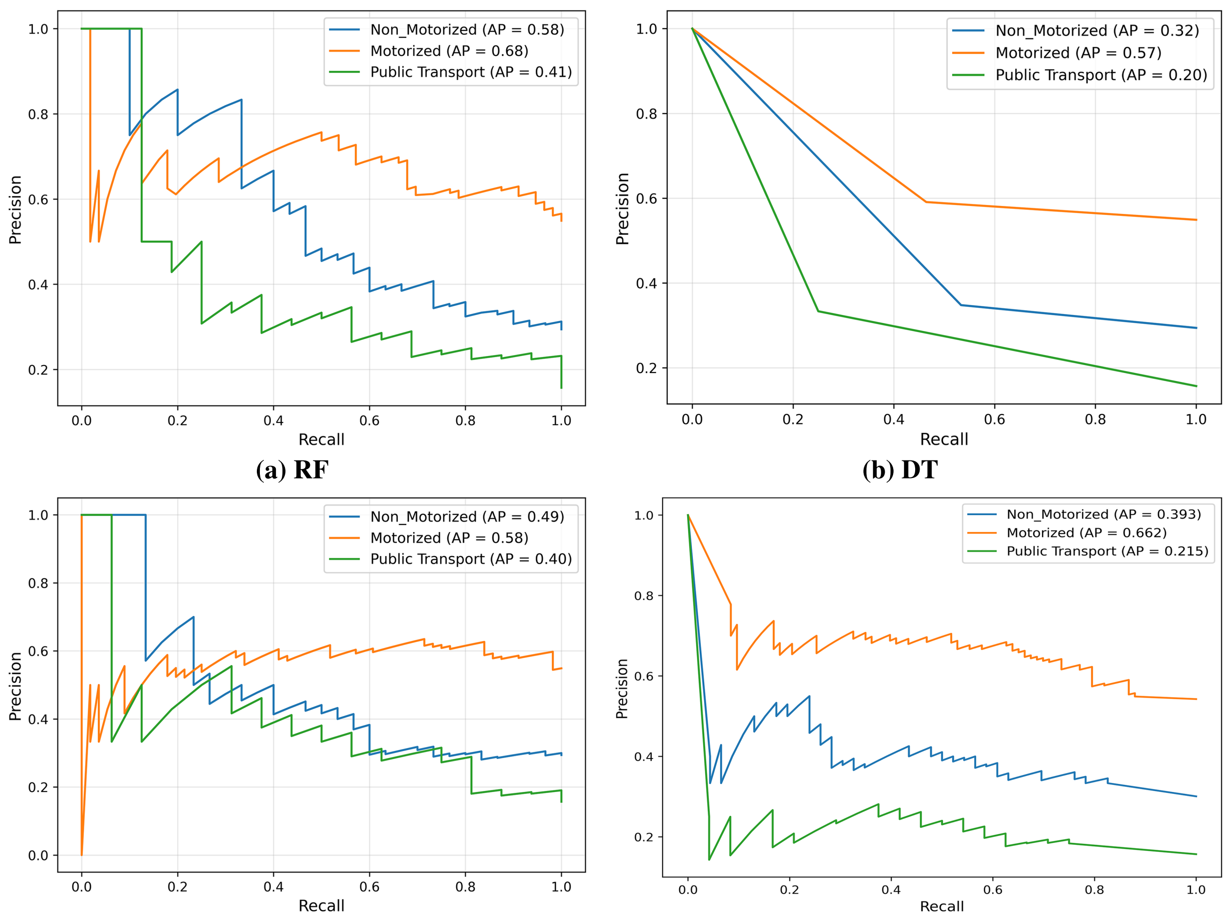


Figure 8: Confusion matrix of ML models.

Fig. 9 presents the PR curves and AP scores for all models across different travel mode classes, illustrating the trade-off between precision and recall in TMC prediction. Among the evaluated models, RF achieved the highest AP scores for all transport mode classes, followed by SVM and KNN, whereas DT demonstrated comparatively lower performance. The superior performance of RF can be attributed to its ensemble learning structure, which combines multiple decision trees to reduce variance, improve generalisation, and effectively capture complex nonlinear relationships within the dataset. SVM also showed strong predictive capability because of its ability to construct optimal decision boundaries in high-dimensional feature space, while KNN provided moderate performance by utilising local neighbourhood information for classification. In contrast, DT underperformed due to its sensitivity to data variation and higher susceptibility to overfitting, particularly in imbalanced datasets. Among the transport mode classes, the MT class achieved the highest AP score, reaching 0.68 for RF, indicating that the model most accurately classified MT observations. Similarly, RF achieved AP scores of 0.58 and 0.41 for NMT and PT, respectively. The consistently higher AP scores for the MT class across all models are primarily associated with the larger number of MT observations in the dataset, which enabled the algorithms to learn more stable and representative classification patterns. Conversely, the relatively lower AP scores for the NMT and PT classes reflect the limited number of observations in these minority classes, which reduced model learning capability despite the application of balancing techniques.



(a) RF **(b) DT**

(c) SVM **(d) KNN**

Figure 9: PR-curves of four ML models.

Fig. 10 illustrates the ROC-AUC values for the RF, SVM, KNN, and DT models, demonstrating the trade-off between the true positive rate and false positive rate for travel mode classification. Consistent with the PR analysis, RF achieved the highest AUC values across all transport mode classes, indicating its superior discriminative capability and overall classification robustness. For the PT class, RF achieved the highest

AUC value of 0.75, followed closely by SVM with an AUC of 0.74, suggesting that SVM also performed effectively in separating PT observations because of its strong capability in constructing optimal nonlinear decision boundaries. Similarly, RF demonstrated the highest AUC values for MT (0.67) and NMT (0.69), followed by KNN with AUC values of 0.64 and 0.611 for MT and NMT, respectively. The comparatively better performance of KNN for MT and NMT may be associated with its instance-based learning approach, which effectively captures local data similarities within relatively dense travel mode classes. In contrast, DT showed lower ROC-AUC performance due to its sensitivity to data variability and limited generalisation capability compared with ensemble and kernel-based methods.

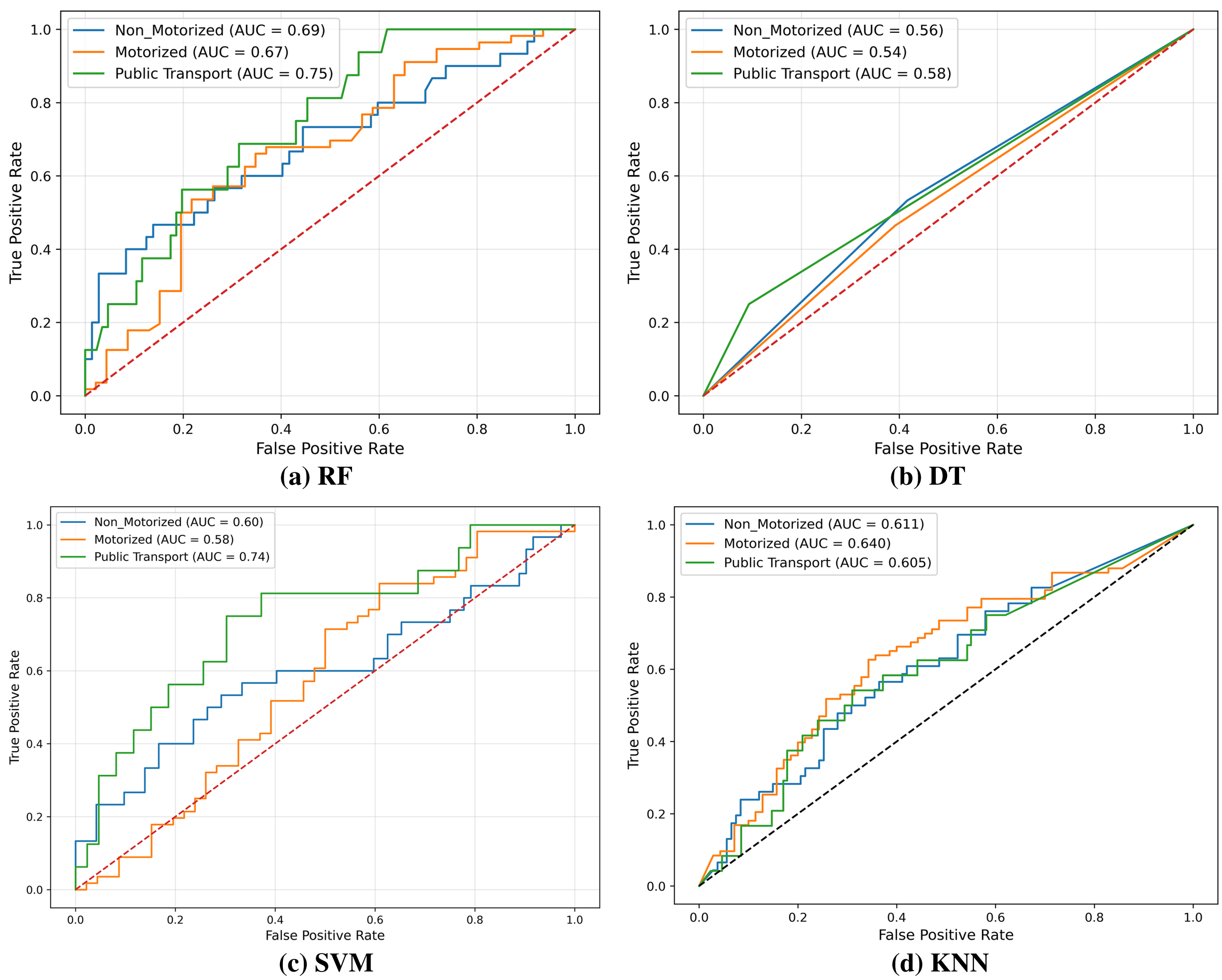


Figure 10: ROC-AUC curves for ML models.

The significance of features with the TMC is the process of determining which elements (features) have the highest impact on a person's decision regarding their transport mode. This is essential for comprehending how people travel, improving transportation systems, and formulating better policy decisions. Researchers, governing bodies, and transportation planners can better understand which factors most affect people's TMC by using feature importance analysis. Using this information, policies, services, and infrastructure may be created to promote more economical and environmentally friendly modes of transportation. The current study applied feature importance analysis to investigate the most significant feature for the TMC prediction. It was found that total travel time is the most significant feature of the work TMC, as shown in Fig. 11, followed by household members. Latest studies claimed that the last mile, total travel time, parking fee, driving license, subsidies, age, and household income significantly influence TMC [22]. For instance, Le and Teng concluded that the "last mile and from and to" plays a vital role in choosing the suburban railway while Abulibdeh concluded that total travel time, parking fee, and sociodemographic variables play a vital role in TMC prediction [13,23].

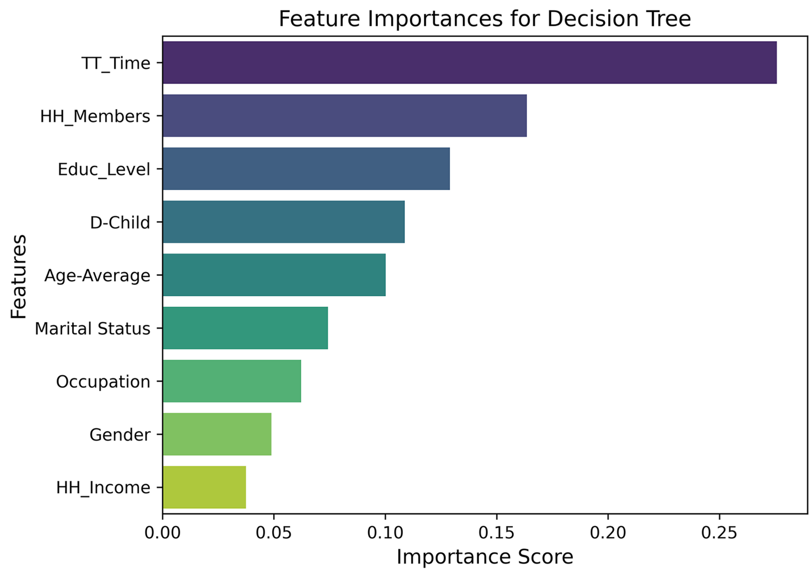


Figure 11: Feature importance.

4 Conclusion

Based on the application of ML algorithms and the hyperparameter optimisation with comparison to the default parameters, the following conclusions can be drawn.

- The results showed that hyperparameter optimisation of ML algorithms significantly enhances the model accuracy and precision both in training and testing, yielding a substantial improvement in predictive accuracy of 16.67% in KNN and 11.15% in SVM. This highlights the necessity of systematic tuning procedures to enhance model robustness and generalisability in TMC modelling.
- The analysis revealed that among evaluated models, RF emerged as the best-performing algorithm for the default parameter, whereas SVM outperformed in optimised parameters, outperforming DT and KNN. This finding reinforces growing evidence that ensemble learning techniques are particularly effective for complex, non-linear TMC prediction problems.
- The results further emphasise that total travel time is the most influential determinant of TMC, followed by household characteristics such as household size and vehicle ownership. These findings confirm that both trip attributes and socio-economic factors play a decisive role in shaping travel behaviour, with longer trips favouring MT over NMT and PT options.
- From a practical perspective, these outcomes provide actionable insights for policymakers, supporting interventions aimed at reducing private vehicle dependency and promoting active and sustainable transport systems.
- However, the study also highlights key limitations. The limited interpretability of ML models and challenges associated with imbalanced datasets restrict their direct applicability in policy contexts. Therefore, future research should focus on integrating interpretable modelling approaches

(SHAP analysis) and conducting comparative analyses with traditional DCMs. Additionally, incorporating multimodal and multitasking travel behaviour would improve the realism and applicability of TMC models.

Acknowledgement: The author would like to thank his colleagues, Dr. Mahdi Aghaabbasi from the University of Central Florida, USA, and Mohammad Tamim Kashifi from Loughborough University, UK, for helping with ML algorithms, especially the hyperparameter optimisation, in the current study.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Mujahid Ali, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AT	Active Transport
PT	Public Transport
MT	Motorized Transport
NMT	Non-Motorized Transport
TMC	Transport Mode Choice
ML	Machine Learning
RF	Random Forest
DT	Decision Tree
SVM	Support Vector Machine
BO	Bayesian Optimisation
PR	Precision-Recall
ROC-AUC	Receiver Operating Characteristic-Area Under the Curve
BMA	Bandung Metropolitan Area

References

1. Cai J, Zhao Z, Zhou Z, Wang Y. Predicting the carbon emission reduction potential of shared electric bicycle travel. Transp Res Part D Transp Environ. 2024;129(23):104107. doi:10.1016/j.trd.2024.104107.
2. Zhang H, Zhang L, Liu Y, Zhang L. Understanding travel mode choice behavior: influencing factors analysis and prediction with machine learning method. Sustainability. 2023;15(14):11414. doi:10.3390/su151411414.
3. Madhuwanthi RAM, Marasinghe A, Rajapakse RPCJ, Dharmawansa AD, Nomura S. Factors influencing to travel behavior on transport mode choice—a case of Colombo metropolitan area in Sri Lanka. Int J Affect Eng. 2016;15(2):63–72. doi:10.5057/ijae.ijae-d-15-00044.
4. Convery S, Williams B. Determinants of transport mode choice for non-commuting trips: the roles of transport, land use and socio-demographic characteristics. Urban Sci. 2019;3(3):82. doi:10.3390/urbansci3030082.
5. Ali M. A machine learning approach to study the effect of parking pricing on transport mode choice to work using 2022 NHTS dataset. Case Stud Transp Policy. 2025;20(12):101478. doi:10.1016/j.cstp.2025.101478.
6. Harbering M, Schlüter J. Determinants of transport mode choice in metropolitan areas the case of the metropolitan area of the Valley of Mexico. J Transp Geogr. 2020;87(2):102766. doi:10.1016/j.jtrangeo.2020.102766.
7. Zhao X, Yan X, Yu A, Van Hentenryck P. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav Soc. 2020;20(1):22–35. doi:10.1016/j.tbs.2020.02.003.

8. Vinayakumar G. A comparison of KNN algorithm and MNL model for mode choice modelling. Eur Transp. 2023;92(92):1–14. doi:10.48295/et.2023.92.3.
9. Salas P, De la Fuente R, Astroza S, Carrasco JA. A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. Expert Syst Appl. 2022;193(1784):116253. doi:10.1016/j.eswa.2021.116253.
10. Bresson G, Dargay J, Madre JL, Pirotte A. The main determinants of the demand for public transport: a comparative analysis of England and France using shrinkage estimators. Transp Res Part A Policy Pract. 2003;37(7):605–27. doi:10.1016/S0965-8564(03)00009-0.
11. Wang F, Ross CL. Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transp Res Rec J Transp Res Board. 2018;2672(47):35–45. doi:10.1177/0361198118773556.
12. Ali M, Fissha Y. Propose adjustable Support Vector Machine approach for classifying imbalanced work travel mode choice data. Transp Res Interdiscip Perspect. 2026;35(1):101786. doi:10.1016/j.trip.2025.101786.
13. Le J, Teng J. Understanding influencing factors of travel mode choice in urban-suburban travel: a case study in Shanghai. Urban Rail Transit. 2023;9(2):127–46. doi:10.1007/s40864-023-00190-5.
14. Liu Y, Yu X, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manag. 2011;47(4):617–31. doi:10.1016/j.ipm.2010.11.007.
15. Ali M. Classification of imbalanced travel mode choice dataset with SMOTE and prediction using interpretable machine learning. Sustain Futur. 2025;10(1):101119. doi:10.1016/j.sftr.2025.101119.
16. Xia Y, Chen H, Zimmermann R. A Random Effect Bayesian Neural Network (RE-BNN) for travel mode choice analysis across multiple regions. Travel Behav Soc. 2023;30(2):118–34. doi:10.1016/j.tbs.2022.08.011.
17. Aghaabbasi M, Ali M, Jasiński M, Leonowicz Z, Novák T. On hyperparameter optimization of machine learning methods using a Bayesian optimization algorithm to predict work travel mode choice. IEEE Access. 2023;11(13):19762–74. doi:10.1109/ACCESS.2023.3247448.
18. Naseri H, Waygood EOD, Wang B, Patterson Z. Application of machine learning to child mode choice with a novel technique to optimize hyperparameters. Int J Environ Res Public Heal. 2022;19(24):16844. doi:10.3390/ijerph192416844.
19. Berrill P, Nachtigall F, Javaid A, Milojevic-Dupont N, Wagner F, Creutzig F. Comparing urban form influences on travel distance, car ownership, and mode choice. Transp Res Part D Transp Environ. 2024;128(9):104087. doi:10.1016/j.trd.2024.104087.
20. Liu Y, Shen R, He M, Li X, Shi Z. Gender differences in commuting travel mode choices among young adults: a spatial heterogeneity perspective. J Transp Geogr. 2025;123(3):104145. doi:10.1016/j.jtrangeo.2025.104145.
21. Zhang R, Yao E, Liu Z. School travel mode choice in Beijing, China. J Transp Geogr. 2017;62(C):98–110. doi:10.1016/j.jtrangeo.2017.06.001.
22. Ali M, Esztergár-Kiss D. Analyzing the interplay between driver's license ownership, household characteristics, socioeconomic factors, and work travel mode choice by using the 2022 NHTS dataset. Transp Res Interdiscip Perspect. 2026;37(12):102053. doi:10.1016/j.trip.2026.102053.
23. Abulibdeh A. Analysis of mode choice affects from the introduction of Doha Metro using machine learning and statistical analysis. Transp Res Interdiscip Perspect. 2023;20(7):100852. doi:10.1016/j.trip.2023.100852.