TRANSPORT PROBLEMS 2025 Volume 20 Issue 2
PROBLEMY TRANSPORTU DOI: 10.20858/tp.2025.20.2.10

Keywords: travel mode choice; machine learning; random forest; school; sustainability

Mujahid ALI[1]*, Choon Wah YUEN[2]

PREDICTION OF SCHOOL TRAVEL MODE CHOICE AND ITS DETERMINANTS – A MACHINE LEARNING APPROACH

Summary. Prediction of travel mode choice (TMC) is crucial for urban planners and policymakers to promote sustainable transportation systems and reduce traffic congestion. In recent decades, the prediction of TMC to schools, which involves daily commuting, has attracted the interest of researchers in green urban planning and a better society. Statistical models are based on many unrealistic premises about the data distribution and are typically used to perform mode choice analysis, which might result in biased model predictions. Moreover, machine learning algorithms that are assumption-free can handle complex, imbalanced, and multiclass datasets with high interpretability and outperform conventional techniques; thus, they have received much attention. Therefore, the present study intends to use modern techniques, such as Naïve Bayes, random forest, gradient boost, support vector machine, and linear regression, to predict the TMC to school (highest level of education) and its influencing factors. The current study contributes to the existing literature through (1) the application of modern techniques for the prediction of school TMC, (2) feature importance to predict the most significant feature of school TMC, (3) a proposal of the best predictive model, and (4) a discussion of the effectiveness of modern techniques over traditional methods. A total of 2756 samples from the NextGen 2022 National Household Travel Survey – California dataset was utilized to predict school TMC and its influencing factors. Based on the predictions, it was found that gradient boost outperformed other machine learning models with an accuracy of 98.9% in training and 83% in testing. Moreover, random forest achieved an accuracy of 77.8% and 71.1% in training and testing. Based on the sensitivity analysis, it was found that age is the most significant factor in determining the TMC to school, followed by the type of school. The findings will help policymakers and can be used to better understand modeling TMCs to schools, promoting sustainable transportation options.

1. INTRODUCTION

Everyday living is fundamentally impacted by transportation, which gives people access to essential services, including social interaction, work, healthcare, and education. However, identifying specific schools' travel mode choices (TMCs) with varying age groups can be a challenging task. In 1969, walking and cycling to school accounted for 47.7% of total trips, which reduced to 12.7% by 2009 and remained at approximately 14% in 2017, according to the National Household Travel Survey (NHTS) in the USA. In 2009, private cars were the predominant TMC to school (45.3%), followed by school buses (39.4%) [1]. As concluded by Lidbe et al., in 2020, cars were the predominant transport mode in school, which contributed about 50% of the total, followed by the school bus at 33%; however, walking contributed 13%, whereas cycling represented less than 1% of total TMC to school in 2017 (NHTS, California dataset) [2]. TMC for school journeys in the European Union (EU) varies by country,

[1] Silesian University of Technology, Faculty of Transport and Aviation Engineering; Krasińskiego 8, 40-019 Katowice, Poland; e-mail: mali@polsl.pl; orcid.org/0000-0003-4376-0459
[2] Centre for Transportation Research, Faculty of Engineering, University Malaya, Wilayah Persekutuan Kuala Lumpur 50603, Malaysia; e-mail: yuencw@um.edu.my; orcid.org/0000-0001-7751-7292
* Corresponding author. E-mail: mali@polsl.pl

influenced by factors such as infrastructure, culture, social context, and policy. Mancini concluded that walking and cycling to educational institutions are less common among adolescents, and they prefer to use motorized transport (MT) [3]. Buehler stated that Germans take four times more trips by foot, bike, and public transport (PT), while driving accounts for 25% fewer trips compared to Americans [4]. The shift from green mobility towards private vehicle use presents a significant challenge for policymakers, as private vehicles not only exacerbate traffic congestion but also contribute to health issues and environmental degradation. Past studies proposed personal, social, demographic, and built environmental factors, such as age, gender, income, vehicle ownership, education, personal preferences, and attitudes that affect schools' TMC, while other studies claimed that weather conditions, total travel time, distance, and spatial variation are determinants of TMC to school [5].

Several studies have been conducted on TMC to schools for children (6–12 years), teenagers (13–18 years) using different spatial conditions such as plane and hilly areas, weather conditions, availability and accessibility, distances to and from residence locations, peak hours, people with disabilities, and household car ownership. Saitluanga and Hmangaihzela investigated the TMC of off-campus college students in hilly areas and concluded that the most influential factors are residence location, vehicle availability, and socio-demographic and economic characteristics. They claimed that high-income households tend to use private vehicles while females tend to live closer to the campus and use active and PT [6]. As concluded by Liu et al., students between the ages of six and 12 years are more inclined to take active transport with lower slopes and longer distances, while students between the ages of 13–18 prefer to use active transport with higher slopes [7]. McDonald et al. claimed that high levels of parent's social support are positively correlated with biking and walking [8]. Moreover, Uddin et al. examined the factors influencing adults with disabilities and concluded that, in the absence of suitable accessibility, people with disabilities tend to experience longer travel times. They asserted that weather has a significant influence on TMC and that individuals with limited financial resources are more likely to rely on PT [9].

In recent decades, the prediction of TMC to schools, a key aspect of daily commuting, has attracted the interest of researchers in sustainable urban planning and improving quality of life. Traditionally, conventional methods have been employed for mode choice analysis; however, these methods often rely on unrealistic assumptions about data distribution, which can lead to biased predictions. In contrast, machine learning (ML) algorithms, which are assumption-free, can handle complex, imbalanced, and multiclass datasets and offer high interpretability while outperforming conventional methods, have garnered increasing interest. Therefore, the current study aims to use several modern techniques, such as naïve Bayes (NB), random forest (RF), gradient boost (GB), support vector machine (SVM), and linear regression to predict school TMC and its influencing factors. Furthermore, the current study suggests a reliable predictive model for school TMC by comparing the performance evaluations of several ML algorithms. Moreover, sensitivity analysis is used to investigate the most influential factor for the school TMC. The current study aids policymakers and urban planners in promoting green mobility and improving school TMCs of different ages by identifying factors that impact decision-making. The analysis carried out in the current study is based on ML algorithms rather than traditional methods.

2. LITERATURE REVIEW

Several studies have been conducted on schools' TMCs, utilizing diverse datasets from around the globe and traditional methods. All forms of transportation, including personal cars, PT, and active transportation like cycling and walking, are frequently utilized to commute to school [10]. For instance, Rothman et al. conducted a comprehensive review of the literature in North America for the period of 1990–2016 on the decline of school active travel to study the influencing factors. They found that the distance from home to school was the most noteworthy factor, whereas child age, low parental education, income, and perception had moderate positive associations with active school travel [11].

Zhang et al. studied the key influencing variables on school TMC for students between the ages of seven and 18 in Beijing, China, using a Beijing Travel survey along with logit-based and tree-based models. They concluded that car ownership, the built environment, and distance were the most influential factors. Longer distances and the availability of household vehicles encouraged the individual to use personal vehicles for school commuting, whereas poor walking and cycling routes encouraged the individual to use MT [12]. Nonetheless, a child's ability to walk to school is hampered by a long commute; on the other hand, enhancing the routes, sidewalks, availability of parks, and traffic patterns seems promising.

Furthermore, several studies have utilized modern techniques such as ML algorithms to predict school TMC and its influencing factors. For instance, Ali conducted a systematic review of discrete choice modeling and ML algorithms for TMC prediction and concluded that non-parametric methods are more accurate than traditional methods [13]. Moreover, the TMC dataset is often imbalanced, with a majority of samples in one class and a minority in another. However, traditional ML models perform poorly on minority classes, leading to biased predictions and suboptimal decision-making. Qian et al. employed the theory of adjustable kernel support vector machine (SVMAK) to classify imbalanced TMC data using a 10-fold cross-validation and found that SVMAK outperformed the standard SVM and enhanced the model's accuracy [14]. Past studies have used hyperparameter optimization with 70:30, 80:20, and 90:10 training and testing ratios, and five-fold cross-validation and interpretable ML algorithms using 90:10 training and testing and 10-fold cross-validation [15] to handle the imbalance in TMC and accurately predict TMC.

Transport, health, and the environment are interlinked, as transport affects health options and the environment [16]. Ali et al. used transport-related physical activities to study the correlations among daily activities, TMC, and health outcomes. Their findings indicated that physical activity intensity mediates the relationship between transport and health, where PT is 0.2% and 1.5%, and active transport is 2.0% positively associated with physical health and 0.2% with social health [17]. However, people with disabilities (health issues) influence transport options. As found by Park et al., those with disabilities take 10–30% fewer trips than other individuals [18].

2.1. Study gap and research contribution

Based on the above literature, numerous studies have been conducted to examine the effect of school TMC on travel behavior and health outcomes; however, limited studies have predicted schools' TMC and its influencing factors. Moreover, past studies utilized traditional methods that rely on assumptions and are unable to handle imbalances and complex datasets. However, due to technological development, recent studies employed modern techniques such as ML algorithms. Therefore, this study aims to use RF, SVM, NB, GB, and LR to predict a school's TMC and its influencing factors using the 2022 National Household Travel Survey (NHTS) – California dataset, which is imbalanced and complex. In addition, the current study compares the performance evaluation of several ML algorithms to identify the most accurate predictive model for school TMC and propose the best predictive model. Based on the outcome of the feature importance, the most influential factor for the school TMC is examined. Furthermore, the present investigation adds to the existing literature by examining the superiority of contemporary approaches over conventional methods, placing particular emphasis on the high precision and accuracy of the ML algorithms. By identifying the factors that influence their choices, the present study enables policymakers to adopt more environmentally friendly transportation and enhance the TMC of schools concerning various interrelated factors.

3. DATA AND METHODS

3.1. Dataset

The NextGen 2022 NHTS – California dataset is an openly accessible resource made accessible through the collaborative efforts of the California Department of Transportation and the U.S. Federal Highway Administration. The dataset contained multidimensional variables such as household information; individual base datasets; trip and travel parameters; vehicle characteristics; built environment variables; TMC for daily leisure, mandatory, and maintenance activities; the intensity (frequency and duration) of TMC; parking conditions; health parameters; living environment; and many other factors that allow researchers and practitioners to use it based on the scopes and aims of their studies. This comprehensive dataset is designed to thoroughly analyze individual travel behavior and personal preferences of U.S. residents regarding their use of TMC for commuting to work, school, and other destinations.

The 2022 NHTS survey was conducted at the national level, comprising a total of 16,997 individual samples from 7893 households, which allows researchers to analyze travel behaviors at both the individual and household levels. After cleaning and normalization of the dataset (removing invalid, missing, and prefer not to answer), the current study used a total of 2756 samples for the statistical analysis of TMC to school. Overall, the dataset contained over 100 variables, including TMC to school, which is further categorized into over 15 classes. The authors only focus on TMC to school as a target variable and socio-demographic variables, educational background, reason for fewer trips, and frequency of transport mode. The target variable contained eight classes of TMC to school: car, SUV, pickup, PT, school bus, bicycle, walked, and others (Table 1). Cars were the most common mode of transportation used to get to school, followed by SUVs and school buses, which accounted for 75.5% of all TMC.

A total of 21 features were selected from the multidimensional dataset that contains socio-demographic variables; educational, travel, and trip characteristics; and intensity variables, such as frequency of TMC to school (Table 2). It is vital to know the type and behavior of the dataset before using it for statistical analysis, as some of the ML modeling deals with categorical data while others can handle numerical datasets; therefore, the type of each variable was studied. Moreover, descriptive statistics of the data were performed to study the mode, mean, standard deviation, median, maximum, and minimum values and, above all, to check for missing values.

Table 1

Description of the 2022 NHTS Dataset and Variable Selection

National Household Travel Survey (NHTS) 2022 – California US Dataset	
Total individual samples	16,997
Total samples used in the current study	2756
Total number of features	21
Target variable	1
Number of classes in the target variable	8
Numerosity	
Car	857
SUV	481
Pickup	98
Public transport	62
School bus	745
Bicycle	49
Walked	254
Others	36

The overall data is split into training and testing (80:20), and the predictions are performed on both datasets. Usually, the training dataset has a higher percentage; therefore, the model accuracy in the training is higher than the model accuracy in the testing. However, some studies claimed higher testing accuracy than training. Past studies used different ratios, such as 70:30, 80:20, and 90:10, with most studies claiming that 80:20 gives the best predictions. Moreover, some researchers separated the data into training, validation, and testing sets; for example, Buijs et al. split the data into 50% for training, 20% for validation, and the remaining 30% for testing [19]. Moreover, past studies used different cross-validation processes, from k-fold to 10-fold cross-validation schemes. The current study also varied the ratio and found the best predictions when using 80:20 with five-fold cross-validation to train the models and prevent overfitting.

Table 2

Descriptive Statistics of Features and Target Variables

Variables	Description	Type	Mean/Mode
Socio-demographic Variables			
R_AGE	Respondent age	Continuous	17.62
R_RACE	Respondent race	Nominal	1.277
R_SEX	Respondent sex	Flag	1.617
Educational Background			
EDUC	Education status	Categorical	1.851
SCHOOL1	Enrolled in a school or academic program		1.00
SCUD	School or academic program description		1.00
SCHTYP	Type of K-12 school enrolled in		1.085
Reason for Fewer Trips			
USAGE2_2	Didn't feel safe	Categorical	2
USAGE2_3	Didn't feel clean		1.979
USAGE2_4	Not reliable		1.936
USAGE2_5	Didn't go where needed		1.979
USAGE2_6	Unaffordable		1.892
USAGE2_7	Health problem		1.83
USAGE2_10	COVID-19		1.766
Frequency of Transport Mode (Last 30 Days)			
LAST30_PT	Used PT in last 30 days	Continuous	1.936
LAST30_MTRC	Used motorcycle in last 30 days		2.00
LAST30_WALK	Walked from place to place in the last 30 days		1.66
LAST30_BIKE	Used bicycle in last 30 days		1.936
*SCHTRN1	Usual transport to school	Categorical	6.128

* Target variable

After the data were split, several ML algorithms, such as GB, RF, NB, SVM, and LR, were applied to predict school TMC. The models were assessed based on their predictions and classification metrics. When the classes are very imbalanced, precision-recall is usually used to measure the success of predictions. In the present study, the data were imbalanced, as cars had the highest number of data samples, whereas walking and cycling had the lowest. Therefore, the precision-recall metric was performed to evaluate the classifier's output quality. A sensitivity analysis was also carried out To assess the feature importance of input factors on the target variable. The classification matrix was used to compare the models and recommend the most accurate predictive model. Fig. 1 shows the overall methodology flowchart of the current study, while Fig. 2 depicts the use of several ML algorithms using Orange software.

GB, RF, NB, SVM, and LR are widely used ML algorithms. GB is an ensemble method that builds models sequentially, with each model correcting the errors of its predecessor. It is often used for classification and regression tasks. RF is another ensemble technique that creates multiple decision trees during the training phase and aggregates their predictions to enhance model accuracy and minimize overfitting. NB is a probabilistic classifier grounded in Bayes' theorem, which assumes that features are independent of each other, making it effective for text classification and problems with high-dimensional data. SVM is a robust classification and regression algorithm that finds the hyperplane that maximizes the margin between data classes, making it well-suited for complex, high-dimensional spaces. Logistic regression (LR) is a straightforward and easily interpretable model commonly used for binary classification tasks. It estimates the probability of an outcome using the logistic (sigmoid) function, assuming a linear correlation between input features and the logarithm of the odds of the target variable.

3.2. Classification of TMC to School

The target variable consists of 15 classes, whereas the current study utilizes eight classes due to the lower number of other classes (less than 30 samples) for TMC to school. The current study utilized the highest level of education and the corresponding transport mode used for it. As depicted in Fig. 3, travelling in a car as a passenger accounts for the highest number of trips to school, representing 31.2% of the total TMC, followed by the school bus at 27%. These two transport modes alone contributed almost 60% of the total TMC, whereas cycling and PT are the least common TMCs to school, contributing only 4% of the total. However, a total of 254 individuals prefer to walk to school, which might be due to the inaccessibility of PT, low-income households, or residing closer to the school.

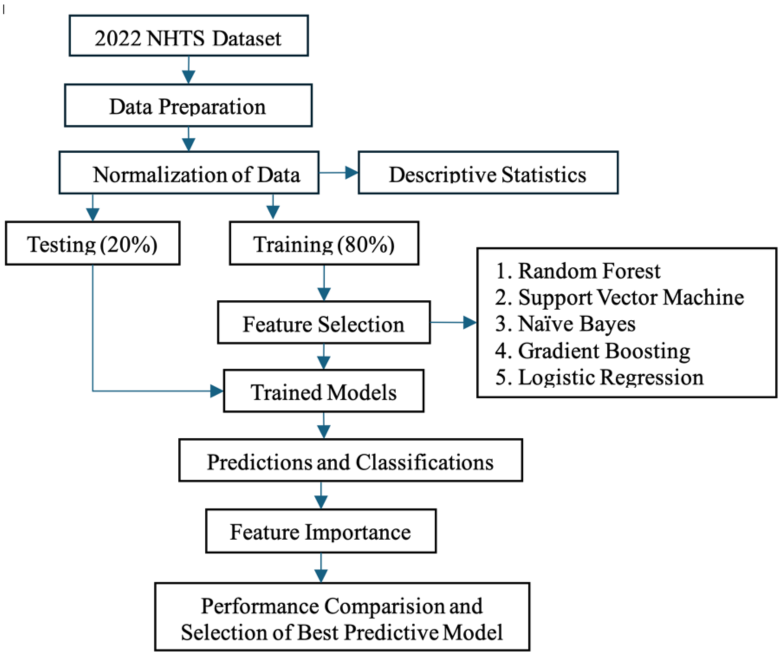

Fig. 1. Methodology flow chart

3.3. Machine Learning Algorithms

Past studies utilized various traditional methods to predict TMC to school; however, these approaches rely on assumptions and often fail to produce accurate results. Moreover, these methods, such as discrete choice models, are usually used for binary classification and multinomial logit analysis, where TMC to school contains multiple classes. Additionally, the NHTS dataset exhibited an imbalance issue, with cars being the dominant and preferred mode of transportation in the US, while bicycles were the least used TMC. This imbalance makes it challenging for conventional methods to produce accurate predictions.

Due to technological development, ML algorithms are widely utilized in several fields, including medicine, economics, and engineering, and have surpassed traditional methods. Recent studies have employed several ML algorithms to predict vehicle crashes, pedestrian fatalities, road safety, and transport mode choice, thereby promoting sustainable transportation systems. Therefore, the current study used the latest version (3.38.1) of Orange and employed RF, SVM, NB, GB, and LR to predict TMC to schools to promote a sustainable transportation system.

3.4. Model Evaluation Performance Metrics

Numerous evaluation metrics have been introduced to gauge the effectiveness of modern methodologies. Various classification metrics—such as area under the curve, precision, F1-score, accuracy, and recall—are employed to evaluate these techniques by measuring model performance based on the relationship between predicted and actual outcomes. The classification metrics are based on the positive and negative predictions of the predicted and actual values, as depicted in Table 3. Typically, accuracy is described as the ratio of positively predicted occurrences to all occurrences, as shown in Equation 1, which ranges from 0–100% (0.00–1.00). Models showing superior accuracy outperformed other models and showed a strong association among the variables.

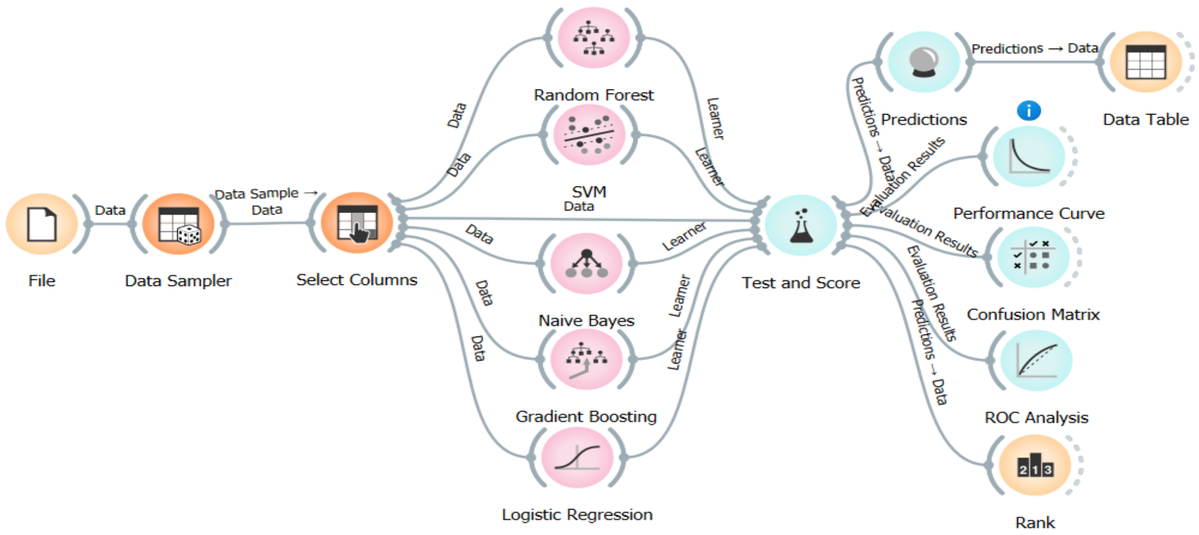

Fig. 2. Utilization of ML algorithms for analysis and predictions

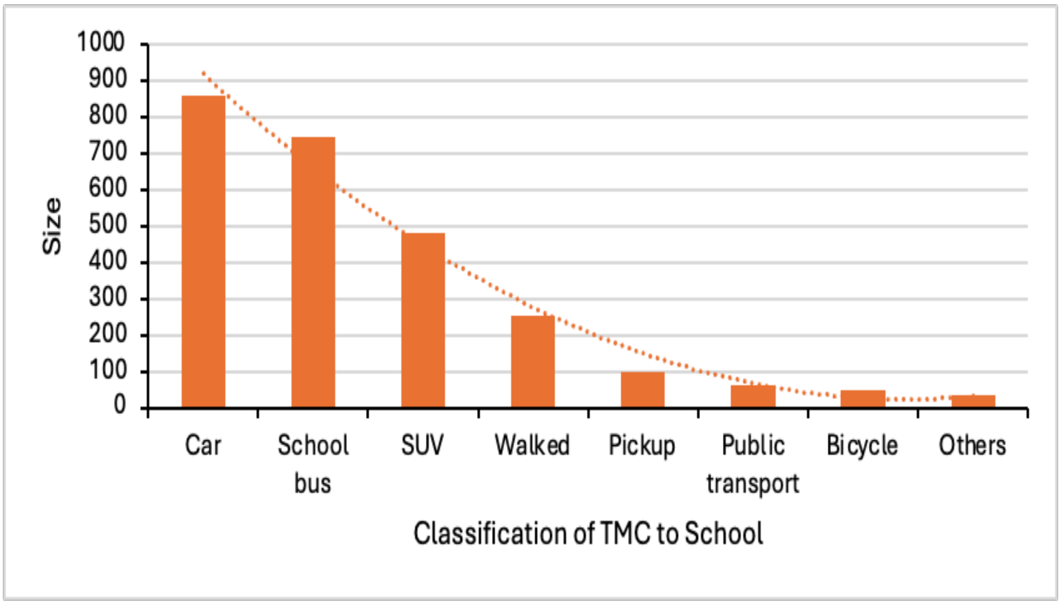

Fig. 3. Classification of TMC to the school

Precision, as shown in Equation 2, is computed as the number of true positives divided by the total number of predicted positives. The capacity of a model to accurately identify every pertinent instance of a positive class is measured by recall, as shown in Equation 3. The F1 score is used to balance the combination of the precision and recall values, as shown in Equation 4.

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

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

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

$$F1 - Score = 2\ x\ \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

4. RESULTS AND DISCUSSION

4.1. Model Estimation Results

Table 4 illustrates the model classification matrix for non-parametric models. With an accuracy of 98.9% in training and 83% in testing, it is evident that GB performed better than any ML model. Moreover, NB has a lower accuracy in training and testing. However, the advanced classifier, coupled with desirable features and computational efficiency, enhances the accuracy of NB and makes it competitive with other ML algorithms. Due to the imbalance in the dataset, NB was biased toward the more frequent class, resulting in poor performance for less frequent choices. In addition, NB often works well with categorical data. However, the dataset contained continuous variables that affected the performance of NB. Fig. 4 illustrates the accuracy of ML models throughout training and testing, with GB achieving the highest accuracy, followed by LR and RF.

Table 3

Confusion Matrix for Classification Performance

		Predictions	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 4

Performance Classification of ML Models

Model	Area under the Curve		CA		F1-Score		Precision		Recall	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
RF	0.871	0.943	0.778	0.711	0.726	0.650	0.741	0.686	0.778	0.711
SVM	0.001	0.111	0.667	0.553	0.582	0.435	0.587	0.586	0.667	0.553
NB	0.983	0.771	0.429	0.340	0.899	0.356	0.944	0.324	0.889	0.158
GB	1.00	0.974	0.989	0.830	0.988	0.809	0.991	0.805	0.999	0.816
LR	1.00	0.907	0.978	0.684	0.840	0.637	0.800	0.668	0.889	0.684

Moreover, the coefficient of determination (R^2) is used to study the associations between the input and the outcome variables using the average values of all models. The R^2 for every ML model for testing and training data was evaluated to predict the best-performing model, as shown in Fig. 5. For the quantitative data analysis, an R^2 over 10% is acceptable, whereas an R^2 over 20% represents a large effect. Even a small effect, if observed from the large and complex dataset examined, is statistically significant. All the models were in the largely acceptable range, as the R^2 values were over 20%. However, GB shows the highest correlation with an R^2 of 77.7%, followed by RF with an R^2 of 50%.

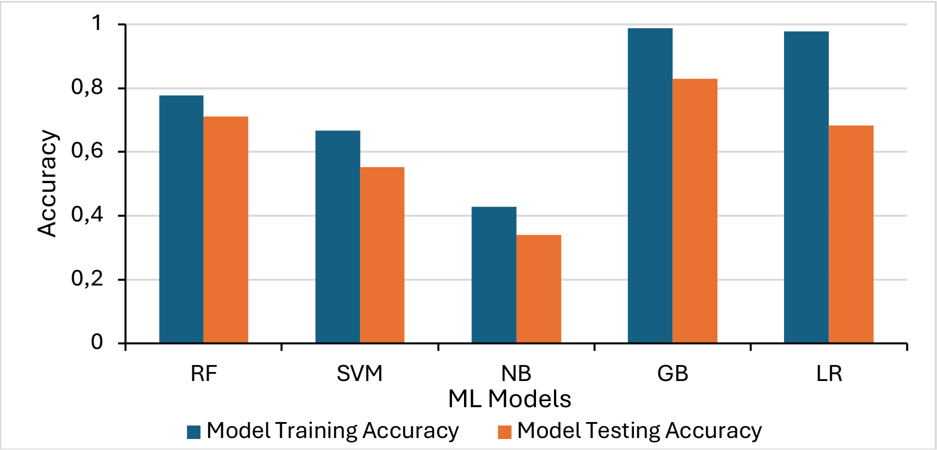

Fig. 4. Accuracy of training and testing models

Predictions from the evaluation tab are used to showcase the model's predictions on input data for both training and testing across all models. The data tab provides a data table that allows the user to view the dataset in a spreadsheet format for further analysis. The points were drawn, and the linear trendline was applied to study the correlation between the data points.

Fig. 6 depicts the model performance of training data for all the models, whereas Fig. 7 depicts the model's performance of the testing data. All the models in the training outperformed those in the testing, as training had 80% of the dataset, while testing contained 20% of the dataset. Among all ML algorithms, GB showed the highest model prediction in both training and testing, with R^2 values of 0.7772 for training and 0.6722 for testing. Although the logistic regression showed the lowest R^2 in both training and testing, it was still in the highest acceptable range of over 20% in both datasets. Overall, GB outperformed other models in terms of classification matrices such as accuracy and precision, coefficient of determination, and model prediction for both the testing and training datasets. Therefore, the current study indicates that GB is the best predictive ML algorithm for the school TMC prediction, followed by RF. The current study is similar to Kashifi et al., who concluded that the GB/LightGBDT model outperformed other models [15].

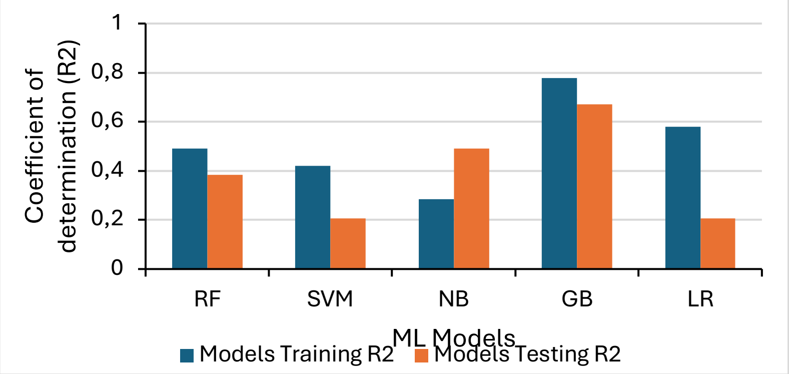

Fig. 5. Coefficient of determination (R^2) of testing and training predictions

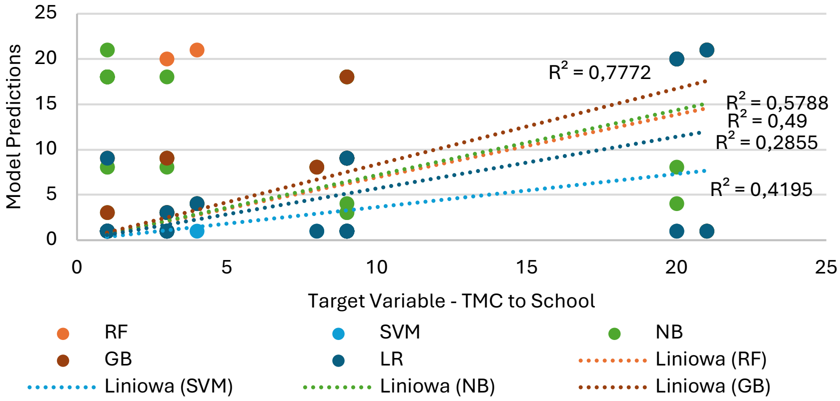

Fig. 6. Training model predictions

In addition, the models were assessed based on average precision (AP)—the area under the precision-recall curve that measures the trade-off. The higher the AP value, the better the model's performance. An area under the curve with a high value indicates both high recall and high precision. High scores for both show that the classifier is generating largely favorable (high recall) and accurate (high precision) results.

The probability thresholds for all the school TMC models were determined and depicted in Fig. 8. The GB model achieved the highest probability threshold, or precision-recall value, of 1.0, followed by the RF model. This indicates that both GB and RF outperformed the other models, demonstrating higher precision and recall values, as confirmed by the performance classification matrix in Table 4. However, the probability threshold of all the models was in the acceptable range, with the precision-recall value for the NB showing the lowest value. A lower precision-recall value means a higher false positive and negative rate, which are the denominators of precision and recall. The AP ranges from 0 to 1, with 1 representing perfect execution and 0.5 showing random guessing. The ROC for school TMC is shown in Fig. 9, which shows that GB and RF outperformed the other models. However, NB shows weak performance. The sensitivity of GB is close to 1, with a low false rate, showing the high accuracy of the model. Moreover, the specificity of the NB model is high and has a low true positive rate (sensitivity), showing the weak accuracy of the model. All the ML algorithms are at acceptable ROC except NB, which is below the linear line, which shows the specific precision level for evaluating classifier performance at this cutoff.

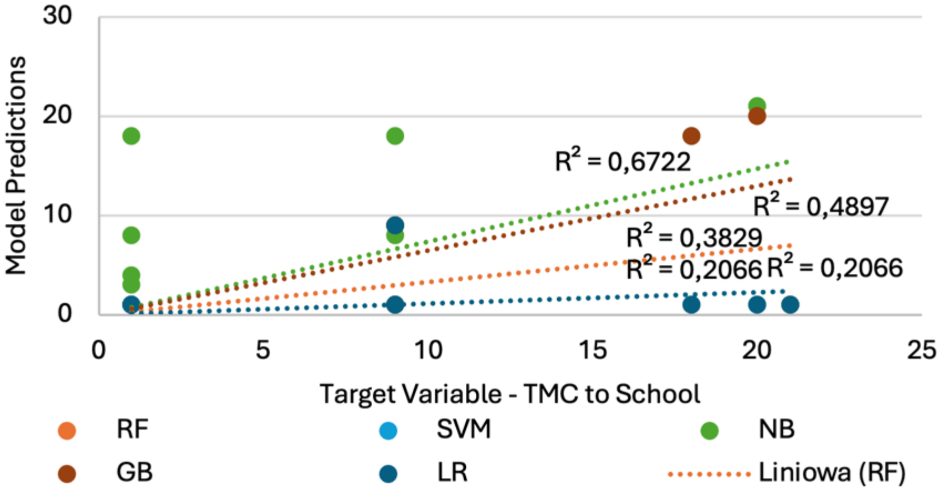

Fig. 7. Testing model predictions

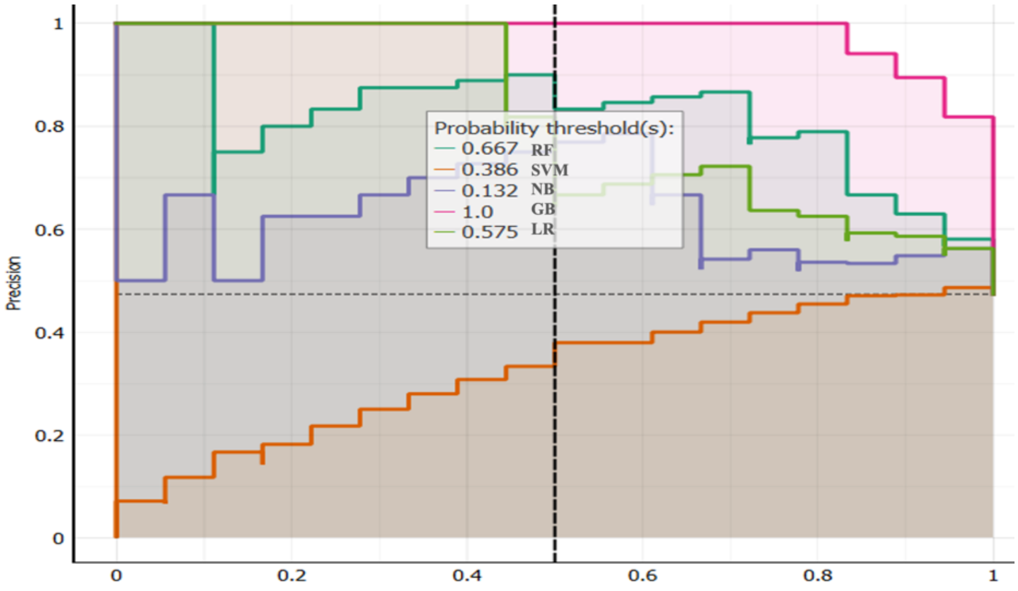

Fig. 8. Precision-recall of TMC to school

4.2. Sensitivity Analysis Results

In addition to the classification metrics, it is essential to identify the most significant features related to the targeted variables. Therefore, feature importance is used to investigate the most significant factor in predicting school TMC. Fig. 10 illustrates the sensitivity analysis used to identify the most significant factor on average of all transport modes—not for each transport option. It was found that age is the most significant factor for the determinants of TMC to school, followed by the type of school. As age changes, the TMC to school also varies. Younger children are typically dependent on their parents, middle-aged individuals tend to use PT, and older individuals often prefer using private vehicles. The current findings align with the outcomes of Le and Teng, who concluded that travel time, age, and gender are ranked high [20].

Merging students above and below the driving license age could influence the results, as access to a car is not equally available across the sample. Using a dummy variable to differentiate between individuals below and above 16 could provide a more nuanced understanding of the data, especially concerning the driving age in the US. This approach would help account for the distinct behaviors between those legally allowed to drive and those reliant on alternative transportation. Moreover, the frequency of transport mode, race, and education level were the most significant factors towards TMC to school. Similar to age, TMC to school also varies with the education level, as those in their secondary school used PT or active transport, while those in high school preferred to use private vehicles. The current findings are similar to the outcomes reported by Saitluanga and Hmangaihzela, who concluded that education level significantly influenced TMC to school [6]. Meanwhile, COVID-19, sex, and health constrain were the lowest significant factors that affect TMC in school. Due to the limitations and restrictions of social distancing, individuals prefer their personal vehicles over PT to pick up or drop off children at school, which significantly influences TMC to school. On the other hand, those who have disabilities or have health problems, which are capability constraints, significantly affect transport options [18].

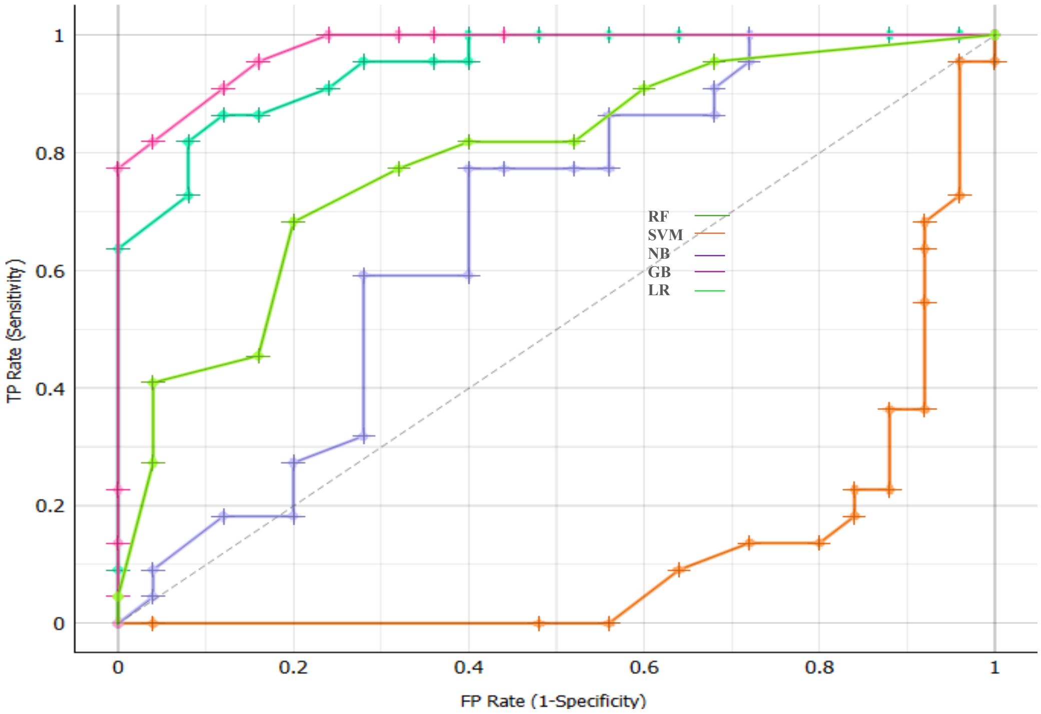

Fig. 9. Mean ROC for TMC to school

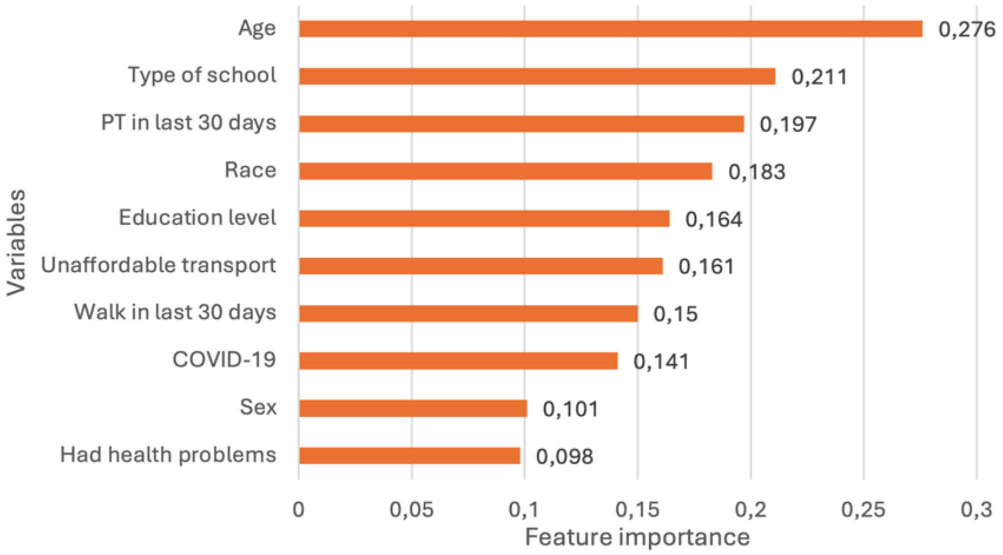

Fig. 10. Sensitivity analysis

5. CONCLUSIONS AND DIRECTIONS FOR FUTURE WORKS

The present study aimed to predict the TMC to school using non-parametric techniques to promote green mobility. Several relevant variables from the 2022 NHTS – California dataset were used to predict the TMC to school. The target variable—TMC to school—has imbalanced classes, such as a high percentage of private vehicles and a low share of active transport to school, which presents a challenging task for conventional techniques, hence the superiority of modern techniques over traditional methods. The current study used ML algorithms to handle the imbalanced classification of data and predict TMC to the school. The ML models were assessed based on the classification metrics that are precision, accuracy, recall, and F1-score. The following conclusions were drawn in light of the findings.

Among all ML algorithms that are applied in the current study, GB outperformed other ML models in both testing and training, with an accuracy of 98.9% in training and 83% in testing. Moreover, RF achieved an accuracy of 71.1% in testing and 77.8% in training. Moreover, NB shows the lowest accuracy in both datasets. The model classification metrics show better performance in training that contained 80% of the dataset rather than in testing. Therefore, the current study claimed that GB and RF are the best predictive models for school TMC prediction. On other datasets, other ML algorithms may outperform GB and RF.

Moreover, the model was assessed by the coefficient of determination (R^2) to investigate the association among the predictors and outcome variables. All models were exceptionally acceptable range, with GB showing the highest R^2 of 77.72% in training and 67.2% in testing, which suggests that GB is an excellent predictive model for the school TMC. Feature importance was conducted to evaluate the most significant features of TMC to school. Age was found to be the most significant determinant, followed by the type of school. Additionally, the frequency of transport mode, race, and education level have been identified as the most significant factors influencing TMC to school. However, COVID-19, sex, and health constraints were the lowest significant factors that affected TMC to school. Several studies have shown that interpretable ML algorithms and hyperparameter optimization outperform typical ML algorithms. Therefore, future researchers should use interpretable ML algorithms for school TMC prediction. Built environments play a crucial role in determining the preferred transport mode to schools, promoting a sustainable transportation system. Therefore, future studies can consider the availability and accessibility of schools in predicting the travel mode choice to schools. Moreover, personal preferences and attitudes towards different transport modes can be considered by future studies to predict the TMC to school.

References

1. McDonald, N.C. & Brown, A.L. & Marchetti, L.M. et al. U.S. School Travel, 2009: An assessment of trends. *American Journal of Preventive Medicine.* 2011. Vol. 41(2). P. 146-151.
2. Lidbe, A. & Li, X. & Adanu, E.K. et al. Exploratory analysis of recent trends in school travel mode choices in the U.S. *Transportation Research Interdisciplinary Perspectives.* 2020. Vol. 6. No. 100146.
3. Mancini, J.M. Census and sustainability: school provision, urban teenagers, and unequal access to active transport in the Republic of Ireland. *Irish Educational Studies.* 2025. Vol. 44(2). P. 1-20.
4. Buehler, R. Determinants of transport mode choice: a comparison of Germany and the USA. *Journal of Transport Geography.* 2011. Vol. 19(4). P. 644-657.
5. Xu, Z. & Aghaabbasi, M. & Ali, M. et al. Targeting sustainable transportation development: the support vector machine and the bayesian optimization algorithm for classifying household vehicle ownership. *Sustainability.* 2022. Vol. 14(17). No. 11094.
6. Saitluanga, B.L. & Hmangaihzela, L. Transport mode choice among off-campus students in a hilly environment: the case of Aizawl, India. *Transport Problems.* 2022. Vol. 17(3). P. 164-172.
7. Liu, Y. & Min, S. & Shi, Z. et al. Exploring students' choice of active travel to school in different spatial environments: A case study in a mountain city. *Journal of Transport Geography.* 2024. Vol. 115. No. 103795.
8. McDonald, N.C. & Deakin, E. & Aalborg, A.E. Influence of the social environment on children's school travel. *Preventive Medicine.* 2010. Vol. 50. P. S65-S68.
9. Uddin, M. & Pan, M.M. & Hwang, H.L. Factors influencing mode choice of adults with travel-limiting disability. *Journal of Transport and Health.* 2023. Vol. 33. No. 101714.
10. Hu, H. & Xu, J. & Shen, Q. et al. Travel mode choices in small cities of China: A case study of Changting. *Transportation Research Part D: Transport and Environment.* 2018. Vol. 59. P. 361-374.
11. Rothman, L. & Macpherson, A.K. & Ross, T. et al. The decline in active school transportation (AST): A systematic review of the factors related to AST and changes in school transport over time in North America. *Preventive Medicine.* 2018. Vol. 111. P. 314-322.
12. Zhang, R. & Yao, E. & Liu, Z. School travel mode choice in Beijing, China. *Journal of Transport Geography.* 2017. Vol. 62. P. 98-110.
13. Ali, M. Discrete Choice models and artificial intelligence techniques for predicting the determinants of transport mode choice – a systematic review. *CMC-Computers, Materials & Continua.* 2024. Vol. 81(2). P. 2161-2194.
14. Qian, Y. & Aghaabbasi, M. & Ali, M. et al. Classification of imbalanced travel mode choice to work data using adjustable SVM model. *Applied Sciences.* 2021. Vol. 11(24). No. 11916.
15. Tamim Kashifi, M. & Jamal, A. & Samim Kashefi, M. et al. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society.* 2022. Vol. 29. P. 279-296.
16. Ali, M. & Macioszek, E. & Onyelowe, K. et al. Interaction of activity travel, GHG emissions, and health parameters using R – A Step towards sustainable transportation system. *Ain Shams Engineering Journal.* 2024. Vol. 15(12). No. 103050.
17. Ali, M. & Macioszek, E. & Yuen, C.W. Health enhancement through activity travel participation and physical activity intensity. *Journal of Transport & Health.* 2024. Vol. 39. No. 101927.
18. Park, K. & Esfahani, H.N. & Novack, V.L. et al. Impacts of disability on daily travel behaviour: A systematic review. *Transport Reviews.* 2023. Vol. 43(2). P. 178-203.
19. Buijs, R. & Koch, T. & Dugundji, E. Using neural nets to predict transportation mode choice: an amsterdam case study. *Procedia Computer Science.* 2020. Vol. 170. P. 115-122.
20. Le, J. & Teng, J. Understanding influencing factors of travel mode choice in urban-suburban travel: a case study in Shanghai. *Urban Rail Transit.* 2023. Vol. 9(2). P. 127-146.

Received 22.11.2023; accepted in revised form 07.06.2025