

MODE CHOICE ANALYSIS OF SCHOOL TRIPS USING RANDOM FOREST TECHNIQUE

Jinit J. M. D'CRUZ¹, Anu P. ALEX², V. S. MANJU³

¹ Office of the Superintending Engineer, Kerala Water Authority, Trivandrum, Kerala, India

^{2,3} Department of Civil Engineering, College of Engineering Trivandrum, Kerala, India

Abstract:

Mode choice analysis of school trips becomes important due to the fact that these trips contribute to the second largest share of peak hour traffic. This scenario is more relevant in India, which has almost 265 million students enrolled in different accredited urban and rural schools of India, from Class I to XII as per the UDISE report of 2019-20. Thus, it becomes necessary to understand what mode of transport will be mostly used for school trips in order to design an efficient transportation system. Modal attributes and socio-economic characteristics are mostly considered as explanatory variables in travel mode choice models. Multinomial Logit (MNL) model is one of the classic models used in the development of mode choice models. These logistic regression models predict outcomes based on a set of independent variables. With the recent advances in machine learning, transportation problems are getting a wide arena of methods and solutions. Among them the method of ensemble learning is finding a prominent place in contemporary modelling. This study explores the potential of using ensembles of random decision trees in mode choice analysis by Random Forest Technique with a comparative analysis on conventional method. It was observed that Random Forest method outperforms MNL method in predicting the mode choice preference of students. The high accuracy of machine learning models is mainly due to its ability to consider complex nonlinear relationship between socio-economic attributes and travel mode choice. These models can learn and identify pattern characteristics extracted from sample data and form adaptive structures through computational process thereby offering insights into the relationships between variables that random utility models cannot recognize. This study considered activity-travel information, personal data and household characteristics of students as attributes for model development and observed that the age of the student and distance of school from home plays a significant role in deciding the mode choice of school trips.

Keywords: Mode choice, Random Forest Technique, Decision trees, Multinomial logit model

To cite this article:

D'cruz, J.J.M., Alex, A.P., Manju, V.S., (2022). Mode choice analysis of school trips using random forest technique. *Archives of Transport*, 63(3), 39-48. DOI: <https://doi.org/10.5604/01.3001.0015.9175>



Contact:

1) jjjinit@gmail.com [<https://orcid.org/0000-0002-9012-7595>]; 2) anualex@cet.ac.in [<http://orcid.org/0000-0001-7172-627X>] – corresponding author; 3) manjuvs@cet.ac.in [<https://orcid.org/0000-0003-2084-561X>]

1. Introduction

Utility of a mode differs from person to person depending on various socio-economic factors. A commuter's choice of mode is primarily based on the utility of available modes and their preferred mode will be the mode with highest utility. Discrete choice models continue to be the widely used method in the past for analysing the mode choice behaviour of commuters (Anas, 1983). They are basically statistical models using regression framework and follows the theory of random utility maximization. Among them, logit models are commonly used due to its capability to realistically represent individual choice behaviour (Koppelman, 1983; McFadden, 1973). Numerous studies during the last couple of decades have investigated the mode choice behaviors using logit models such as Binary logit, MNL, Nested logit and Mixed logit models. Multinomial Logit (MNL) model is one of such classical models which assume a predefined underlying relationship between independent and dependent variables.

There are several variables that affect the travel mode choice of a commuter (Ortuzar and Willumsen, 2011; Rahman, 2020; Babu et al., 2018). Literature shows that socio-demographics play an important role in travel mode choice decisions. This mainly includes gender, education, employment, income, and vehicle ownership (Bhat and Sardesai, 2006). In addition to that, the land use and accessibility are two interlinked factors that determines the favoured travel mode for a commuter (Pinjari et al., 2007). Extent of willingness to pay to reduce their travel time by a unit is another factor which has been studied in many cases. Certain recent research studies (Li et al., 2010) incorporated reliability of travel time as a factor in mode choice behaviour. However, in the case of school trips, the mode choice will also be influenced by the preference of the elders in the family. A poor walking environment will increase automobile dependence and discourage elders to support walking or cycling to school. "Poor walking environment" means a built environment of low densities, little mixing of land uses, long blocks, incomplete sidewalks, and other hallmarks of sprawl. The study on significance of the relationship between mode choice and perceived distance from home to school shows that the probability of traveling by automobile instead of foot increases from 20% at a 0.8 km distance to 50% at 2 km and 80%

at 3.2 km (Black et al., 2004). Students with shorter walk and bike times to school were more likely to walk and cycle (Ewing et al., 2004).

Mode choice behaviors for the school trips were explained using MNL method in many studies (Wilson et al., 2010; McDonald, 2008; Yarlagaadda and Srinivasan, 2008; Ewing et al., 2004; Ashalatha et al., 2013). These MNL models have their own model assumptions and properties. It assumes that the ratio of probabilities of any two alternatives is independent of the choice set (Ben-Akiva and Lerman, 1985). MNL method is widely accepted in mode choice modelling because of its mathematic structure which eases parameter estimation. However, violation of these assumptions results in biased predictions. The MNL model is also criticized for its Independent of Irrelevant Alternatives (IIA) property which makes it difficult to account for choice variations among different individuals. In order to address the aforementioned discrepancies that affect the accuracy of the models, machine learning methods can be adopted as a promising alternative.

Machine-learning classifiers are effective in modeling travel behaviour of an individual (Rasouli and Timmermans, 2014; Alex et al., 2019). Classification and Regression Trees (CART), Naive Bayes (NB), Boosting Trees (BOOST), Bagging Trees (BAG), Support Vector Machine (SVM), Neural Networks (NN) and Random Forest (RF) are some of the machine learning classifiers (Zhao et al., 2020). These classifiers do not depend on the underlying data structure and hence eliminates biased predictions. It is seen as a subset of artificial intelligence. Some of the prominent machine learning models are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Tress (DT), Bayesian Network and Genetic Algorithm (GA) (Abduljabbar et al., 2019; Alex et al., 2016; Alex et al., 2021). Machine learning allows flexible model structures and represents complex relationship with its extraordinary ability to learn and recognize patterns. Machine-learning method considers mode choice prediction as a classification problem and identifies the single best performing model. This can also cause inconsistent predictions when the input data contains errors or bias. Advanced classifiers such as Random Forest (RF) uses novel approach for data exploration and analysis and can minimize these problems with their tree-structured non parametric classification technique.

RF method is proposed by Breiman (2001) and has its base in CART (Classification and Regression Tree). RF is a collection of individual decision trees that operates as an ensemble. Ensemble algorithms are more powerful for predictions and modelling because it produces results by combining predictions from multiple models. This optimizes the predictive performance. The issue of behavioral heterogeneity can be addressed with the help of random forest approach. The error in model forecasting can be reduced by applying model ensembles rather than a single model. Bagging, Boosting, Adaboost, Stacking and Random forest are some of the most popular ensemble learning algorithms. Random forest is actually an extension of bootstrap aggregation or bagging. Random Forest creates uncorrelated decision trees using bagging and feature randomness for each tree creation. Random Forest method develops individual decision trees by combining a subset of the explanatory variables. This reduces the generalization error. Bagging is a statistical estimation technique where a statistical quantity like a mean is estimated from multiple random samples from the data with replacement. In the case of bagging, the algorithm selects the best split point at each step in the tree building process. On the other hand, RF algorithm creates individual trees using the split points that are selected from a random subset of the input attributes. So here each tree is created via bagging without replacement. Random forest can also be used to rank the importance of a variable by measuring the out-of bag error in data points. The test set error will be estimated internally during the run and hence there is no need for a separate test set or cross-validation (Ben-Akiva and Lerman, 1985). Recent studies on RF method shows its effectiveness in solving transportation prediction and classification problems (Gong et al., 2018). A study on the driving behaviour of travellers at signalized intersection shows that RF method performs well with mixed types of data and exhibits high prediction power in multi category classification problem. Certain studies on prediction of travel time using RF method shows that RF method doesn't require data pre-processing and can fit complex nonlinear relationships. This study explores the potential of using ensembles of random decision trees in mode choice analysis by Random Forest Technique along with a comparative analysis on conventional MNL method. The following three points are of major interest. (a) Ascertain

the variables that determine the mode choice of school trips. (b) Develop mode choice models using random forest technique and MNL method. (c) Compare the results of ensemble learning RF method with conventional MNL method. It helps to predict the modes used by the students for making school trips, which will be helpful in the transportation planning process.

2. Study area and data collection

The study analyses various factors commonly affecting mode choice for school trips in Thiruvananthapuram city (Fig. 1.) and modelled it. This capital city of Kerala is an educational hub and is endowed with a galaxy of educational institutions, most of which are located in the centre of the city. Home-to-school trips and school-to-home trips were considered in this study. These trips increase the congestion at peak hours. As per the 2011 Census report, the district has a population of 33, 01,427 with 55.75% urban population. Among that the student population is 3, 85,464. The projected population for the year 2020 is found as 39,77,428 at a population growth rate of 2.07%. It is necessary to analyse relevant characteristics of traffic as well as students to study the mode choice behaviour of students. A total of 364 household samples were collected from the study area. Data were collected randomly from each ward of the study area.

Data collected for this study includes socio-economic and highway network characteristics. Socio-economic data was collected by home interview survey. It consisted of household information, personal information and activity- travel information. Household information included location of the household, type of dwelling unit, household size and vehicle ownership. Personal information included gender, age, occupation and monthly income. Activity-travel information consisted of details of the activities of students, school start time, duration of school activity, mode of travel, travel cost, travel distance, travel time and willingness to shift to public/ shared transit modes. The gender wise distribution of the sample data shows that the study area has almost equal distribution of male and female student population. About 71% of the students fall under the age group of 5-17, 28% is between 18-25 and only 1% between the ages of 26-40. Vehicle ownership details showed that 45% of households have two-wheeler as well as car, followed by 35% having two

wheelers only. Major share of school trips was made using bus/van, two-wheeler, cycle/auto/walk. Thus, these three categories of modes were considered for developing the mode choice models. 91% of students have their school starting time between 8.00 a.m. and 10.00 a.m. which coincides with the office starting hours leading to an increased peak hour traffic.

3. Methodology

Two modelling formulations namely, MNL model and Random Forest method, were applied to investigate the key factors in mode choice for school trips. Correlation of variables were determined and highly correlated variables were eliminated for model development. The significant variables that affect the mode choice were found as two-wheeler availability, gender, age group and distance. Development of the models is explained in the following.

3.1. Multinomial Logit Model

In conventional mode choice modelling methods, the basic assumption is that each individual is attempting to maximize his utility. The concept of utility assumes that there is a method of combining

the various features of all the alternatives to give one measure of utility which is consistent across all the alternatives within the choice set. MNL is one of such classic utility-based models confining to the method of logistic regression of classification. The individual, who is the decision maker will choose an option from the choice set if and only if the utility of that option is greater than or equal to the utility of all other options in the choice set. The utility of an alternative can be decomposed into two components from the perspective of decision maker. One component of the utility function is called the deterministic (or observable) portion, which can be observed by the analyst. The other component is the difference between the unknown utility considered by the individual and the utility estimated by the analyst. It can be represented using eq. (1).

$$U_{iq} = V_{iq} + \varepsilon_{iq} \quad (1)$$

Where U_{iq} is the true utility of the alternative i to the decision maker q , V_{iq} is the deterministic portion of the utility estimated by the analyst, and ε_{iq} is the error portion of the utility unknown to the analyst.

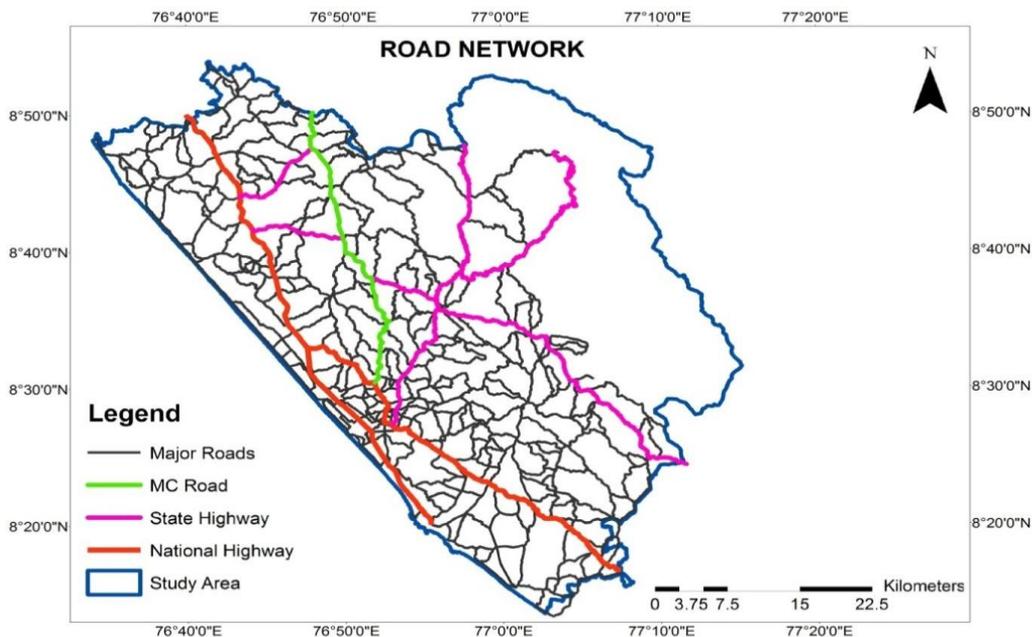


Fig. 1. Study Area

The deterministic portion V_{iq} depends mainly on the attributes of the alternatives in the choice set. This measure may vary across alternatives for the same individual and also among individuals due to difference in their preferences. Thus, it is formulated as a linear combination (Koppelman and Bhat, 2006) as in eq. (2).

$$V_{iq} = V(S_q) + V(X_i) + V(S_q, X_i) \quad (2)$$

Where $V(S_q)$ is the portion of utility associated with characteristics of individual q , $V(X_i)$ is the portion of utility associated with the attributes of alternative i and $V(S_q, X_i)$ is the portion of utility which results from the interaction between the attributes of alternative i and the characteristics of individual q . SPSS software was used for MNL mode choice modelling, which gives the choice probabilities of each mode as a function of the systematic portion of the utility of all the modes. Based on the hypothesis of rational choice, Probability of alternative i chosen by student q can be formulated as given in eq. (3).

$$P_n = e^{V_n} / \sum_{m=1}^M e^{V_m} \quad (3)$$

Where P_n is probability that the individual selects the mode n , V_n is utility of mode n , V_m is utility of any mode and M is set of all available student mode. Estimation results of MNL model is shown in Table 1. The likelihood ratio tests indicating the contribution of each variable to the model were obtained as shown in Table 2. The likelihood ratio test is a hypothesis test that the variable contributes to the reduction in error measured by the -2-log likelihood statistic. All the variables presented had significant parameter estimates and logical signs. Table 3 shows the statistical comparison of intercept-only model and final model.

'Intercept only' model does not include any predictor variables and simply fits an intercept to predict the output variable. 'Final' model describes a model that includes the specified predictor variables and has been arrived at through an iterative process that maximizes the log likelihood of the outcomes seen in the output variable. The final model was an improvement on the intercept-only model by including the predictor variables and maximizing the log likelihood of the outcomes seen in the data. The likelihood ratio test showed the contribution of each variable to the model. The chi-square statistic is the

difference between the -2-log -likelihoods of the null or intercept-only and final models. As the significance level of the test was less than 0.05, it can be concluded that the final model outperforms the null model. In this model, chi-square value of 195.019 has significance (0.000) which is less than 0.001, so there is a significant relationship between the dependent variable and the set of independent variables. In the MNL model, the age and distance were found as significant variables affecting the school mode choice behaviour. Both the variables have statistically significant ($P < 0.05$) contribution to the explanation of mode choice behaviour of students. The coefficients for gender were positive for van/bus, which implied that males were more likely to use buses or van than two wheelers for school trips. However, in the case of walk/auto/cycle, females were more likely to use those modes than two wheelers. As the coefficient of distance is positive, it indicates that when distance increases students prefer bus as their mode for school trips to two wheelers.

3.2. Random Forest Method

Random Forest method uses decision tree as base classifier. This method encompasses tree predictors and two randomization principles: - bagging and random feature selections. The approach in building a tree is to split the data using the explanatory variables at each node until the final level, the target variable is obtained. During this process at each node, the algorithm finds the best independent variable to be used for splitting the data using the eq. (4) in which p_i is the proportion of entropy for each variable belonging to class i , and c is the number of values in each variable.

$$\text{Entropy} = \sum_{i=1}^c (-p_i \log p_i) \quad (4)$$

The algorithm splits the data and continues to the next step. In classification random forest, number of trees are generated and then based on the vote of each tree, the category of the target variable is determined with the highest vote. All trees are different from each other and they are built on randomly drawn subset of data which then results in selection of different explanatory variables. Fig. 2 illustrates the sample tree structure and a closer view of selected area.

Table 1. Parameter Estimates

	Mode	B	Standard Error	Wald	DF	Sig.	Exp (B)
Walk/Cycle/Auto	Intercept	7.307	2.527	8.362	1	0.004	
	2W availability	2.556	1.307	3.824	1	0.051	12.890
	Gender	-0.262	0.603	0.188	1	0.665	0.770
	Age Group	-2.553	0.673	14.370	1	0.000	0.078
	Distance	-2.144	0.429	24.958	1	0.000	0.117
Van/ Bus	Intercept	-2.436	1.559	2.440	1	0.118	
	2W availability	2.794	1.119	6.233	1	0.013	16.354
	Gender	0.820	0.421	3.798	1	0.051	2.270
	Age Group	-2.318	0.425	29.783	1	0.000	0.098
	Distance	0.830	0.221	14.102	1	0.000	2.293

Note: The reference category is: Two-wheeler

Table 2. Model Fitting Information

Effect	Model Fitting Criteria		Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi- Square	df	Sig.	
Intercept	154.267	32.753	2	.000	
Two wheeler availability	133.581	12.067	2	.002	
Gender	127.913	6.399	2	.041	
Age Group	163.463	41.949	2	.000	
Distance	264.192	142.678	2	.000	

Table 3. Values of Likelihood Ratio Tests

Effect	Model Fitting Criteria		Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi- Square	df	Sig.	
Intercept Only	316.533	195.019	8	.000	
Final	121.514				

One of the most important results of a random forest regarding variable selection is the measure of effectiveness which each predictor has while describing the data. The importance of variables comes from a permutation process in the original database. To examine the importance level of each variable separately, the quantity of the mentioned variable changes randomly to a different value and the results are re-estimated. If the number of misclassified data is considerable, it is concluded that this variable has a significant role in the process of random forest modelling and predicting. The average of variable importance measures is then calculated for all the

classes in the total data and it is called mean decrease accuracy index. Variable importance of X_j is calculated using eq. (5)

$$VI(X_j) = \frac{1}{ntree} \sum_{f=1}^{ntree} (\text{error } OOB_f - \text{error } OOB_{fj}) \quad (5)$$

Where $ntree$ is the number of built trees in the forest, $errorOOB_f$ is the error rate before permutation and $errorOOB_{fj}$ is the error rate after the permutation of variable X_j .

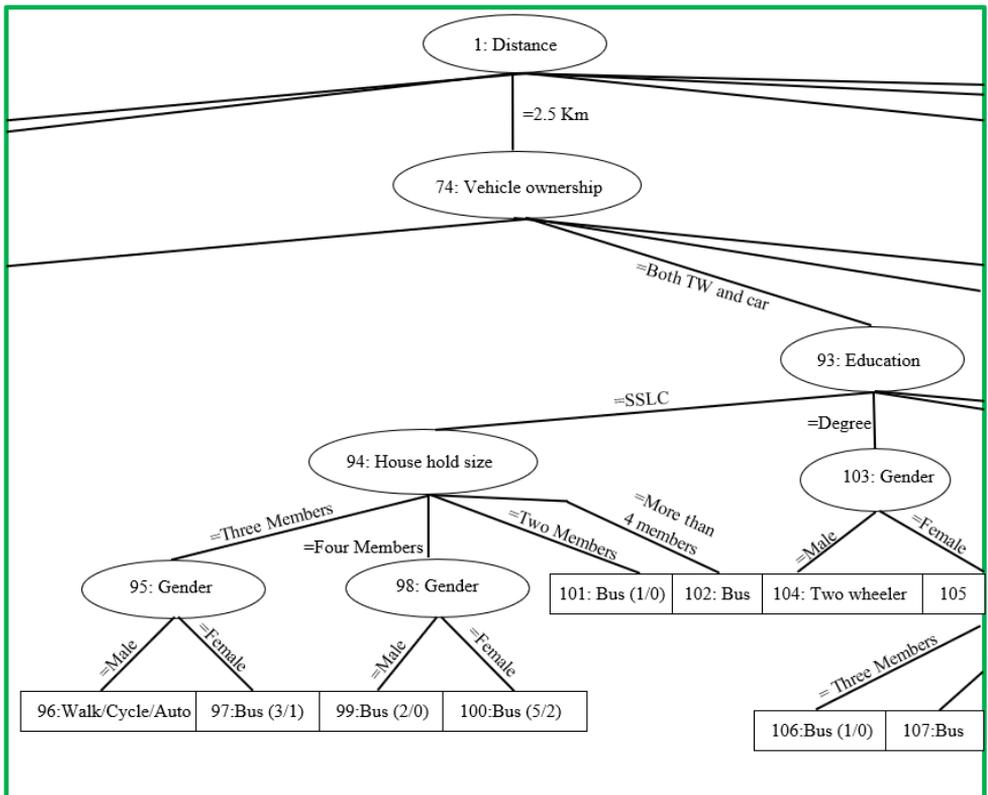
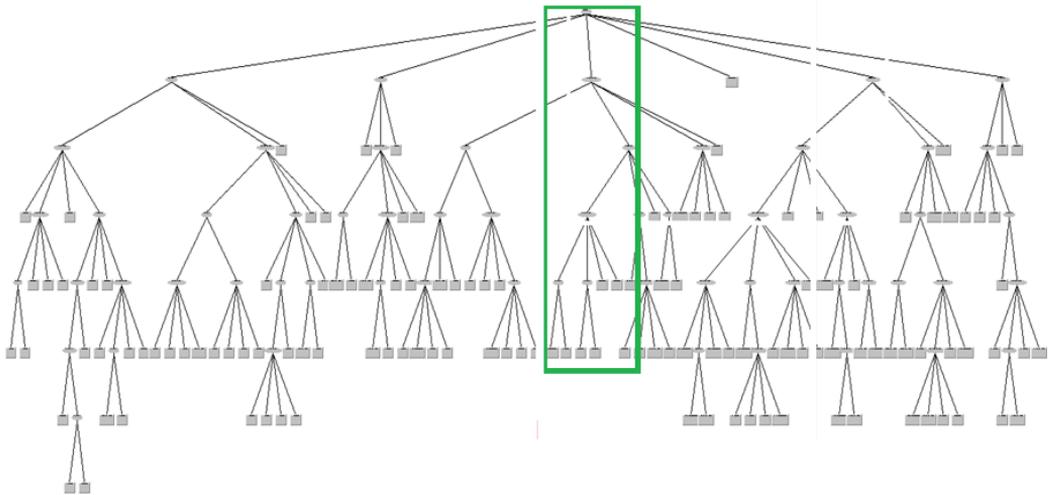


Fig. 2. Random Forest Decision Tree

WEKA software (WEKA 3.9) has been considered for performing the mode choice analysis using Random Forest Decision Tree (RFDT) model and the results obtained by these models are discussed in the following section. The result of mode choice analysis in terms of prediction accuracy of RFDT model is presented in Table 4. The various statistical measures that are employed to measure the statistical significance of the RFDT model are Kappa statistic, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative Absolute Error (RAE). Kappa statistics compares observed accuracy with expected accuracy. It is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves by taking into account random chance.

Landis and Koch (1977) proposed kappa values 0.00-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and 0.81-1.00 as almost perfect. Thus the obtained RFDT model represents almost perfect results with kappa statistics value of 0.8244. The mean absolute error is a quantity which is used to measure how close predictions are to the eventual outcomes. A lower value of mean absolute error implies better predictability on part of model. RFDT model attains a much lower value of 0.0571. RFDT model depicts a lower RMSE value of 0.1481. RAE is calculated by taking the total absolute error and normalizing it by dividing by the total absolute error of the RF classifier. Its value ranges from 0 to infinity, with 0 corresponding to the ideal. The obtained RFDT model depicts a value of 0.3122 which is closer to ideal value. After developing the models, both multinomial logit model and random forest mode choice model were validated using 20% of samples collected, which was not taken for the purpose of modelling. The models were also compared with prediction accuracies.

Table 4. Statistical Measures of Random Forest Model

Evaluation Parameter	RFDT Mode Choice Model
Correctly classified samples	92.47%
Incorrectly classified samples	7.53%
Kappa statistic	0.8244
Mean Absolute Error (MAE)	0.0571
Root Mean Squared Error (RMSE)	0.1481
Relative Absolute Error (RAE)	31.22 %

4. Model validation

The developed models were validated in two phases. The first phase was measuring model statistics and second was measuring the prediction success table of the models. The training set used for developing the model consists of 292 school trips and the test set used for validation consists of 72 school trips. A ten-fold iteration technique has been used for RFDT model for best possible solution. The models which qualified model statistics measurements were only considered for prediction success table. Model statistics for MNL model was done by inspecting sign and significance of coefficient, checking standard error of coefficient, measuring log likelihood function, predictive ability of the model, and McFadden pseudo R square value. In the case of random forest model, the kappa statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE) were considered to measure the statistical significance of the model. Prediction results for MNL and RFDT model is given in Table 5. The prediction ability of the RFDT model is 83.34%, whereas that of MNL model is 76.38%.

Table 5. Prediction results of MNL and RFDT models

TransportMode	MNL model		RFDT model	
	Correctly Predicted	Wrongly Predicted	Correctly Predicted	Wrongly Predicted
Bus	45	9	50	4
Two- Wheeler	3	4	2	5
Walk/Cycle/Auto	7	4	8	3
Prediction Accuracy	76.38%		83.34%	

5. Conclusion

Mode choice behavior among school children draws more attention among policy makers and urban planners nowadays. There are various studies conducted during the last decades, which examined the mode choice behavior of school children. Various descriptive and analytical studies were conducted, where influential variables were identified in the first and power of variables were recognized using various data mining methods in the latter. The present study focused on examining the feasibility of applying an emerging data mining technique to school trip mode

choice modelling. Random Forest method, an advanced data mining approach was used to develop the model and the results were compared with widely used MNL model. This study also investigated the influence of various socio-economic attributes on mode choice of school trips. The attributes that are found relevant are vehicle availability, household size, age and gender of the student, level of education and distance of school from home. It is found that age of the student and distance of school from home plays significant roles in the mode choice and observed that as distance increases, students prefer bus to two wheelers.

It is found from the present study that out of the two models developed, Random Forest technique significantly outperforms the logit model in prediction accuracy. This high performance is mainly due to its adaptability in handling big data and making predictions using majority voting among all classification trees. Variable interactions can be efficiently analysed in RF method to model complex nonlinear relationships. On the other hand, logit models are explicitly formulated against a theoretical framework which neglects the nonlinear effects among the variables. The prediction accuracy of RF method can further be increased by incorporating more detailed and specific attribute data set. The novel concept of meta random forest wherein the random forest themselves were used as base classifiers for making ensembles can provide more accurate results. Future scope of the study will be such improvised RF methods in which more diverse and less correlated base classifiers are utilized for improved accuracy. From a practical point of view, these kinds of machine learning methods establish opportunities for developing enhanced analytical techniques to interpret high dimensional human travel behaviour.

Acknowledgment

This work was supported by Kerala State Council for Science, Technology and Environment under ETP 09/12 Grant.

References

[1] Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S., (2019). Applications of Artificial Intelligence in Transport: An Overview, *Sustainability*, 11(1): 189.
[2] Anas, A., (1983). Discrete choice theory, information theory and the multinomial logit and

gravity models, *Transportation Research Part B: Methodological*, 17(1), 13-23.

- [3] Alex, A. P., Manju, V.S., Isaac, K. P., (2016). Modelling Of Daily Activity Schedule of Workers Using Unsupervised Machine Learning Technique, *International Journal for Traffic and Transport Engineering*, 6(1), 77 – 91.
[4] Alex, A. P., Manju, V.S., Isaac, K. P., (2019). Modelling of travel behaviour of students using artificial intelligence. *Archives of Transport*, 51, 7-19.
[5] Alex, A. P., Manju, V.S., Isaac, K. P., (2021). Modelling of Activity-Travel Pattern with Support Vector Machine, *European Transport \ Trasporti Europei*, 82, 1825-3997. <https://doi.org/10.48295/ET.2021.82.2>.
[6] Ashalatha, R., Manju, V. S., Zacharia, A. B., (2013). Mode Choice Behavior of Commuters in Thiruvananthapuram City. *Journal of Transportation Engineering*, 139(5), 494-502.
[7] Babu, D., Balan, S., Anjaneyulu, M. V. L. R., (2018). Activity-travel patterns of workers and students: a study from Calicut city, India. *Archives of Transport*, 46(2), 21-32.
[8] Ben-Akiva, M., Lerman, S.R., (1985). Discrete choice analysis: theory and application to travel demand. (18th ed.). *Cambridge: MIT Press*, (Chapter 11).
[9] Bhat, C.R., Sardesai, R., (2006). The Impact of Stop-Making and Travel Time Reliability on Commute Mode Choice. *Transportation Research Part B: Methodological*, 40 (9), 709-730.
[10] Black, C., Collins, A., Snell, M., (2004). Encouraging Walking: The Case of Journey-to-School Trips in Compact Urban Areas. *Urban Studies*, 38 (7), 1121-1141.
[11] Breiman, L., (2001). *Random Forests. Machine Learning*, 45, 5-32.
[12] Ewing, R., Schroeer, W., Greene, W., (2004). School Location and Student Travel Analysis of Factors Affecting Mode Choice. *Transportation Research Record*, 1895(1), 55-63.
[13] Gong, L., Kanamori, R., Yamamoto, T., (2018). Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. *Travel Behaviour and Society*. 11, 131-140.

- [14] Koppelman, F. S., (1983). Predicting transit ridership in response to transit service changes. *Journal of Transportation Engineering*, 109(4), 548–564.
- [15] Koppelman, F. S. Bhatt, C., (2006). A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models. U.S. Department of Transportation, *Federal Transit Administration*, (Chapter 3).
- [16] Li, Z., Hensher, D. A., Rose, J.M., (2010). Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation Research Part E: Logistics and Transportation Review*, 46(3): 384-403.
- [17] Mcdonald, N. C., (2008). “Children’s mode choice for the school trip: the role of distance and school location in walking to school. *Transportation* , 35, 23–35.
- [18] Mcfadden, D., (1974). Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P., (Eds.), *Frontiers in Econometrics* 105-142. Academic Press, NY.
- [19] Ortuzar, J.D., Willumsen, L.G., (2011). *Modeling Transport*. (4th ed.). Chichester: Wiley, (Chapter 6).
- [20] Pinjari, A. R., Pendyala, R. M., Bhat, C.R., Waddell, P.A., (2007). Modelling residential sorting effects to understand the impact of the built environment on commute mode choice. *Transportation*, 34(5), 557- 583.
- [21] Rahman, F.I., (2020). Analysing the factor influencing travel pattern and mode choice based on household interview survey data: a case study of Dhaka city, Bangladesh. *Scientific Journal of Silesian University of Technology, Series Transport*, 109, 153-162.
- [22] Rasouli, S., Timmermans, H.J., (2014). Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research*, 14 (4), 412-424.
- [23] Wilson, E.J., Marshall, J., Wilson, R., Krizek, K. J. (2010). By foot, bus or car: children's school travel and school choice policy. *Environment and Planning A*, 42(9), 2168-2185.
- [24] Yarlagadda, A.K., Srinivasan, S. (2008). Modelling Children’s School Travel Mode and Parental Escort Decisions. *Transportation*, 35(2), 201-218.
- [25] Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., (2020). Prediction and behavioural analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 20, 22–35.