International Journal of GEOMATE, Oct., 2019 Vol.17, Issue 62, pp. 219 - 230
ISSN: 2186-2982 (P), 2186-2990 (O), Japan, DOI: https://doi.org/10.21660/2019.62.94618
Special Issue on Science, Engineering & Environment

CAR OWNERSHIP DEMAND MODELING USING MACHINE LEARNING: DECISION TREES AND NEURAL NETWORKS

Patiphan Kaewwichian [1], * Ladda Tanwanichkul [1] and Jumrus Pitaksringkarn [2]

[1]Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand; [2]Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

*Corresponding Author, Received: 28 Feb. 2019, Revised: 28 March. 2019, Accepted: 15 April. 2019

ABSTRACT: A household car ownership modeling is crucial in understanding the impact on an individual’s or a family’s travel behavior in traveling demand analysis. Trips or tours as a unit of analysis can be used in the modeling of car ownership demand for analyzing travel needs. Machine learning is widely used to describe a car owner’s decision since the machine learning model was specifically designed to give more accurate predictions through a variety of mechanisms. This research presents car ownership modeling using two types of machine learning models, including decision trees and neural networks. The impacts of socio-demographic attributes on household car ownership demand are discussed and compared against these two models after adding the main attributes of variables from tour-based models. Data was collected from 2,015 households surveyed in Khon Kaen Province, Thailand, conducted in 2015. The outcomes indicate that the machine learning model can be used to predict household car ownership. It also found that when using the default parameters across all datasets, whereas the neural networks provide a more accurate result than the decision tree algorithm. However, in cases where the household car ownership prediction from the dataset with add attributes of the key variable used in tour-based models. In that case, the neural networks algorithm would give a prediction accuracy that corresponded to results as found from the prediction using a dataset with only the household's socio-demographic attributes.

Keywords: Discrete choice models; Primary destination; Tour-based model; Tour generation; Tour type

1. INTRODUCTION

A household car ownership model is essential for traveling demand analysis since it plays a significant role in the process of transportation planning. The model has been widely used as a critical factor to understand the impact on an individual’s or a family’s traveling behavior [1], such as a trip-based model, which is part of the aggregate models. In a trip-based model, this household car ownership model has impacts on a choice of trip frequency and destination choice for each trip – e.g., a home-based trip (HB) or a non-home-based trip (NHB). Moreover, the model also impacts on the trip mode choice, which leads to either work-activity destinations or non-work activity destinations [2,3]. In the case of tour-based and activity-based models, the household car ownership model encourages a tour generation related to a primary destination choice of a tour type and a trip mode choice of a tour [4,5].

Higher demand for car ownership demand encourages more trips and also generally proficiency in creating more numbers of trips chaining. Particularly in the case of a developing or expanding town, many individuals seem to travel by private cars, since it is more convenient and flexible for them [6]. Subsequently, increased use of private vehicles impacts to air quality and can lead to health problems [7,8]. In 2010, 14% of all air pollution was from the transport sector; therefore, the transport sector is a significant aspect of creating air pollution in the world [9]. As a result, an understanding of the number of cars that each household want to own and the operating variables on the car ownership in each family is necessary for analysis on the individual’s traveling demand and the planning of transportation investment and policy.

Generally, a household uses as a unit of decision-making in disaggregate auto ownership model. The model assembles all the homes and uses them to predict car ownership demand at a traffic analysis zone (TAZ). In particular, this model contains more behaviors structurally compared to the aggregate models [10], and it can better detect a causal relationship between the ownership demand factors and the car ownership level [11,12]. Accordingly, this disaggregate model is a popular method used to create a car ownership demand model. However, the variables to show household auto ownership are mostly categorical variables, so the disaggregate auto ownership models widely use a discrete choice analysis (DCA) technique, to estimate car ownership [13].

Meanwhile, a multinomial logit (MNL) model is a model with a discrete choice analysis technique based on a random utility maximization principle that has been broadly used to estimate car ownership demand [14]. This model is the most applicable due to its simple structure and the short period of estimation [15]. However, the MNL model is unable to be used when the IIA (Independence from Irrelevant Alternatives) does not exist. Namely, the ratio of choice probabilities between any two choices should not depend on other options or not depend on increasing or decreasing the number of options [16,17].

Limitations of a discrete choice model can overcome by machine learning models, such as decision trees (DT) and neural networks (NN) algorithms, which are accessible and state-of-the-art approaches [18,19]. Machine learning models were created to provide a more accurately predicted result, decreasing prediction error instead of estimation error. That is, the model becomes more effective when predicting unseen data. Additionally, the machine learning models provide more advantages than discrete choice techniques, e.g., an ability to create a non-linear system as an extremely complicated relationship within an individual's behavior. In terms of model development, it is unnecessary to define a model structure in advance. There is no need to establish an Independence of Irrelevant Alternative (IIA), so it is possible to reduce an incompatibility between the model structure and the analytical data [20,21]. Therefore, this research investigates the capability and performance of car ownership demand modeling of two widely used data mining methods: decision trees (DT) and neural networks (NN), with three answer classes: 0, 1, and 2+. Moreover, this research aims to provide an increase in the efficiency of predictions for car ownership modeling, by adding the main attributes of variables of each area types from tour-based and activity-based models, such as primary destination location and tour type, etc., in addition to the household engineering attributes alone.

This research was prepared in the fifth section (the first being this introduction). The second portion is an explanation of the machine learning models. The third section presents the area of study and experimental design. The fourth section provides outcomes and discussion. Lastly, the conclusion and suggestions for further research will be discussed in the fifth section.

2. METHODS

The modeling procedure began by converting raw data into variables or predictors. During this stage, data were categorized into socio-demographic engineering attributes, tour attributes, and zone attributes (household income size, tour type, and a primary destination, etc.). The variables were then used to create and test the performance of the car ownership demand model with three answer classes (0, 1, 2+) using a cross-validation method. The first ten variables with optimized weight were primarily considered.

The key indicator of the model's performance was the accuracy derived from the confusion matrix table, a square table with dimensions equal to the number of answer classes. For instance, if there were two classes ("Yes" or "No"), the table would have a dimension of 2x2, where the class in the columns included the actual answers while the class in the rows were the predicted answers. From the table, the accuracy value was a ratio between the correct predicted value (on a diagonal cell) and the total predicted values. However, there were other vital indicators available, such as precision, recall or individual match rate, and F-measure.

2.1 Machine Learning Models

In the case of the disaggregate models, the discrete choice theory was widely applied to make a transportation plan. By this conceptual framework, the individual's travel decision-making would be simulated from the utility-maximizing choice of the available choice set. This method was also used for and applied to design a transportation plan, especially for the trip mode choice and destination choice. This could additionally be applied to the car ownership demand model. Nevertheless, there are still some advanced problem-solving techniques purposively developed to the problem of the car ownership demand model, such as decision trees and neural networks that are algorithms of the machine learning models (ML).

The machine learning method, as previously mentioned, was developed to provide more accurate predictions, e.g., the car ownership demand prediction. From a statistical perspective, the machine learning method is an automated exploratory analysis on a large dataset. The engineering attribute of the machine learning method does not need changing the data format, as is required for the discrete choice modeling. Indeed, the machine learning model (ML) is commonly a non-linear model that uses cross-validation to create and test the performance of the model to obtain a high-performing model that is applicable in predicting unseen data.

2.1.1 Decision Trees

Decision trees (DT) are one method of machine learning that has been widely used since they were able to interpret and simplify data. This model is created by averaging repeated attribute partitioning.

At each level of the tree or the root node, the algorithm calculates the Information Gain Ratio (IG) of each attribute or feature, comparing them with the answer classes to find the highest IG as the root of the decision tree. To be exact, those attributes can classify the data sample for the modeling to derive similar answer classes as much as possible. This process continues to the last node, or leaf node, and classifies all data samples into different sub-sets with similar answers before completion. Once finished, the decision tree can finally be built.

Alternatively, this decision tree is like an order of the data classification to create a suitable “If-Then” rule from the root node to the leaf node. This rule will describe all data samples. To avoid an overfitting problem, the tree will be typically pruned to prevent leaf growth, where more leaves mean more possible errors in the prediction procedure. This pruning method makes the tree more useful in predicting a decision-making structure.

2.1.2 Neural Networks

Neural networks (NN) were simulated from the human brain and function as complex, nonlinear and parallel computers. This neural network is a processing element that transfers data between the input and output through either linear or nonlinear mapping. This neural network is being broadly used to solve transportation problems, such as traffic forecasting, traffic control, evaluation of traffic parameters, maintenance of transport infrastructure, transport policy and economics, driver behavior and autonomous vehicles.

When compared with discrete choice analysis (DCA), a neural network uses the relationship and error correction as a critical mechanism to identify a problem or relationship. Alternatively, the multinomial logit (MNL) model with a discrete choice analysis is based on the random utility maximization principle. In practical terms, the neural network works with simple steps in which a node or neuron receives an input signal from other factors or nodes and then assembles those signals. Those input signals entering this cell will be inhibited or activated with the weight of each connecting line, with either a positive or negative charge, respectively. After the input signals are assembled, the result is an input for the activation function and is later sent out as an output.

A neural network contains several hidden layers; each layer is hidden with a node or neuron with different activation functions. The neural network structure for the data classification problem of transportation is in the topology or architecture form, and the most popular style is the multilayer feedforward network. The advantage of this neural network is that it can solve data classification problems and non-linear multi-dimensional pattern recognition.

In this research, the researchers decided to use the multi-layer feed-forward neural network with one hidden layer for the household car ownership prediction; the node number on the hidden layer is: (number of attributes + number of classes) / 2 + 1).

3. DATA

The Faculty of Engineering at Khon Kaen University launched a survey on the traveling behavior of households living in the Khon Kaen Municipality Zone, under the study on the suitability of the engineering, economic, financial, and environmental impacts of the 2015 Khon Kaen expressway master plan.

The population was randomized as the questionnaire was concerned with household travel information. The data collection was conducted with face-to-face interviews with family members from 2,015 households of 73 traffic analysis zones (2% of the total homes in the target area). These traffic analysis zones were classified based on the GIS database; the area types were based on residential density [22] and could be classified into four types including 1) Central Business District (CBD); 2) urban areas; 3) suburban areas; and 4) rural areas, as illustrated in Fig.1 These area types represented the attributes of each traffic zone and were used to define the kinds of tour origin and the primary destinations of each tour.

3.1 Pre-Processing

The survey data consisted of both the household and the individual data while their socioeconomic data was used to define the engineering attributes concerned with the socio-demographic attributes and zone attributes. Personal travel information from each household types around the target area was used to create a tour [23] and to define the variables of the tour and accessibility attributes, as presented in Table 1.

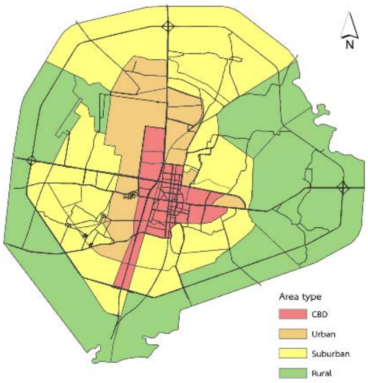


Fig.1 Area types in the Khon Kaen municipality zone (Thailand).

Table 1 Explanatory variables.

Socio-demographic attributes		
No.	Variable[1]	Values
1.	Gender	Male; female
2.	Type of dwelling	Detached house; commercial buildings; townhouse; condominium/flat, etc.
3.	HHT#	Two categories of worker vs. non- worker variable and two groups of dwelling type
4.	Income (Baht*1000)	No income; 0.001-2.5; 2.501-5.0; 5.001-7.5; 7.501-10; 10.001-15; 15.001-20; 20.001-25; 25.001-30; 30.001-40.; 40.001-50; 50.001-75; 75.001-100; 100.001-150; >150
5.	Employment status	A number of full-time workers in the HH; Non-workers; Self-employed; Students; etc.
6.	Apptype	Number of CEOs; white collar; blue collar; red collar; pink collar; student, etc.
7.	Kids	Number of children in HH
8.	HHS	Household size
9.	Age	<6; 7-19; 20-39; 40-59; 60-79; >80
Zone attributes		
10.	Areatype	CBD (central business district); urban; suburban; rural
11.	PoDwell	Percentage of detached houses
12	PHHSlow[2]	PHHS1to2; PHHS3to4; PHHS5to6
Tour attributes		
13.	Tour type	1: 1 (more) stop; **0:** no stop
14.	Numb trip	Number of trip segments within each tour
15.	Main Mode	Car; motorcycle; motor tricycle; minibus; train; bicycle; walking; other
Accessibility attributes		
16.	Accessibility[3]	Acci; aij time

Note: [1]The variables comprising many groups, such as socio-demographic attributes, area type are a measure of the intensity of development in an area and reflection of the density of employment and residents in each zone, tour attributes, and accessibility attributes. These variables were used to create and test the performance of the models. [2]Percentage of the household size (HHS) with 2, 4, 6 family members and an average income was lower than 15,000, 30,000, and 50,000, respectively. [3]Accessibility measure indicates the effort of a commuter to overcome the spatial and timing separation between zones. Acci represents the average travel time between location *i* and a random location within the area; aij is the free flow travel time between traffic zones *i* and *j* [24].

Based on the data analysis, 33.6% of the households started their tour from a suburban area, 28.0% started from CBD, 21.7% started from an urban area, and 16.7% started from a rural area. From the target area, it was found that 18.0% of the households did not own a car, 47.8% owned one car, and 34.2% owned two or more cars.

When considered separately by household types (HHT#) based on two variable groups, a number of employed compared to unemployed household members, and another two variable groups from the kinds of dwelling - commercial buildings, townhouses, condominiums, or flats and detached houses - they could be classified into four types. The household type with the maximum number was families with the number of employed members equal to or higher than the number of unemployed members. In terms of the dwelling type, these households lived in detached houses (HHT3). Nearly 50% of the members in this household type owned one car, and the remainder owned two or more cars or zero cars, respectively. The next household type had the number of employed members equal to or higher than the number of unemployed members, but these households lived in commercial buildings, townhouses, condominiums, or flats (HHT1). More than 50% of the members in this household type owned one car, and the rest owned two or more cars or zero cars, respectively. When considering the data based on tour origin, suburban areas were the tour origin of most families with the number of employed members equal to or higher than the number of unemployed members, and all members lived in detached housing with one or more cars. Rural areas were where family members owned zero vehicles. In terms of suburban areas, most households had two family members with an average yearly income of 25,000-75000 baht, which could increase to 100,000-150,000 baht if the households had additional employed family members. Still, most of the households with three family members in the tour origin earned, on average, 15,000-100,000 baht.

Meanwhile, the central business district (CBD) was the starting area for most of the households with the number of employed members equal to or higher than the number of unemployed members but lived in commercial buildings, townhouses, condominiums, or flats and owned zero, one, and two or more cars. In the case of the CBD, most households had four family members with an average income of 15,000-150,000 baht that could increase to 200,000 if they had an additional member. However, most of the households with five family members in this household type earned an average of 20,000-200,000 baht per year.

In general, the data analysis indicated that the households with the number of employed members higher than unemployed members were significant for the traveling demand analysis and the prediction on the household car ownership demand. To be exact, when the family increases in size, the household income increases according to the number of employed members in that family.

3.2 Experiment

The experiment included the use of the surveyed data to test whether the machine learning models, including the decision tree (DT) and neural network (NN), could provide precise and accurate results for the household car ownership prediction when Zone attribute variables, Tour attribute variables, and Accessibility attribute variables correctly added for each area types, in addition to socio-demographic attributes alone. To avoid, an overfitting problem, k-folds cross-validation method was applied to create and test the performance of those models.

The data set was collected from an individual's information and later converted into household level data. Specifically, the machine learning automatically selects the variables and assembles them as much as possible. However, in this research, only the first ten variables with optimized weights were considered.

Accordingly, this research explored not only the demographic parameters, which mostly reflected the household's car ownership of both decision trees and neural network algorithms, it also compared the data to indicate the contrast between those two models after the addition of more attributes of critical parameters. In this research, two datasets were built, including ML1, or the household car ownership data based on the household's attributes and socio-demographic attributes; and ML2 dataset, which included Zone attribute variables, Tour attribute variables, and Accessibility attribute variables. These two datasets have shown in Table 1, where the unavailable value (NA) was already validated and deleted.

During the processing step of ML1, the data normalized, and their optimized weights selected the first ten variables. This data was used to create and test the model by a cross-validation method where the accuracy was the key indicator of the model's performance. For the second phase of the experiment, ML1 has used again, but only the first ten variables with the optimized weights combined into ML2 including Zone attribute variables, Tour attribute variables, and Accessibility attribute variables. Later, these data were normalized, and the first ten variables with optimized weights were selected again for ML2 to create and test the model's performance, still using cross-validation. Finally, the accuracy of each household's car ownership model derived from the decision tree, and neural network algorithms from both sets (ML1 and ML2) was compared using the t-statistic with a significance level set at $\alpha = 0.05$.

At every stage, RapidMiner Studio Educational 8.2.000 used as a software tool for all machine learning models with default parameters from both the decision tree and neural network algorithms.

4. RESULTS ANALYSIS

The two datasets were used to create and test the performance of the household car ownership model using machine learning models. The accuracy of those models was compared using a t-statistic value with a significance level set at $\alpha = 0.05$.

4.1 Prediction from ML1

The results from ML1 were based on the first ten variables of optimized weights, selected from household's socio-demographic attributes. These variables similarly corresponded to previous studies on household car ownership demand models [7] [12], such as dwelling type, household income, number of employed and unemployed members in a household, occupation, number of children and students in a household, household sizes, number of people within an age category, and type of household. These were used to analyze the relationship between the variables and the results of the modeled prediction.

There was still an imbalanced answer class amongst those household data within the target area. To be precise, this included 18.0% with zero car ownership, 47.8% with ownership of one car, and 34.2% with ownership of two or more cars. In detail, the CBD households contained 16.4% of zero car ownership, 49.1%, of ownership of one car, and 34.5% of ownership of two or more cars; the urban area household contained 16.8% of zero car ownership, 44.8%, of ownership of one car, and 38.4% of ownership of two or more cars; suburban area households contained 16.1% of zero car ownership, 49.6%, of ownership of one car, and 34.3% of ownership of two or more cars; rural area households contained 26.1% of zero car ownership, 46.1%, of ownership of one car, and 27.8% of ownership of two or more cars.

The results presented in Fig.2 and Tables 2 and 3 indicate the prediction accuracy from each model and the individual and aggregate match rates of the prediction in each answer class, derived from the confusion matrix table.

The individual and aggregate match rates are used to evaluate and compare the car ownership modeling performance of the machine learning models, on individual and aggregate levels. The individual match rate is the ratio of the number of correctly predicted individual observations for one answer class to the total number of the actual observations chosen in this class. The aggregate match rate is defined as the ratio of the number of correctly and incorrectly predicted observations for one answer class to the total number of actual observations chosen in the answer class.

In the case of ML1, when the accuracy of the household car ownership prediction used as a key indicator, the neural network algorithm gave more accurate results than did the decision tree algorithm in all datasets, as presented in Fig.2 and Table 2, with an accuracy of the neural network and decision trees algorithms of 0.553 to 0.660 and 0.482 to 0.622, respectively.

This result was an exception for the households with the tour origin from rural areas, where the accuracy was not significantly different (NN with accuracy 0.623; DT with accuracy 0.602). Nevertheless, according to the confusion matrix with the answer class of zero cars and two or more cars in Table 3, the neural network algorithm provided more accurate (37.0 % and 65.3%) results than did the decision tree algorithm (24.2 % and 49.8%) based on the individual match rate. Even though the answer of one car for the neural network algorithm seemed to show a lower individual match rate than the decision tree algorithm, this answer class of the decision tree algorithm still contained a higher aggregate match rate. Indicated that, at this answer class, the decision tree algorithm could not separate the 1 car answer from the other answers class (0 Car and 2+ Cars).

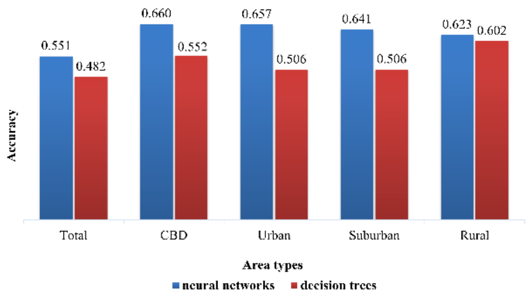


Fig.2 Performance measures for ML1 in each area types by NN and DT algorithms.

4.2 Prediction from ML2

ML2 was a dataset which added the main attributes of variables used in both tour-based and activity-based models from Table 1 in order to derive more accurate prediction result, such as origin or destination area types, percentage of detached houses in each zone, percentage of the household with low average income within each zone; these were the variables involved with zone attributes. Moreover, the tour attribute variables included the number of trip segments within each tour, the primary mode of the tour, and tour type, as well as the accessibility variables including travel time between zones within the tour. These variables added to ML1 with only the variables used for the modeling. When these data were combined, the first ten variables with the optimized weights were selected to create and test the performance of those models again.

Table 2 Comparison of performance vectors from NN and DT algorithms for ML1 using significance test (T-test).

Total dataset		NNML1 0.551	DTML1 0.482
NNML1	0.551		🟧
DTML1	0.482		
CBD is tour origin		NNML1 0.660	DTML1 0.552
NNML1	0.660		🟧
DTML1	0.552		
Urban is tour origin		NNML1 0.657	DTML1 0.506
NNML1	0.657		🟧
DTML1	0.506		
Suburban is tour origin		NNML1 0.641	DTML1 0.506
NNML1	0.641		🟧
DTML1	0.506		
Rural is tour origin		NNML1 0.623	DTML1 0.602
NNML1	0.623		0.323
DTML1	0.602		

Note: Values with colored background are less than α=0.05, which indicates a likely significant difference between the actual mean values. NNML1 = Neural networks algorithm with ML1, DTML1 = Decision trees algorithm with ML1

The results from ML2 are illustrated in Fig. 3 and Table 4, with an accuracy of the neural network and decision trees algorithms of 0.548 to 0.654 and 0.478 to 0.504, respectively. Table 4 with the default parameter, the neural network algorithm demonstrated more accurate results than the decision tree algorithm in all datasets. When comparing the accuracy of ML1 and ML2 from the neural network algorithm, the result found as presented in Table 5. In terms of the total dataset, the prediction performance on the household car ownership demand from those two datasets was similar. When individually considering the tour origin area types, urban area types similarly demonstrated accurate results. Namely, the model could identify the answer class in a similar amount from any dataset used for the modeling. Still, the confusion matrix (Tables 6 and 7) of ML2 still shows some classes where the individual match rate was higher than that of ML1. Therefore, it can be concluded that some variables of zone attributes, tour attributes, and accessibility attributes could identify the answer classes better.

Table 3 Confusion matrices generated by the NN and DT algorithms for ML1 (a rural area is tour origin).

NN Accuracy (62.3 %)		Actual class			I (%)	A (%)
		0 Car	1 Car	2+ Cars		
		211	373	225		
Predicted class	0 Car 135	78	37	20	**37.0**	64.0
	1 Car 416	79	279	58	74.8	**111.5**
	2+ Cars 258	54	57	147	**65.3**	114.7
DT Accuracy (60.2 %)		Actual class			**I (%)**	**A (%)**
		0 Car	1 Car	2+ Cars		
		211	373	225		
Predicted class	0 Car 65	51	12	2	**24.2**	30.8
	1 Car 591	156	324	111	86.9	**158.4**
	2+ Cars 153	4	37	112	**49.8**	68.0

Note: I = Individual match rate, A = Aggregate match rate, NN = Neural networks algorithm, DT = Decision trees algorithm

In terms of the prediction accuracy of CBD, suburban, and rural area types (Table 5) showed significant differences in their prediction performance, indicating that adding more attributes of the variables used for both tour-based and activity-based models into ML1 did not improve the predicted result. Despite the unimproved result, some variables of zone attribute, tour attribute, and accessibility attribute surprisingly demonstrated a significant weight and became the key variables that perfectly reflect the household car ownership demand, rather than the variables selected from ML1 for the modeling.

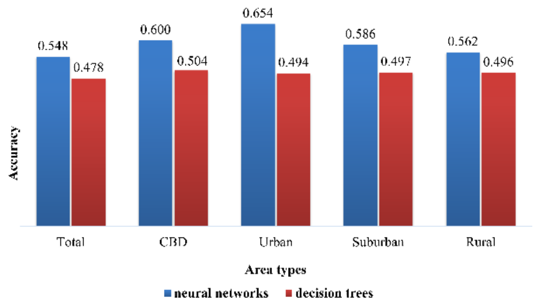


Fig.3 Performance measures for ML2 in each area type by NN and DT algorithms.

Table 4 Comparison of performance vectors from NN and DT algorithms for ML2 using significance test (T-test).

Total dataset		NNML2 0.548	DTML2 0.478
NNML2	0.548		▨
DTML2	0.478		
CBD is tour origin		NNML2 0.600	DTML2 0.504
NNML2	0.600		▨
DTML2	0.504		
Urban is tour origin		NNML2 0.654	DTML2 0.494
NNML2	0.654		▨
DTML2	0.494		
Suburban is tour origin		NNML2 0.586	DTML2 0.497
NNML2	0.586		▨
DTML2	0.497		
Rural is tour origin		NNML2 0.562	DTML2 0.496
NNML2	0.562		▨
DTML2	0.496		

Note: Values with colored backgrounds are less than α=0.05, which indicates a likely significant difference between the actual mean values.
NNML2 = Neural networks algorithm with ML2, DTML2 = Decision trees algorithm with ML2

Table 5 Comparison of performance vectors from NN algorithms for ML1 and ML2 using significance test (T-test).

Total dataset		NNML1 0.551	NNML2 0.548
NNML1	0.551		**0.782**
NNML2	0.548		
CBD is tour origin		NNML1 0.660	NNML2 0.600
NNML1	0.660		▨
NNML2	0.600		
Urban is tour origin		NNML1 0.657	NNML2 0.654
NNML1	0.657		**0.873**
NNML2	0.654		
Suburban is tour origin		NNML1 0.641	NNML2 0.586
NNML1	0.641		▨
NNML2	0.586		
Rural is tour origin		NNML1 0.623	NNML2 0.562
NNML1	0.623		▨
NNML2	0.562		

Note: Values with colored background are less than α=0.05, which indicates a likely significant difference between the actual mean values.
NNML1 = Neural networks algorithm with ML1, NNML2 = Neural networks algorithm with ML2

Table 6 Confusion matrices generated by the NN algorithms for ML1 and ML2 (Total dataset).

NN Accuracy ; ML1 (55.11 %)		Actual class			I (%)	A (%)
		0 Car	1 Car	2+ Cars		
		874	2321	1657		
Predicted class	0 Car 213	125	75	13	14.3	24.4
	1 Car 2595	548	1476	571	63.6	111.8
	2+ Cars 2044	201	770	1073	64.8	123.4
NN Accuracy ; ML2 (54.78 %)		Actual class			**I (%)**	**A (%)**
		0 Car	1 Car	2+ Cars		
		874	2321	1657		
Predicted class	0 Car 373	208	140	25	**23.8**	42.7
	1 Car 2880	528	1585	767	**68.3**	124.1
	2+ Cars 1599	138	596	865	52.2	96.5

Note: I = Individual match rate, A = Aggregate match rate, NN = Neural networks algorithm

Considering the household level data of each target area, the household type (HHT#), a variable involved with the socio-demographic attributes of the households living in urban areas and rural areas, had some impact on the model's performance for the household car ownership prediction on both ML1 and ML2. Hence, this research purposively discussed only the household level data (HHT1 and HHT3) from urban and rural areas. These household data were used to create the model, and the results presented in Fig. 4.

Table 7 Confusion matrices generated by the NN algorithms for ML1 and ML2 (urban is tour origin).

NN Accuracy ; ML1 (65.75 %)		Actual class			I (%)	A (%)
		0 Car	1 Car	2+ Cars		
		177	472	405		
Predicted class	0 Car 62	19	29	14	10.7	35.0
	1 Car 627	132	389	106	82.4	132.8
	2+ Cars 365	26	54	285	70.4	90.1
NN Accuracy ; ML2 (65.37 %)		Actual class			I (%)	A (%)
		0 Car	1 Car	2+ Cars		
		177	472	405		
Predicted class	0 Car 84	39	29	16	**22.0**	47.5
	1 Car 593	104	375	114	79.4	125.6
	2+ Cars 377	34	68	275	67.9	93.1

Note: I = Individual match rate, A = Aggregate match rate, NN = Neural networks algorithm

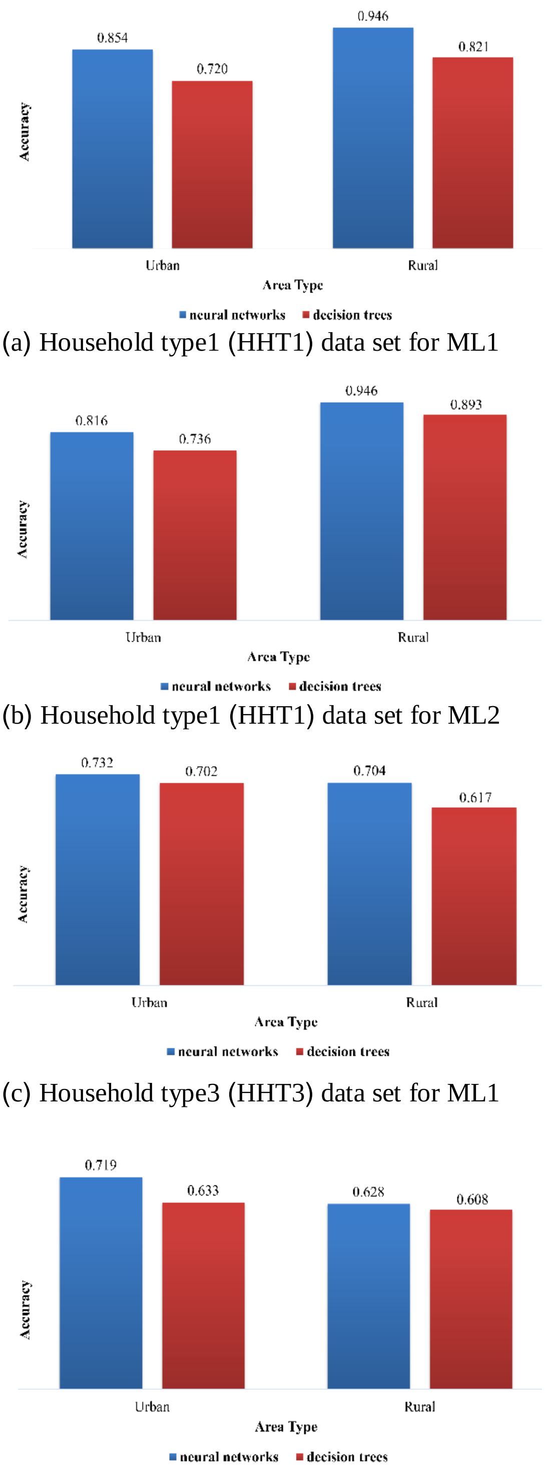
Obviously, in the case of HHT1 in urban areas, the neural networks algorithm demonstrated more accurate results than did the decision tree algorithm for both datasets (ML1 and ML2). However, both algorithms showed a similar effect for the same data from rural areas. In the case of HHT3 in urban areas, the neural network algorithm showed more accurate results than did the decision tree algorithm for ML2. In contrast, for ML1, both algorithms showed a similar result. Notably, in the case of HHT3 in rural areas, both the neural network and decision tree algorithms showed the same effect for ML2, but the neural network algorithm significantly demonstrated more accurate prediction than did the decision tree algorithm for ML1. By comparing the accuracy results of both datasets by the neural network algorithm, HHT1 and HHT3 from urban areas showed no difference in the data accuracy when ML1 and ML2 were used to create a model.

The prediction results based on two datasets (ML1, ML2) show, on both accuracy and the individual match rate, that the NN model outperforms the DT model. This is because the DT model has problems with estimation processing in continuous data; data have to first be grouped into ranges manually or automatically by a software tool. The selection of the fields may unwittingly hide useful patterns. Meanwhile, the NN model does not suffer from the estimation algorithm, due to the existence of the backpropagation technique, though the NN model has an interpretation problem. However, considering its higher prediction performance, we believe that it has executed value more so than the decision trees model.

Comparing the DT with the NN model, Xie, Lu, and Parkany (2003) found that the NN provided a better representation of the travel mode decision-making of households and suggested it as more appropriate. The comparative evaluation by the NN model shows that the two datasets, ML1 and ML2, are comparable but the ML1 dataset has slightly better prediction capability on the household car ownership modeling. These results indicate that machine learning has differential explanatory variables with weights optimized per attribute, which causes slightly different prediction capability.

5. CONCLUSION & FUTURE WORK

In conclusion, the machine learning models, the decision tree algorithm (DT) and neural network algorithm (NN), were applicable for household ownership demand prediction, whereas the neural network algorithm performs a more accurate prediction, based on accuracy, than did the decision tree algorithm for all datasets. Nonetheless, when comparing both ML1 and ML2 and when the neural network algorithm was used for the household car ownership modeling, the predicted results based on accuracy did not show any significant differences.



(a) Household type1 (HHT1) data set for ML1

(b) Household type1 (HHT1) data set for ML2

(c) Household type3 (HHT3) data set for ML1

(d) Household type3 (HHT3) data set for ML2

Fig.4 Performance measures for ML1 and ML2 in two area types by DT and NN algorithms

This research discusses a 2015 data from the study on the suitability of the engineering, economic, financial, and environmental impacts of the 2015 Khon Kaen expressway master plan and uses it for creating and testing the performance of the machine learning models, using a k-folds cross-validation method. The data comparison on the household car ownership using the decision tree and neural network algorithms on these two datasets, ML1 and ML2, where ML1 is based on the household's attributes and socio-demographic attributes, and ML2 includes the data of ML1 with the addition of zone attribute variables, tour attribute variables, and accessibility attribute variables. As a result, the neural network algorithm gives more accurate prediction than the decision tree algorithm for both datasets and for all types of test areas. When comparing the accuracy of ML1 and ML2 from the neural network algorithm, it was found that after adding more attributes to the key variables used for both tour-based and activity-based models, the overall accuracy was similar. The results from the target areas indicate that it is unnecessary to use the variables from the tour-based or activity-based models preceding the processing step of the household car ownership prediction.

However, the individual match rate from the confusion matrix table suggests that some answer classes contain a higher individual match rate after adding more attributes to the variables used for the tour-based and activity-based models. This means that despite the similar accuracy, recall or individual match rate becomes more accurate. Specifically, when the number of data in the class of interest is much smaller than the other class. (i.e. zero cars) The individual match rate of the total dataset has a higher percentage than the original 66.4%. This finding affirms that the household ownership demand prediction is necessary for transportation; the machine learning models can provide useful information for urban design and transportation planning. In this regard, a future study may aim to further develop the model with higher performance. Indeed, after adding the variables with some impacts on the answer class, it is necessary to solve the answer class imbalance by assigning different weights to each class, or Cost-Sensitive Learning (CSL). This will be tested to make it more possible to suitably improve the data prediction method and policy issuance for urban transportation planning via the use of machine learning models.

6. ACKNOWLEDGMENTS

The researcher would like to thank Mr. Sorasak Siewsirikul, a full-time lecturer in the Civil

Engineering Department, Rajamangala University of Technology (Khon Kaen Campus) for his great support in decoding the data for this research. I would also like to express gratitude to the Faculty of Engineering, Khon Kaen University for the household's travel data acquired from the study on the suitability of the engineering, economic, financial, and environmental impacts of the 2015 Khon Kaen expressway master plan. Finally, I would like to thank RapidMiner Studio Educational 8.2.000 for their useful software, which was an essential resource for the success of this research.

7. REFERENCES

[1] Paredes M, Hemberg E, O'Reilly U-M, Zegras C., Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?, Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017, pp.780-785.

[2] Bhat CR., Work Travel Mode Choice and Number of Non-Work Commute Stops, Transportation Research Part B: Methodological, Vol. 31, Issue 1, 1997, pp.41-54.

[3] Putu Hermawati SAA, Muhammad Isran Ramli, Sumarni Hamid., Choices Models of Trip Chain and Transportation Mode for International Tourists in Tourism Destination Island, International Journal of GEOMATE, Vol. 16, Issue 55, 2019, pp.195-203.

[4] Sener IN, Ferdous N, Bhat CR, Reeder P., Tour-Based Model Development for TxDOT : Evaluation and Transition Steps, Tech Report, 2009, pp.1-227.

[5] Fischer MM, Nijkamp P, Papageorgiou YY., Spatial Choices and Processes: Elsevier, 2013, pp.1-370.

[6] Horowitz AJ., Statewide Travel Forecasting Models, Transportation Research Board, Vol. 358, 2006, pp.1-116.

[7] Handy S, Cao X, Mokhtarian P., Correlation or Causality Between The Built Environment and Travel Behavior? Evidence from Northern California., Transportation Research Part D: Transport and Environment, Vol. 10, Issue 6, 2005, pp.427-444.

[8] Van Acker V, Witlox F., Car Ownership as a Mediating Variable in Car Travel Behaviour Research Using a Structural Equation Modelling Approach to Identify its Dual Relationship, Journal of Transport Geography, Vol. 18, Issue 1, 2010, pp.65-74.

[9] Roxas Jr NR, Fillone AM, Roquel KID., Estimating the Environmental Effects of the Car Shifting Behavior Along EDSA, International Journal of GEOMATE, Vol. 14, Issue 44, 2018, pp.8-14.

[10] Potoglou D, Susilo Y., Comparison of Vehicle-Ownership Models, Transportation Research Record: Journal of the Transportation Research Board, Vol. 2076, 2008, pp.97-105.

[11] Bunch DS, Chen B., Automobile Demand and Type Choice, Handbook of Transport Modelling, 2nd Edition, Emerald Group Publishing Limited, 2007, pp.541-557.

[12] Matas A, Raymond J-L., Changes in The Structure of Car Ownership in Spain, Transportation Research Part A: Policy and Practice, Vol. 42, Issue 1, 2008, pp.187-202.

[13] Ben-Akiva ME, Lerman SR., Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Vol. 9, 1985, pp.1-397.

[14] Potoglou D, Kanaroglou PS., Modelling Car Ownership in Urban Areas: A Case Study of Hamilton, Canada, Journal of Transport Geography, Vol. 16, Issue 1, 2008, pp.42-54.

[15] Bhat CR, Pulugurta V., A Comparison of Two Alternative Behavioral Choice Mechanisms for Household Auto Ownership Decisions, Transportation Research Part B: Methodological, Vol. 32, Issue 1, 1998, pp.61-75.

[16] Zhang Y, Xie Y., Travel Mode Choice Modeling with Support Vector Machines, Transportation Research Record: Journal of the Transportation Research Board, Vol. 2076, 2008, pp.141-150.

[17] Xie C, Lu J, Parkany E., Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks, Transportation Research Record: Journal of the Transportation Research Board, Vol. 1854, 2003, pp.50-61.

[18] Karlaftis MG, Vlahogianni EI., Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights, Transportation Research Part C: Emerging Technologies, Vol. 19, Issue 3, 2011, pp.387-399.

[19] Pamuła T., Neural Networks in Transportation Research–Recent Applications, Transport Problems, Vol. 11, 2016, pp.27-36.

[20] Witten IH, Frank E, Hall MA, Pal CJ., Data Mining: Practical Machine Learning Tools and

Techniques, Morgan Kaufmann, 2016, pp.1-621.

[21] Pitombo CS, de Souza AD, Lindner A., Comparing Decision Tree Algorithms to Estimate Intercity Trip Distribution, Transportation Research Part C: Emerging Technologies, Vol. 77, 2017, pp.16-32.

[22] kkn-DPT. City Development Plan, [Online]. Available:http://pllu.dpt.go.th/pllu/Default.aspx [Accessed: 1 June 2018], 2018.

[23] Kawwichian P, Tanwanichkul L., Method of Trip-Chaining and Tour Formation for Travel Demand Model Development, RMUTI Journal, Vol. 11, Issue 1, 2018, pp.57-68.

[24] Allen WB, Liu D, Singer S., Accessibility Measures of U.S. Metropolitan Areas, Transportation Research Part B: Methodological, Vol. 27, Issue 6, 1993, pp.439-449.