MEKATRONIKA
e-ISSN: 2637-0883
VOL. 2, ISSUE 1, 73 – 78
DOI: https://doi.org/10.15282/mekatronika.v2i1.6745

ORIGINAL ARTICLE

Mode Choice Prediction using Machine Learning Technique for A Door-to-Door Journey in Kuantan City

Nur Fahriza Mohd Ali[1*], Ahmad Farhan Mohd Sadullah[1], Anwar P. P. Abdul Majeed[2], Mohd Azraai Mohd Razman[2] and Rabiu Muazu Musa[3]

[1]School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Pulau Pinang, Malaysia.
[2]Innovative Manufacturing, Mechatronics and Sports Laboratory, Faculty of Manufacturing and Mechatronics Engineering Technology, Universiti Malaysia Pahang, 26600, Pekan, Pahang, Malaysia.
[3]Center for Fundamental and Continuing Education, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia.

*Email corresponding author:

ABSTRACT – A door-to-door journey in a public transportation system is a notable concept that is practically being promoted among users to consider public transport as an important alternative. The door-to-door journey will integrate the travel segments starting from home to destination, including all visible amenities. Users' preferences on the time travel of these key segments are necessary to be understood. In this case, Machine Learning technique has been seen as a robust computational advancement to forecast their travel mode choice. However, the most convenient model as the best predictor is still questionable. To address this issue, we employed some pre-eminent machine learning models, specifically Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), k-Nearest Neighbor (kNN) as well as Support Vector Machine (SVM), to compare their travel mode choice prediction performance of users in the city of Kuantan. The data collection was conducted in Kuantan City via Revealed/Stated Preferences (RPSP) Survey between 8:00 AM to 5:00 PM on weekdays. The data collected was split into a ratio of 80:20 for training and testing before evaluating them between the aforesaid models. The results depicted that the Random Forest could provide satisfactory classification accuracies for both training and testing data up to 68.3% and 61.3%, respectively, compared to the other evaluated machine learning models. In summary, Random Forest provides a good result in the training and testing data and is considered as the best predictor in this research to forecast users' mode choice in the city of Kuantan.

ARTICLE HISTORY
Received: 3rd May 2020
Revised: 25th May 2020
Accepted: 14th June 2020

KEYWORDS
Mode Choice
Door-to-Door Journey
Machine Learning Models
Classification Accuracy

INTRODUCTION

The door-to-door approach can be defined as combining all segments of the travel chain for both public transport and private vehicles modes that are counted in the total travel time. The door-to-door approach ensures that both travel modes are comparable in the real-time required to travel with these modes. The door-to-door journey will integrate the travel segments starting from home to destination, including all visible amenities. Under the door-to-door journey framework, more detailed geographic information including essential traveling facilities, the locations of users' trip origins, users' desired destinations, as well as time travel information of daily trips can be integrated for mode choice analysis, which is necessary for policymakers to improve the provision of public transportation services. The total travel time of a journey is crucial. The door-to-door journey framework is also adequate to provide comparisons on the performance between modes: public transport and its closest competitors, which are private vehicles, especially cars. A door-to-door journey in a public transportation system is a notable concept that is practically being promoted among users to consider public transport as an important mode alternative. Prediction on users' travel mode choice can be assisted via the door-to-door journey concept to manage high daily traffic volume on the roads that occurred especially during peak hours. Forecasting travel demand is necessary to be understood to allow public transportation planners to provide adequate services that are sufficient for users and improve the performance of the services so that users will be attracted to keep travelling with public transport in their daily routine.

Recently, the travel time concept of door-to-door journey is gaining more attention in many scopes of studies, such as accessibility analysis conducted by Zhao and Yu in 2018 [1]. Some other research conducted by Benenson and colleagues in 2011 and Salonen and Toivonen in 2013 applied the travel time measures for a door-to-door journey to study users' travelling within intracity [2], [3]. The capability of the door-to-door concept to measure and distinguish each component of a journey starting from its origin until destination allows a more prudent evaluation and comparison on travelling time using different travel modes, either public transport or private vehicles. Other researchers revealed that door-to-door study had been implemented in High-Speed Rail (HSR), yet it is an under-represented area [4]. Current research on a door-to-

*CORRESPONDING AUTHOR |Nur Fahriza Mohd Ali | ✉ fahriza90@gmail.com

door journey based on travel time approach is employed in HSR accessibility studies to explore its effectiveness to join the transport intersection of intracity to intercity access between cities [5]. Their research depicted that the total travel time via a door-to-door journey for an intercity trip takes a significant portion of intracity travel and affects a lot of benefits for serving cities. These studies, as mentioned earlier, approved the significance of the employment travel time approach via a door-to-door journey to enhance the interpretation of the complete cycle of a daily trip and urge further studies to explore more detailed of its impacts on public transportation services[4], [6]. The largest benefits of exploration on a travel time approach via a door-to-door journey is it will allow new infrastructures to be built as a bridging system at the inter-and intra-city trips, and most likely to assist further studies of their extensive effects on developing a connection between cities and provide some visions on transport planning at many levels [7], [8]. The introduction of a door-to-door framework is somehow important, and the machine learning models will emphasize the rationale of which models will accurately predict users' daily mode in Kuantan City.

Machine learning technique has been seen as a robust computational advancement to forecast users' travel mode choice. Machine learning development has become a revolution, and the application has widened in many fields, including transportation. In this paper, some machine learning models are applied to estimate their performance in modelling by observing their prediction quality. The introduction of these models is an alternative to display complex behaviour modelling and pattern recognition. Recent studies have indicated the power of the machine learning technique in transportation research studies. Zhang and Xie, in 2008, explained that the Support Vector Machine (SVM) is an impressive classifier of its ability to predict the mode share among other machine learning models and multinomial logit models (MNL) [9]. Meanwhile, a study conducted by Hagenauer and Helbich in 2017 manages to present the superiorities of Random Forest (RF) in making predictions related to transportation studies [10]. At the end of this article, some finest machine learning models, such as Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), *k*-Nearest Neighbour (*k*-NN), and SVM are evaluated to identify users' interest on mode choice while making daily traveling to Kuantan City.

METHODOLOGY

Flow Chart

In this section, the data collection, as well as data analysis, are described. Figure 1 outlines the flows on the research framework starting from data collection at the selected areas until analysis of data using the machine learning technique.

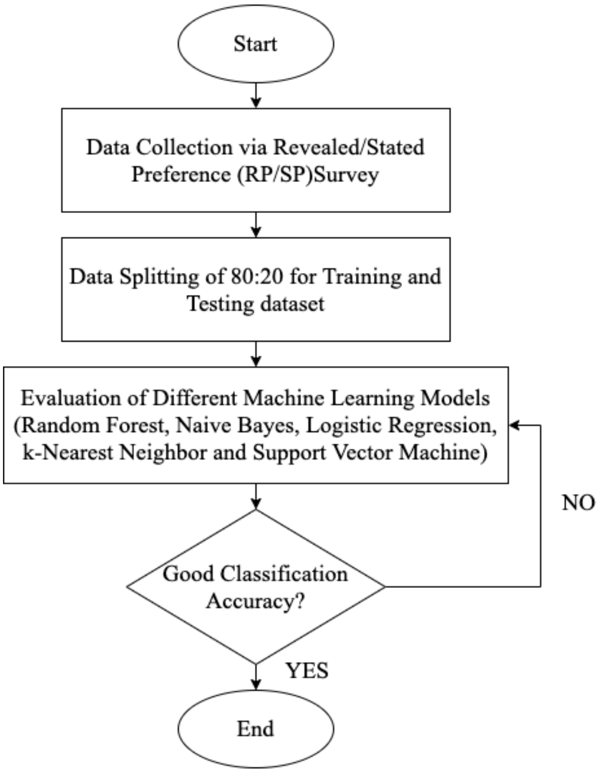

Start
Data Collection via Revealed/Stated Preference (RP/SP)Survey
Data Splitting of 80:20 for Training and Testing dataset
Evaluation of Different Machine Learning Models (Random Forest, Naive Bayes, Logistic Regression, k-Nearest Neighbor and Support Vector Machine)
Good Classification Accuracy?
NO
YES
End

Figure 1. Research methodology for mode choice forecasting.

Data collection

The most notable survey in transportation studies, known as Revealed Preference Survey and Stated Preference Survey, is being adopted for data collection conducted during weekdays starting from 8:00 AM until 5:00 PM for the proposed methodology. The data was collected in Kuantan City as in Figure 2, were from shopping malls, recreational areas, public parks, and academic institutions. The total number of respondents who took part in this survey was 386 respondents, considering those who travelled to Kuantan City as a daily routine.

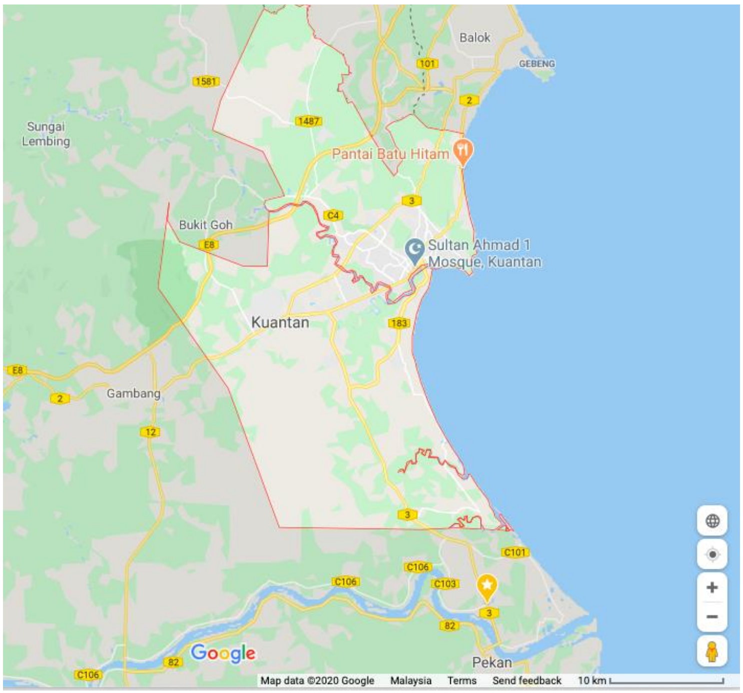


Figure 2. The selected area for data collection.

An accurate and precise methodology is necessary to perform remarkable prediction models that achieve high accuracy to predict users' daily mode choice preferences in Kuantan City. The survey form consists of three main steps as presented in Figure 1, including data collection, data pre-processing, and evaluation of machine learning models. The mode choices preferred by users were categorized mainly as P or N. P choices can be expressed as public transport as users' mode choice as depicted in Figure 3, whilst N choices are expressed as users' mode choice as private vehicles (including soft mode) illustrated in Figure 4. The questionnaire is divided into two main sections as following:

a) Personal information

- The distance from users' origin to destination (Region) including Region 1 (Users who lived ≤5km), Region 2 (Users who lived between 6 to 20km), Region 3 (Users who lived between 21 to 40 km), and Region 4 (Users who lived >40km).
- Age
- Gender
- Income
- Employment status

b) Travel information

- Walking distance from home to a nearest bus stop (WD1) in minutes,
- Waiting time at the bus stop (WT) in minutes,
- Sitting time in the public transport (IVT) in minutes,
- Walking distance from last stop to destination (WD2) in minutes,
- Total travel time of all travel components from origin to destination (TT) in minutes, and
- Users' preferences on ticket prices (SP Ticket) in Ringgit Malaysia (RM).

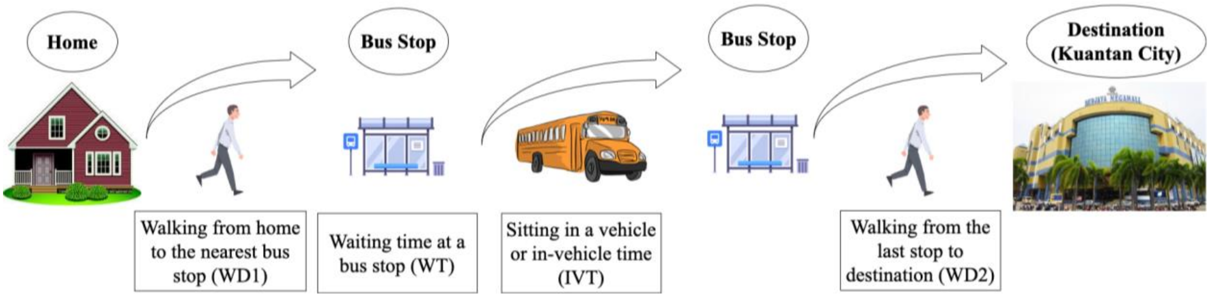

Home
Bus Stop
Bus Stop
Destination (Kuantan City)
Walking from home to the nearest bus stop (WD1)
Waiting time at a bus stop (WT)
Sitting in a vehicle or in-vehicle time (IVT)
Walking from the last stop to destination (WD2)

Figure 3. A door-to-door journey concept using public transport.

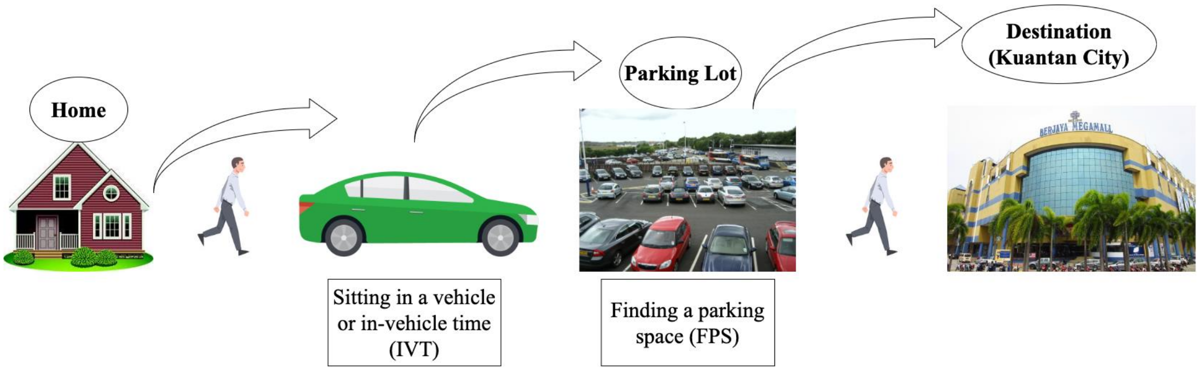

Home
Parking Lot
Destination (Kuantan City)
Sitting in a vehicle or in-vehicle time (IVT)
Finding a parking space (FPS)

Figure 4. A door-to-door journey concept using private vehicles.

Pre-processing Phase

The proportion of training and testing datasets was split into 80:20, respectively, to be employed by the selected machine learning models for training and testing purposes before evaluating their performance in mode choice prediction.

Classifier

The machine learning models that have been selected to be evaluated of their performance in classifying travel mode choice, including Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), k-Nearest Neighbour (kNN), and Support Vector Machine (SVM). The activation function of each classier utilized the default settings of the Orange platform. The evaluation measurement to compare the performance of these machine learning is Classification Accuracy (CA).

$$CA = (TP + TN) / (TP + FP + TN + FN) \qquad (1)$$

The evaluation of machine learning performance is indicated by FP and FN, which can be defined as the false positive and false negative, where TP and TN are the true positive and negative rates.

EXPERIMENTAL RESULTS

Results and Discussion

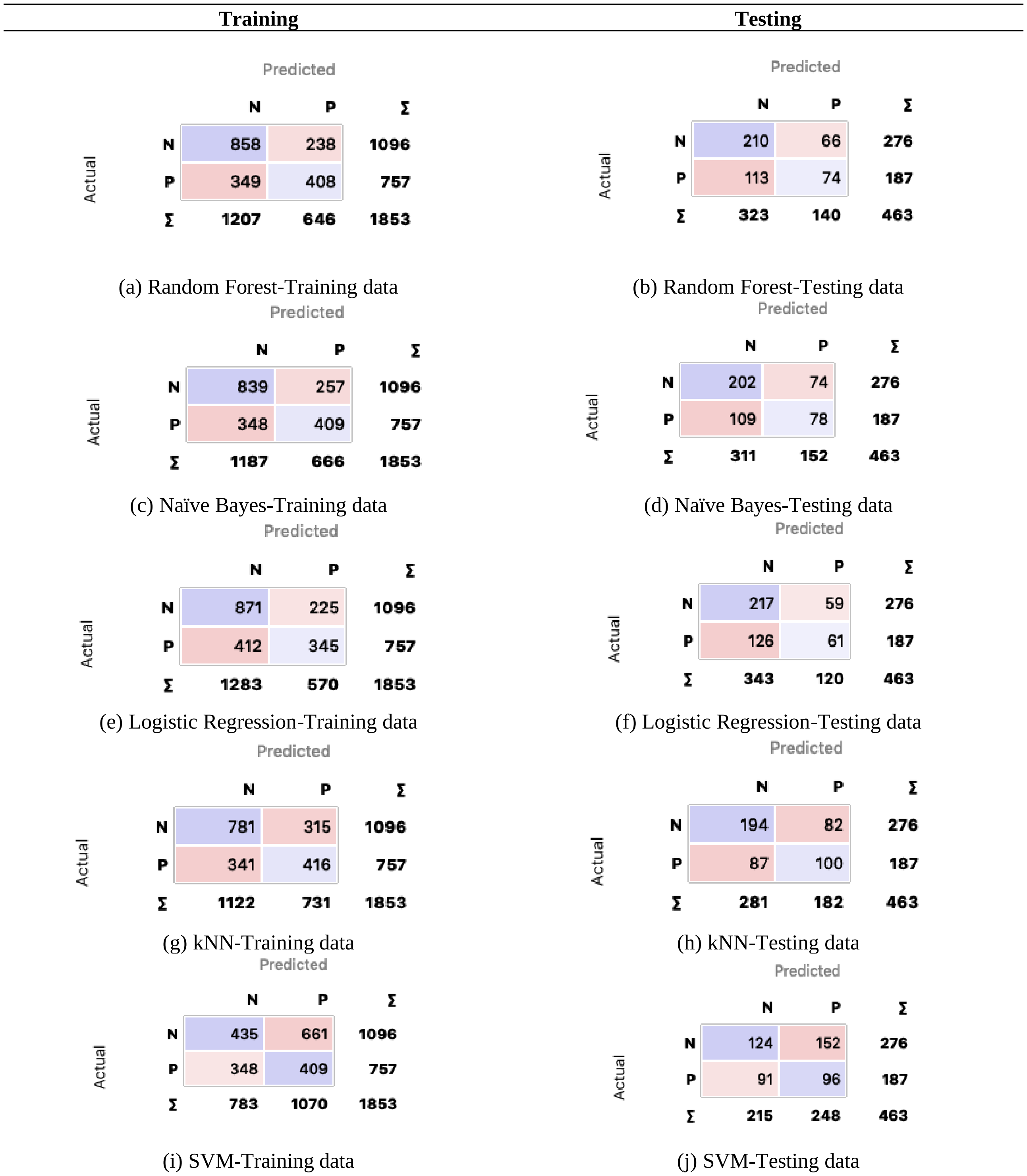
The results computed for evaluation are based on the classification accuracy (CA) of the training and testing phase. The classification results were compared to estimate the capability and the effectiveness of the proposed classifiers for modelling travel mode choice. These five classifiers of RF, NB, LR, *k*-NN, and SVM were evaluated, and the results were depicted in Table 1.

Table 1. The evaluation of the classification accuracy of each classifier.

Machine Learning Models	CA %	
	Training	Testing
RF	68.3	61.3
NB	67.4	60.5
LR	65.6	60.0
k-NN	64.6	63.5
SVM	45.5	47.5

From Table 1, shows that the RF classifier achieves the highest accuracy result of 68.3% compared to all other classifiers. The results also show that the NB classifier achieves the second-highest accuracy result with 67.4%, against the other classifiers. Meanwhile, LR, *k*-NN, and SVM is the least, attained accuracy result of 65.6%, 64.6%, and 45.5%, respectively. The testing accuracy results for RF, NB, LR, *k*-NN, and SVM are 61.3%, 60.5%, 60%, 63.5%, and 47.5%, respectively. The SVM classifier is unsuitable for predicting in this dataset, which in future, researchers are advised to perform a rigorous study to justify this impairment of results.

The efficacy of the RF classifier in predicting mode choice was in line with earlier researchers, although they were subjected to different scopes. In a study conducted by [11], the accuracy result of RF achieved 71.89% while predicting the household travel mode. In another study conducted by [12] to study transportation mode detection, including mode choices made by people with disabilities, the accuracy achieved by the RF classifier was 79% and 67% for both crutches and wheelchair users, respectively.



Training
Testing
Predicted
Actual

Actual	N	P	Σ
N	858	238	1096
P	349	408	757
Σ	1207	646	1853

(a) Random Forest-Training data

Actual	N	P	Σ
N	210	66	276
P	113	74	187
Σ	323	140	463

(b) Random Forest-Testing data

Actual	N	P	Σ
N	839	257	1096
P	348	409	757
Σ	1187	666	1853

(c) Naïve Bayes-Training data

Actual	N	P	Σ
N	202	74	276
P	109	78	187
Σ	311	152	463

(d) Naïve Bayes-Testing data

Actual	N	P	Σ
N	871	225	1096
P	412	345	757
Σ	1283	570	1853

(e) Logistic Regression-Training data

Actual	N	P	Σ
N	217	59	276
P	126	61	187
Σ	343	120	463

(f) Logistic Regression-Testing data

Actual	N	P	Σ
N	781	315	1096
P	341	416	757
Σ	1122	731	1853

(g) kNN-Training data

Actual	N	P	Σ
N	194	82	276
P	87	100	187
Σ	281	182	463

(h) kNN-Testing data

Actual	N	P	Σ
N	435	661	1096
P	348	409	757
Σ	783	1070	1853

(i) SVM-Training data

Actual	N	P	Σ
N	124	152	276
P	91	96	187
Σ	215	248	463

(j) SVM-Testing data

Figure 5. The confusion matrix of machine learning models.

The overall classification accuracy using the proposed method is illustrated in the Figure 5 confusion matrix. Further inspection of the confusion matrix diagram of RF as well as the rest of the designated classifiers is shown in Figure 6. For the RF classifier, the confusion matrix revealed that the number of misclassifications is about 238 choices; recorded by the model which is came from mode choice "N" that was misclassified as mode choice "P". Meanwhile, misclassifications occurred by the model came from mode choice "P" was greater, recorded as 349 choices that were misclassified as mode choice "N". For the other classifiers including NB, LR, *k*-NN, and SVM, the confusion matrix revealed that the number of misclassifications is about 257, 225, 315, 661 choices, respectively; recorded by the model which is came from mode choice "N" that was misclassified as mode choice "P". Misclassifications also occurred by the model came from mode choice "P" for NB, LR, *k*-NN, and SVM classifiers, recorded as 348, 412, 341, 348 choices, respectively, that was misclassified as mode choice "N".

The features employed in this study can be used to forecast users' daily travel mode choices. These features, including walking distance from home to the nearest bus stop (WD1), waiting time at the bus stop (WT), sitting time in the public transport (IVT), walking distance from last stop to destination (WD2), total travel time of all travel components from origin to destination (TT), users' preferences on ticket prices (SP Ticket), the distance from users' origin to destination (Region), Age, Gender, Income, and Employment status were effective in predicting users' daily travel mode choice.

.

CONCLUSION AND FUTURE WORK

It can be concluded that the selected classifiers including RF, NB, LR, *k*-NN, and SVM, are capable of performing prediction on a daily travel mode choices of users in Kuantan city centre for a door-to-door journey. It was shown from the study that the Random Forest classifier performs significantly better than any other of the investigated classifiers. However, since human decision-making is complex, further investigation of the significant features is recommended. The evaluation of each classifier was based on the default setting of the Orange platform. Therefore, in future research, we recommended evaluating machine learning models by tuning some applicable hyperparameters to robust their performance for training and testing data analysis and adding more classifiers to investigate the effectiveness of the machine learning models in classifying mode choice.

ACKNOWLEDGMENT

The authors would like to thank Universiti Sains Malaysia for supporting this work.

REFERENCES

[1] Y. Zhao and H. Yu, "A door-to-door travel time approach for evaluating modal competition of intercity travel: A focus on the proposed Dallas-Houston HSR route," *J. Transp. Geogr.*, vol. 72, no. July, pp. 13–22, 2018.

[2] I. Benenson, K. Martens, Y. Rofé, and A. Kwartler, "Public transport versus private car GIS-based estimation of accessibility applied to the Tel Aviv metropolitan area," *Ann. Reg. Sci.*, vol. 47, no. 3, pp. 499–515, 2011.

[3] M. Salonen and T. Toivonen, "Modelling travel time in urban networks: Comparable measures for private car and public transport," *J. Transp. Geogr.*, vol. 31, pp. 143–153, 2013.

[4] J. Marti-Henneberg, "Challenges facing the expansion of the high-speed rail network," *J. Transp. Geogr.*, vol. 42, pp. 131–133, 2015.

[5] M. Diao, Y. Zhu, and J. Zhu, "Intra-city access to inter-city transport nodes: The implications of high-speed-rail station locations for the urban development of Chinese cities," *Urban Stud.*, vol. 54, no. 10, pp. 2249–2267, 2017.

[6] S. Peer, J. Knockaert, P. Koster, Y. Y. Tseng, and E. T. Verhoef, "Door-to-door travel times in RP departure time choice models: An approximation method using GPS data," *Transp. Res. Part B Methodol.*, vol. 58, pp. 134–150, 2013.

[7] A. M. E. Lopez, E. Ortega, "www.econstor.eu," in *50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy", 19-23 August 2010, Jönköping, Sweden*, 2010.

[8] J. M. Ureña, P. Menerault, and M. Garmendia, "The high-speed rail challenge for big intermediate cities: A national, regional and local perspective," *Cities*, vol. 26, no. 5, pp. 266–279, 2009.

[9] Y. Zhang and Y. Xie, "Travel mode choice modeling with support vector machines," *Transp. Res. Rec.*, no. 2076, pp. 141–150, 2008.

[10] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Syst. Appl.*, vol. 78, pp. 273–282, 2017.

[11] L. Liang, M. Xu, S. Grant-Muller, and L. Mussone, "Household travel mode choice estimation with large-scale data—an empirical analysis based on mobility data in Milan," *Int. J. Sustain. Transp.*, vol. 0, no. 0, pp. 1–16, 2019.

[12] T. Bantis and J. Haworth, "Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics," *Transp. Res. Part C Emerg. Technol.*, vol. 80, pp. 286–309, 2017.