Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 8789067, 15 pages
https://doi.org/10.1155/2022/8789067

Research Article

Using a Machine Learning Approach to Predict the Thailand Underground Train's Passenger

Wuttipong Kusonkhum,[1] Korb Srinavin,[1] Narong Leungbootnak,[1] and Tanayut Chaitongrat[2]

[1]Department of Civil Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand
[2]Construction and Project Management Center, Faculty of Architecture, Urban Design and Creative Arts, Mahasarakham University, Maha Sarakham 44150, Thailand

Correspondence should be addressed to Korb Srinavin; korbsri@kku.ac.th

Received 9 November 2021; Revised 27 March 2022; Accepted 1 April 2022; Published 10 May 2022

Academic Editor: Yuan Gao

In today's world, data has become an asset for businesses. Many sectors use data technology to advance their businesses. Building management is one of the processes on which numerous studies have been conducted to assist building users. Thailand has progressed in terms of transportation infrastructure and public transportation. The Metropolitan Rapid Transit (MRT) system has more than one hundred million users per year. However, crowding is a concern in the present since crowding creates a problem and reduces customer pleasure. The goal of this research is to create a machine learning model for forecasting passenger demand over time. In addition, standard data collecting equipment was used to collect data from the Metropolitan Rapid Transit (MRT) Purple Line. This line has a total of 16 stations. Station name, date, day, month, period, number of passengers, holidays, weekends, and weather are among the nine factors. Analysis approaches included the analysis phase, classification, and regression algorithm. However, the regression algorithm's accuracy is poor and therefore cannot be used. Before using machine learning classification methods, the K-means was used to cluster the types of passengers. In addition, for this investigation, three classification methods were used: artificial neural network, random forest, and decision tree. Furthermore, the findings revealed that the artificial neural network has a high predicting accuracy. The accuracy value stated is more than 0.85 for demand over time.

1. Introduction

The fact that infrastructure is improving is rarely news in a growing country. A beneficial influence of efficiency on infrastructure output has been shown empirically in numerous research investigations [1]. The essential systems and citizen services that a country or organization requires to function efficiently, such as transportation and electricity supply, are referred to as "infrastructure." Infrastructure is a component of the national economy's territorial structure, which is made up of the transportation, not just railway transport [2]. Infrastructure is a vast area with many distinct components; however, they may all be divided into two categories. Transportation infrastructure is an essential component of every city's or state's transportation system. As a consequence of the societal expansion and the increase of international relations as a result of globalization processes, transportation has become a more important component for economic and social development [3]. However, because virtually all infrastructure projects are developed for public transit, they must be managed properly to ensure project success.

As a result, project management skill differs from that of other industries that influence project types like hospital or railway construction. Project management's aims are to complete a project within its scope, budget, quality, and schedule restrictions [4, 5]. Railway engineering is a big system project with the following features [6] when contrasted to regular industrial and civil building. Currently, project management focuses on postdelivery project management, such as zero-waste building management or crowd control in public facilities such as hospitals and railway stations [7]. To enhance building management and determine the trend's prospects, many technologies were used. An approach to analyzing vast volumes of data is machine learning. It is one of the technologies that have been used to enhance operations by analyzing data and forecasting user behaviors [8–11].

Thailand's 20-Year Transportation System Development Strategy (2017–2036) [12] is a project that focuses on building transportation infrastructure, particularly rail transportation. The reason for this shift is to escape traffic congestion and travel with ease, as shown by the growing number of passengers who use the electric train system in metropolitan areas each year. Parasuraman et al. argued that passengers' or users' perceived service quality can be assessed by comparing their needs or expectations to the actual service received, with perceived quality as an indicator of passenger satisfaction [13]. Public authorities are now playing an important role in encouraging sustainable development policies and in promoting sustainable urban mobility practices that aim to minimize the use of private automobiles and promote the use of sustainable modes of transportation such as public transportation. This transportation plan will confront a variety of problems in urban and peri-urban regions. These factors include public transportation's regularity, quality of service, and congestion. Estimating and predicting travel demand constitute a key challenge in this setting [14, 15]. One of the most common uses of smart card data analysis is to estimate and anticipate travel demand. Forecasting can help with both service and travel planning. Prediction can produce average travel demand depending on the time period examined. Forecasting can help match transportation supply to demand in real time [16]. Given the volatility and complexity of passenger flow changes in urban rail transportation, using a prediction model to obtain a more accurate forecast of short-term passenger flow is both critical and challenging [17]. The railway is a vital artery for the country's economic development. At the moment, demand for railway passenger transportation is multi-structured, multi-leveled, and multi-segmented. A key difficulty is to ensure coordinated growth of the railway companies and the economy. The demand for passenger transportation is diversifying and individuating [18].

Since the development of intelligent transportation systems in recent decades, forecasting short-term traffic flow and projecting traffic conditions in the near future in a quantitative manner [19] have become a major topic in transportation research [20]. Accurate short-term traffic forecasting might, in fact, aid proactive dynamic traffic control by monitoring existing traffic and estimating its immediate future. Scholars consider the problem of minimizing road traffic congestion [21, 22]. All of these objectives and benefits include informing travelers or drivers about traffic conditions [23, 24], as well as providing real-time traffic monitoring and management [24]. In fact, forecasting short-term traffic flow in metro transportation is substantially more challenging, as metro traffic flow is highly influenced by the heterogeneity and unpredictability of individual travel behavior, and AFC data does not reflect traffic conditions promptly. It has previously been attempted to anticipate short-term passenger flow using AFC data. For example, Leng et al. [25] suggested a metro-net oriented probability tree technique for passenger prediction based on origin and destination (OD) information. Sun et al. [26] developed a wavelet and support vector machines (SVM) hybrid method to predict Beijing subway passenger flow, particularly during morning and evening peak hours.

As a result, service providers assess passenger happiness in order to improve quality and service standards for sustainable urban electric trains, as these factors can increase passengers' quality of life and contentment. Furthermore, the Thailand Transport System Development Plan considers and mentions the development of an urban electric train system that will cover Bangkok and counties, as well as important cities in every province of Thailand [27]. One of the essential factors for Smart Cities is the Thai government's goals. Urban railroads have recently received a lot of interest since they are practically the only mode of transportation in the city that can travel without being stuck in traffic. It also helps us to predict passenger demand using data technologies [8, 28].

Consequently, we target collected data that can be used to train a machine learning model to anticipate passenger demand at any given time. In addition, in this study, we look at several machine learning methods that can be determined with great accuracy. The work's contribution was the developed model of passenger transportation behaviors, which took into consideration the availability of new urban railroads, such as the MRT Purple Line [29].

2. Background of Study

2.1. Construction Project Management. Construction project management is different from other industries [4]. Because there are a lot of dangers and elements that are up for debate. As well as a variety of project parameters, such as project kinds. [30, 31]. They are challenging and one-of-a-kind in terms of specifics. A construction project for a hospital, for example, where the building type provides complexity and purpose building of a railway and a stadium [5, 32]. There are 10 knowledge areas of project management, namely, project integration, project scope management, project time management, project cost management, project quality management, project human resource management, project communication management, project risk management, project procurement management, and project stakeholder management. Efforts are now being made to enhance building management following user operations. Except for public benefit projects, in order for the building to be efficient and constantly developed, there are numerous things to consider, including user demand in the facility. The research of rail building projects in Thailand, on the other hand, represents a significant potential and growth for Thailand's development. The trend of technology and innovation, such as Big Data technology [8], is one of the most important factors driving the development of railway construction. These have an impact on numerous infrastructure studies that focus on building management using data technology [9].

2.2. Data-Driven Transportation. In the last decade, data-driven approaches have become an alternate strategy for a number of studies in transportation and building management, such as train delay estimation. Gorman [33] used linear regression for a first-class freight train (BNSF) in order to identify the components that cause delays. The model is run on nine different districts, each with its own set of traffic patterns and track layouts. Train delays are calculated for each of the eight districts, taking into account parameters like horsepower per ton, track geometry, train priorities, meets, passes, overtakes, and train spacing variations. It is the first time that regression algorithms have been used to forecast delays in US freight train data. Moreover, the number of factors considered in the regression is significantly smaller, since data on passenger trail is significantly more limited compared to data available (internally) to the freight railroad. Kecman and Goverde [34] offer a microscopic model for railroad networks to forecast train travel time and delay. To predict train delays, historical track occupancy data is utilized to train the parameters in the microscopic model. Hansen et al. are another group that utilizes a data-driven technique to estimate train delays [35], where an online model is trained using historical track occupancy data and then applied to a section of the Dutch railway route. Railroad track occupation data is not publicly accessible in the United States. To predict train delays, the regression models presented in this article employ train departure time information at stations. Google and Amtrak have collaborated on a program to track Amtrak trains and estimate arrival times [36]. There is, however, no research on the algorithms or their correctness. As a result, this study is the first quantitative and data-driven investigation of strategies for estimating passenger train delays in the United States [37]. Consequently, there are a few studies on demand of passengers with ML algorithm for prediction in a short period.

2.3. Machine Learning Model for Public Transportation. In the United Kingdom, passenger train services are experiencing a renaissance. Approximately 200 new stations have opened on publicly operated railroads in the United Kingdom since 1970. The key to organization throughout the operating phase is the quantity of passengers who utilize it. Many public transportation networks are experiencing increased congestion and crowding as metropolitan populations grow. Growing urban populations cause many public transit systems to experience increasing congestion and crowding. Crowding is associated with negative effects on traveler satisfaction and well-being, including stress, anxiety, threat to personal safety and security, and loss of productivity due to lack of seating space [38, 39]. According to studies, the perceived journey time of passengers increases when congestion increases [40, 41]. Vehicle stay durations at stations, as well as passenger waiting times, are affected by crowding, which increases headway unpredictability and decreases dependability [42, 43]. As a result, additional trucks are necessary to meet demand, resulting in considerable operating expenses for the operator. Even during peak hours, passenger loads on trains and metros can be extremely unequally distributed across cars, contributing to crowding concerns [44, 45]. Because of the uneven passenger loads, the trains' effective capacity is substantially lower than the stated capacity predicated on all cars being used equally. The periods, dates, capacity of each waiting position, crowding distribution across train cars, and exit placement at the destination station are all elements that impact passenger loads [46–48].

Currently, with the advancement of technology, gadgets are becoming smaller and more powerful, and Internet access is becoming more affordable and widespread. This has resulted in a profusion of linked gadgets on the Internet, resulting in the fascinating Internet-of-Things (IoT) movement [49]. The fundamental goal of IoT is to connect smart devices and things, which are critical components of the Internet. The fusion of these interesting physical and digital worlds is providing fascinating development prospects. Logistics, transportation, asset monitoring, smart homes, smart buildings, energy, defence, and agriculture are just a few of the prominent sectors where IoT applications have been effectively proven across industries. The availability of data technology may be able to alleviate train crowding. Researchers have paid a lot of attention to the use of user data to analyze mobility in public transit. The initial study focused on data completeness and enrichment with the goal of identifying transfers and passenger demand [14, 15]. Recently, a significant amount of data-driven research on passenger flow forecasting has been conducted utilizing data mining and machine learning techniques. In order to anticipate passenger flow in railways, a prediction model was created [50].

2.4. Algorithm of Machine Learning. Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on enabling computer systems to learn from data about a given job automatically. Rule-based learning approaches [51], artificial neural network methods [52–54], case-based reasoning strategies [55, 56], and hybrid methodologies [57, 58] are all used in building to model judicial reasoning and forecast litigation outcomes.

2.4.1. Regression Algorithm. The supervised machine learning approach of regression is concerned with estimating the numerical value of a target variable based on input data, for example, estimating the cost of a design based on design specifications. There are several forms of regression. The connection between a dependent variable y and one explanatory variable x is modeled using basic linear regression. The logistic regression is used to evaluate the probability of a specific class or event occurring, such as pass/fail, win/lose, alive/dead, or healthy/sick. This may be used to represent a wide range of events, such as determining if a photograph contains a cat, a dog, a lion, or other animals. A probability of 0 to 1 would be assigned to each detected object in the image, with a total of one. This is a common regression method [8]. Equation (1) should be used to refer to them.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p \beta_p + \epsilon. \tag{1}$$

2.4.2. Artificial Neural Network (ANN) Algorithm. Artificial neural networks (ANNs) are a type of machine learning analysis. The methods of artificial neural networks (ANNs) are ideally adapted to classification and function estimation issues. These algorithms have been widely employed in tackling difficult industrial issues since their inception. The most popular kind of ANN is the multilayer perceptron (MLP). An input layer, a hidden (intermediate) layer, and an output layer are the three layers that make up an ANN. Through deep learning, ANN algorithms have lately revolutionized machine learning. New ANN algorithms are being developed to learn from data with large dimensionality (i.e., Big Data), seeking special attention in all the construction industry applications where ANN is employed [8]. The neural network model has hidden units as shown in Figure 1, and they should be referred to as (2)-(3) [59].

$$\begin{aligned} f(x) &= \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X) \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k g\left(W_{k0} + \sum_{j=1}^{p} W_{kj} X_j \right). \end{aligned} \tag{2}$$

It is built up here. The K activations A_k, $k = 1,\ldots, K$, in the hidden layer are computed as functions of the input features X1,..., Xp:

$$A_k = h_k(X) = g\left(W_{k0} + \sum_{j=1}^{p} W_{kj} X_j \right). \tag{3}$$

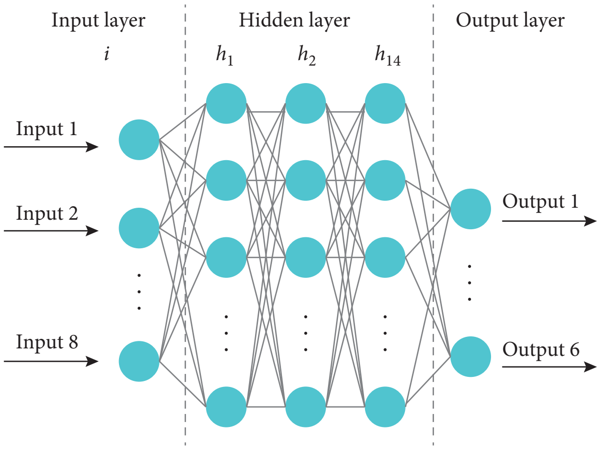

Figure 1: The structure of ANN.

2.4.3. Random Forest Algorithm. The random forest classifier is made up of many tree classifiers, each of which is produced using a random vector sampled separately from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector [60].

Random forests outperform bagged trees thanks to a tiny change in the way the trees are decorated. On bootstrapped training samples, we create numerous decision trees, similar to bagging. When creating these decision trees, however, a random sample of m predictors is picked as split candidates from the whole set of p predictors each time a split in the tree is examined. Only one of the m predictors can be used in the split. At each split, a new sample of m predictors is selected, and we usually pick m p—that is, the number of predictors examined at each split is about identical.

In other words, while creating a random forest, the algorithm is not even permitted to examine a majority of the available predictors at each split in the tree. This may appear absurd, but there is a good reason behind it. Assume the dataset contains one extremely strong predictor and a few more somewhat strong predictors. The majority, if not all, of the trees in the bagged tree collection will employ this strong predictor in the top split. As a result, all of the bagged trees will have a similar appearance. As a result, the bagged tree forecasts will be strongly connected. Unfortunately, averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities. In particular, this means that bagging will not lead to a substantial reduction in variance over a single tree in this setting.

By forcing each split to evaluate only a subset of the predictors, random forests are able to avoid this difficulty. As a result, the strong predictor will be ignored in the vast majority of splits (p m)/p, giving other forecasters a better opportunity. This procedure may be thought of as decorating the trees, resulting in a less variable and hence more trustworthy average of the generated trees. The size of the predictor subset m is the primary distinction between bagging and random forests. For example, if $m = p$ is used to construct a random forest, then bagging [59] is the result.

2.4.4. Decision Tree Algorithm. The contemporary machine learning approach to predicting qualitative and quantitative target attributes is decision trees (DTs). The first step in creating DT is to locate the decision node, which is followed by recursively splitting nodes until no further divisions are allowed. The robustness of DT is determined by the logic used to divide nodes, which is measured using terms like information gain (IG) and entropy reduction [8]. A simple decision tree model with a single binary goal variable Y (0 or 1) and two continuous variables X1 and X2, all of which span from 0 to 1, is shown in Figure 2 [59]; the primary components of a decision tree model are nodes and branches, and the most significant processes in developing a model are splitting, halting, and pruning.

3. Research Methodology

The data for this study came from the Metropolitan Rapid Transit (MRT) which provided the authorization to use the system for data gathering. The needed data included

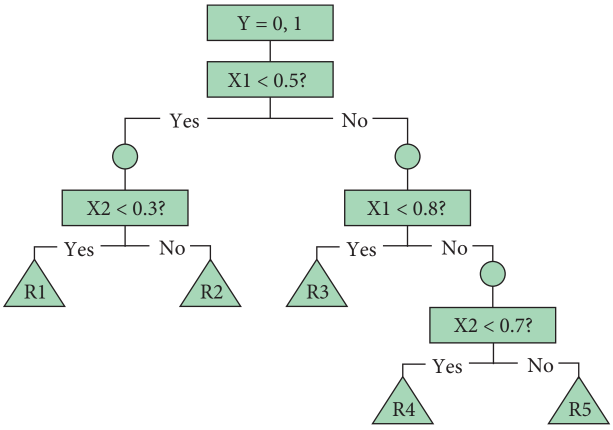

Figure 2: Sample decision trees based on binary target variable Y.

information on nine factors in 2017–2019. There are 4 processes of study: regression algorithm modeling, K-means clustering, classification algorithm modeling, and validation data with confusion matrix. The data was separated into two sections based on the gathering of the essential data: 80% of the data was used for model training, and 20% of the data was used for model validation [61, 62].

3.1. Population of Study. The MRT Purple Line is Bangkok's fifth rapid transit line, which is the population in this study. There are 16 stations of MRT Purple Line: Khlong Bang Phai (101), Talad Bang Yai (102), Sam Yaek Bang Yai (103), Bang Phlu (104), Bang Rak Yai (105), Bang Rak Noi Tha It (106), Sai Ma (107), Phra Nang Klao Bridge (108), Yaek Nonthaburi 1 (109), Bang Krasor (110), Nonthaburi Civic Center (111), Ministry of Public Health (112), Yaek Tiwanon (113), Wong Sawang (114), Bang Son (115), and Tao Poon (116). This railway line has opened in August 2016. The data was collected from 2017 to 2019. The data was collected with paper that should be prepared in a CSV file to ensure that there was no missing value and unknown category. There are nine factors for input data collected from the government, namely, station name, date, day, month, period, number of passengers, holidays, weekends, and weather, as shown in Tables 1 and 2.

3.2. Regression Algorithm Model Development. The collected data needed to be prepared in a CSV file to ensure that there was no missing value and unknown category. Moreover, a computer program was necessary to perform linear regression algorithm and logistic regression [64]. The computer program was written in Python language and ran on Anaconda software.

Table 1: Input data.

Input name	Detail	Reference
Station name	There are 13 stations of Metropolitan Rapid Transit (MRT) Purple Line in Thailand.	[63]
Date	There are seven days for a week.	[16, 27]
Day	This is the number of each month.	[16, 27]
Month	There are 12 months for each year.	[16]
Period	The system of the MRT collects the data of passenger number every 15 minutes since its operation on each day.	[16, 63]
Number of passengers	This is the number of passengers in each period.	[16, 18]
Holidays	These are the holidays in Thailand calendar.	[27]
Weekends	These are Sundays.	[27]

Table 2: Attributes in input data.

No.	Attributes	Factor	Unit
1	Station name	16	Station
2	One day of a week	7	Day
3	Date	30–31	Day
4	Month	12	Month
5	Period	75	Period
6	Number of passengers	6	Group
7	Holidays	2	Case
8	Weekends	2	Case

3.3. Clustering with K-Means Technique. Even with huge datasets, K-means clustering is simple to use, especially when utilizing heuristics like Lloyd's method. It has been utilized successfully in a variety of fields, including market segmentation, computer vision, and astronomy. It is also frequently used as a preprocessing step for other algorithms, such as finding a starting configuration. The K-means technique may be used in cluster analysis to split the input dataset into k parts (clusters). However, the pure K-means method is not particularly versatile, as it has limitations in terms of application (except when vector quantization as above is the desired use case). In particular, the parameter k is known to be hard to choose (as discussed above) when it is not given by external constraints. Another limitation is that it cannot be used with arbitrary distance functions or on nonnumerical data. For these use cases, many other algorithms are superior [65]. They should be referred to as follows:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_{i1} - m_{j1})}, \tag{4}$$

where $i = x_i$ and $j = y_i$ are two n-dimensional data objects.

3.4. Machine Learning Model Development with Classification Algorithm. The collected data needed to be prepared in a CSV file to ensure that there was no missing value and unknown category. Moreover, a computer program was necessary to perform KNN, SVM, ANN, and decision tree algorithm. The hidden layer size of ANN is 14 [64]. The computer program was written in Python language and ran on Anaconda software.

Table 3: Percentage of day of a week in this study.

No.	Station	Frequency	Percent	Cumulative percent
1	Monday	151,245	14.4	14.4
2	Tuesday	150,937	14.5	28.9
3	Wednesday	148,653	14.2	43.1
4	Thursday	151,162	14.4	57.5
5	Friday	148,833	14.2	71.7
6	Saturday	145,655	13.9	85.6
7	Sunday	150,407	14.4	100.0
	Sum	1,046,892	100	

Table 4: Percentage of month in this study.

No.	Month	Frequency	Percent	Cumulative percent
1	January	73,199	7.0	7.0
2	February	100,783	9.6	16.6
3	March	108,373	10.4	27.0
4	April	106,097	10.1	37.1
5	May	111,223	10.6	47.7
6	June	106,823	10.2	57.9
7	July	74,399	7.1	65.0
8	August	74,397	7.1	72.1
9	September	72,000	6.9	79.0
10	October	73,198	7.0	86.0
11	November	72,000	6.9	92.9
12	December	74,400	7.1	100.0
	Sum	1,046,892	100	

3.5. Verifying the Model. The classification model was verified for its accuracy, precision, and recall by constructing a confusion matrix and using the following equations [66]:

$$\text{accuracy} = \frac{(TP - TN)}{(TP - TN - FP - FN)}, \quad (5)$$

$$\text{precision} = \frac{TP}{(TP - FP)}, \quad (6)$$

$$\text{recall} = \frac{TP}{(TP - FN)}. \quad (7)$$

A confusion matrix is also a table that displays the numbers of true positives, false positives, true negatives, and false negatives, as stated below:

True positive (TP) is a class label that has been accurately anticipated.

False positive (FP) occurs when a label does not belong to a class yet is projected to be positive by the classifier.

The label true negative (TN) does not belong to the class and is properly predicted.

The label false negative (FN) belongs to the class, but it is anticipated to be negative [67, 68].

The accuracy of a model is defined as the ratio of the total number of accurate classifications to the total number of projected classifications. Precision is also described as the capacity to get consistent findings from several measurements. Random error, a type of observational mistake in information retrieval, causes precise values to vary from one another. Recall is sometimes defined as the percentage of relevant documents successfully recovered [69].

Table 5: Percentage of date in this data.

No.	Date	Frequency	Percent	Cumulative percent
1	1st	13,956	3.3	3.3
2	2nd	13,959	3.3	6.6
3	3rd	13,957	3.4	10.0
4	4th	13,961	3.2	13.2
5	5th	13,957	3.3	16.6
6	6th	13,961	3.3	19.9
7	7th	13,956	3.3	23.2
8	8th	13,960	3.3	26.5
9	9th	13,956	3.3	29.8
10	10th	13,960	3.3	33.1
11	11th	13,957	3.1	36.2
12	12th	13,960	3.2	39.4
13	13th	13,957	3.3	42.7
14	14th	13,959	3.3	46.1
15	15th	13,958	3.2	49.3
16	16th	13,959	3.2	52.5
17	17th	13,959	3.1	55.6
18	18th	13,960	3.3	58.9
19	19th	13,957	3.2	62.1
20	20th	13,960	3.4	65.5
21	21st	13,957	3.1	68.6
22	22nd	13,960	3.3	71.9
23	23rd	13,957	3.2	75.1
24	24th	13,961	3.3	78.4
25	25th	13,957	3.3	81.8
26	26th	13,960	3.3	85.1
27	27th	13,958	3.4	88.5
28	28th	13,958	3.4	92.0
29	29th	13,958	3.1	95.1
30	30th	13,958	3.1	98.2
31	31st	13,960	1.8	100.0
	Sum	1,046,892	100	

Table 6: Percentage of holidays in this data.

No.	Station	Frequency	Percent	Cumulative percent
1	Holidays	913,993	87.3	87.3
2	Non-holidays	132,899	12.7	100.0
	Sum	1,046,892	100	

Table 7: Percentage of weekend in this data.

No.	Station	Frequency	Percent	Cumulative percent
1	Weekends	881,366	84.2	84.2
2	Weekdays	165,526	15.8	100.0
	Sum	1,046,892	100	

4. Results and Discussion

The result of this study has included six parts: 1. general information; 2. linear regression of machine learning model; 3. K-means clustering; 4. classification of machine learning model; 5. verifying the model; 6. evaluation of forecasting.

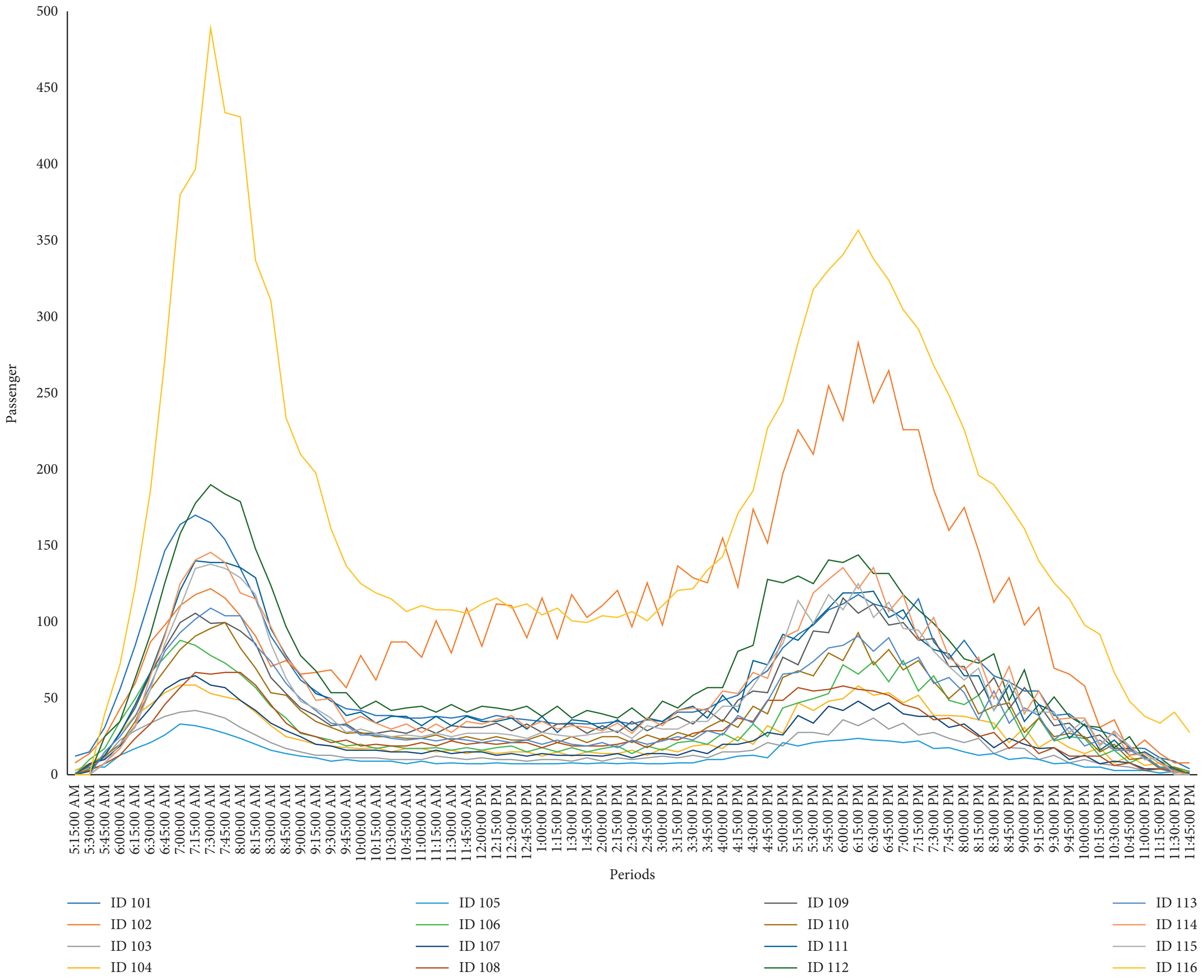

Figure 3: The average of passengers in each station.

4.1. General Information. There are eight parameters for input data: station name, one day of a week, day, month, period, number of passengers, holidays, and weekends. The most preferable day of a week for this study is Tuesday, which accounts for 14.5% but the percentage is near to those of other days as shown in Table 3. The most preferable month of data is the May, which accounts for 10.6% as shown in Table 4. The percentage of each date is nearly as shown in Table 5. The holidays account for 12.7% and the weekends account for 15.8% as shown in Tables 6 and 7.

There are 16 stations of MRT Purple Line, and we collected data of every 15 minutes of people usage for each station. The data of this study show that the most crowded station is Tao Poon station (116), with a maximum average of 484 passengers. The second station is Khlong Bang Phai (101), with a maximum average of 283 passengers. The following station is Talad Bang Yai (102), with a maximum average of 173 passengers as shown in Figure 3. The 283-person average represents a high level of in-station railway crowding, as measured by the use of available standee spaces, which is common for trains, metros, and buses [63]. For nearly every station, the busiest times are 6.15–8.30 AM and 4.00–8.00 PM. According to another study, the morning peak occurs between the hours of 6 : 00 AM and 9 : 00 AM [63].

4.2. Regression Model. A linear regression algorithm was used to develop a model for forecasting the number of passengers. Table 8 shows that the accuracy of the algorithm is 55.55 percent, but this accuracy is low and could not be useful, as shown in detail in Table 9. Accordingly, the stated accuracy has at most a very small effect on people's trust in the model [70].

4.3. K-Means Clustering for Passenger Type. The results of K-means clustering show that the passenger behavior could be separated into six groups. In addition, the initial cluster center of cluster 1 is zero people, that of cluster 2 is 959 people, that of cluster 3 is 480 people, that of cluster 4 is 720 people, that of cluster 5 is 1,327 people, and that of cluster 6 is 240 people, as shown in Table 10. The final cluster center of cluster 1 is 25.87 people, that of cluster 2 is 598.28 people, that of cluster 3 is 225.12 people, that of cluster 4 is 392.23

Table 8: Accuracy of linear regression.

Input	Accuracy
Linear regression	52.54

Table 9: Validation of statistic model.

Parameter	Result
R-squared	0.520
Adj. R-squared	0.520
Skewness	2.815
Kurtosis	26.104

Table 10: K-means result.

Cluster	Initial cluster center	Final cluster center	Number of passengers
1	0.00	25.87	796,104
2	959.00	598.28	2,563
3	480.00	225.12	27,608
4	720.00	392.23	9,667
5	1327.00	891.80	277
6	240.00	110.15	190,673

people, that of cluster 5 is 891.80 people, and that of cluster 6 is 110.15 people, as shown in Table 10. Finally, the results indicate the number of passengers for each cluster as shown in Table 10. The ANOVA test is shown in Table 11. The length of each cluster is shown in Table 12.

4.4. Classification of Machine Learning Model. Three algorithms were used to develop a model for passenger behavior classification in each period: ANN, random forest, and decision tree. Table 13 shows that the accuracy values of the algorithms are close to each other, but the highest is that of the ANN algorithm, being 89.80 percent. In order for the accuracy to be useful, it has to be more than 80 percent [70].

4.5. Verifying the Model. The confusion matrix [68] is used to calculate the model's classification accuracy. The matrix of ANN model showed that the model made correct prediction for 188,036 out of 209,379 cases. Therefore, the gray box is misclassified and the white box is correctly classified as shown in Figure 4, and the number zero in confusion matrix table means that the model did not make a mistake in prediction for each case. Similarly, the ANN model's precision can also be calculated by using the confusion matrix. The precision can be divided into six cases of passenger volume (i.e., cluster 1, cluster 2, cluster 3, cluster 4, cluster 5, and cluster 6), as shown in Table 14. For the first case, cluster 1, the model achieved a precision of 95%. For cluster 2, the model achieved a precision of 73%. For cluster 3, the model achieved a precision of 70%. For cluster 4, the model achieved a precision of 68%. For cluster 5, the model achieved a precision of 74%. For cluster 6, the model achieved a precision of 75%.

Table 11: ANOVA test.

	DF	Error mean square	F test	Sig.
Total	5	515.302	1,817,531.084	0.000

Table 12: Number of passengers in each group.

Cluster	Number of passengers
1	0–70
2	513–756
3	176–324
4	325–512
5	>757
6	71–175

Table 13: Accuracy of each algorithm.

Input	Accuracy
Artificial neural network	89.80
Random forest	88.21
Decision tree	86.49

The matrix of random forest model showed that the model made correct prediction for 184,714 out of 209,379 cases. Therefore, the gray box is misclassified and the white box is correctly classified as shown in Figure 5. The random forest model's precision can also be calculated by using the confusion matrix. The precision can be divided into six cases of passenger volume (i.e., cluster 1, cluster 2, cluster 3, cluster 4, cluster 5, and cluster 6), as shown in Table 15. For the first case, cluster 1, the model achieved a precision of 94%. For cluster 2, the model achieved a precision of 68%. For cluster 3, the model achieved a precision of 63%. For cluster 4, the model achieved a precision of 68%. For cluster 5, the model achieved a precision of 38%. For cluster 6, the model achieved a precision of 70%.

The confusion matrix is used to calculate the model's classification accuracy. The matrix of decision tree model showed that the model made correct prediction for 181,092 out of 209,379 cases. Therefore, the gray box is misclassified and the white box is correctly classified as shown in Figure 6.

The decision tree model's precision can also be calculated by using the confusion matrix. The precision can be divided into six cases of passenger volume (i.e., cluster 1, cluster 2, cluster 3, cluster 4, cluster 5, and cluster 6), as shown in Table 16. For the first case, cluster 1, the model achieved a precision of 93%. For cluster 2, the model achieved a precision of 66%. For cluster 3, the model achieved a precision of 56%. For cluster 4, the model achieved a precision of 61%. For cluster 5, the model achieved a precision of 34%. For cluster 6, the model achieved a precision of 67%.

The precision of confusion matrix has shown that the ANN algorithm could show the highest accuracy in all the cases; however, for cluster 4 behavior prediction, random forest might outperform ANN with great efficiency. for cluster 4 (case 4). This point could prove that the traditional data have a relation for application data technology [8].

		Actual class					
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Predicted class	Cluster 1	153, 321	0	128	4	0	5, 811
	Cluster 2	0	286	54	138	8	50
	Cluster 3	554	0	5, 428	316	0	3, 285
	Cluster 4	16	62	412	1, 040	1	387
	Cluster 5	0	42	0	6	25	1
	Cluster 6	8, 335	3	1, 708	22	0	27, 936

Figure 4: Confusion matrix table of ANN model.

Table 14: Confusion matrix of ANN model.

	Precision	Recall	F1-score	Support
Cluster 1	0.95	0.96	0.95	159,264
Cluster 2	0.73	0.53	0.62	536
Cluster 3	0.70	0.57	0.63	9,583
Cluster 4	0.68	0.54	0.60	1,918
Cluster 5	0.74	0.34	0.46	74
Cluster 6	0.75	0.74	0.74	38,004
Accuracy			0.90	209,379
Macro avg.	0.76	0.61	0.67	209,379
Weighted avg.	0.89	0.90	0.90	209,379

		Actual class					
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Predicted class	Cluster 1	152, 232	7	244	9	0	6, 887
	Cluster 2	2	256	58	120	6	52
	Cluster 3	647	9	4, 806	302	0	3, 789
	Cluster 4	22	69	470	1, 016	9	387
	Cluster 5	0	30	1	8	10	1
	Cluster 6	9, 392	4	2, 091	48	1	26, 394

Figure 5: Confusion matrix table of random forest model.

Table 15: Confusion matrix of random forest model.

	Precision	Recall	F1-score	Support
Cluster 1	0.94	0.96	0.95	159,379
Cluster 2	0.68	0.52	0.59	494
Cluster 3	0.63	0.50	0.56	9,553
Cluster 4	0.68	0.51	0.58	1,973
Cluster 5	0.38	0.20	0.26	50
Cluster 6	0.70	0.70	0.70	37,930
Accuracy			0.88	209,379
Macro avg.	0.67	0.56	0.61	209,379
Weighted avg.	0.88	0.88	0.88	209,379

		Actual class					
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Predicted class	Cluster 1	150, 731	3	363	11	0	8, 192
	Cluster 2	5	249	76	134	26	40
	Cluster 3	811	7	4, 732	384	0	3, 602
	Cluster 4	35	92	467	930	4	370
	Cluster 5	0	17	0	14	17	2
	Cluster 6	10, 678	11	2, 876	64	3	24, 433

Figure 6: Confusion matrix table of decision tree.

Table 16: Confusion matrix of decision tree model.

	Precision	Recall	F1-score	Support
Cluster 1	0.93	0.95	0.94	159,300
Cluster 2	0.66	0.47	0.55	530
Cluster 3	0.56	0.50	0.52	79,536
Cluster 4	0.61	0.49	0.54	1,898
Cluster 5	0.34	0.34	0.34	50
Cluster 6	0.67	0.64	0.65	38,065
Accuracy			0.86	209,379
Macro avg.	0.63	0.56	0.59	209,379
Weighted avg.	0.86	0.86	0.86	209,379

Table 17: Precision accuracy of each algorithm.

Cases	Machine learning algorithm		
	ANN (%)	Random forest (%)	Decision (%)
Cluster 1	95	94	93
Cluster 2	73	68	66
Cluster 3	70	63	56
Cluster 4	68	68	61
Cluster 5	74	38	34
Cluster 6	75	70	67

Classification algorithm performance is normally measured by assessing classification accuracy. Artificial neural networks may be used to produce good results from classification algorithms [71] as shown in Table 17.

4.6. Evaluation of Forecasting. In this section, we evaluate the forecasting performance in terms of forecasting step. Forecast step refers to granularity of data aggregation, and so far we use 6 cases of behaviors for train station. Here, we compare the performance with different forecasting process. There are forecasting for each station and new behaviors from K-means analysis (three cases for each station, namely, low, medium, and high). The behavior of passenger for each station is shown in Table 18 [59]. We employed an absolute error metric, i.e., the mean absolute percentage error (MAPE), defined by (8), to objectively evaluate model performance [72].

Table 18: Performance of model with rain data.

Station ID	Cases of passengers in station		
	Low	Medium	High
ID01	<439	440–879	>880
ID02	<292	293–586	>587
ID03	<65	66–131	>132
ID04	<107	108–215	>216
ID05	<145	146–331	>332
ID06	<140	141–282	>283
ID07	<184	185–447	>448
ID08	<97	98–196	>197
ID09	<132	133–278	>279
ID10	<131	132–263	>264
ID11	<171	172–345	>346
ID12	<192	193–387	>388
ID13	<116	117–233	>234
ID14	<187	188–376	>377
ID15	<189	190–380	>381
ID16	<662	663–1326	>1327

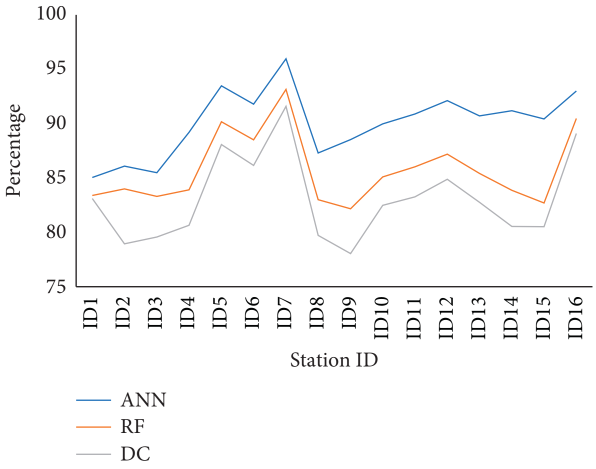

Figure 7: Accuracy of model for each station.

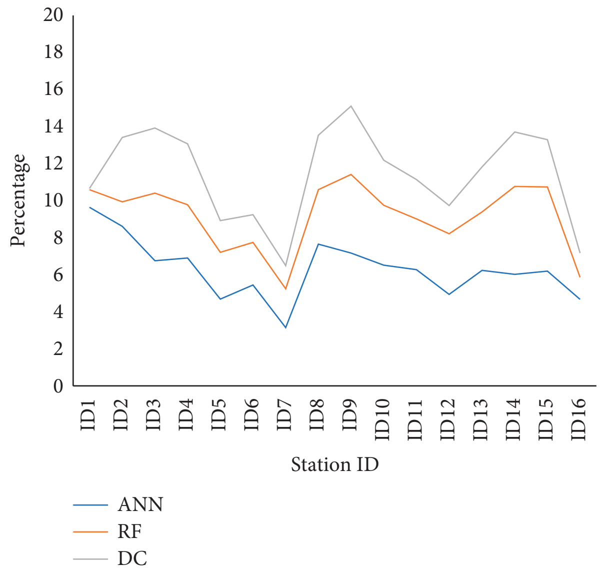

Figure 8: MAPE of model for each station.

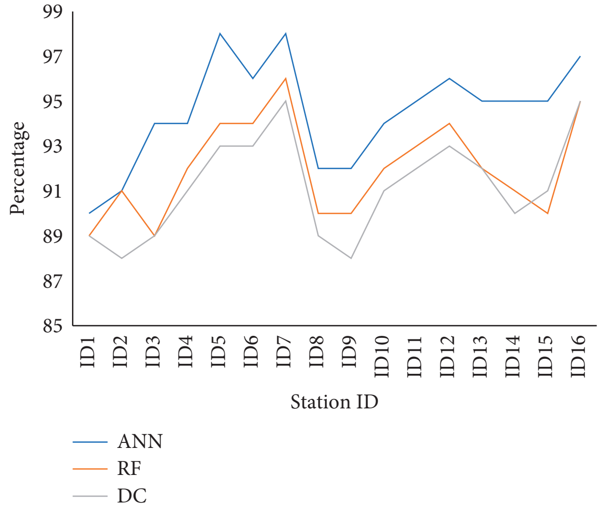

Figure 9: Precision of model for low passenger in each station.

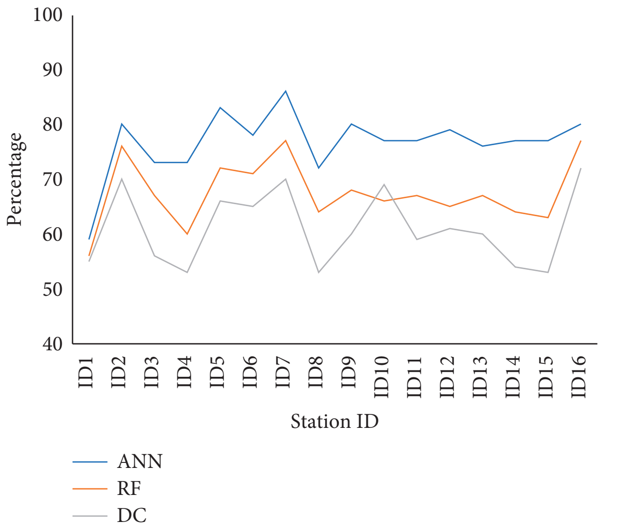

Figure 10: Precision of model for medium passenger in each station.

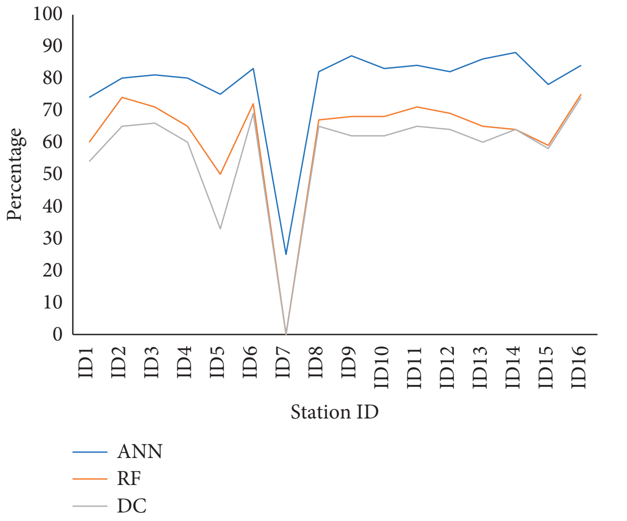

Figure 11: Precision of model for high passenger in each station.

$$MAPE = \frac{100}{N}\sum_{k=1}^{N}\left|\frac{\widehat{y}_k - y_k}{y_k}\right|. \quad (8)$$

The accuracy of each model increases as it is processed in each station, according to the model's performance. The ANN model, on the other hand, has been processed with greater precision than that of previous methods. The ANN model has an accuracy of more than 85% as seen in Figure 7. Figure 8 shows that The MAPE of a forecasting model boosts prediction accuracy, the output of MAPE processing achieves a number less than 10, and the superior performance in each station is that of the ANN model. ANNs are capable of extracting high degrees of abstraction from raw data, making them a popular and accurate tool in computer vision [73]. The precision of confusion matrix has shown that the ANN algorithm could show the highest accuracy in each station. However, because this station has limited data on this circumstance, only station ID7 could be predicted with high accuracy with a high passenger case as shown in Figures 9–11. The rain becomes the factor for improving model [74]. Figure 12 shows a comparison of model performance using nonrain data versus rain data. However, because this study employed a large amount of data from the user's everyday activities, this element may boost performance a little, as shown in Tables 18 and 19.

The performance of each algorithm is summarized in this section. The ANN processes data with more precision than previous methods. Station name, date, day, month, period, number of passengers, holidays, weekends, and weather are among the ten input variables that have been chosen. This finding implies that factor qualities are important in determining the prevalence and intensity of passenger behavior [74].

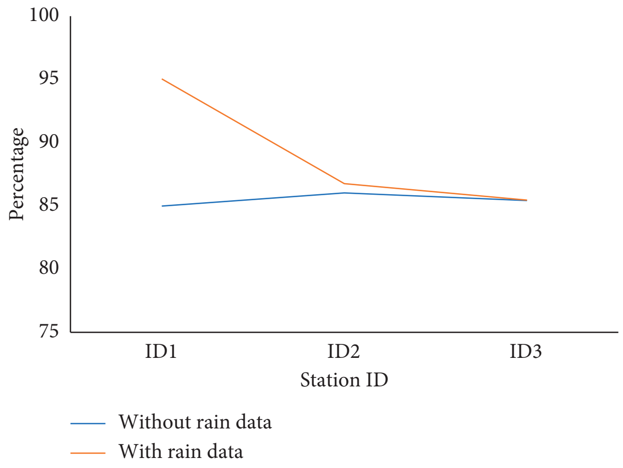

Figure 12: Comparison performance with rain factor.

Table 19: Performance of model with rain data.

Station ID	Accuracy	MAPE	Precision		
			Low	Medium	High
ID01	95.05	3.14	0.97	0.86	0.88
ID02	86.76	8.27	0.92	0.80	0.79
ID03	85.46	8.81	0.90	0.71	0.78

The prediction performance for this study, a multiclass classification, is encouraging when compared to the binary classification studies by Zhang et al. [75] and Chou & Lin [76]. The accuracy of Zhang et al.'s SVM in predicting whether a project is of "excellent profitability" or is "less profitable" ranged from 0.74 to 0.91. However, their dataset had a class imbalance issue (i.e., the majority of the firms are "less profitable"), and their simulation results revealed that the majority class accounted for all of the expected values. In another study, Chou & Lin [76], using their ensemble model, were able to attain a prediction accuracy of 0.84 in forecasting Public-Private Partnership project conflict, i.e., "dispute" or "no disagreement." In this study, three algorithms are used to assess and forecast performance. "Low passenger, medium passenger, and high passenger" are the three classes predicted by the ML models. In this investigation, the ANN performed exceptionally well, with a relatively high accuracy of 0.95 in Station ID7 and a lesser model with an accuracy of 0.85 in Station ID1. This research looked at several forms of prediction mistakes as well as the dataset's unbalanced distribution of classes. The ANN model with previous time variables can be used as a reference variable in the predictive control system [77].

Furthermore, the study demonstrates that passenger behavior does not occur at random. Furthermore, it is shown that when there is a large amount of passenger data, depending just on timing data may efficiently anticipate the amount of passenger flow for each station. Nearly half of the 9 input variables come from passenger daily life indicators, indicating that project management issues have an impact on accident occurrence and severity. This is said to be comparable to how some experts would judge passenger behavior.

5. Conclusion

Based on nine characteristics acquired from conventional data, we present a model for passenger prediction for the MRT Purple Line using ANN, decision tree, and random forest. The government provided eight criteria for evaluating the machine learning study, namely, station name, date, day, month, period, passenger number, holidays, and weekends. These markers can be classified with high accuracy using ANN, decision tree, and random forest. In other circumstances, however, the Purple Line prediction model has a low accuracy. The procedure of upgrading the prediction model is carried out for each station using the generated model.

In each station, the clustering algorithm was used once again. Three examples of passenger behavior are shown in this paper, all of which are based on past research. The procedure could be completed with great precision in each station, and we do it with the weather on a daily basis. Finally, because the data in this study is large and has an impact on the prediction model, the rain data is ineffective for this framework. The contribution of this study, data from previous Thai government work, might be used with data technologies; however, traditional data collection should be enhanced.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to sincerely thank the Graduate School, Khon Kaen University, for their funding support. This research work was also supported by the Research Fund of the Faculty of Engineering, Khon Kaen University, under the Research Scholarship for Ph.D. Students project, Contract No. Ph.D. 4/2562. In addition, the authors would like to thank the Metropolitan Rapid Transit for allowing their data to be used in this study. Without any of them, our research would not be accomplished.

References

[1] D. A. Aschauer, "Is public expenditure productive?" *Journal of Monetary Economics*, vol. 23, no. 2, pp. 177–200, 1989.

[2] S. M. Rinaldi, J. P. Peerenboom, and T. K. Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, 2001.

[3] O. Skorobogatova and I. Kuzmina-Merlino, "Transport infrastructure development performance," *Procedia Engineering*, vol. 178, pp. 319–329, 2017.

[4] K. K. Chitkara, *Construction Project Management*, Tata McGraw-Hill Education, New York, NY, United States, 1998.

[5] A. Guide, *Project Management Body of Knowledge (Pmbok® Guide),* Project Management Institute, Pennsylvania, United States, 2013.

[6] W. Lin, L. Yaqi, and W. Enmao, "Research on risk management of railway engineering construction," *Systems Engineering Procedia*, vol. 1, pp. 174–180, 2011.

[7] J. L. Hensen and R. Lamberts, *Building Performance Simulation for Design and Operation*, Routledge, England, UK, 2012.

[8] M. Bilal, L. O. Oyedele, J. Qadir et al., "Big Data in the construction industry: A review of present status, opportunities, and future trends," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 500–521, 2016.

[9] K. Srinavin, W. Kusonkhum, B. Chonpitakwong, T. Chaitongrat, N. Leungbootnak, and P. Charnwasununth, "Readiness of applying big data technology for construction management in Thai public sector," *Journal of Advances in Information Technology*, vol. 12, no. 1, pp. 1–5, 2021.

[10] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution that Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt, Boston, Massachusetts, United States, 2013.

[11] E. Siegel, *Predictive analytics: The power to predict who will click, buy, lie, or die*, John Wiley & Sons, Hoboken, New Jersey, United States, 2013.

[12] Ministry of Transport, *The 20 Years' Thailand Transport System Development Strategy (2017–2036)*, Ministry of Transport, Bangkok, Thailand, 2016.

[13] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "A conceptual model of service quality and its implications for future research," *Journal of Marketing*, vol. 49, no. 4, pp. 41–50, 1985.

[14] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[15] W. Wang, J. P. Attanucci, and N. H. M. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, pp. 131–150, 2011.

[16] F. Toqué, M. Khouadjia, E. Come, M. Trepanier, and L. Oukhellou, "Short & long term forecasting of multimodal transport passenger flows with machine learning methods," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 560–566, Yokohama, Japan, October 2017.

[17] R. Liu, Y. Wang, H. Zhou, and Z. Qian, "Short-term passenger flow prediction based on wavelet transform and kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 158025–158034, 2019.

[18] F. Zhu, "Factor analysis of railway passenger transport demand in Eastern China," in *Proceedings of the 2nd International Conference on Education, Economics and Management Research (ICEEMR 2018)*, Atlantis Press, Singapore, June 2018.

[19] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *Journal of Transportation Engineering*, vol. 123, no. 4, pp. 261–266, 1997.

[20] M. Alam, J. Ferreira, and J. Fonseca, "Introduction to intelligent transportation systems," in *Intelligent Transportation Systems*, pp. 1–17, Springer, Berlin/Heidelberg, Germany, 2016.

[21] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 557–569, 2016.

[22] Z. Cao, S. Jiang, J. Zhang, and H. Guo, "A unified framework for vehicle rerouting and traffic light control to reduce traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1958–1973, 2017.

[23] J. Ma, B. L. Smith, and X. Zhou, "Personalized real-time traffic information provision: Agent-based optimization model and solution framework," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 164–182, 2016.

[24] H. Hashemi and K. F. Abdelghany, "Real-time traffic network state estimation and prediction with decision support capabilities: Application to integrated corridor management," *Transportation Research Part C: Emerging Technologies*, vol. 73, pp. 128–146, 2016.

[25] B. Leng, J. Zeng, Z. Xiong, W. Lv, and Y. Wan, "Probability tree based passenger flow prediction and its application to the beijing subway system," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 195–203, 2013.

[26] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-svm short-time passenger flow prediction in beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[27] N. Wonglakorn, V. Ratanavaraha, A. Karoonsoontawong, and S. Jomnonkwao, "Exploring passenger loyalty and related factors for urban railways in Thailand," *Sustainability*, vol. 13, no. 10, p. 5517, 2021.

[28] D. Namiot, Z. Kutuzmanov, E. Fedorov, and O. Pokusaev, "On the assessment of socio-economic effects of the city railway," *International Journal of Open Information Technologies*, vol. 6, no. 1, pp. 92–103, 2018.

[29] A. Misharin, D. Namiot, and O. Pokusaev, "On passenger flow estimation for new urban railways," *IOP Conference Series: Earth and Environmental Science*, vol. 177, no. 1, Article ID 012012, 2018.

[30] T. Chaitongrat, S. Liwthaisong, P. Aksorn, W. Kusonkhum, and N. Leungbootnak, "Causal relationship model of problems in public sector procurement," *Geomate Journal*, vol. 20, no. 80, pp. 52–58, 2021.

[31] B. Teanngen, W. Kusonkhum, S. Liwthaisong, T. Chaitongrat, and K. Srinavin, "Risk factors affecting conflict management for construction government project in Thailand," *Multidisciplinary Technologies for Industrial Applications*, vol. 1, pp. 94–105, 2020.

[32] B. M. Jervis and P. Levin, *Construction Law, Principles and Practice*, McGraw-Hill College, New York, US, 1988.

[33] M. F. Gorman, "Statistical estimation of railroad congestion delay," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 3, pp. 446–456, 2009.

[34] P. Kecman and R. M. Goverde, "Online data-driven adaptive prediction of train event times," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 465–474, 2014.

[35] I. A. Hansen, R. M. Goverde, and D. J. van der Meer, "Online train delay recognition and running time prediction," in *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, pp. 1783–1788, Funchal, Portugal, September 2010.

[36] R. Wang and D. B. Work, "Data driven approaches for passenger train delay estimation," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 535–540, Gran Canaria, Spain, September 2015.

[37] J. Preston, "Demand forecasting for new local rail stations and services," *Journal of Transport Economics and Policy*, vol. 25, no. 2, pp. 183–202, 1991.

[38] G. Beirão and J. S. Cabral, "Understanding attitudes towards public transport and private car: A qualitative study," *Transport Policy*, vol. 14, no. 6, pp. 478–489, 2007.

[39] A. Tirachini, D. A. Hensher, and J. M. Rose, "Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand," *Transportation Research Part A: Policy and Practice*, vol. 53, pp. 36–52, 2013.

[40] M. Wardman and G. Whelan, "Twenty years of rail crowding valuation studies: Evidence and lessons from British experience," *Transport Reviews*, vol. 31, no. 3, pp. 379–398, 2011.

[41] S. Raveau, Z. Guo, J. C. Muñoz, and N. H. M. Wilson, "A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics," *Transportation Research Part A: Policy and Practice*, vol. 66, pp. 185–195, 2014.

[42] K. M. Kim, S.-P. Hong, S.-J. Ko, and D. Kim, "Does crowding affect the path choice of metro passengers?" *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 292–304, 2015.

[43] W. H. K. Lam, C.-Y. Cheung, and C. F. Lam, "A study of crowding effects at the Hong Kong light rail transit stations," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 5, pp. 401–415, 1999.

[44] Q. Zhang, B. Han, and D. Li, "Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 5, pp. 635–649, 2008.

[45] Parsons Brinckerhoff, *Transit Capacity and Quality of Service Manual*, 2013.

[46] S. Peftitsi, E. Jenelius, and O. Cats, "Determinants of passengers' metro car choice revealed through automated data sources: A Stockholm case study," *Transportmetrica: Transportation Science*, vol. 16, no. 3, pp. 529–549, 2020.

[47] H. Kim, S. Kwon, S. K. Wu, and K. Sohn, "Why do passengers choose a specific car of a metro train during the morning peak hours?" *Transportation Research Part A: Policy and Practice*, vol. 61, pp. 249–258, 2014.

[48] Z. Christoforou, P.-A. Collet, B. Kabalan et al., "Influencing longitudinal passenger distribution on railway platforms to shorten and regularize train dwell times," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2648, no. 1, pp. 117–125, 2017.

[49] J. A. Stankovic, "Research directions for the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, 2014.

[50] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1623–1632, 2016.

[51] J. E. Diekmann and T. A. Kruppenbacher, "Claims analysis and computer reasoning," *Journal of Construction Engineering and Management*, vol. 110, no. 4, pp. 391–408, 1984.

[52] D. Arditi, F. E. Oksay, and O. B. Tokdemir, "Predicting the outcome of construction litigation using neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 13, no. 2, pp. 75–81, 1998.

[53] M. P. Kim, "US Army Corps Engineers construction contract claims guidance system," in *Excellence in the Constructed Project*, pp. 203–209, ASCE, Reston, Virginia, United States, 1989.

[54] K. Chau, "Predicting construction litigation outcome using particle swarm optimization," in *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 571–578, Springer, Berlin, Heidelberg, 2005 June.

[55] O. B. Tokdemir, "Using case-based reasoning to predict the outcome of construction litigation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 14, no. 6, pp. 385–393, 1999.

[56] K. W. Chau, "Prediction of construction litigation outcome - a case-based reasoning approach," in *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 548–553, Springer, Dalian, China, June 2006.

[57] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using boosted decision trees," *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp. 387–393, 2005.

[58] J.-H. Chen and S. C. Hsu, "Hybrid ANN-CBR model for disputed change orders in construction projects," *Automation in Construction*, vol. 17, no. 1, pp. 56–64, 2007.

[59] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, p. 18, Springer, New York, 2013.

[60] L. Breiman, "Random Forests-Random Features (# 567)," Technical report, Dept, of Statistics, Univ. of California, Berkeley, 1999.

[61] A. Gondia, A. Siam, W. El-Dakhakhni, and A. H. Nassar, "Machine learning algorithms for construction projects delay risk prediction," *Journal of Construction Engineering and Management*, vol. 146, no. 1, Article ID 04019085, 2020.

[62] J. Hurwitz and D. Kirsch, *Machine Learning for Dummies*, p. 75, IBM Limited Edition, 2018.

[63] E. Jenelius, "Data-driven metro train crowding prediction based on real-time load data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2254–2265, 2019.

[64] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with Java implementations," *Acm Sigmod Record*, vol. 31, no. 1, pp. 76-77, 2002.

[65] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[66] Z. M. Yaseen, Z. H. Ali, S. Q. Salih, and N. Al-Ansari, "Prediction of risk delay in construction projects using a hybrid artificial intelligence model," *Sustainability*, vol. 12, no. 4, p. 1514, 2020.

[67] M. Ali, D.-H. Son, S.-H. Kang, and S.-R. Nam, "An accurate CT saturation classification using a deep learning approach based on unsupervised feature extraction and supervised fine-tuning strategy," *Energies*, vol. 10, no. 11, p. 1830, 2017.

[68] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, p. 218, 2016.

[69] J. Hernández-Torruco, J. Canul-Reich, J. Frausto-Solis, and J. J. Méndez-Castillo, "Towards a predictive model for Guillain-Barré syndrome," in *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7234–7237, Milan, Italy, (August2015).

[70] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, Glasgow, Scotland Uk, May2019.

[71] M. M. Saritas and A. Yasar, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.

[72] P. Huang, C. Wen, L. Fu, Q. Peng, and Z. Li, "A hybrid model to improve the train running time prediction ability during high-speed railway disruptions," *Safety Science*, vol. 122, Article ID 104510, 2020.

[73] P. Samui, S. S. Roy, and V. E. Balas, *Handbook of Neural Computation*, Academic Press, Cambridge, Massachusetts, United States, 2017.

[74] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. L. Tsui, "Forecasting short-term passenger flow: An empirical study on shenzhen metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, 2018.

[75] H. Zhang, F. Yang, Y. Li, and H. Li, "Predicting profitability of listed construction companies based on principal component analysis and support vector machine-Evidence from China," *Automation in Construction*, vol. 53, pp. 22–28, 2015.

[76] J.-S. Chou and C. Lin, "Predicting disputes in public-private partnership projects: Classification and ensemble models," *Journal of Computing in Civil Engineering*, vol. 27, no. 1, pp. 51–60, 2013.

[77] S. Park, M. Kim, M. Kim et al., "Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN)," *Journal of Hazardous Materials*, vol. 341, pp. 75–82, 2018.