Feature Engineering Algorithms for Traffic Dataset

Akibu Mahmoud Abdullah¹, Raja Sher Afgun Usmani², Thulasiammal Ramiah Pillai³, Ibrahim Abaker Tarigio Hashemi⁴, Mohsen Marjani⁵
School of Computer Science and Engineering, Taylor’s University¹,²,³,⁵
Selangor, Malaysia
College of Computing and Informatics, Department of Computer Science⁴, University of Sharjah, 27272 Sharjah, UAE

Abstract—As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. The emergence of an automatic traffic management system allows us to record and monitor motor vehicles’ movement in a road segment. One of the challenges researchers face is when the historical traffic data is given as an annual average that contains incomplete data. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am. Hence there is a need to estimate hourly traffic volume for the remaining hours. Feature engineering can be used to overcome the issue of incomplete data. This paper proposed feature engineering algorithms that can efficiently estimate hourly traffic volume and generate features from the existing dataset for all traffic census stations in Malaysia using queuing theory. The proposed feature engineering algorithms were able to estimate the hourly traffic volume and generate features for three years in Jalan Kepong census station, Kuala Lumpur, Malaysia. The algorithms were evaluated using the Random Forest model and Decision Tree Models. The result shows that our feature engineering algorithms improve machine learning algorithms’ performance except for the prediction of $NO_2$ using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction. The algorithm is applied in one of the traffic census stations in Kuala Lumpur, and it can be used for the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes average hourly data.

Keywords—Feature engineering algorithm; queuing theory; Road Traffic Volume Malaysia (RTVM); machine learning algorithms

I. INTRODUCTION

As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. This reason results in the emergence of an automatic traffic management system for recording and monitoring the hourly and daily movement of motor vehicles. Several studies reported that motor vehicles are primary sources of air pollution in the urban area worldwide [1]. The concentration and increase of air pollution depend on the increase of traffic volume, speed of the vehicle, type of vehicle and many more factors. Researchers are looking for creative solutions like smart cities and GIS systems to avoid traffic congestion and volume [2, 3]. A study conducted reveals that traffic volume has a significant impact on $PM_{10}$, $NO_x$, $NO$, and $NO_2$ concentrations [4]. [5]’s study shows that the increase of the vehicle increases the concentrations of air pollution during peak hours in the morning and evening. Traffic volume, traffic congestion, and low speed increase level of $PM$ and $NO_x$ emissions [6]. A study in Kula Lumpur shows that air pollution concentration strongly depends on traffic volume, waiting time on the road, speed of the vehicle, and fuel consumption [7]. Speed of the vehicle, composition, traffic volume, intensity, and acceleration influenced the concentration of air pollution [8].

One of the challenge researchers are facing is when the recorded traffic data is given as an annual average that contains incomplete data- and most of these researchers are conducting multidisciplinary studies- [9]. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. AADT data are collected using surveillance cameras to count and monitor passing vehicles on a 24-hour, 16-hour, 5-hour, or 1-hour basis. These data are used mostly in road transport studies, such as estimation of fuel consumption, roadway planning, emission prediction, traffic operation, travel behavior, accident predictions, and many more [10]. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am, as shown in Fig. 1. The total daily traffic volume and peak hour (hourly traffic volume from 8.00

![Fig. 1. Summary of Daily and Hourly Traffic Volume.](image-url)
to 9.00 am) were highlighted with red color in the Figure. The highlighted blue color indicates the volume of type of the vehicle. The not available (N/A) in the Figure shows that the remaining hourly traffic volume and volume of the type of the vehicles were missing. There is a need to estimate the hourly traffic for the remaining hours. In this study, feature engineering was applied to overcome the issue of incomplete hourly traffic volume.

Feature engineering is one of the most challenging and significant tasks in data science. Extracting and generating new variable from the existing dataset is a difficult task, and also consume time, and effort to process variable in dataset before applying them in the model. Feature engineering is the process of extracting and generating new features or variables from the existing dataset which helps in improving the performance of Machine Learning Algorithms. It also helps to understand the data deeply and gives more valuable insights. Data scientist spend more than 80% of their time on cleaning the dataset [11].

The structure of the paper is presented as follows, Section II presents the related works, Section III discusses the methodology, and it is divided into two sections; namely, Section III-A presented the data and how it was collected, and Section III-B discusses the proposed feature engineering algorithms. In Section IV, the results were presented, it has two sections, Section IV-A is the feature engineering algorithms result and Section IV-B is the prediction of traffic emissions with and without traffic dataset. Discussion was presented in Section V. Lastly, the conclusion is discussed in Section VI.

II. RELATED WORK

Traffic volume is one of the important variable, which contribute for increasing air pollution level produced by automobiles. Many studies were conducted to estimated hourly traffic volume, for example [12] applied extreme gradient boosting tree (XGBoost) and graph theory to estimate hourly traffic volume at location without traffic sensor in Utah United States of America (USA). The developed model was able to estimate hourly traffic volume. Study of [13] propose deep learning algorithm and image processing method to estimate traffic volume, vehicle type, and vehicle speed using recorded traffic video. The model was found good with 90% of accuracy for traffic volume estimation. Estimation of hourly traffic volume was conducted using Artificial Neural Network (ANN) [14]. The applied ANN model was able to estimated hourly traffic volume.

Time spent on the road and speed vehicles were responsible for the variability and trend of air pollution. Several studies were conducted to estimate vehicle speed and time spent on the road. These studies include [15], [16], [17], [18], [19], [20], [21], [22], [23], and [24]. These studies can be divided into three. Firstly, most of the studies have speed parameters in their data, so they developed models to estimate the vehicle speed in a location that they do not have traffic sensor stations. Some researchers proposed an algorithm to estimate the dataset’s missing values, while others focus on estimating the speed to evaluate their models using the historical dataset. The second category is having a recorded video of the moving automobiles on the road, so they developed methods to estimate the vehicle’s speed. Lastly, some studies installed sensor devices on the road to calculate and estimate the vehicles’ speed.

In general, all these studies used four types of traffic datasets, namely, sensor device data, video-based data, image-based data, and vehicle data. Fig. 2 presents the four types of datasets used for the estimation of vehicle speed and time spent on the road. All of the presented studies none of them estimate or generate vehicle speed and time spent using AADT dataset. To the best of our knowledge, we could not find a study that generates the cars’ speed and time spent on the road using AADT dataset.

The RTVM data have been used by many studies in Malaysia. Table V presents the researches were conducted in Malaysia using the RTVM dataset. These studies mostly used the data to explore the Level of Service (LOS), the number of registered vehicles, and traffic density. There is a lack of study using hourly traffic data due to unavailability and incomplete hourly traffic volume dataset. In this paper, we proposed feature engineering algorithms which can efficiently estimate hourly traffic volume and generate features from existing traffic dataset (RTVM dataset). Queuing model is proposed to generate vehicle speed and time spent on the road features.

III. METHODOLOGY

A. Data

There is a total number of 554 traffic monitoring stations all over Malaysia. The traffic was recorded hourly for the state and federal roads in Malaysia by the State Public Works Department (JKR Negeri) and organized by Road Traffic Volume Malaysia (RTVM). In 1982, the first printed copy of the organized national traffic census was published by the RTVM. From early 1999, the data were available in Compact Disc (CD). In contrast, the online version started from 2014 to date. The traffic sensor is conducted twice a year during March-April or September-October. The data collection is categorized into three types, type 0, type 1, and type 2. The type 0 data is recorded for 24 hours in 7 days, while type
1 for 16 hours in 7 days, and type 2 for 16 in 1 day. The recording for 16 hours started from 6.00 am to 10.00 pm. The vehicles are divided into six classes, as presented in Table I. We also utilize the air pollution dataset from the AQM stations provided by the Department of Environment (DOE), Malaysia and feature engineered in our previous work [11, 25, 26].

<table>
<thead>
<tr>
<th>No</th>
<th>Class</th>
<th>Types of Vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Class 1</td>
<td>Motor Cars</td>
</tr>
<tr>
<td>2</td>
<td>Class 2</td>
<td>Small Vans</td>
</tr>
<tr>
<td>3</td>
<td>Class 3</td>
<td>Lorries</td>
</tr>
<tr>
<td>4</td>
<td>Class 4</td>
<td>Large Vans (Heavy 2-axes)</td>
</tr>
<tr>
<td>5</td>
<td>Class 5</td>
<td>Buses</td>
</tr>
<tr>
<td>6</td>
<td>Class 6</td>
<td>Motorcycles</td>
</tr>
</tbody>
</table>

The RTVM divides the carriageway into two, single and dual carriageway. Table II describes the carriageway types, their code, and how many lanes each road includes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Single Carriageway</th>
<th>Code</th>
<th>Dual Carriageway</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1-1</td>
<td>Two-Lane</td>
<td>K1+2</td>
<td>Three Lane - One way</td>
</tr>
<tr>
<td>T1-2</td>
<td>Three-Lane</td>
<td>K2+2</td>
<td>Four Lane - One way</td>
</tr>
<tr>
<td>T2</td>
<td>Two-Lane - One way</td>
<td>K2-2</td>
<td>Four Lane</td>
</tr>
<tr>
<td>T2-2</td>
<td>Four-Lane</td>
<td>K1+(1+2)</td>
<td>Six Lane</td>
</tr>
<tr>
<td>T3</td>
<td>Three-Lane - One way</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Fig. 3 summarizes the total number of traffic census stations in each region in Malaysia. Johor is recorded as the highest region with 75 stations, while Kuala Lumpur with the lowest number of stations with 5.

![Traffic Census Stations](image)

B. Feature Engineering Algorithms

1) Feature Estimation: The feature engineering algorithm is proposed to estimate hourly traffic volume. The algorithm performs two tasks; EstimationOfData and DataDistribution. The EstimationOfData is performed by selecting the station, peak hour, and the normal hour. Six peak hours (peak hours 7.00 to 8.00 am, 8.00 to 9.00 am, 9.00 to 10.00 am, 10.00 to 11.00 am, 17.00 to 18.00 pm, and 18.00 to 19.00 pm) were selected and distributed randomly from the daily average traffic volume for six months, while the remaining amount of the daily traffic volume were distributed randomly to the ten hours (normal hours 11.00 am to 12.00 pm, 12.00 to 13.00 pm, 13.00 to 14.00 pm, 14.00 to 15.00 pm, 15.00 to 16.00 pm, 16.00 to 17.00 pm, 19.00 to 20.00 pm, 20.00 to 21.00 pm, 21.00 to 22.00 pm, and 22.00 to 23.00 pm) for six months as given in the following equations:

\[ p = h \] (1)

\[ n = d - p \] (2)

The \( p \) is the peak hour, \( h \) stand for hourly data given by RTVM, \( n \) is the normal hour, \( d \) daily traffic volume. The percentage of the vehicle type was distributed from the total daily traffic volume using the below equation. The \( v \) is the vehicle type, \( d \) is the daily traffic volume, and \( c \) is the percentage of vehicle type.

\[ v = d * p \] (3)

The DataDistribution is the distribution of the estimated traffic data obtained from EstimationOfData. Since the data is based on a six-month average, we create three-years data with hourly rows for sixteen hours, because the RTVM data is based on sixteen hours. We first distribute the amount of peak...
TABLE V. PREVIOUS STUDIES THAT USED RTVM DATASET

<table>
<thead>
<tr>
<th>Author</th>
<th>Objective</th>
<th>Location</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>[27]</td>
<td>Condition of vehicle engine based on driving behavior</td>
<td>Kuala Lumpur</td>
<td>RTVM_2015</td>
</tr>
<tr>
<td>[28]</td>
<td>Effect of traffic-related air pollution on traffic policemen</td>
<td>Klang Valley</td>
<td>RTVM_2017</td>
</tr>
<tr>
<td>[29]</td>
<td>Association between traffic-related air pollution with respiratory symptoms and DNA damage</td>
<td>Kajang, Hulu Langat, Selangor</td>
<td>RTVM_2016</td>
</tr>
<tr>
<td>[31]</td>
<td>The autonomous emergency braking system was proposed for the primary accident that is occurring</td>
<td>Jalan Butterworth, Penang</td>
<td>RTVM_2016</td>
</tr>
<tr>
<td>[32]</td>
<td>Proposed a preventive model for road maintenance</td>
<td>Petaling</td>
<td>RTVM</td>
</tr>
<tr>
<td>[33]</td>
<td>Exploring the effect contributing to the vehicle accident</td>
<td>Sabah</td>
<td>RTVM</td>
</tr>
<tr>
<td>[34]</td>
<td>Investigating the influence of rubbernecking towards vehicle deceleration rate due to primary accident in the urban area</td>
<td>Jalan Butterworth and Jalan George Town, Penang</td>
<td>RTVM_2016</td>
</tr>
<tr>
<td>[35]</td>
<td>Proposed Malaysia driving cycle (MDC) for light-duty test</td>
<td>Terengganu</td>
<td>RTVM_2015</td>
</tr>
<tr>
<td>[36]</td>
<td>Investigates driving cycle which contributes to producing air pollution and fuel consumption</td>
<td>Kuala Lumpur</td>
<td>RTVM</td>
</tr>
<tr>
<td>[37]</td>
<td>Estimation of particulate matter from non-exhaust and exhaust vehicle</td>
<td>Klang Valley</td>
<td>RTVM_2014</td>
</tr>
<tr>
<td>[38]</td>
<td>Development of the driving cycle for route selection along with mountainous areas</td>
<td>Penang</td>
<td>RTVM_2015</td>
</tr>
<tr>
<td>[39]</td>
<td>Investigates single-vehicle accident along with mountainous areas</td>
<td>Sabah</td>
<td>RTVM_2013</td>
</tr>
<tr>
<td>[40]</td>
<td>The concentration of air pollution near a primary school</td>
<td>Pahang</td>
<td>RTVM_2014</td>
</tr>
<tr>
<td>[41]</td>
<td>Driving behavior among heavy track drivers due to pavement damage</td>
<td>Ampang, Kuala Lumpur</td>
<td>RTVM</td>
</tr>
<tr>
<td>[42]</td>
<td>Proposed a model for estimating annual daily traffic for single carriageway</td>
<td>Johor</td>
<td>RTVM_2012</td>
</tr>
<tr>
<td>[43]</td>
<td>Level of CO and CO2 due to the increase of motor vehicles in industrial areas</td>
<td>Shah Alam, Seremban, and Kuantan</td>
<td>RTVM</td>
</tr>
<tr>
<td>[45]</td>
<td>Vehicle compassion in Malaysia and Indonesia</td>
<td>Malaysia</td>
<td>RTVM_2009</td>
</tr>
<tr>
<td>[47]</td>
<td>The trend of public transport in Perak</td>
<td>Perak</td>
<td>RTVM_2005</td>
</tr>
</tbody>
</table>

hour for the six hours randomly and then insert and distribute the normal hours randomly (the remaining ten hours), which is the remaining ten hours. Lastly, we distribute the amount to the type of vehicles based on the percentage given in RTVM data. The algorithm is presented in Fig. 4. The six peak hours were distributed randomly using range with minimum and maximum values (so that the values would not be same). Similarly, normal hours were distributed randomly using range with minimum and maximum values.

2) Feature Generation: In this study, queuing theory is applied to calculate average speed of vehicle and time spent on the road. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing system considers arrival time, number of server and service time. The arrival time is considered as the arrival of motor vehicles on the particular road segment, the server is the installed camera that recorded the passing vehicles, while the service rate is the time after the vehicle leaves where the camera was installed. The Queuing model structure is presented in Fig. 5, and the notations used in the model.

The arrival rate of the vehicles is distributed using poison distribution. Similarly, the service rate is exponentially distributed. The time spent or waiting time is calculated using the following equation:

\[
W = \frac{\lambda^2}{\mu(\mu - \lambda)}
\]  

(4)

Algorithm 1: Estimation of Daily Traffic Volume

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```

```
for i in range(6):  
    value = random(min, max)  
    print(value)
```
The average speed of the motor vehicles is calculated using the below equation:

\[ s = \frac{L}{t} \]  

IV. RESULT

A. Feature Engineering Algorithms

Motor vehicle has a significant impact on traffic emissions concentration, and human health; moreover, it causes accidents, traffic congestion, and fuel consumption. The emergence of an automatic traffic management system allows us to record and monitor every vehicle passing on the road. The recorded data are used primarily in transportation studies. In Malaysia, the RTVM provided annual average traffic data. There is a need to estimate the hourly traffic volume. This study proposed a feature engineering algorithm that will efficiently estimate hourly traffic volume and generate features from the existing dataset in Malaysia’s traffic census stations. The RTVM gives an average hourly (for peak hour) and daily traffic volume in specific stations. The proposed algorithm was able to estimate the hourly traffic volume for three years based on the yearly average provided by RTVM and distributed the types of vehicles based on the percentage given in the data. Furthermore, the queuing theory was able to generate the vehicle’s average speed and time spent on the road. The output of the feature engineering algorithms (estimated and generated features) was shown in Fig. 6.

B. Prediction of Traffic Emission With and Without Traffic Dataset

Due to the lack of studies that generate and estimate features from the RTVM dataset. To justify the claim that feature engineering improves machine learning models’ performance, we proposed Random Forest and Decision Three machine learning algorithms to predict traffic emissions concentrations using the estimated and generated features (traffic dataset). Additional dataset of air quality and meteorological variables in [11] study were used. The input and output variables were presented in Table VI.

<table>
<thead>
<tr>
<th>Input Variables</th>
<th>Meteorological Variables</th>
<th>Output Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic Volume, Types of vehicle (Car and Taxi, Van and Utilities, Medium Lorries, Heavy Lorries, Buses, and Motorcycles)</td>
<td>Wind Speed, Wind Direction, Temperature, and Relative Humidity.</td>
<td>Carbon Monoxide (CO), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO₂), Nitrogen Oxides (NOₓ).</td>
</tr>
</tbody>
</table>

Evaluation metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) were used to evaluate the performance of the models. First of all, we predict the level of the CO, NO, NO₂, and NOₓ pollutants using meteorological features but without traffic data. Lastly, we included the traffic dataset (estimated and generated features) for prediction. The result shows that our feature engineering algorithms improve the accuracy of the machine learning models (Random Forest and Decision Tree models) by predicting the level of traffic pollutants except for the NO₂, which shows no improvement using the Random Forest model as presented in Table VII and VIII. We can also visualize the results in Figures 7, 8, and 9 for the Random Forest Model, while Figures 10, 11, and 12 for the Decision Tree Model.
Fig. 8. MSE for the Random Forest Model.

Fig. 9. RMSE for the Random Forest Model.

Fig. 10. MAE for the Decision Tree Model.

Fig. 11. MSE for the Decision Tree Model.

Fig. 12. RMSE for the Decision Tree Model.

**TABLE VII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING RANDOM FOREST MODEL**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>MAE (Without TD)</th>
<th>MSE (Without TD)</th>
<th>RMSE (Without TD)</th>
<th>MAE (With TD)</th>
<th>MSE (With TD)</th>
<th>RMSE (With TD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>0.1294</td>
<td>0.0294</td>
<td>0.1715</td>
<td>0.1146</td>
<td>0.0244</td>
<td>0.1564</td>
</tr>
<tr>
<td>NO</td>
<td>0.1348</td>
<td>0.0452</td>
<td>0.2127</td>
<td>0.1169</td>
<td>0.0347</td>
<td>0.1862</td>
</tr>
<tr>
<td>NO₂</td>
<td>0.1874</td>
<td>0.0646</td>
<td>0.2543</td>
<td>0.2736</td>
<td>0.1567</td>
<td>0.3959</td>
</tr>
<tr>
<td>NOₓ</td>
<td>0.2481</td>
<td>0.1223</td>
<td>0.3497</td>
<td>0.2205</td>
<td>0.09789</td>
<td>0.3128</td>
</tr>
</tbody>
</table>

**TABLE VIII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING DECISION TREE MODEL**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>MAE (Without TD)</th>
<th>MSE (Without TD)</th>
<th>RMSE (Without TD)</th>
<th>MAE (With TD)</th>
<th>MSE (With TD)</th>
<th>RMSE (With TD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>0.1663</td>
<td>0.0491</td>
<td>0.2216</td>
<td>0.151</td>
<td>0.0349</td>
<td>0.2096</td>
</tr>
<tr>
<td>NO</td>
<td>0.1874</td>
<td>0.0846</td>
<td>0.2543</td>
<td>0.2736</td>
<td>0.1567</td>
<td>0.3959</td>
</tr>
<tr>
<td>NO₂</td>
<td>0.2394</td>
<td>0.1073</td>
<td>0.3275</td>
<td>0.1954</td>
<td>0.0775</td>
<td>0.2785</td>
</tr>
<tr>
<td>NOₓ</td>
<td>0.3129</td>
<td>0.2031</td>
<td>0.4506</td>
<td>0.2205</td>
<td>0.09789</td>
<td>0.4252</td>
</tr>
</tbody>
</table>

Fig. 7, 10, 8, 11, 9, and 12 show that the performance of RF model for prediction with traffic data were better than the DT with lower MAE, MSE, and RMSE for CO, NO, and NOₓ, except NO₂ which show higher MAE, MSE, and RMSE. Similarly, for the prediction without traffic dataset, the RF performed better than DT with lower MAE, MSE and
RMSE for \( CO, NO, NO_2 \) and \( NO_2 \). We can conclude that the Random Forest Model was a good choice for the prediction of traffic emission concentrations using traffic dataset and without traffic dataset also.

V. DISCUSSION

Motor vehicles become one of the primary concern for the government and agencies. Automobiles create many problems such as accident and emissions of pollutants to atmosphere. The air pollution produced by motor vehicles have significant impact on human health and the environment as well. Several studies have been conducted for estimation and prediction of traffic emissions. The variability and increase of air pollution depend on traffic characteristics. One of the issue researchers are facing is when the traffic data was provided as an annual average daily traffic (ADDT). The RTVM provides ADDT dataset. The data were incomplete as shown in the Fig. 1. We proposed feature engineering algorithms to estimated and generate missing values in the RTVM dataset. Our feature engineering algorithms were able to generate and estimate the missing features as presented in the Fig. 6. Our feature engineering algorithm is applied in one of the stations in Kuala Lumpur traffic census stations, and this algorithm can be used on the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes an average hourly dataset. There some limitations in this study, firstly, the estimated hourly traffic volume is proposed due to insufficient hourly traffic volume, which may not provide the exact traffic volume hourly. The RTVM does not provide speed of the vehicle, we calculate vehicle speed as an average basis (the speed of the vehicle is constant). This study could not extract acceleration/deceleration. Some studies suggested that different types of fuel have different emissions, but consideration of fuel type is not provided in this study. Jalan Kepong traffic census station was the selected station in this study, the remaining stations were not studied.

VI. CONCLUSION

Motor vehicles are the primary source of air pollution in metropolitan globally. Air pollution has a significant effect on human health with diseases such as asthma, cardiovascular, and respiratory. Motor vehicle also causes accidents and create congestion at road segments. Due to these reasons, the government and agencies introduce an automated traffic management system to record the passing vehicles on the road. The recorded data has been used in various studies by researchers. The Road Traffic Volume Malaysia provides incomplete traffic data. We proposed a new feature engineering algorithm to overcome the issue of incomplete traffic data. The proposed feature engineering algorithms could estimate the hourly traffic volume and generate features for three years in Jalan Kepong, Kuala Lumpur, Malaysia. The algorithm was evaluated by predicting four traffic pollutants \( CO, NO, NO_2 \), and \( NO_2 \) using Random Forest and Decision Tree models. The prediction was conducted in two phases, phase one is prediction without traffic dataset (estimated and generated features), and phase two is the prediction with traffic dataset. The result shows that our feature engineering algorithms improve machine learning models’ performance except for the prediction of \( NO_2 \) using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction.

ACKNOWLEDGMENT

This research is funded by Taylor’s University under the research grant application ID (TUF/R/2017/004/04) entitled as “Modeling and Visualization of Air-Pollution and its Impacts on Health”. We are also thankful to Department of Environment, Malaysia for providing the AQM station dataset and Ministry of Works, Malaysia for providing the RTVM dataset.

REFERENCES

[17] Z. Zhang and X. Yang, “Freeway traffic speed estimation by regression machine-learning techniques using probe vehicle and sensor detector


