DOI: https://doi.org/10.55981/mipi.2023.1734
E- ISSN 2541-1233

Majalah Ilmiah Pengkajian Industri
(Journal of Industrial Research and Innovation)

Vol.17, No 1, April 2023 - (16-23)

Journal homepage: https://ejournal.brin.go.id/MIPI

Modeling Indonesian Motor Vehicle Tax Coefficients Based on Machine Learning Emission Data

Fitra Hidiyanto[1*], Kurnia Fajar Adhi Sukra[1], Rizqon Fajar[1], Nilam Sari Octaviani[1]
Dhani Avianto Sugeng[2]

[1]Research Center for Transportation Technology, National Research and Innovation Agency, Indonesia
[2]Research Center for Energy Conversion and Conservation, National Research and Innovation Agency, Indonesia
**Correspondence E-mail: fitr009@brin.go.id*

ABSTRACTS

This study utilized machine learning-based modeling to predict motor vehicle tax coefficients in Indonesia based on vehicle emission data. Three machine learning algorithms, namely Random Forest (RF), AdaBoost (AB), and Neural Network (NN), were employed to develop regression models for the tax coefficients. The research process involved data pre-processing, exploratory data analysis, feature ranking, and regression modeling. Model evaluation was performed using metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). The findings revealed that all three algorithms produced tax coefficient models for diesel vehicles with R^2 values approaching 1. Among them, NN achieved the highest R^2 value of 0.987, followed by RF with 0.986 and AB with 0.985. NN also performed the best in terms of MSE (0.023), RMSE (0.152), but MAE (0.076) achieved by RF for diesel vehicles. For gasoline vehicles, the NN algorithm yielded an R^2 value of 0.970, while RF and AB algorithms resulted in R^2 values of 0.969 and 0.946, respectively. NN also obtained the best MSE (0.086), RMSE (0.293), and MAE (0.122) values achieved by RF for gasoline vehicles. These results indicate that the tax coefficient models developed using RF, AB, and ANN algorithms effectively fit the measurement data. These models can support policymakers in formulating taxation regulations based on emission levels and vehicle fuel types, encouraging the adoption of environmentally friendly vehicles. Furthermore, they have the potential to reduce vehicle emissions and improve air quality through more effective taxation regulations.

ARTICLE INFO

Article History:
Received 21 Jan 2023
Revised 20 Feb 2023
Accepted 25 Mar 2023
Available online 28 Apr 2023

Keyword:
Carbon monoxide
Coefficient tax
Hydrocarbons
Machine learning
Neural network
Nitrogen oxides
Orange data mining
Vehicle emission

INTRODUCTION

The emissions from motor vehicles, including carbon monoxide (CO), hydrocarbons (HC), nitrogen oxides (NOx), and particulate matter, have caused negative impacts on human health and the environment. Carbon monoxide[1], hydrocarbons [2], and nitrogen oxides (NOx) [3][4] are hazardous substances that pollute the environment and have detrimental effects on the human body, while particulate matter consists of small particles emitted from vehicles and other sources. These substances can be deeply inhaled into the lungs and cause various health problems, including asthma, heart disease, and cancer [5][6].

***Corresponding Author | Fitra Hidiyanto|✉ fitr009@brin.go.id**

Table 1. Draft KLHK Regulation, classification basis for 4W (four-wheeled) vehicle emission tax rating

VEHICLE CATEGORY M, N, and O	Parameter				
CATEGORY	Made Year	CO (%)	HC (ppm)	Opacity (%HSU)	Test Method
Gasoline fuel	< 2007	4.5	1200		idle
	≥ 2007	1.5	200		idle
Diesel fuel					Free acceleration
GVW < 3.5 ton	< 2010			70	
	≥ 2010			40	
GVW > 3.5 ton	< 2010			70	
	≥ 2010			50	

Exhaust gas emissions testing is conducted to ensure that vehicles operate efficiently and produce fewer emissions. Furthermore, the emissions test results, including CO, hydrocarbons, NOx, and particulate matter, are utilized to calculate the applicable tax rates for vehicle owners. The collected data can then be fed into a machine learning model trained to predict the tax coefficients of vehicles based on their emission levels and fuel types.

The utilization of machine learning for prediction purposes in relation to motor vehicle emissions has been explored for various objectives. Some of these include using ML to predict vehicle CO_2 emissions, employing methods such as Lasso Regression, Multiple Linear Regression, XGBoost, Support Vector Regressor (SVR), Random Forest, and Ridge Regression [7], Gaussian regression has been employed for CO_2 emission analysis [8], and a comparison between ML methods and deep learning for predicting vehicle CO_2 emissions has been conducted [9]. Additionally, the prediction of vehicle emissions (CO, CO_2, HC, NOx) has been performed using Artificial Neural Networks (ANN) and Genetic Algorithm (GA), with ANN being widely used for engine emission modeling [10].

In this study, the determination of rating/coefficient tax classification is conducted by structuring the scope of motor vehicle emission values based on the latest draft of the Ministry of Environment and Forestry (KLHK) regulation, as presented in **Table 1**.

In this paper, the authors conducted modeling and prediction using motor vehicle emission data in Indonesia through machine learning techniques, specifically Adaboost, Random Forest, and Neural Network methods. They compared the performance of these models to assess the quality of the obtained results. The machine learning tools used in this study were provided by Orange data mining [11], which is a widely used application for data processing and facilitates the application of machine learning techniques [12].

Furthermore, the authors saved the machine learning models obtained from the training of existing data. These models can be utilized for predicting the tax coefficient based on new motor vehicle emission data. This would provide valuable insights for the Ministry of Environment and Forestry (KLHK) as a research partner in determining policies related to motor vehicle emission taxes.

In this study, the determination of rating/coefficient tax classification is conducted by structuring the scope of motor vehicle emission values based on the latest draft of the Ministry of Environment and Forestry (KLHK) regulation, as presented in **Table 1**. This environmental tax is applied in several countries in Asia and Australia region. In Singapore and Israel, the tax increase due to the emission of nitrogen oxides (NOx) and particulate matter (PM). In Japan, vehicle fuel efficiency determines the country's environmental performance tax. Singapore, on the other hand, has a payback rate according to the emission tiers of each vehicle [13].

METHODS

The data used in this research consists of emission test results from 4W motor vehicles in several major cities across Indonesia. The data includes carbon emissions, including CO, HC, CO_2, O2, opacity, year of manufacture, fuel type, vehicle category, engine displacement, combustion system, tonnage, mileage, and lambda. These data were obtained from the research partner, the Ministry of Environment and Forestry (KLHK), totaling 353,224 records.

Subsequently, the data was filtered and processed by the authors to be suitable for machine learning analysis. Not all the data points were utilized, resulting in the creation of two separate datasets: one for gasoline-fueled vehicles and the other for diesel-fueled vehicles. Based on existing regulations, such as Ministerial Regulation No. 5/2006 and the latest draft from KLHK, the vehicles were grouped according to their type, year of manufacture, and emission levels. These groupings served as the basis for determining the rating/coefficients for taxation purposes

using the vehicle emission data, for which the authors developed the machine learning models.

For the machine learning modeling, the authors employed CO, HC, and age as input variables, with the target variable being the emission rating for gasoline-fueled vehicles. In the case of diesel-fueled vehicles, opacity and year of manufacture were used as input variables, with the target being the emission rating.

Rating/Coefficient Values Determination Method

To determine the coefficients values based on the KLHK regulation in **Table 1**, this research establishes three rating groups for four-wheeled diesel vehicles (4W) and four emission groups for gasoline vehicles, consisting of CO and HC gases, as shown in **Table 2**. In this research, the coefficient values are determined based on the data obtained from KLHK, which includes the vehicle's manufacturing year, emission test year, engine capacity (CC), and opacity value for 80,102 diesel 4W vehicles. For gasoline 4W vehicles, the data consists of the vehicle's manufacturing year, emission test year, engine capacity (CC), CO value, and HC value, totaling 273,116 data points. As for gasoline K2 vehicles, the data includes the vehicle's manufacturing year, emission test year, engine capacity (CC), CO value, and HC value, with a total of 90,000 data points.

Machine Learning Methods

The stages in the author's machine learning method include data pre-processing, exploratory data analysis, feature ranking, regression modeling of coefficients/rating based on emissions, followed by prediction to determine the model's performance quality using parameters such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). Additionally, the obtained machine learning models are stored, and prediction testing is performed using the saved models to determine the coefficient values for the tested data. In this study, machine learning modeling is conducted using the Orange data mining tool developed by the Bioinformatics Laboratory at the University of Ljubljana in Slovenia [14].

Table 2. Rating Determination for (a) Gasoline and (b) Diesel Emissions in 4W (four-wheeled) Vehicles

GASOLINE.

YEAR	CO (%)				HC [ppm]			
	< 0.5	0.5-1	1 - 4	> 4	< 100	100 - 150	150 - 1000	> 1000
< 2007	1	2	3	4	1	2	3	4
2007 - 2018	0.5	1.5	2.5	3.5	0.5	1.5	2.5	3.5
> 2018	0	1	2	3	0	1	2	3

(a)

DIESEL

YEAR	OPACITY [%]			
	< 30	30 - 40	40 - 65	> 65
< 2010	1	2	3	4
2010 - 2021	0.5	1.5	2.5	3.5
> 2021	0	1	2	3

(b)

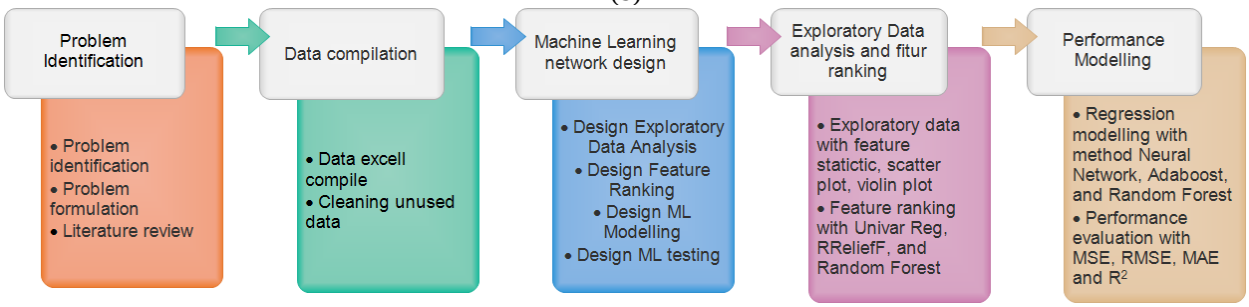

Figure 1. Machine learning method for coefficient/rating tax based on emission data.

Orange is an open-source software for data mining and machine learning, which includes visualization, modeling, and data analysis capabilities. In summary, the machine learning modeling method in this research for coefficients/ratings of emissions can be seen in **Figure 1.**

Data pre-processing

Data pre-processing in this study utilized the pre-processing widget to filter the input features and output targets, as well as to handle missing or unknown values [15]. Furthermore, outliers with values significantly deviating from their distribution were removed to avoid misinterpretation or measurement errors.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in data analysis. This method is conducted to discover patterns, anomalies, test hypotheses, and validate assumptions based on data summaries and statistical graphs [16]. In this research, EDA was performed using various widgets available in the "visualize" tab of Orange, including scatter plots, box plots, violin plots, mosaic displays, as well as feature statistics in the "data" tab and pie charts in the "educational" tab. The feature statistics widget [17] was utilized to observe data distributions, basic statistics such as minimum and maximum values, means, medians, and the presence of unknown values. The data used in this research description as in feature statistic can be seen as in Error! Reference source not found. and **Table 4**.

Feature Selection / Ranking

Feature selection in machine learning is used to reduce the number of input features and eliminate irrelevant or less influential inputs towards the output/target [16]. This is done to improve accuracy, reduce overfitting and training time. However, in this study, only a few features are used, so feature ranking is employed to determine the rank value of each feature with respect to the output/target, indicating their level of influence on the output/target value.

In this research, we applied feature ranking to both the diesel and gasoline datasets in order to identify the influential features in the model and determine their degree of impact. For this purpose, we utilized three feature ranking methods: Univariate Regression (Univar.Reg.), RReliefF, and Random Forest. The results of these methods can be observed in **Figure 2** and **Figure 3.** The graphs indicate that for diesel-fueled vehicles, the rating value is predominantly influenced by opacity, accounting for approximately 97.7% of the rating value, followed by age at 2.3% (RF). On the other hand, for gasoline-fueled vehicles, the rating value, which determines the tax amount, is primarily influenced by HC at around 79.7%, followed by CO at 17%, and age at 3.3% (RF).

Machine Learning Modeling (ML)

The ML regression modeling in this study utilizes three algorithms: Random Forest (RF), AdaBoost (AB), and Neural Network (NN). The models are trained using 100% of the data and then tested using two prediction methods: prediction on the full dataset using the prediction widget and cross-validation with five folds using the test and score widget [15][18]. The sequence of widgets used can be seen in **Figure 4.** The trained ML models are saved for future use in predicting new data.

Performance Evaluation

To evaluate the models, four regression scoring metrics are used: MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) [20], and R^2 (R-Squared) [20], [21], calculated using the following formulas:

$$MSE = \frac{1}{n}\sum(y - y\_pred)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n}\sum|\, y - y\_pred| \quad (3)$$

where n is the number of samples, y is the actual value, and y_pred is the predicted value.

$$R^2 = 1 - (SS\_res / SS\_tot) \quad (4)$$

$$= 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \bar{y}_i\right)^2} \quad (5)$$

where SS_res is the sum of squared residuals, and SS_tot is the total sum of squares.

All these evaluation scores are crucial for assessing the performance of regression models and selecting the appropriate model for the data.

Table 3. Description of Input Data Diesel Fuel

Name	Count	Mean	Median	Std	Minimum	Maximum
Opacity	80103	50.587	46	28.816	0	100
Vehic. Age	80103	7.76	6	6.528	0	94

Table 4. Description of Input Data Gasoline Fuel

Name	Count	Mean	Median	Std	Minimum	Maximum
CO	273117	0.531	0.03	1.510	0	10
HC	273117	89.748	20	260.387	0	10000
Vehic. Age	273117	5.3	4	5.252	0	71

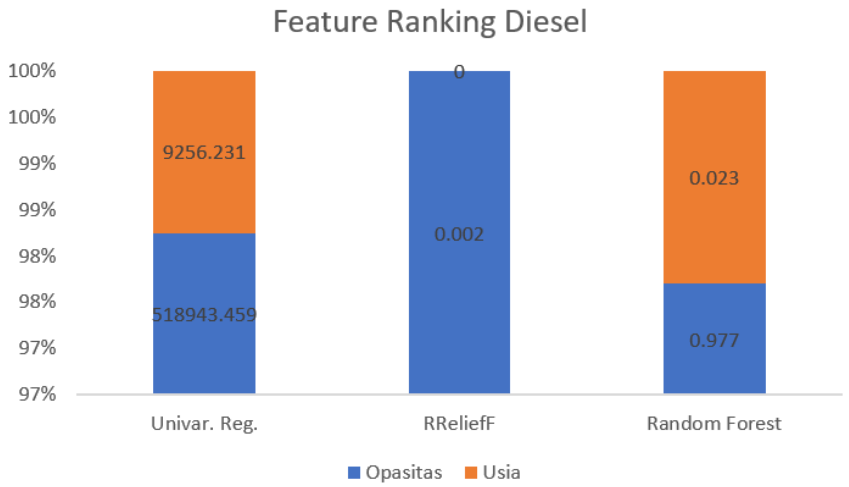

Figure 2. Feature ranking graph for Diesel Fuel.

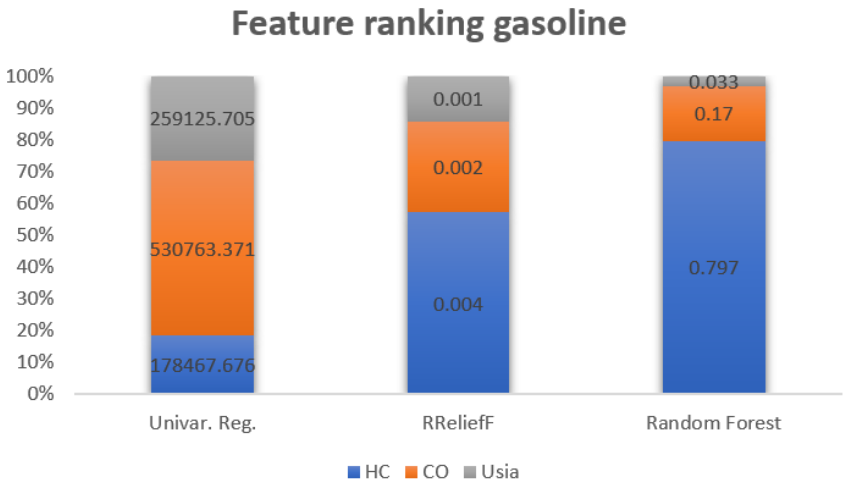

Figure 3. Feature ranking graph for Gasoline Fuel.

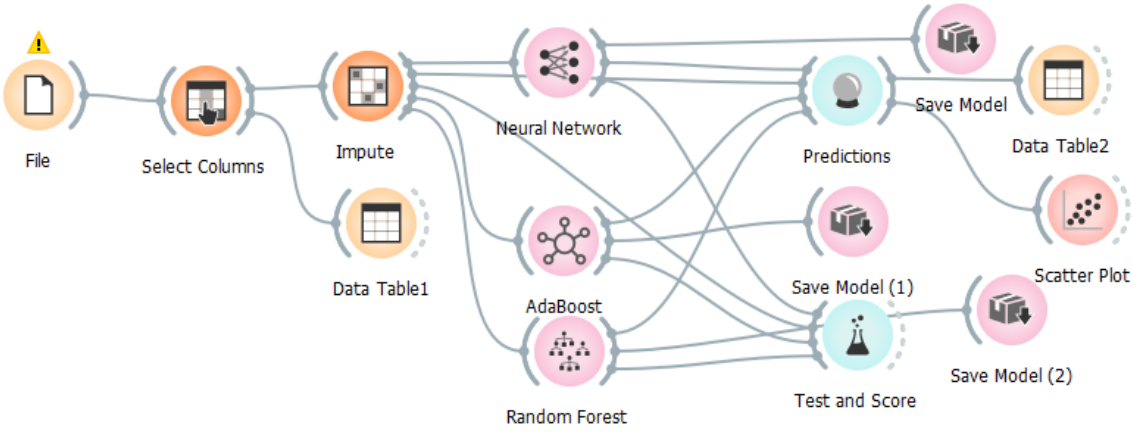

Figure 4. The sequence of widgets for modeling, evaluation, and saving machine learning models

RESULT AND DISCUSSION

Modeling the classification of motor vehicle exhaust gas emission test data is beneficial for the following reasons:

a. Assisting in determining the emission gas values according to the desired emission standards for policymakers based on emission reduction targets, implementation of environmentally friendly fuel and vehicle technologies, as well as tax collection.
b. Assisting in determining the factors and their respective influences on motor vehicle exhaust gas emissions, such as the type of fuel used, vehicle age, engine technology, and others.
c. Assisting in developing strategies to reduce the level of exhaust gas emissions.
d. Assisting in determining the remaining lifespan of tested motor vehicles based on emission test results and predefined thresholds.

The performance of the ML modeling for 4W vehicles is as follows:

Table 5 presents the modeling of diesel vehicle emission data using three different algorithms, where the performance metric used by the author is the R^2 score, which indicates the closeness to 1. The highest scores obtained were 0.987 (NN), 0.986 (RF), and 0.985 (AB), with an average R^2 score of 0.986 for diesel vehicles. It is evident that all three scores are highly competitive, with minimal differences between them.

On the other hand, **Table 6** displays the performance of ML predictions for gasoline-fueled vehicles, with the highest scores achieved being 0.97 (NN), 0.969 (RF), and 0.946 (AB). The average R^2 score for gasoline vehicles is 0.961, which is close to 1. Therefore, we can conclude that the model can effectively explain the data variation and is a good fit for the modeled data.

Figure 5 and **Figure 6** depict the comparison between the ML modeling results and the actual ratings, showing a high degree of similarity in the data with a regression coefficient of 0.99. Although there are still some errors in the data, they are tolerable due to their small magnitude.

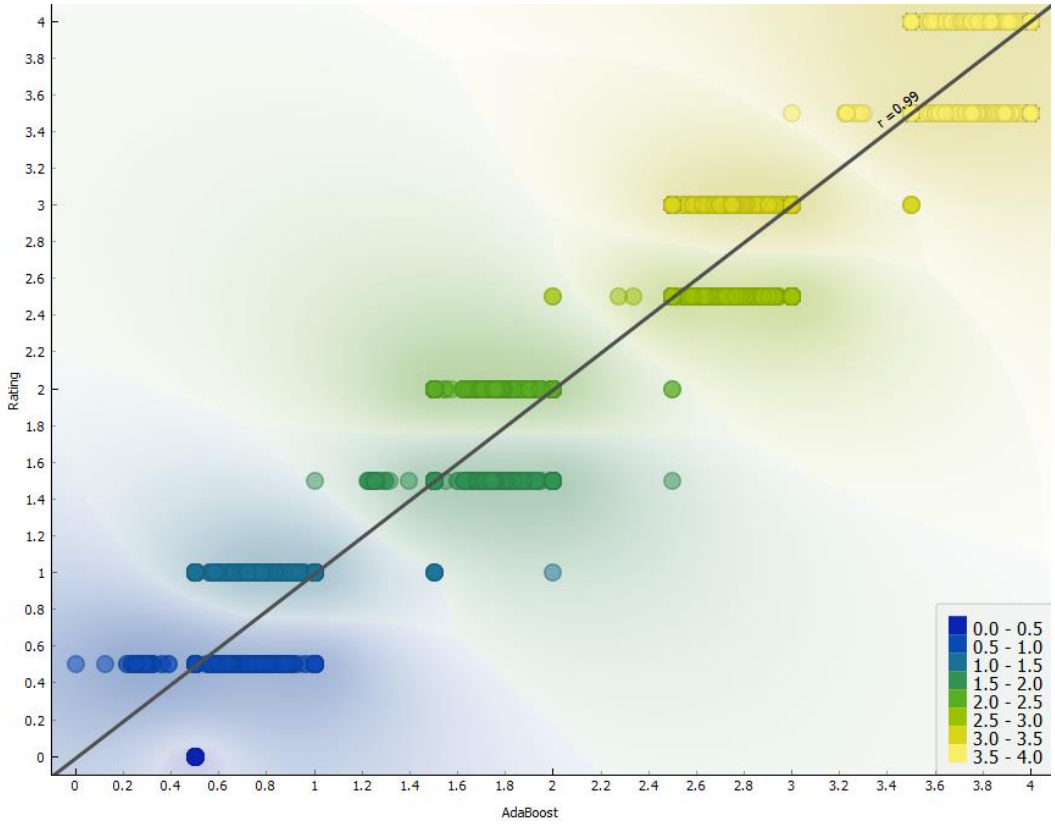

Figure 5. Comparison graph between AdaBoost ML model and diesel vehicle ratings.

Table 5. ML modeling results for diesel vehicles, Cross-validation method (folds = 5).

Model	MSE	RMSE	MAE	R^2
Random Forest	0.024	0.153	0.076	0.986
Neural Network	0.023	0.152	0.096	0.987
AdaBoost	0.027	0.164	0.087	0.985

Table 6. ML modeling results for gasoline vehicles, Cross-validation method (folds = 5).

Model	MSE	RMSE	MAE	R^2
Random Forest	0.088	0.297	0.122	0.969
Neural Network	0.086	0.293	0.144	0.97
AdaBoost	0.154	0.392	0.213	0.946

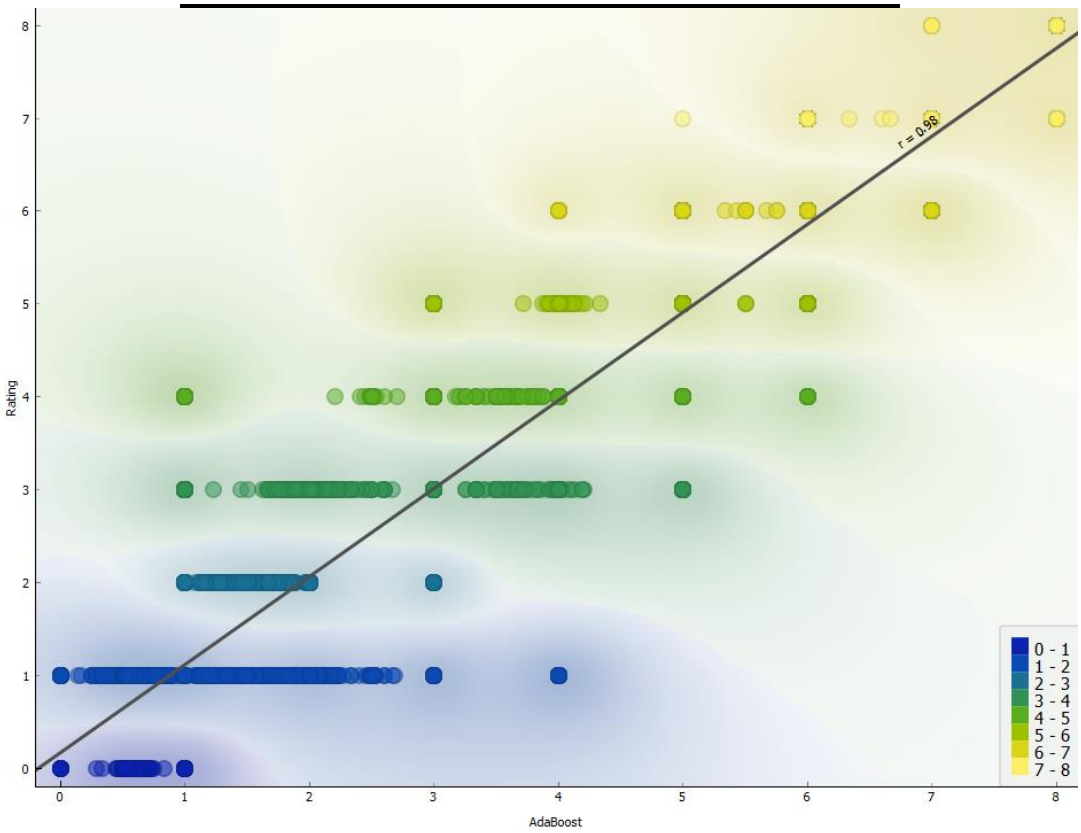

Figure 6. Comparison graph between AdaBoost ML model and gasoline vehicle ratings

CONCLUSION AND RECOMMENDATION

The conclusions drawn from this study are as follows:

- The modeling of motor vehicle tax coefficients based on motor vehicle emission data can be efficiently and effectively performed using Machine Learning (Artificial Intelligence) methods with open-source software (free)
- The classification modeling of motor vehicle emission data has resulted in a mapping between input and output parameters that can be utilized by policymakers for emission reduction programs, setting emission thresholds, and imposing taxes to finance environmental damages.
- Tax coefficient modeling can be performed using various available algorithms according to the required accuracy. In this study, Neural Network, Adaboost, and Random Forest algorithms were used, and the highest performance value (R^2) was obtained with Neural Network for diesel-fueled vehicles with R^2 = 0.987 and for gasoline-fueled vehicles with R^2 = 0.982. However, the average R^2 for diesel vehicles was 0.986, and for gasoline vehicles, the average R^2 was 0.961. These scores approach 1, indicating a strong correlation between input and output variables, resulting in small errors. Therefore, it can be concluded that the predictions from the machine learning model are highly accurate, and the model can explain the variation in target data with respect to the input, making it suitable for the modeled data.
- Tax coefficient modeling can be used to simulate the additional tax amount for vehicle owners based on the emitted exhaust gases in an objective and fair manner. It is objective because it relies on real-time emission measurements and government policies that consider the social and economic conditions of the surrounding community.

The recommendations derived from this study are as follows :

- It is necessary to conduct a simulation and validation of the generated tax coefficient model before

implementation to ensure its reliability and acceptance by the community.
- The results of the modeling need to be disseminated to stakeholders such as regulators, academics, associations, and motor vehicle users.
- There is a need for regular updates of emission test data and IKU values to improve the accuracy of the machine learning tax coefficient model in predictions.

Author Contributions

The First three Authors have contributed equally to this work, while the last two authors are a member.

REFERENCES

[1] M. A. Rizaldi, R. Azizah, M. T. Latif, L. Sulistyorini, and B. P. Salindra, "Literature Review: Dampak Paparan Gas Karbon Monoksida Terhadap Kesehatan Masyarakat yang Rentan dan Berisiko Tinggi," *J. Kesehat. Lingkung. Indones.*, vol. 21, no. 3, pp. 253–265, 2022, doi: 10.14710/jkli.21.3.253-265.

[2] M. Ince and O. K. Ince, "Introductory Chapter: Sources, Health Impact, and Environment Effect of Hydrocarbons," M. Ince and O. K. Ince, Eds. Rijeka: IntechOpen, 2019, p. Ch. 1. doi: 10.5772/intechopen.89039.

[3] W. de Vries, "Impacts of nitrogen emissions on ecosystems and human health: A mini review," *Curr. Opin. Environ. Sci. Heal.*, vol. 21, no. x, p. 100249, 2021, doi: 10.1016/j.coesh.2021.100249.

[4] S. Shaw and B. Van Heyst, "Nitrogen Oxide (NOx) emissions as an indicator for sustainability," *Environ. Sustain. Indic.*, vol. 15, no. x, p. 100188, 2022, doi: 10.1016/j.indic.2022.100188.

[5] K. H. Kim, E. Kabir, and S. Kabir, "A review on the human health impact of airborne particulate matter," *Environ. Int.*, vol. 74, pp. 136–143, Jan. 2015, doi: 10.1016/J.ENVINT.2014.10.005.

[6] R. B. Hamanaka and G. M. Mutlu, "Particulate Matter Air Pollution: Effects on the Cardiovascular System," *Front. Endocrinol. (Lausanne).*, vol. 9, no. November, pp. 1–15, 2018, doi: 10.3389/fendo.2018.00680.

[7] A. S. Chadha, Y. Shinde, N. Sharma, and P. K. De, "Predicting CO2 Emissions by Vehicles Using Machine Learning BT - Data Management, Analytics and Innovation," 2023, pp. 197–207. doi: https://doi.org/10.1007/978-981-19-2600-6_14.

[8] N. Ma, W. Y. Shum, T. Han, and F. Lai, "Can Machine Learning be Applied to Carbon Emissions Analysis : An Application to the CO 2 Emissions Analysis Using Gaussian Process Regression," vol. 9, no. September, pp. 1–8, 2021, doi: 10.3389/fenrg.2021.756311.

[9] S. Shah, S. Thakar, K. Jain, B. Shah, and S. Dhage, "A Comparative Study of Machine Learning and Deep Learning Techniques for Prediction of CO2 Emission in Cars," *Lect. Notes Networks Syst.*, vol. 587, pp. 749–758, 2023, doi: 10.1007/978-981-19-7874-6_55/COVER.

[10] S. Khurana, S. Saxena, S. Jain, and A. Dixit, "Materials Today : Proceedings Predictive modeling of engine emissions using machine learning : A review," *Mater. Today Proc.*, vol. 38, pp. 280–284, 2021, doi: 10.1016/j.matpr.2020.07.204.

[11] I. Popchev and D. Orozova, "Algorithms for Machine Learning with Orange System," *Int. J. online Biomed. Eng.*, vol. 19, no. 4, pp. 109–123, 2023, doi: 10.3991/ijoe.v19i04.36897.

[12] U. Thange, V. K. Shukla, R. Punhani, and ..., "Analyzing COVID-19 Dataset through Data Mining Tool 'Orange,'" *2021 2nd Int. …*, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9357754/

[13] V. O'Riordan, F. Rogan, B. Ó'Gallachóir, and H. Daly, "Impact of an emissions-based car tax policy on CO2 emissions and tax revenue from private cars in Ireland," *Int. J. Sustain. Transp.*, vol. 0, no. 0, pp. 1–13, 2022, doi: 10.1080/15568318.2022.2132562.

[14] Z. B. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, "Orange: Data Mining Toolbox in Python," *J. Mach. Learn. Res.*, vol. 14, p. 2349−2353, 2013, [Online]. Available: https://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf

[15] A. Ishak, K. Siregar, Asfriyati, R. Ginting, and M. Afif, "Orange Software Usage in Data Mining Classification Method on the Dataset Lenses," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1003, no. 1, 2020, doi: 10.1088/1757-899X/1003/1/012113.

[16] F. Hidiyanto, S. Leksono, Sigit Tri Atmaja, and R. Fajar, "Data Exploratory Analysis and Feature Selection of Low-Speed Wind Tunnel Data for Predicting Force and Moment of Aircraft," *Maj. Ilm. Pengkaj. Ind.*, vol. 16, no. 2, pp. 87–94, 2022, doi: 10.29122/mipi.v16i2.5285.

[17] L. Irawan, L. H. Hasibuan, and F. Fauzi, "Analisa Prediksi Efek Kerusakan Gempa Dari Magnitudo (Skala Richter) Dengan Metode Algoritma Id3 Menggunakan Aplikasi Data Mining Orange," *J. Teknol. Inf. …*, 2020, [Online]. Available: http://e-journal.upr.ac.id/index.php/JTI/article/view/1079

[18] I. Indriyanti, N. Ichsan, H. Fatah, T. Wahyuni, and...,"IMPLEMENTASI ORANGE DATA MINING UNTUK PREDIKSI HARGA BITCOIN," *… Ris. Sains dan …*, 2022, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti/article/view/762