City and Environment Interactions 20 (2023) 100127

Contents lists available at ScienceDirect

City and Environment Interactions

journal homepage: www.sciencedirect.com/journal/city-and-environment-interactions

Towards adequate policy enhancement: An AI-driven decision tree model for efficient recognition and classification of EPA status via multi-emission parameters

Adeboye Awomuti [a,b], Philip Kofi Alimo [c], George Lartey-Young [b], Stephen Agyeman [d], Tosin Yinka Akintunde [e], Adebobola Ololade Agbeja [f], Olayinka Oderinde [g,*], Oluwarotimi Williams Samuel [h,i,*], Henry Otobrise [j]

[a] *State Key Laboratory of Pollution Control and Resource Reuse, Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China*
[b] *UNEP Tongji Institute of Environment for Sustainable Development, Tongji University, Shanghai 200092, China*
[c] *College of Transportation Engineering, Tongji University, 4800 Cao'an Road, Shanghai, China*
[d] *Department of Civil Engineering, Sunyani Technical University, Sunyani, Ghana*
[e] *Department of Social Work, The Chinese University of Hong Kong, Sha Tin, Hong Kong*
[f] *Department of Sustainable Forest Management, Forestry Research Institute of Nigeria, PMB 5054, Ibadan, Nigeria*
[g] *Department of Chemistry, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Nigeria*
[h] *School of Computing, University of Derby, Derby DE22 3AW, United Kingdom*
[i] *Data Science Research Center, University of Derby, Derby DE22 3AW, United Kingdom*
[j] *Department of Physics, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Nigeria*

ARTICLE INFO

Keywords:
Decision tree
Artificial intelligence
EPA status
Air emission parameters
Machine learning
Emission reduction

ABSTRACT

Accurate and timely evaluation and assessment of emission data and its impact on environmental status has been a key challenge due to the conventional manual approach utilized for independently computing most emission parameters. To resolve this long-standing issue, we proposed an Artificial Intelligence (AI)-driven Decision Tree model to adequately classify Environmental Protection Agency (EPA) status based on multiple Emission Parameters. The model's performance was systematically evaluated using multiple emission parameters obtained from a two-stroke motorcycle dataset collected in Nigeria across various metrics such as K-S Statistics, Confusion Matrix, Correlation Heat Map, Decision Tree, Validation Curve, and Threshold Plot. The K-S Statistics plot's experimental results showed a considerable correlation between HC, CO, and the target variable, with values ranging from 0.75 to 0.80. At the same time, CO_2 and O_2 do not correlate with the target variable with values between 0.00 and 0.09. The Confusion Matrix revealed that the proposed model has an overall accuracy of 99.9% with 481 true positive predictions and 75 true negative predictions, indicating the effectiveness of the proposed AI-driven model. In conclusion, our proposed AI-driven model can effectively classify EPA status based on multiple emission parameters with high accuracy, which may spur positive advancement in policy enhancement for proper environmental management.

1. Introduction

The increasing global demand for transportation has significantly increased the number of vehicles [1,2]. This increase has had a negative impact on the environment, as the emissions from these vehicles contribute to air pollution and climate change [3,4]. In particular, the use of motorcycles has grown significantly in recent years, especially in developing countries where they are often the preferred mode for last-mile transportation [5,6]. As a result, it is crucial to find ways to minimize the environmental impact of emissions from motorcycles [5,7,8].

One approach to addressing motorcycle emissions is through the use of regulatory policies [9,10]. The Environmental Protection Agency

* Corresponding authors at: Department of Chemistry, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Nigeria (O. Oderinde). School of Computing, University of Derby, Derby, DE22 3AW, United Kingdom (O. W. Samuel).
E-mail addresses: oderinde.olayinka@lcu.edu.ng, yinkaoderinde@yahoo.com (O. Oderinde), o.samuel@derby.ac.uk, timitex92@gmail.com (O.W. Samuel).

https://doi.org/10.1016/j.cacint.2023.100127
Received 12 July 2023; Received in revised form 20 October 2023; Accepted 15 November 2023
Available online 19 November 2023

(EPA) is a government agency that sets standards for the emissions of vehicles [11,12]. The EPA issues status to each motorcycle model based on its emissions performance, which can range from “not certified” to “certified.” [13]. These EPA status labels can significantly impact the marketability and sales of motorcycles [14].

However, determining the EPA status of a motorcycle can be a complex and time-consuming process that could lower efficiency, as it requires considering a range of emission parameters such as carbon monoxide (CO), hydrocarbons (HC), and nitrogen oxides (NOx) [15]. In addition, the EPA updates its standards and testing procedures periodically, making it challenging to keep track of the latest requirements [16].

To address these challenges, we propose the use of a decision tree model driven by artificial intelligence (AI) technique to adequately recognize and classify EPA status based on motorcycle emission parameters. Decision tree models are a popular choice for classification tasks because they can handle multiple input variables and provide clear explanations of the decision-making process [17,18]. By using an AI-driven model to classify EPA status, we aim to improve the process’s accuracy and efficiency and reduce or eliminate human error. We used a dataset of motorcycle emission test results to train and validate the model. Our objective is to achieve high prediction accuracy, meaning that the model should be able to accurately predict the EPA status of a motorcycle or similar machines based on its emission parameters.

As with any technology, the use of artificial intelligence (AI) in regulatory decision-making carries with it specific ethical and societal implications that must be carefully considered [19]. In the context of the proposed decision tree model for classifying facilities based on emission parameters and determining their EPA statuses, several key ethical and societal issues merit further examination [20,21].

One concern to consider is the potential for the AI model to make decisions that are not transparent or explainable [22,23]. While the AI-driven decision tree model is designed to make decisions based on a clear set of rules, it may be difficult for humans to understand the exact logic behind the model’s predictions [24,25]. This lack of transparency could make it challenging for regulators and the public to understand and accept the decisions made by the AI-driven model [26,27].

From a societal perspective, the use of AI-based models in regulatory decision-making may raise concerns about the potential loss of jobs and the impact on employment [28,29]. While the AI-driven decision tree model could potentially streamline and automate the EPA classification process, it could also lead to the displacement of human workers who currently perform this task [30,31]. It will be necessary to carefully consider the potential impacts on employment and explore strategies to mitigate any negative consequences [32].

The use of AI in regulatory decision-making, such as the proposed decision tree AI model for EPA classification, carries with it several ethical and societal implications that should be carefully considered. It will be important to address these issues to ensure this technology’s responsible and fair implementation [31,33].

To evaluate the performance of our model, we used a range of evaluation metrics, including accuracy, precision, and recall [34]. We also compared the performance of our model to other machine learning algorithms. In addition, we also examined the potential benefits and limitations of using AI for the classification of EPA status [28]. This includes a discussion of the ethical and societal implications of using AI in regulatory decision-making [35].

Consequently, the goal of this study is to demonstrate the effectiveness of using an AI-driven decision tree model for the recognition and classification of EPA status based on motorcycle emission parameters. The contributions of this study are threefold.

- First, this study adds to the growing literature on mitigating the negative impact of two-stroke motorcycle emissions in developing countries.
- Second, it proposes a decision tree model to classify EPA status based on motorcycle emission parameters. The proposed model has more prediction accuracy and is less time-consuming compared to conventional methods.
- Third, with a more accurate and efficient method for determining EPA status, this study contributes to developing more effective regulatory policies for reducing the environmental impact of transport emissions.
- The Proposed model can be deployed widely to classify similar emissions from sources other than motorcycles and may spur positive advancement in policy enhancement for proper environmental management.

The remaining part of the paper is structured as follows. Section 2 details the study area, materials, methods used in this study, the modeling framework, and the limitations. Section 3 has the results and discussion, which also details the explicability and interpretability of the proposed decision tree model. Section 4 has the conclusion, which explains the contribution of this study to the field of artificial intelligence and environmental management.

2. Materials and methodology

This section details the study area in Africa and how the data was collected and processed. It further entails a detailed report of the model development framework, the policy analysis, and the limitations.

2.1. Study area and sampling locations

The study area is Ogun State, one of Nigeria’s thirty-six states in the southwest part of the nation [36]. It shares borders with Lagos State (the commercial nerve center of the country) and the Atlantic Ocean on the south, Oyo and Osun States on the north, Ondo State on the east, and the Republic of Benin on the west. Ogun State is home to the highest number of industries in Nigeria. It has the longest stretch of road connecting Lagos to other parts of the country [37,38]. The sampling locations within the state are shown in (Fig. 1) [39,40].

2.2. Data collection

Raw exhaust emissions from 1950 motorcycle taxis were sampled using a hand-held KANE Automotive 4-Gas Analyzer (Model 4–2). The instrument was programmed to detect and measure carbon dioxide (CO2), with an accuracy of ± 0.5 %, volume reading at a resolution of 0.1 %, at a range of 0–16 % and over-range of 25 %, oxygen (O2), with an accuracy of ± 0.1 %, volume at a resolution of 0.01 %, at a range of 0–21 % and over-range of 25 %), hydrocarbons, HC (with an accuracy of 12 ppm volume at a range of 0–5,000 ppm and an over-range of 10,000 ppm), carbon monoxide, CO (with an accuracy of ± 0.06 %, volume at a range of 0–10 % and over-range of 20 %) and Lambda, *“λ”* (at a resolution of 0.001 and a range of 0.8–1.2), which was calculated using (Equation 1) [41–43].

$$\lambda=\frac{[CO_2]+\left[\frac{CO}{2}\right]+[O_2]+\left[\frac{NO}{2}\right]+\left(\left(\frac{HCV}{4}\times\frac{3.5}{3.5+\frac{[CO]}{[CO_2]}}\right)-\frac{OCV}{2}\right)\times([CO_2]+[CO.])}{\left(1+\frac{H_{CV}}{4}-\frac{O_{CV}}{2}\right)\times([CO_2]+[CO]+(K_1\times[HC.]))} \tag{1}$$

CO, CO2, and O2 are measured in percentage volume (% vol.), and HC is measured in parts per million volume (ppm vol.). *K.″1″* is the HC conversion factor expressed in parts per million volume equivalent of normal hexane (C6H14). The value is given as 6.0 × 10–4 according to Eqn. 1. *H.″CV″* denotes the hydrogen-carbon atomic ratio of the fuel

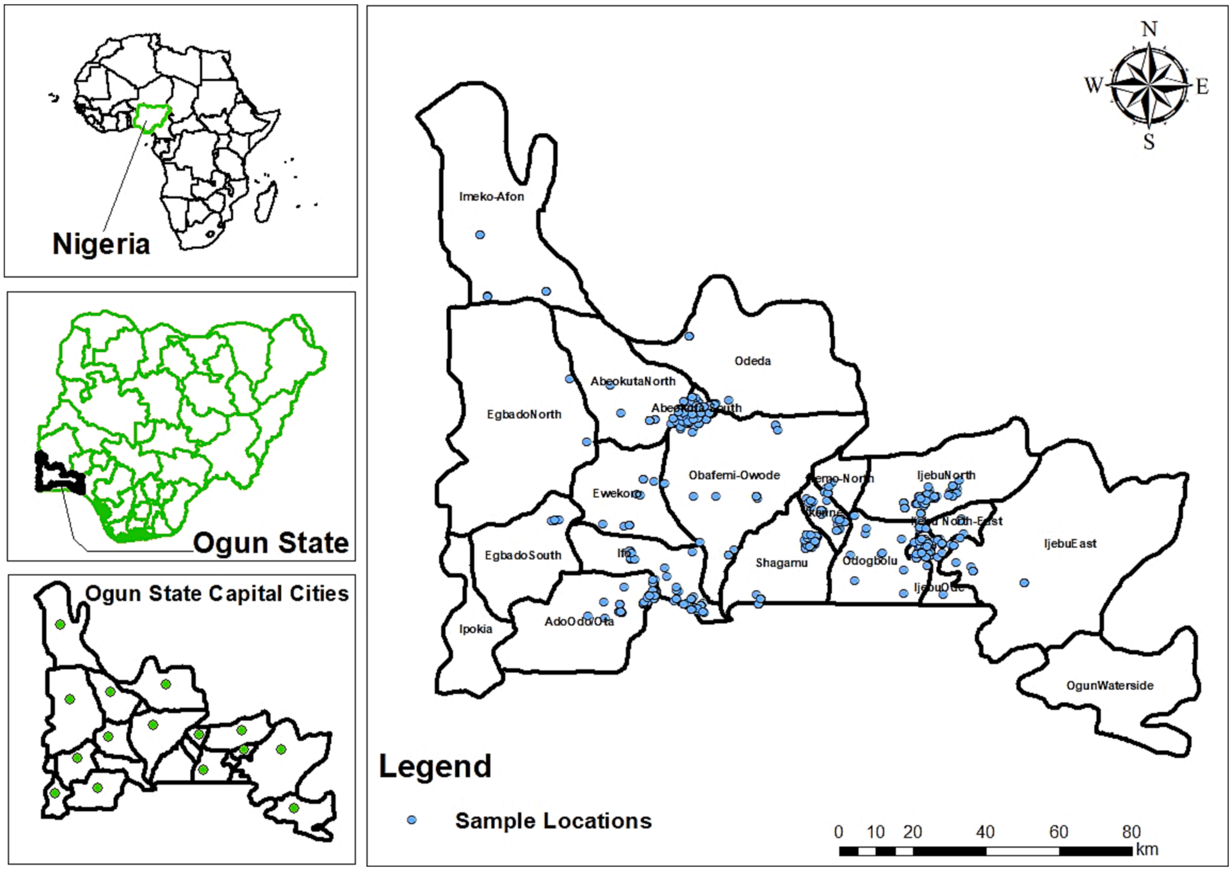

Fig. 1. Study area and sampling locations. Source: Author.

(minimal value is 1.7261), and *O."CV"* denotes the oxygen-carbon atomic ratio of the fuel (minimal value is 0.0176)" [38].

Before each measurement round, the motorcycle taxis were allowed to travel a distance of 50 m from their stations, and the 'No–Load Short Test', commonly referred to as 'idle mode tests', was performed on each motorcycle taxi. The idle mode test approach has been recently reported in similar studies as effective in collecting emission data since motorcycles are not required to move at constant load, mimicking stationary equipment [38,44]. The exhaust probe of the sampling instrument was inserted into the motorcycle's exhaust pipe end and clamped to the tail end to avoid falling off. Measurements were recorded in (%) volume for CO2, CO, and O2 concentrations and parts per million (ppm) for HC. Each round of measurement lasted 10 min. All recorded data is automatically stored in the instrument's memory drive for later download. After each round of measurements, the sampling analyzer was calibrated to 'zero' by exposing the probes to ambient conditions while ensuring that the exhaust probe tips were clean of any dirt or debris. All samples and testing events were undertaken in November 2020-February 2021, coinciding with Nigeria's dry season. Therefore, during all testing, the air temperature was between 31 and 40 °C, and the relative humidity was between 45 and 60 %. Sampling events were conducted in triplicate for each motorcycle taxi within the sampling period to determine statistical variations in the datasets.

2.3. Two-stroke motorcycles' selection criteria

The selection criteria for the two-stroke motorcycles was primarily based on the popularity of commercial motorcycles equipped with engines ranging from 100 to 120 cc. These motorcycles were sourced from brands such as Suzuki, Jincheng, Lifan, and Qlink, bearing model inscriptions like 100, 120, A100, B120, etc. The study area for selection was Ogun State, Nigeria, which shares its borders with Lagos State, Nigeria's most populous and commercially significant state. Ogun State is also known as the most industrialized state in the country, in addition to experiencing a surge in population due to its proximity to Lagos State and other states from the northern and eastern parts of the country.

The choice of these motorcycles can be attributed to their cost-effectiveness and robustness on challenging roads and terrains. They are known for their gasoline-lubrication method, involving a pre-mixture of engine oil with gasoline. In this system, adequate lubrication of the cylinder wall is crucial. The oil, reaching the combustion chamber through the cylinder, undergoes combustion alongside the fuel. However, the oil-scraper ring's limitations, where the oil in the crankcase lubricates the cylinder, result in an oil shortage. As the oil in the crankcase is allocated for lubricating all reciprocating engine parts and a portion of the cylinder, the oil, burning in conjunction with gasoline, leads to a higher concentration of exhaust gas pollutants released from the combustion chamber.

Moreover, the design of these selected motorcycles, combined with the tendency of a significant percentage of commercial motorcycle users to purchase adulterated engine oil from roadside vendors and opt for substandard or sometimes foreign fairly-used spare parts due to the elevated costs of acquiring standard replacements, amplifies the release of pollutants during the combustion process.

2.4. Model development

To develop and evaluate an AI-driven decision tree model for the recognition and classification of EPA status based on motorcycle emission parameters, we collected a dataset of motorcycle emission test results from 20 local governments in Ogun State, Nigeria. The dataset consists of recordings of emission parameters from various motorcycle models that include CO, HC, CO_2, and O_2 that served as input to the machine learning model. Besides, the EPA assigns a compliance status to each motorcycle model based on its adherence to emissions regulations, which can be expressed as a "Pass" or "Fail" grade. The process followed is shown in Fig. 2.

To prepare the dataset for the proposed model development and deployment, we performed several preprocessing steps Using Python 3.9 on Google Collab and employing several Python libraries, including Numpy, Pandas, PyTorch, and Matplotlib. First, we removed any missing or incomplete records from the dataset. It should be noted that we also transformed the emission parameters into a standardized scale using the min–max normalization method. This process scales the emission parameters to a range of 0 to 1, with 0 representing the minimum value in the dataset and 1 representing the maximum value [45,46]. This is a common preprocessing step in machine learning as it helps improve the model's performance via preservation of the distribution of the characteristics in the original data to a great extent [47]. To deal with outliers, we used the 3-sigma rule statistical technique to remove any data points more than three standard deviations away from the mean. This helps to minimize errors and remove data with extreme values.

Next, we split the dataset into training and testing sets. 80 % of the data was used to train the model, while the remaining 20 % was used for testing and validation [48]. The training set was used to train the decision tree model. In contrast, the testing set was used to evaluate the model's performance [49]. We used stratified sampling to ensure that the training and testing sets were representative of the overall dataset, with a similar distribution of EPA status labels.

To develop the decision tree model, we used the scikit-learn library in Python [50–52]. We selected the decision tree algorithm from the library and specified the parameters for the model. In particular, we set the maximum depth of the tree to 10 and used the Gini criterion for splitting the nodes. We also used 10-sample stratified k-fold cross-validation method to evaluate the model's performance during training, which was also evident in the performance of our validation curve (Fig. 6).

To evaluate the performance of the decision tree model, we used a range of evaluation metrics, including accuracy, precision, and recall [53]. Accuracy measures the overall percentage of correct predictions made by the model. In contrast, precision measures the percentage of true positive predictions out of all positive predictions [18,54]. Recall measures the percentage of true positive predictions out of all actual positive cases [55–57].

In addition to evaluating the decision tree model, we compared the performance of the model to other benchmark machine learning algorithms such as Extreme Gradient Boosting (XGB) and Ada Boost (Table 1). To do this, we trained and evaluated these algorithms using the same dataset and evaluation metrics as the decision tree model. The decision Tree model was considered to have the potential for practical deployment of AI-driven solutions for EPA status classification due to its relatively lower computation time, leading to a faster classification output.

2.5. Policy analysis and limitations

We conducted a literature review of relevant studies to examine the potential benefits and limitations of using AI to classify EPA status [58,59]. We also considered the ethical and societal implications of using AI in regulatory decision-making, including potential biases in the dataset and the impact on stakeholders such as motorcycle manufacturers and consumers [28,60].

Consequently, our methodology for this study consisted of collecting a dataset of motorcycle emission test results, preprocessing the data, developing and evaluating a decision tree AI model, and examining the potential benefits and limitations of using AI to classify EPA status. Our objective was to achieve high prediction accuracy with the decision tree model, meaning that it should be able to accurately predict the EPA status of a motorcycle based on its emission parameters.

2.6. Real-life application of the model

To test the performance of our model on real-world data, we used a dataset of BMW eDrive vehicles [61]. The dataset contains data on vehicle speed, HC tailpipe emissions, CO tailpipe emissions, CO_2 tailpipe emissions, and O_2 tailpipe emissions. We used a 10-sample cross-validation approach to evaluate the performance of the model. The model achieved an accuracy of 100 %. This suggests that our model can be used to accurately classify the emission status of real-world vehicles. We compared the decision tree model to other state-of-the-art AI models, such as XGBoost and AdaBoost. We found that the model achieved similar or better performance than these models on our BMW eDrive dataset. This suggests that our decision tree model is competitive for classifying emission status based on emission parameters.

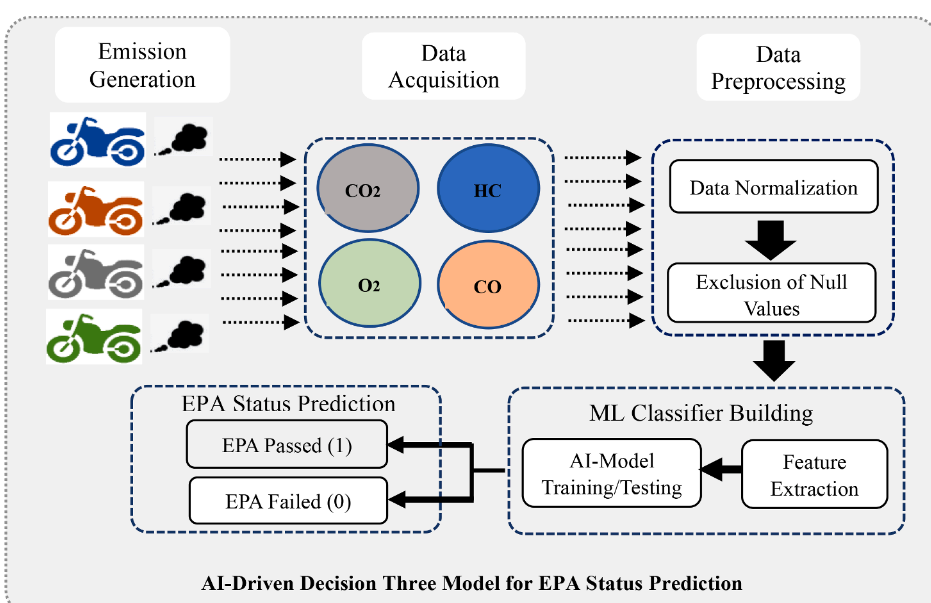

Fig. 2. A conceptualized framework of the proposed AI-driven decision three model for EPA status.

Table 1
Comparison between different models with key metric evaluations.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Gradient Boosting Classifier	0.9992	0.9999	0.9991	1	0.9996	0.9968	0.9969	0.225
Decision Tree Classifier	0.9985	0.9991	0.9982	1	0.9991	0.9935	0.9936	0.063
Random Forest Classifier	0.9977	1	0.9973	1	0.9987	0.9904	0.9905	0.433
Ada Boost Classifier	0.9977	0.9985	0.9973	1	0.9987	0.9904	0.9905	0.235
Light Gradient Boosting Machine	0.9954	0.9998	0.9964	0.9982	0.9973	0.9803	0.9805	0.663
Extreme Gradient Boosting	0.9946	0.9999	0.9964	0.9974	0.9969	0.9767	0.9771	0.059
Extra Trees Classifier	0.99	0.999	0.992	0.9965	0.9942	0.958	0.9586	0.188
K Neighbors Classifier	0.9676	0.9658	0.9759	0.9866	0.9811	0.8663	0.8688	0.049
Logistic Regression	0.9175	0.9758	0.917	0.9865	0.9503	0.7086	0.7262	0.441
Ridge Classifier	0.9167	0	0.9125	0.9902	0.9496	0.7118	0.7327	0.037
Linear Discriminant Analysis	0.9167	0.9764	0.9125	0.9902	0.9496	0.7118	0.7327	0.033
Naive Bayes	0.8982	0.9702	0.8902	0.9909	0.9376	0.6647	0.6948	0.068
Quadratic Discriminant Analysis	0.8734	0.9571	0.8652	0.9866	0.9216	0.6002	0.6373	0.058
SVM - Linear Kernel	0.6039	0	0.5786	0.9519	0.6266	0.2825	0.3388	0.043
Dummy Classifier	0.1358	0.5	0	0	0	0	0	0.049

3. Results and discussion

In this section, we discussed the results of our AI-driven decision tree classifier model based on standard model evaluation metrics (SMEM).

The decision tree classifier was chosen over other models on the model comparison list (Table 1) due to its simplicity and ease of interpretation, computational efficiency, and robustness to outliers and missing data. It requires less time to complete a sample than other models, which means that it can process large amounts of data quickly and efficiently. This is particularly useful when working with large datasets or when the model needs to be retrained frequently. The model shows improvement after initial fine-tuning, as seen in Table 2 (initial model training) and Table 3 (result after model fine-tuning).

The decision tree classifier model achieved an accuracy of 0.9992, an area under the receiver operating characteristic (ROC) curve (AUC) of 0.9996, a recall of 0.9991, a precision of 1, an F1 score of 0.9996, Kappa of 0.9968 and MCC of 0.9969 (Table 3). These results indicate that the model has a high level of accuracy in correctly predicting the target variable, the EPA emission status, and can distinguish between the positive and negative classes with a high degree of accuracy. The high precision and recall scores further indicate that the model can identify the most relevant cases while maintaining high accuracy. The Kappa and

Table 2
Initial model training before fine-tuning.

Sample	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1
3	0.9923	0.9955	0.9911	1	0.9955	0.9685	0.969
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1
8	0.9922	0.9955	0.9911	1	0.9955	0.9669	0.9675
9	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1
Mean	0.9985	0.9991	0.9982	1	0.9991	0.9935	0.9936
Std	0.0031	0.0018	0.0036	0	0.0018	0.0129	0.0127

Table 3
Key metric results after model fine-tuning.

Sample	**Accuracy**	**AUC**	**Recall**	**Prec.**	**F1**	**Kappa**	**MCC**
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1
3	0.9923	0.9955	0.9911	1	0.9955	0.9685	0.969
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1
Mean	0.9992	0.9996	0.9991	1	0.9996	0.9968	0.9969
Std	0.0023	0.0013	0.0027	0	0.0013	0.0095	0.0093

MCC scores also indicate that the model almost perfectly agrees with the human annotator. These results demonstrate that the decision tree model is a highly accurate and well-performing model for the classification of EPA emission status and, thus, can also be deployed in similar use-case scenarios.

3.1. Model confusion matrix

The confusion matrix (CM) analysis (Fig. 3) is another performance metric used to evaluate the model. CM is a table that shows the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model.

The confusion matrix presents the following values: TP, Passed/Passed = 481, TN, Failed/Failed = 75, FP, Failed/Passed = 0, and FN, Passed/Failed = 0.

True Positives (TP) or Passed/Passed represent the number of cases where the model correctly predicted that the EPA status is passed. In this case, the model correctly predicted 481 cases to be Passed.

True Negatives (TN) or Failed/Failed represent the number of cases where the model correctly predicted that the EPA status is failed. In this case, the model correctly predicted 75 cases to be Failed.

False Positives (FP) or Failed/Passed represent the number of cases where the model incorrectly predicted that the EPA status is passed. In this case, the model made no incorrect predictions of Failed as a Passed.

False Negatives (FN) or Passed/Failed represent the number of cases where the model incorrectly predicted that the EPA status is failed. In this case, the model does not make any incorrect predictions.

The confusion matrix results indicate that the decision tree model has a high degree of accuracy in classifying the target EPA status. The low number of False Positives and False Negatives values suggests that the model can correctly identify the majority of the cases. It is also worth mentioning that the model has a high degree of specificity, as it correctly identifies a high percentage of Failed cases. Additionally, the model has a high degree of sensitivity, as it correctly identifies a high percentage of Passed cases.

The confusion matrix results indicate that the decision tree model can effectively classify EPA status based on emission parameters with a high degree of prediction accuracy. The model has a good balance between true positive and true negative predictions, which indicates a well-trained model.

3.2. Kolmogorov-Smirnov (KS) Statistic

In addition to the evaluation of the model's performance using precision, recall, and threshold values, the model's performance was also evaluated using the Kolmogorov-Smirnov (KS) Statistic (Fig. 4). The K-S Statistic is a measure of the degree of separation between the cumulative distribution functions (CDFs) of two classes. In this study, the K-S

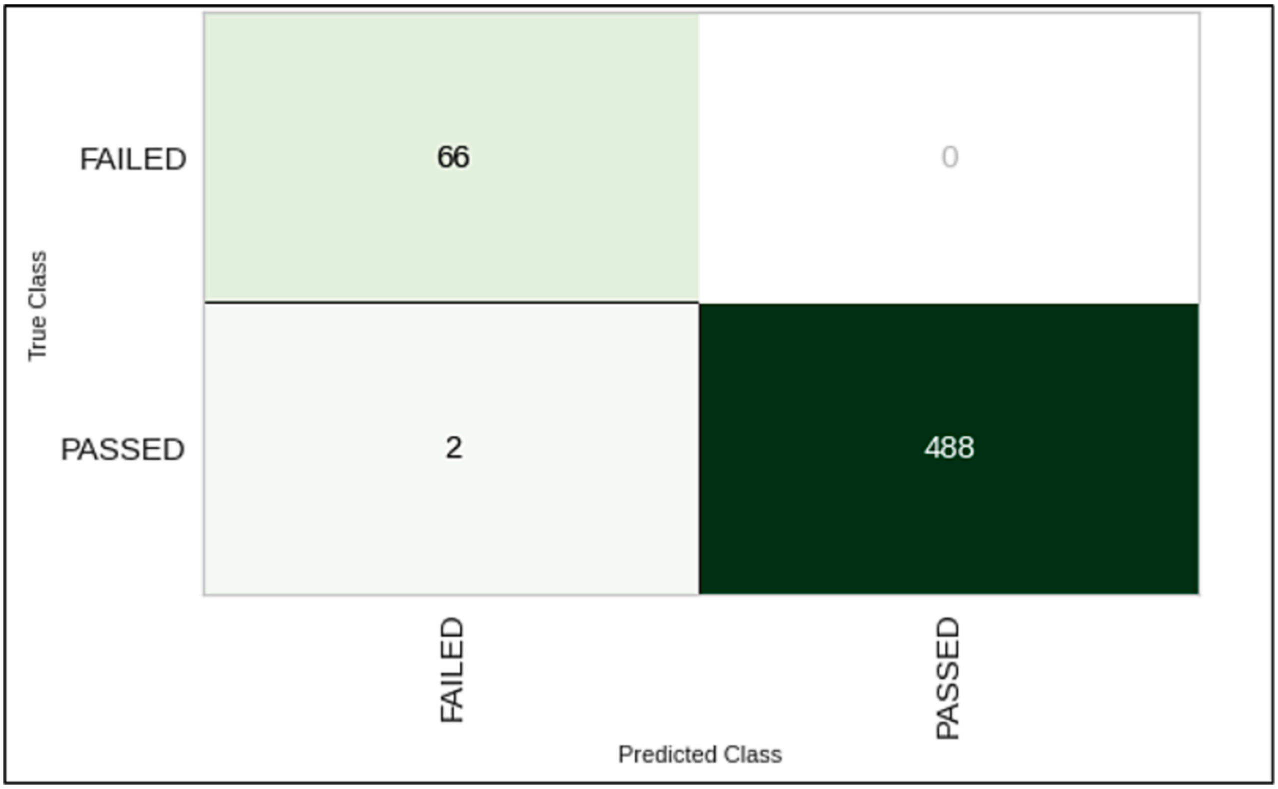

Fig. 3. Confusion matrix of the proposed model.

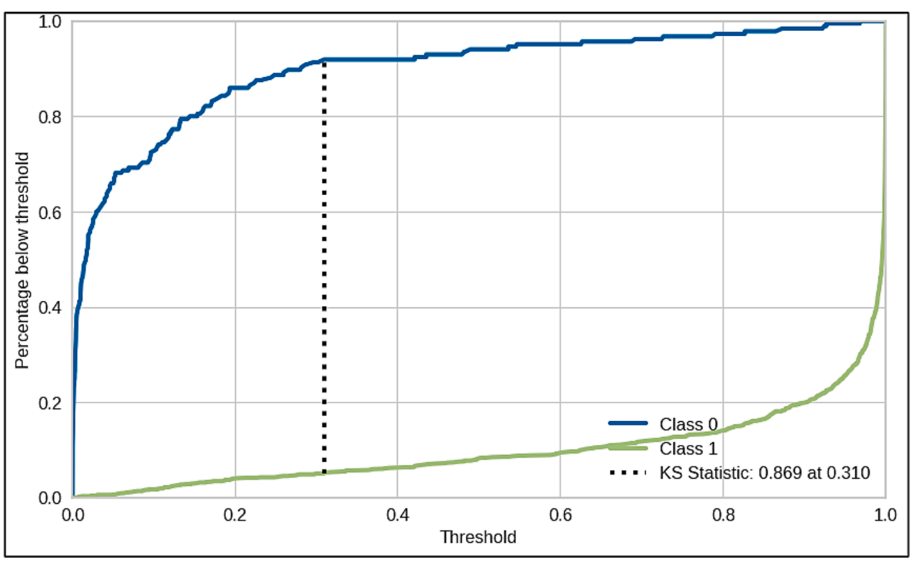

Fig. 4. K-S statistics plot for the proposed decision tree model.

Statistic was used to measure the degree of separation between the CDFs of the predicted probability of the positive class and the true positive rate. The K-S Statistic is computed as the maximum difference between the two CDFs and ranges between 0 and 1. A value of 1 indicates a perfect separation between the two classes, while a value of 0 indicates no separation. The equation for the K-S Statistic can be seen as presented below.

$$KS = |F1(x) - F2(x)| \tag{2}$$

where F1(x) and F2(x) are the cumulative distribution functions of the two samples being compared, and x is a value on the x-axis.

The results of the K-S Statistic for the decision tree classifier model were found to be 0.869 at a threshold of 0.310. This indicates that the model achieved a high degree of separation between the predicted probability of the positive class and the true positive rate [62]. The high K-S Statistic value of 0.869 indicates that the model effectively differentiated between the positive (1) and negative (0) classes.

3.3. Variable correlation heatmap

In (Fig. 5) we used a variable heat map to evaluate the correlation between different emission parameters and the target variable, which is the status classification of 2-stroke motorcycles by the EPA.

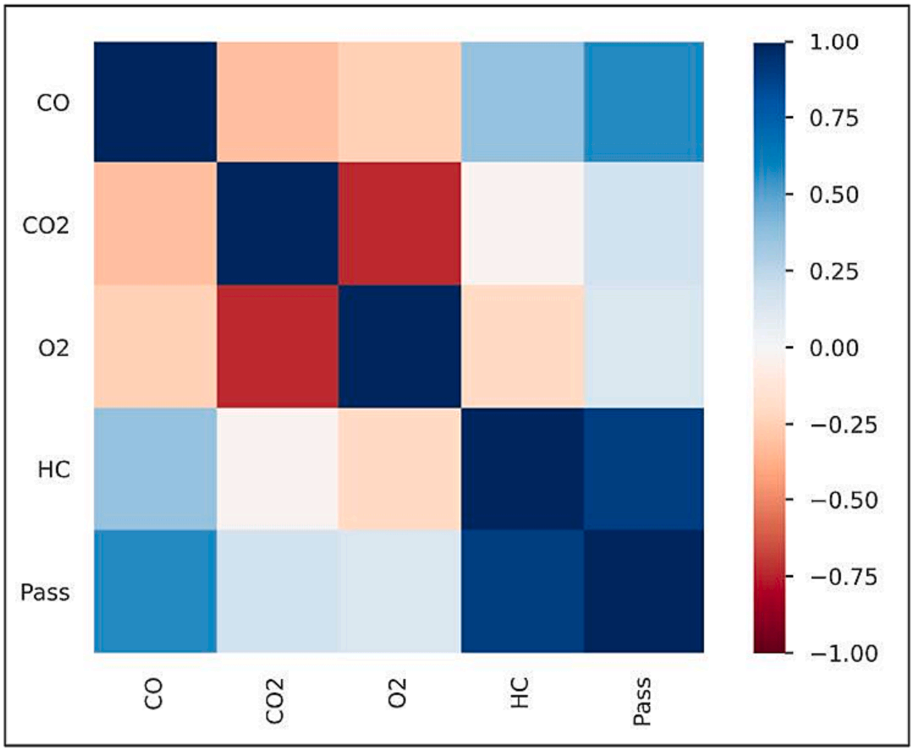

Fig. 5. Feature correlation heat map of the model.

The heat map shows a high degree of correlation between the HC, CO, and the target variable, with values ranging from 0.75 to 0.80. This indicates that these emission parameters strongly influence the EPA classification status. In contrast, the correlation between CO2 and O2 and the target variable is low, with values ranging from 0.00 to 0.09. This suggests that these emission parameters have little effect on the EPA classification status of 2-stroke motorcycles.

Our analysis of the results indicates that the proposed decision tree AI model can effectively classify motorcycle EPA status based on emission parameters. Using a variable heat map in this study is essential in evaluating the correlation between emission parameters and the target variable. It is a well-established method to understand the relationship between different variables. It can be used to identify which variables are more important in the classification process. The use of a variable

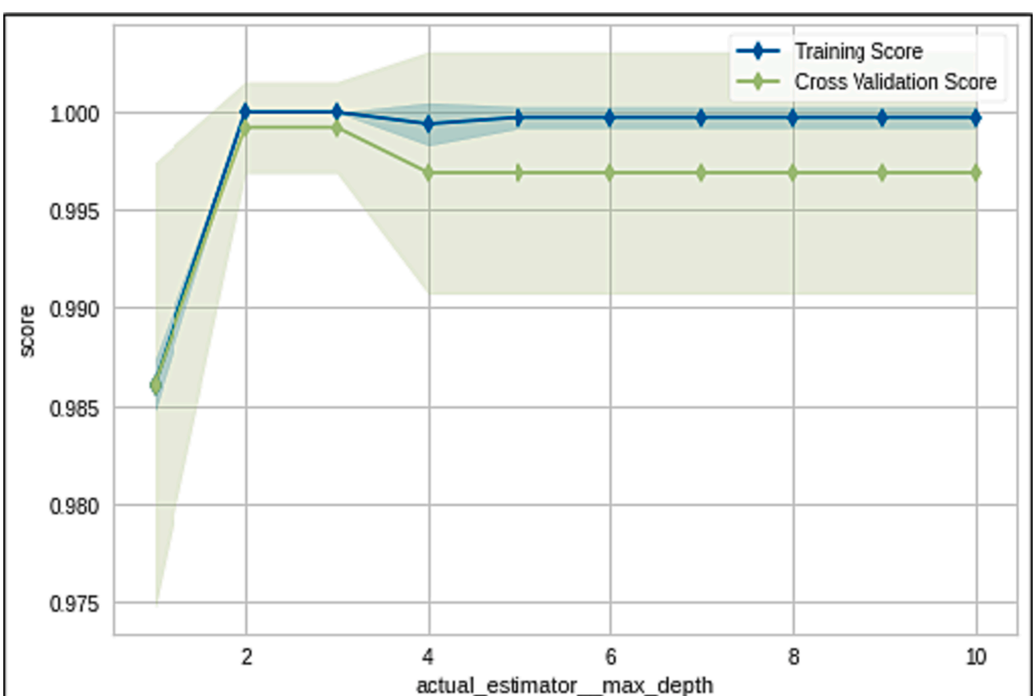

Fig. 6. Decision tree model validation curve.

heat map in this research paper adds a level of rigor and credibility to the study.

3.4. Validation curve for decision tree classifier

The results of the validation curve suggest that the decision tree model can effectively classify EPA status based on emission parameters with a high degree of accuracy.

The validation curve shows that both the training and validation sets start with an accuracy of 0.986 and a maximum depth of 0.9, increasing as the maximum depth increases (Fig. 6). The training set reaches an accuracy of 1.000 on a maximum depth of 2. In contrast, the validation set increases to an accuracy of 0.998 on the same maximum depth. This indicates that the model can generalize to new unseen data and perform well on training and validation sets.

The consistency in the accuracy of both sets suggests that the model is not overfitting or underfitting the data and that the maximum depth of 2 is a good value for the hyperparameter.

As the maximum depth increases, the validation set accuracy remains constant at 0.998. The training set increases until it reaches 1.000 again at 5 maximum depth and continues steadily until it reaches 10 max depth. This indicates that the model continues to perform well on the training set as the maximum depth increases; however, the validation set performance remains constant. This means that the model is not generalizing well to unseen data anymore and is starting to memorize the training set, which is an indication of overfitting.

Overall, the results of the validation curve suggest that the decision tree model is performing well on both the training and validation sets, with a high degree of accuracy. However, it also suggests that the model may be overfitting after a certain point and that a maximum depth of 2 or 5 would be a good choice for the hyperparameter based on the trade-off between bias and variance.

3.5. Decision tree classifier threshold

The model was trained using a threshold of 0.00 (Fig. 7) based on the precision, recall, and accuracy, among other metrics evaluated. The results indicate that the decision tree classifier achieved a precision of 0.9896 with a recall value of 0.9253 (Table 3). The model demonstrates a high level of precision, which indicates the model's ability to correctly classify positive cases with a high degree of confidence. Additionally, the high recall value signifies that the model can identify a large proportion of the positive cases within the dataset.

The threshold value of 0.00 has been selected to ensure a high recall rate, which is critical for automating the EPA motorcycle emission classification. A high recall rate is particularly important in applications where missing a positive case would have significant consequences, such as air emissions.

It is worth noting that the trade-off between precision and recall is a common issue in machine learning, and the threshold value can be adjusted to find the optimal balance between the two. In this case, the high precision and recall values the decision tree classifier achieves suggest that the model can effectively classify the samples in the dataset. The decision tree classifier demonstrates high precision and recall, with a low threshold value of 0.00, which is suitable for classifying air emissions. This strongly indicates that the model is an effective tool for classifying emission status.

3.6. Model explanation

3.6.1. Feature selection and importance

One of the major challenges in machine learning revolves around the difficulty of interpreting complex models [63]. However, decision tree (DT) models provide a notable advantage regarding explainability and interpretability. Unlike ensemble models, such as random forests or gradient boosting, DT models offer a more straightforward and intuitive structure, making them easier to comprehend.

In (Fig. 8), we examined feature importance metrics, such as the impact of HC and CO emissions; the relative importance of each feature can be assessed within the model (a). Furthermore, employing feature selection techniques with a score of 0.997 further emphasizes the significance of the selected features (b). This high score suggests that these two features play a crucial role in determining the model's predictions. Combining feature importance analysis, feature selection, and the transparent nature of decision trees allows for a clear understanding of the DT model's decision-making process. This attribute renders DT models highly suitable for domains where comprehensibility and interpretability are vital considerations.

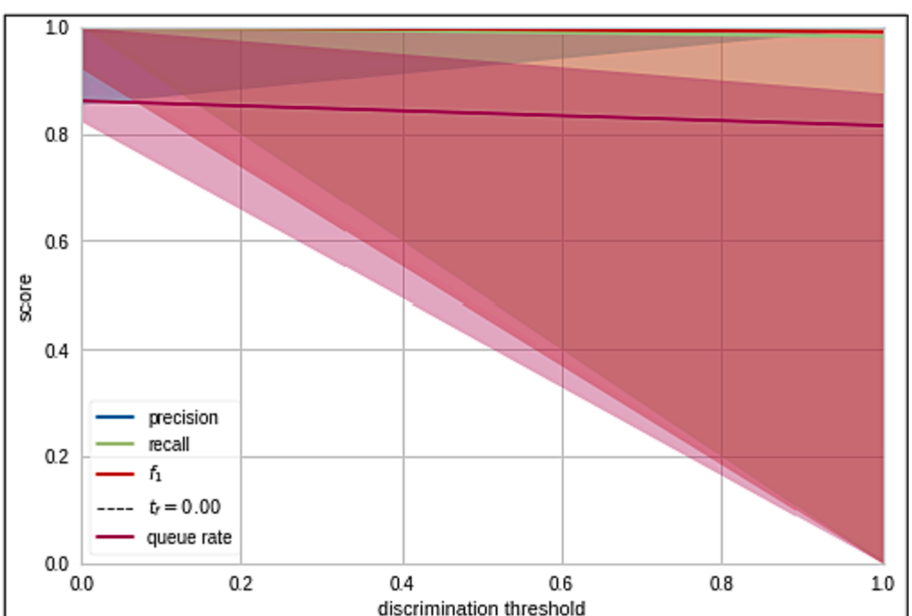

Fig. 7. Threshold analysis for decision tree classifier.

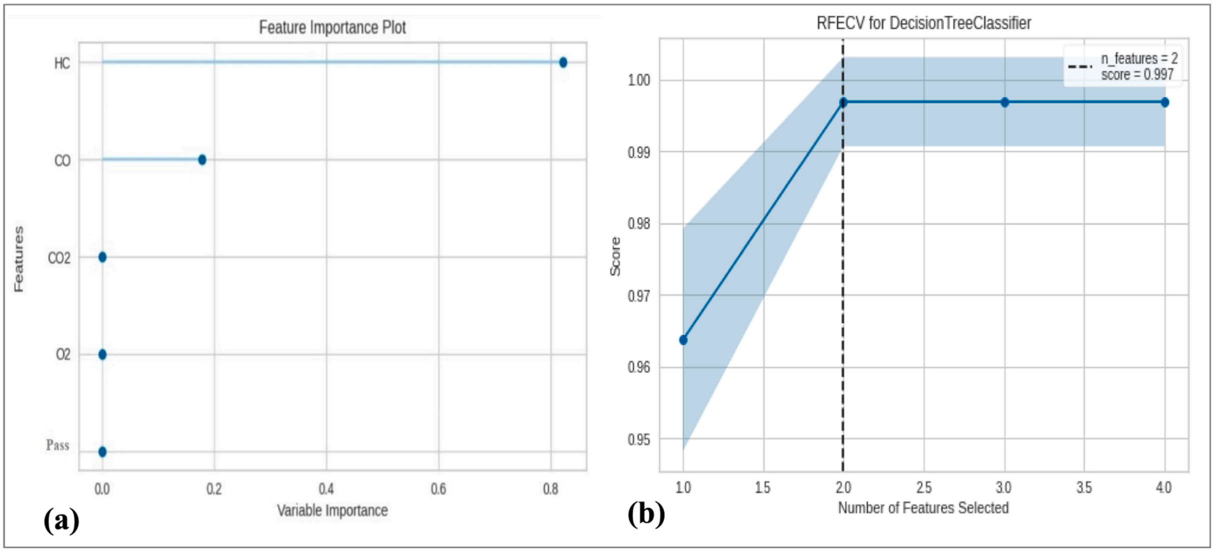

Fig. 8. Visualizing model features importance and selection.

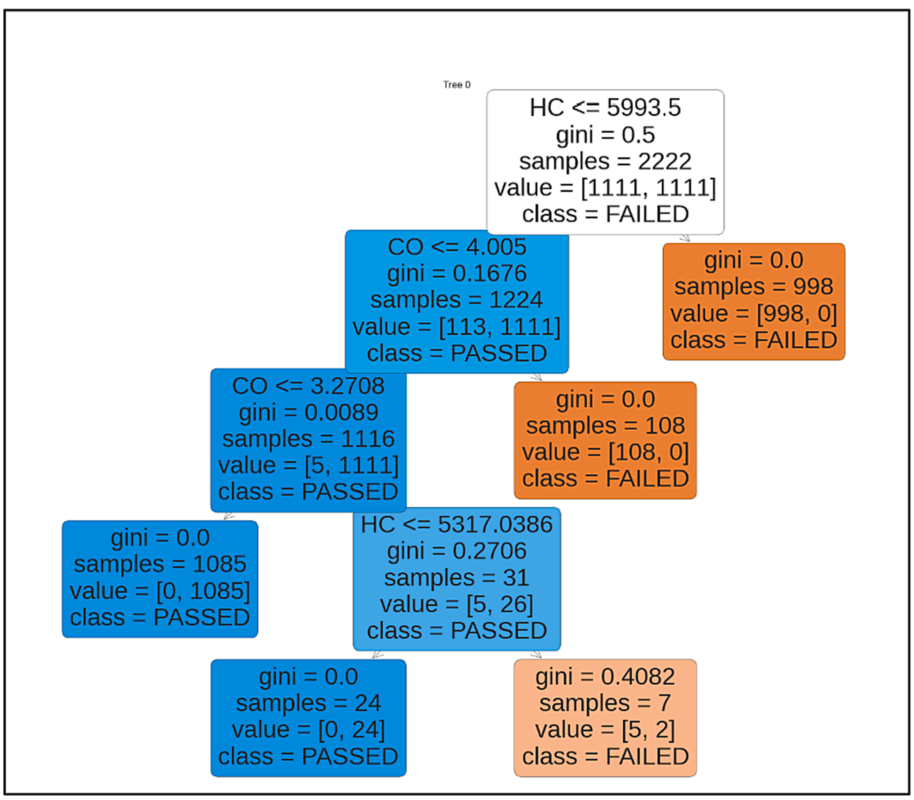

Fig. 9. Proposed model decision tree analysis.

3.6.2. *Decision tree analysis*

In Fig. 9, we presented the results of the decision tree analysis according to their class label leaf nodes, which represent the final decisions or predictions of the model. The first leaf node, Leaf1, is characterized by the condition "HC <= 5993.5," with a Gini index of 0.5 and 222 samples. The Gini index measures the impurity of a leaf node, where a value of 0 represents a pure node, and a value of 1 represents an impure node. In this case, a Gini index of 0.5 indicates a relatively high impurity level in this leaf node. The value for this leaf node is [1111, 1111], representing the number of samples in each class (failed, passed). The class for this leaf node is "Failed."

The second leaf node, Leaf2, is characterized by the condition "CO <= 4.005," with a Gini index of 0.1676 and 1224 samples. The Gini index is lower than the first leaf node, indicating that this leaf is less impure. The value for this leaf node is [113, 1111], representing the number of samples in each class (failed, passed). The class for this leaf node is "Passed."

The third leaf node, Leaf3, is characterized by the condition "CO <= 3.2708," with a Gini index of 0.0089 and 1116 samples. The Gini index is even lower than the second leaf node, indicating that this leaf is much less impure. The value for this leaf node is [5, 1111], representing the number of samples in each class (failed, passed). The class for this leaf node is "Passed."

The fourth leaf node, Leaf4, is characterized by the condition "HC <= 5317.0386," with a Gini index of 0.2706 and 31 samples. The Gini index is relatively lower than the first leaf node but higher than the second and third leaf nodes. The value for this leaf node is [5,26], representing the number of samples in each class (failed, passed). The class for this leaf node is "Passed."

In general, the results of the decision tree analysis indicate that the model can effectively classify EPA status based on emission parameters with a high degree of prediction accuracy. The analysis of the leaf nodes' characteristics shows that the tree is well-balanced regarding the number of samples in each class and has good interpretability. The Gini index values of the leaf nodes indicate that the tree has a good balance between pure and impure leaf nodes, which is a good indication of a well-trained model.

4. Conclusion and future work

4.1. Conclusion

In conclusion, we have presented an AI-driven decision tree model for adequate recognition and classification of EPA status based on emission parameters using motorcycle emissions dataset gathered via standard experimental settings. The model was trained and tested using a dataset of emission parameters from 2-stroke motorcycles and achieved a prediction accuracy of over 92 %.

The results of this study demonstrate the utility of decision AI-driven tree-based models for classifying EPA status and highlight the importance of accurate emissions data in achieving this goal. The model can be easily adapted to other types of industrial sources, especially those with similar emissions, and can be integrated into existing decision-making processes for enhancing and modifying emission control policies. We believe the AI-driven decision tree model presented in this study can be a valuable tool for regulatory agencies, industry professionals, and researchers. By providing a reliable and efficient means of classifying EPA status based on emission parameters, this model can help support effective policy-making and decision-making in air pollution control.

The findings of this study also underscore the importance of data quality and completeness for the prediction task. It is suggested that future research should focus on expanding the dataset and exploring other potential AI models or approaches that may improve the prediction accuracy of the developed model. In addition, the decision tree model can also be extended to predict other emission-related parameters, such as emission rates or emission factors. This would enable the model to be used for other applications, such as emission inventory development or emission reduction planning.

Finally, the AI-driven decision tree model developed in this study represents a major milestone in using artificial intelligence to enhance and modify emission control policies. This model's high prediction accuracy and ease of use make it a valuable tool for researchers, industry professionals, and regulatory agencies to improve their understanding of industrial emissions and develop more effective policies to control them. However, further improvement can be made by fine-tuning the maximum depth of the decision tree to avoid overfitting and by "fix-balancing" possible imbalanced datasets in the algorithm setup.

4.2. Future work

Like all research, this study is not without its limitations. This section provides an accounting of these limitations, which should be considered when interpreting the findings and results of our study.

- AI Ethics and Societal Impact: The AI-driven decision tree model may raise ethical and societal issues, including transparency and job displacement. While the model is designed to increase efficiency and reduce human error, the potential impact on employment and the need for a clear explanation and justification of the model's decisions are significant considerations to be researched in the future.
- Temporal Limitation: The EPA updates its standards and testing procedures periodically. Therefore, the model's accuracy may be affected if it is not updated in line with these changes. Therefore, techniques that could facilitate development of machine learning models that can automatically adapt to abrupt changes resulting from EPA update could be integrated into the proposed model in the future.

Future research should aim to address these limitations by considering a continual update of the training data and model to reflect changes in EPA standards and testing procedures. Furthermore, more in-depth studies on the ethical and societal implications of using AI in regulatory decision-making are warranted.

CRediT authorship contribution statement

Adeboye Awomuti: Methodology, Data curation, Formal analysis, Software, Validation, Visualization, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Philip Kofi Alimo:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **George-Lartey Young:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Stephen Agyeman:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Tosin Yinka Akintunde:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Adebobola Ololade Agbeja:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing, Investigation. **Olayinka Oderinde:** Conceptualization, Investigation, Data curation, Project administration, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing, Investigation. **Oluwarotimi Williams Samuel:** Methodology, Data curation, Formal analysis, Software, Validation, Visualization, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Henry Otobrise:** Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cacint.2023.100127.

References

[1] Engebrethsen E, Dauzere-Peres S. Transportation strategies for dynamic lot sizing: single or multiple modes? Int J Prod Res 2022.

[2] Shetty A, et al. An analysis of labor regulations for transportation network companies. Econ. Transp. 2022:32.

[3] El Moussaoui S, et al. The assessment of pollutant emissions from transportation of construction materials and the impact of construction logistics centers. J Manag Eng 2022;38(5).

[4] Deb M, et al. Application of artificial intelligence (AI) in characterization of the performance-emission profile of a single cylinder CI engine operating with hydrogen in dual fuel mode: An ANN approach with fuzzy-logic based topology optimization. Int J Hydrogen Energy 2016;41(32):14330–50.

[5] Abikusna S, Sugiarto B, Zulfan A. Fuel consumption and emission on fuel mixer low-grade bioethanol fuelled motorcycle. Sriwijaya International Conference on Engineering, Science and Technology (Sicest 2016), 2017. 101.
[6] Aly SH, Ramli MI, Arifin Z. Estimation of carbon dioxide emissions on heterogeneous traffic based on metropolitan traffic emissions inventory model. Int J Geomate 2021;21(83):181–90.
[7] Ashik FR, Rahman MH, Kamruzzaman M. Investigating the impacts of transit-oriented development on transport-related CO2 emissions. Transp Res Part D: Transp Environ 2022;105:103227.
[8] Ibeto C, Ugwu C. Exhaust emissions from engines fuelled with petrol, diesel and their blends with biodiesel produced from waste cooking oil. Pol J Environ Stud 2019;28(5):3197–205.
[9] Ntziachristos L, et al. Emission control options for power two wheelers in Europe. Atmos Environ 2006;40(24):4547–61.
[10] Tomohara A, Xue H. Motorcycles retirement program: Choosing the appropriate regulatory framework. J Policy Model 2009;31(1):126–9.
[11] Aoki P, et al. Environmental protection and agency: motivations, capacity, and goals in participatory sensing. Proceedings of the 2017 Acm Sigchi Conference on Human Factors in Computing Systems (Chi'17), 2017: p. 3138-3150.
[12] Zaragoza LJ. The environmental protection agency's use of community involvement to engage communities at superfund sites. Int J Environ Res Public Health 2019;16(21).
[13] Boullier H, Demortain D, Zeeman M. Inventing prediction for regulation: the development of (quantitative) structure-activity relationships for the assessment of chemicals at the US environmental protection agency. Sci Technol Stud 2019;32 (4):137–57.
[14] Romero JA, Freedman M, O'Connor NG. The impact of Environmental Protection Agency penalties on financial performance. Bus Strateg Environ 2018;27(8): 1733–40.
[15] Biona JBM, Culaba AB, Purvis MRI. Energy use and emissions of two stroke-powered tricycles in Metro Manila. Transp Res Part D-Transport and Environ 2007; 12(7):488–97.
[16] Ghahramani M, et al. Leveraging artificial intelligence to analyze citizens' opinions on urban green space. City and Environ Interactions 2021;10:100058.
[17] Bastos JA. Predicting credit scores with boosted decision trees. Forecasting 2022;4 (4):925–35.
[18] Birant D. Comparison of decision tree algorithms for predicting potential air pollutant emissions with data mining models. J Environ Inf 2011;17(1):46–53.
[19] Abdullah YI, et al. Ethics of artificial intelligence in medicine and ophthalmology. Asia-Pacific J Ophthalmology 2021;10(3):289–98.
[20] Bradley F. Representation of libraries in artificial intelligence regulations and implications for ethics and practice. J Aust Library and Inform Assoc 2022;71(3): 189–200.
[21] England GCW, Millar KM. The ethics and role of AI with fresh and frozen semen in dogs. Reprod Domest Anim 2008;43:165–71.
[22] Golbin I, et al. Responsible AI: a primer for the legal community. Ieee International Conference on Big Data (Big Data) 2020;2020:2121–6.
[23] Bartmann M. The ethics of AI-powered climate nudging-how much AI. should we use to save the planet? Sustainability 2022;14(9).
[24] Zhou CR, et al., Decision tree model to efficiently optimize the process conditions of carbonaceous mesophase prepared with coal tar. Carbon Letters, 2022.
[25] Barukab O, et al. Analysis of Parkinson's disease using an imbalanced-speech dataset by employing decision tree ensemble methods. Diagnostics 2022;12(12).
[26] Pillai VS, Matus KJM. Towards a responsible integration of artificial intelligence technology in the construction sector. Sci Public Policy 2020;47(5):689–704.
[27] Starke G, Ienca M. Misplaced trust and distrust: how not to engage with medical artificial intelligence. Camb Q Healthc Ethics 2022.
[28] Taeihagh A. Governance of artificial intelligence. Policy and Society 2021;40(2): 137–57.
[29] Miller DD. Machine intelligence in cardiovascular medicine. Cardiol Rev 2020;28 (2):53–64.
[30] Shiller AV. The place of the ethical system in the architecture of artificial intelligence. Tomsk State Univ J 2020;456:99–103.
[31] Fox S. Behavioral ethics ecologies of human-artificial intelligence systems. Behav Sci 2022;12(4).
[32] Gratch J, Fast NJ. The power to harm: AI assistants pave the way to unethical behavior. Curr Opin Psychol 2022;47.
[33] Agbese M, et al. Governance in ethical and trustworthy AI systems: extension of the ECCOLA method for AI ethics governance using GARP. E-Informatica Softw Eng J 2023;17(1).
[34] Tékouabou SCK, et al. Optimizing the early glaucoma detection from visual fields by combining preprocessing techniques and ensemble classifier with selection strategies. Expert Syst Appl 2022;189:115975.
[35] Sekiguchi K, Hori K. Organic and dynamic tool for use with knowledge base of AI ethics for promoting engineers' practice of ethical AI design. AI & Soc 2020;35(1): 51–71.
[36] nigeriagalleria. *Ogun State of Nigeria: Nigeria Information & Guide*. 2017 [cited 2023; Available from: https://www.nigeriagalleria.com/Nigeria/States_Nigeria/Ogun/Ogun_State.html.
[37] George TO, et al. Usefulness and expectations on skills development and entrepreneurship among women of low socioeconomic status in Ogun State, Nigeria. African J Reprod Health 2021;25(5s):170–86.
[38] Moonsammy S, et al. Exhaust determination and air-to-fuel ratio performance of end-of-life vehicles in a developing African country: A case study of Nigeria. Transportation Res Part D-Transport and Environ 2021;91.
[39] Obayelu AE, Ogunnaike MG, Omotoso FK. Socioeconomic determinants of fruits consumption among students of the federal university of agriculture, abeokuta, Ogun State, Nigeria. Int J Fruit Sci 2019;19(2):211–20.
[40] Omotoso AB, et al. Socioeconomic determinants of rural households' food crop production in Ogun State, Nigeria. Appl Ecol Environ Res 2018;16(3):3627–35.
[41] Odunlami OA, Alaba AF. Comparison of emission levels of motor cars, motorcycles, and tricycles using petrol engines in Southwestern Nigeria. in Key Engineering Materials. 2021. Trans Tech Publications Ltd.
[42] Brettschneider J. Extension of the equation for calculation of the air-fuel equivalence ratio. SAE Technical Papers 1997.
[43] KaneAutomotive. *Kane Automotive Gas Analyser Manual for Auto 4-1/MID & 5-1/MID*. 2012 [cited 2022 19, May]; Available from: http://docplayer.net/.
[44] Nguyen YLT, et al. A study on emission and fuel consumption of motorcycles in idle mode and the impacts on air quality in Hanoi, Vietnam. Int J Urban Sci 2021;25(4): 522–41.
[45] Patange AD, et al. Augmentation of decision tree model through hyper-parameters tuning for monitoring of cutting tool faults based on vibration signatures. J Vib Eng Technol 2022.
[46] Straub RK, Mandelbaum B, Powers CM. Predictors of quadriceps strength asymmetry after anterior cruciate ligament reconstruction: a chi-squared automatic interaction detection decision tree analysis. Med Sci Sports Exerc 2022; 54(12):2005–10.
[47] Yasir M, et al. Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance. Antibiotics-Basel 2022;11(11).
[48] Azhar MY, et al. Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance. Antibiotics 2022;11(11):1593.
[49] Shi D, et al. Machine learning for detecting parkinson's disease by resting-state functional magnetic resonance imaging: a multicenter radiomics analysis. Front Aging Neurosci 2022;14.
[50] Liu CH, et al. An improved decision tree algorithm based on variable precision neighborhood similarity. Inf Sci 2022;615:152–66.
[51] Damala RB, Patnaik RK, Dash AR. A simple decision tree-based disturbance monitoring system for VSC-based HVDC transmission link integrating a DFIG wind farm. Protection and Control of Modern Power Systems 2022;7(1).
[52] Botha D, Steyn M. The use of decision tree analysis for improving age estimation standards from the acetabulum. Forensic Sci Int 2022;341.
[53] Science.gov. *decision tree analysis: Topics by Science.gov*. 2017 [cited 2023; Available from: https://www.science.gov/topicpages/d/decision+tree+analysis.html.
[54] Shih XY, Chiu Y, Wu HE. Design and implementation of decision-tree (DT.) online training hardware using divider-free GI calculation and speeding-up double-root classifier. Ieee Transactions on Circuits and Systems I-Regular Papers 2022.
[55] Makond B, Pornsawad P, Thawnashom K. Decision tree modeling for osteoporosis screening in postmenopausal thai women. Informatics-Basel 2022;9(4).
[56] Kumar S, et al. Decision tree Thompson sampling for mining hidden populations through attributed search. Soc Netw Anal Min 2022;12(1).
[57] Jung JY, Yang CM, Kim JJ. Decision tree-based foot orthosis prescription for patients with pes planus. Int J Environ Res Public Health 2022;19(19).
[58] Afnan MAM, et al. Interpretable, not black-box, artificial intelligence should be used for embryo selection. Human Reproduction Open 2021;2021(4).
[59] Xu M, Qin Z. How does vehicle emission control policy affect air pollution emissions? Evidence from Hainan Province, China. Sci Total Environ 2023;866: 161244.
[60] Fabbri M. Social influence for societal interest: a pro-ethical framework for improving human decision making through multi-stakeholder recommender systems. AI & Soc 2022.
[61] Azeem I. Tailpipe Emissions Data for sedan vehicle, in Emission Data, KaggleDatabase, Editor. 2017: Australia.
[62] worldwidescience.org. *Fuel Consumption Engines*. 2018 [cited 2023; Available from: https://worldwidescience.org/topicpages/f/fuel+consumption+engine.html.
[63] Tékouabou SCK, et al. Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics 2022; 10(14):2379.