Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

MACHINE LEARNING FOR CONCRETE SUSTAINABILITY IMPROVEMENT: SMART FLEET MANAGEMENT

Coralia TANASUICA (ZOTIC)*[1]
Mihai Daniel ROMAN*[2]

Abstract: *In the dynamic landscape of modern business operations, ensuring economic security through efficient and intelligent fleet management is imperative. This necessitates a dual focus on safeguarding revenue streams and optimizing operational costs. The aim of this study centers on two main objectives: first, to identify driving behaviors that have a substantial impact on vehicle maintenance costs; second, to ensure the sustainability of the fleet is managed effectively. To achieve these objectives, the research employs unsupervised Machine Learning (ML) techniques for segmenting driving styles based on diverse parameters collected from Internet of Things (IoT) devices. Furthermore, the Long Short-Term Memory (LSTM) algorithm is used for forecasting fuel consumption, offering a predictive glance into future expenditures. The methodology is based on the analysis of data gathered from sensors installed on the vehicle's Controller Area Network (CAN), collected over a span of five months. The findings spotlight a subset of drivers whose aggressive driving significantly influences maintenance costs and highlight optimal indicators for drivers to monitor to minimize CO_2 emissions. Additionally, the study identifies key performance indicators that drivers should monitor to reduce CO_2 emissions, contributing to the environmental sustainability of the fleet. This investigation not only elucidates the financial and environmental implications of driving behaviors but also showcases the transformative potential of ML technologies in enhancing the strategic management of vehicle fleets. Through this exploration, the research advocates for the integration of advanced analytics and sustainable practices as foundational elements for businesses striving to achieve economic security and operational resilience.*

Keywords: *driver behavior, fleet sustainability, unsupervised machine learning, clustering analysis, CO_2 forecasting.*

UDC : *[005.:656.01]+[004.8:656.052]*

JEL Code: *C38, D01, C15, F64.*

Introduction

In the recent period, there has been a notable increase in research and studies focused on identifying driver behavior. This surge is evident in efforts to devise optimal solutions for autonomous vehicles as well as to protect individuals and society by reducing traffic incidents and promoting eco-friendly driving habits that decrease harmful emissions. Many of these studies rely on data samples from driving simulators, examining specific aspects of driver behavior, such as

*[1] *Coralia TANASUICA (ZOTIC), PhD Student, The Bucharest University of Economic Studies, Romania, Email: tanasuicacoralia22@stud.ase.ro, ORCID: 0009-0003-6570-4375*
*[2] *Mihai Daniel ROMAN, Ph.D. Professor, The Bucharest University of Economic Studies, Romania, Email: mihai.roman@ase.ro, ORCID: 0000-0002-3859-7629*

Received: 14.03.2024 Accepted: 01.04.2024

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

overtaking maneuvers, in controlled conditions. The objective of the current research was to explore how Machine Learning (ML) can be utilized to discern real-world driver behaviors in traffic—beyond simulator data—to identify patterns that adversely affect a company's budget. Another aim of this study is to estimate future emission reductions achievable by the company's vehicle fleet, providing insights for management to keep emissions below a certain threshold and steering the company towards sustainability in terms of CO_2 emissions.

The data utilized in this study were collected in real time using sensors installed on a sample of nine vehicles from the fleet. The substantial volume of data gathered enabled a multifaceted analysis of driving behavior by integrating various dimensions and utilizing a mix of indicators and input variables. These included parameters such as cornering speed, the frequency of undercarriage impacts, the number of braking events at speeds exceeding 60 km/h, and the volume of CO_2 emissions over a specified time interval. This comprehensive approach facilitated a detailed examination of driver behavior patterns and their environmental impact.

Moreover, the sensors equipped on the vehicles were capable of recording a wide range of impacts that occurred on various parts of the vehicle, including frontal collisions, side impacts, overhead strikes, and impacts from below. The granularity of this data, capturing not just the occurrence of such impacts but also their intensities, added another layer of depth to the analysis. This nuanced data collection enabled the study to pursue multiple segmentation directions, categorizing vehicles not only based on driving style but also according to the volume and intensity of incidents impacting the vehicle. This dual-axis segmentation offered a more holistic view of driver behavior and vehicle usage patterns, shedding light on potential areas for operational improvements and risk mitigation strategies.

The findings of this research are promising regarding the identification of driver patterns that negatively impact company profitability. However, in terms of forecasting emissions, the study is inconclusive. Despite the application of advanced forecasting methodologies, such as Long Short-Term Memory (LSTM) networks, known for their capacity to handle sequence prediction problems effectively, the study does not succeed in establishing a robust model for emissions forecasting. This outcome suggests a gap in the predictive capability of the current model, underscoring the necessity for additional research and refinement of forecasting methodologies. Expanding the volume of data could provide a more comprehensive foundation for the model to learn from, capturing a wider array of driving conditions and emissions patterns. Additionally, refining the granularity of the data might allow for a more nuanced understanding of the relationship between driving behaviors and emissions, facilitating the development of a more accurate and predictive model.

This research progresses through several structured sections. Initially, in Section Two, a comprehensive review of relevant literature is provided. This section sets the stage for understanding the context and significance of the current research within the broader academic landscape.

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

In section Three, the methodologies and models used in the study are detailed. This part explains the choice of Machine Learning (ML) techniques, including unsupervised learning for driver segmentation and neural networks for fuel consumption prediction. The choice of unsupervised learning is highlighted due to its effectiveness in identifying patterns within the data without the need for pre-labeled instances. This method facilitates the segmentation of drivers into distinct groups based on their driving styles, enabling targeted interventions for improving fuel efficiency and reducing emissions. Strategies such as data preprocessing, feature selection, and normalization are implemented to prepare the dataset for analysis, ensuring that the ML models can learn effectively from the data.

Section Four offers a detailed presentation of the dataset, including the calculated indicators, accuracy rates, and findings of the case study. It provides a comprehensive look at the dataset, describing the data collection process through sensors, the variables measured, and the methods used to maintain data accuracy. This section transitions from the technical aspects of data handling to the analysis of calculated indicators, accuracy rates, and case study findings, offering a clear view of the research outcomes. In addition to these aspects, significant attention was placed on testing the models and strategies to identify the optimal model by analyzing performance indicators. This involved a detailed examination of various models to evaluate their effectiveness in capturing the nuances of the dataset and accurately predicting outcomes. A key focus of this evaluation was on combating overfitting and underfitting, ensuring that the models neither overspecialized on the training data nor remained too generalized to make accurate predictions.

The pivotal results of this study reveal a distinct segment of drivers characterized by a less cooperative driving style, leading to increased fuel consumption, an increased frequency of car incidents, and a greater likelihood of requiring premature car maintenance. These findings have significant implications for operational efficiency and environmental impact.

To address these challenges, the study suggests the implementation of specific Key Performance Indicators (KPIs) aimed at changing driver behavior and reducing company losses. This recommendation is based on the analysis, having the final scope to improve driver cooperation, decrease fuel consumption, and lower incident rates. The discussion on KPIs not only suggests solutions to the observed problems but also encourages future research and practical applications to enhance safety, sustainability, and efficiency in fleet management.

Literature review

The literature reviewed in this research is organized into three categories: Driver Behavior Analysis and Safety, Machine Learning in Transportation and Logistics, and Innovative Technologies for Traffic and Vehicle Management. This structure aids in systematically examining the existing studies and approaches within these areas. By organizing

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

the literature this way, the review sets out to cover the broad spectrum of knowledge, from understanding driver behavior and ensuring safety to the application of machine learning for optimizing transportation and logistics, and finally, to exploring innovative technologies for better traffic and vehicle management.

Papers in the category of Driver Behavior Analysis and Safety focus on understanding and classifying driver behavior using machine learning and statistical analysis. They aim to enhance road safety and driving efficiency, examining factors like aggressive driving, fuel consumption, and driving patterns in different scenarios such as highways, roundabouts, and signalized intersections. One of the papers (Raman & Anuj, 2022) focuses on using machine learning to analyze and classify driving behavior based on vehicle On-Board Diagnostic (OBD) data. It explores how various driving parameters can indicate different driving styles and their impact on vehicle performance and safety. The study utilizes Support Vector Machine (SVM) and AdaBoost algorithms for classification and demonstrates high accuracy in distinguishing among diverse driving behaviors. This research contributes to understanding driver behavior patterns through telematics and machine learning, offering potential applications in vehicle safety systems and driver coaching tools.

A novel framework is proposed, designed to detect road rage and aggressive driving using GPS and heart rate data (Subramanian & Bhargavi, 2023). The proposed framework integrates GPS data to track vehicular movement and heart rate monitoring to assess the driver's physiological state, indicative of stress or aggression levels. The approach utilizes machine learning classifiers to categorize drivers and demonstrates the potential of this system in real-world applications. The findings from this research underscore the potential of integrating technology with behavioral science to improve road safety and tailor insurance premiums more accurately to individual driving habits.

A novel approach to urban traffic planning and safety improvements (Tasnim & Huthaifa, 2022) focuses on analyzing the impact of roundabout design on driver behavior using machine learning techniques. The study aims to categorize drivers into different behavior profiles (reckless, conservative, and casual) and examines how roundabout design influences these behaviors. Using data collected from drones, the research applies unsupervised machine learning, specifically K-means clustering, to classify driver behavior and assess roundabout design effectiveness. This research contributes to the understanding of driver behavior at roundabouts and offers insights for urban traffic planning and safety improvement. By identifying specific design elements that influence driving styles, this research provides valuable insights for urban planners and traffic safety experts aiming to design roundabouts that promote safe and efficient driving behaviors. The use of machine learning techniques, particularly K-means clustering, proves effective in dissecting complex driver behavior patterns.

An approach to traffic safety (Mouhammed et al., 2022) explores the classification of driving behavior at signalized intersections under different signal conditions using vehicle

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

kinematics data. The study employs the K-means clustering algorithm to categorize driving behaviors and investigates the impact of signal conditions on driver behavior. The research findings indicate that driving behavior is more influenced by individual habits and personality than by signal conditions, although the flashing green signal condition tends to induce more conservative driving behavior. This has important implications for the design of traffic signals and the development of safety measures, highlighting the need for traffic management strategies that consider the diversity of driver behaviors. By demonstrating the predominant influence of individual characteristics over external signal conditions, the study suggests avenues for further research in traffic safety and signal design, aiming to create safer road environments.

The driver behavior is also investigated in working zones (Huthaifa et al., 2022) using unsupervised machine learning. The study employs the K-means algorithm to classify drivers into aggressive, conservative, and normal categories based on vehicle kinematic data. The data was collected from 67 participants using a driving simulator, simulating various work zone scenarios to accurately capture driver responses. The findings reveal a higher number of aggressive and conservative drivers compared to normal drivers, suggesting that drivers either cautiously navigate through work zones or display aggressive behavior. The study provides insights for policymakers to enhance road safety in work zones. By identifying specific behavior patterns prevalent in such environments, the research contributes to the broader effort of enhancing safety measures and informing the development of more effective traffic management strategies in work zones.

The category of Machine Learning in Transportation and Logistics includes studies that apply machine learning techniques to challenges in transportation and logistics. Topics cover predictive maintenance in facility management, fuel consumption prediction for trucks, and real-time supply chain risk mitigation. These studies emphasize the role of AI and machine learning in improving efficiency and sustainability in these sectors.

One of the analyzed papers (Sheunesu et al., 2023) focuses on developing models for predicting truck fuel consumption. The study uses logistic regression and neural networks, comparing their effectiveness in forecasting fuel usage based on various driving parameters. This research addresses the challenge of managing fuel costs in the logistics industry, particularly in the context of rising oil prices and economic recovery post-Covid-19. The findings demonstrate the viability of these methods for efficient fuel management, contributing to cost reduction and environmental sustainability in transportation.

There are researchers (Nawal et al., 2023) who discuss the integration of machine learning in predictive maintenance within the Industry 4.0 paradigm. It reviews the evolution of maintenance strategies, emphasizing the transition from reactive to predictive maintenance using machine learning. The paper identifies the challenges in implementing machine learning-based predictive maintenance, particularly in small and medium enterprises, and provides a comprehensive overview of this approach. It highlights the importance of data quality, algorithm

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

selection, and the necessity of a preliminary phase in predictive maintenance projects. The paper aims to guide researchers and practitioners in successfully implementing machine learning in predictive maintenance, contributing to the advancement of smart manufacturing.

The last category of papers, included in the Innovative Technologies for Traffic and Vehicle Management class, explores the use of technologies like animation-based variable message signs and driver monitoring systems to improve traffic management and vehicle safety. These studies investigate the impact of such technologies on driver behavior and road safety, particularly in work zones and public transportation.

There is a paper (Mustafa et al., 2020) that investigates the effectiveness of innovative variable message signs (VMS) in enhancing road safety in work zones. It compares traditional static signs with graphical and animation-based VMS in a driving simulator study. The research aims to determine if these VMS systems can reduce drivers' traveling speeds, encourage early lane-changing, and improve spacing between vehicles. The study's findings suggest that animation-based VMS can effectively influence driver behavior, contributing to safer road conditions in work zones.

A drowsiness detection system (Ziryawulawo et al., 2023) is developed for the Kayoola EVS bus. It employs machine learning techniques using facial recognition algorithms to monitor driver alertness and triggers an alarm in cases of detected drowsiness. The study highlights the effectiveness of the developed system in real-time operation and its potential impact in reducing road accidents caused by driver fatigue. This research contributes to the field of advanced driver assistance systems and transportation safety.

A study was conducted (Delussu et al., 2021) focusing on minimizing fuel consumption in public buses using sensor data and Bayesian networks. The research methodology centers around the collection and analysis of data from various sensors installed on public buses. These sensors monitor numerous aspects of bus operation, such as acceleration and braking patterns, as well as fuel usage. The study employs Bayesian networks to analyze the complex relationships between these variables, aiming to identify patterns and factors that significantly impact fuel consumption. The results suggest potential strategies for fuel efficiency improvements, contributing to environmental sustainability and cost reduction in public transportation. This study not only advances the field of transportation research but also offers practical solutions for reducing the environmental impact of public buses, aligning with broader goals of sustainable urban development

Each category represents a significant contribution to the fields of transportation engineering, logistics, and road safety, showcasing the diverse applications of machine learning and advanced analytics in these areas.

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

Data and Methodology

The study takes a comprehensive approach, not solely focusing on CO_2 emissions estimation but also on discerning drivers' behavioral patterns and identifying less sustainable routes in terms of emissions and terms of the impact of external events of varying intensities on car level. Additionally, it encompasses forecasting future emissions and fuel consumption, while pinpointing emission parameters that can be minimized, specifically targeting the reduction of supplementary CO_2 emissions in the analysis.

The dataset for this research was sourced from IoT sensors fitted on a select sample of nine diverse vehicles within a company's fleet, encompassing three vans and trucks used for freight transport, three service vehicles operated by company staff, and three rental cars leased out to individuals or other businesses. These sensors were installed in April, with data collection starting with May 2023 and channelled into a data lake. From this repository, data were extracted via APIs, initially aggregated, and subsequently utilized in various stages of the analysis. Spanning from May 2023 to September 2023, the data accumulated to a considerable volume of approximately 50 GB, providing a substantial foundation for a comprehensive five-month analytical period.

I will explore the four methodological directions used to gain a clearer understanding of their interconnections, with the ultimate goal of identifying points where company management can alter variables and strategies. This will enable the company to benefit not only in terms of its image as a sustainable entity but also through reduced costs.

Supervised Machine Learning for CO_2 Estimation

Linear regression was employed due to its predictive capabilities for continuous data, allowing to model a functional relationship between independent variables (vehicle type, distance, and fuel type) and the dependent variable (CO_2 emissions).

The linear regression model was developed using Python and the Scikit-learn library, streamlining the process of estimating CO_2 emissions for a vast dataset. While initial emission calculations for specific routes were performed using the myclimate calculator, the extensive volume of data points, spanning five months, necessitated a more efficient computational approach. Consequently, the linear regression model provided a systematic method to estimate emissions for the entire dataset, thereby optimizing the analysis process in our comprehensive study.

The dataset included distinct categories of vehicles and varied distances traveled, using different types of fuel. Emissions calculations were done using the myclimate (https://www.myclimate.org/en/) calculator, which is recognized for its accuracy and reliability in environmental impact assessment.

The model proved to be a robust predictor of CO_2 emissions, considering it had a coefficient of determination, denoted as R^2, higher than 85%. The coefficient is a statistical

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

measure in regression analysis that represents the proportion of variance for the dependent variable that's explained by the independent variables in the model. It provides a sense of the goodness of fit of the model to the data. An R^2 value of 1 indicates that the regression predictions perfectly fit the data, while an R^2 of 0 indicates that the model does not explain any of the variability in the outcome data around its mean.

Unsupervised Machine Learning for Vehicle Segmentation

Clustering was selected as the analytical method to segment the dataset based on driver behavior, external events of different intensities at the car level, or route characteristics without using predefined labels. This technique allowed for the exploration of the underlying structures in the data, revealing distinct groupings among drivers' behaviors, the diversity in their driving styles, and the various route patterns. These clusters help to identify behavioral trends and route patterns, which are important in developing personalized driver indicators for reducing the maintenance costs of the car and optimizing route planning.

In the study, Hierarchical Clustering was utilized for a detailed and layered analysis. The clustering model was built using Python and the Scikit-learn (sklearn) library. This approach facilitated the segmentation of routes and the categorization of drivers based on their driving styles, as well as the routes they frequently used and various external events impacting the vehicles. This method proved instrumental in delineating distinct patterns in driving behavior and route selection, offering nuanced insights into the diverse driving dynamics and external influences encountered.

The elbow method helped to determine the optimal number of clusters by finding the point where the rate of decrease in within-cluster variation sharply changes. Subsequently, the Silhouette score, ranging from -1 to +1, was calculated for each sample, offering a measure of how similar that object is to its cluster and compared to other clusters. The formula for the Silhouette coefficient in a dataset is represented as:

$$\frac{b-a}{\max(a,b)} \quad (1)$$

Where:

b = the mean distance between a sample and all other points in the next nearest cluster

a = the mean intra-cluster distance, or the average distance of a sample to all points in its cluster.

This coefficient's final value is the mean of all individual sample coefficients. A value close to 1 indicates clear segmentation, while a value near 0 suggests overlapping clusters and less effective segmentation.

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

Descriptive Cluster-Level Analysis

This analysis sought to uncover common patterns and noteworthy differences between the formed clusters, providing a granular understanding of the segmented groups.

ANOVA (Analysis of Variance) was employed to rigorously assess the statistical significance of mean differences across the identified clusters. This method allowed us to ascertain whether the observed variations in cluster characteristics were statistically substantial. For visual representation, we utilized box plots. These plots provided a concise and informative overview of the data distribution within each cluster, highlighting the range, central tendency, and any potential outliers, thereby offering a clear understanding of the data's spread and variability. This dual approach of ANOVA and box plots enriched our analysis, ensuring both statistical rigor and visual clarity in our findings.

LSTM for Future Emission Forecasting

A bidirectional LSTM (Long Short-Term Memory) network was implemented, which is capable of learning long-term dependencies and recognizing patterns across time-series data. This advanced neural network architecture is designed to capture both forward and backward dependencies in the data, enabling a comprehensive understanding of temporal patterns and trends. It is very good at recognizing complex, long-term relationships within time-sequenced data, making it an ideal choice for accurately forecasting future emissions and consumption trends, based on the historical data patterns it learns.

The LSTM's architecture was optimized for the specific data characteristics, including adjusting the number of layers to balance the model's complexity and computational efficiency, and the batch size to improve learning stability. Additionally, the learning rate was adjusted to promote convergent behavior, ensuring that the model consistently moved toward optimal performance during training.

These methods combined provide a robust framework for both analyzing current vehicle emissions and forecasting future trends, which is critical for developing effective environmental policies and strategies.

The Model and Findings

The data flow for this case study can be described as follows: Data is collected from IoT sensors installed on a fleet of vehicles, encompassing a variety of types including Vans, job cars, and rentals. This data is then transmitted and consolidated into a centralized data lake beginning May 1st. This data is extracted via APIs for subsequent analysis stages, providing a robust dataset spanning five months and aggregating to 50 GB in volume for comprehensive insights. Once data is extracted from the data lake via APIs, it is loaded into a MySQL database where further aggregation takes place to create data marts. These data marts serve as the

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

training ground for the machine learning models. The outcomes of these models are then fed back into the MySQL database. This integration allows for extensive data analysis, where the interplay between model results and raw data can be examined in-depth to derive actionable insights. The described flow is represented into the next figure (Figure 1):

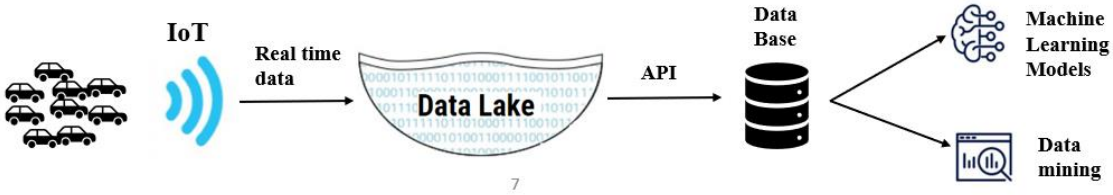

Figure 1. Flow of data

Source:authors designed

For the CO_2 emissions estimation component of the study, conducted from May 1, 2023, to September 30, 2023, a specific approach was undertaken. Initially, a manual calculation was performed using the myclimate calculator for a subset of the data to establish a baseline for emissions based on specific vehicle parameters such as route distance, fuel type (diesel/E10), and vehicle category (SUV, van, truck, luxury, mid-range car). This manual approach provided a detailed sample set of calculated emissions.

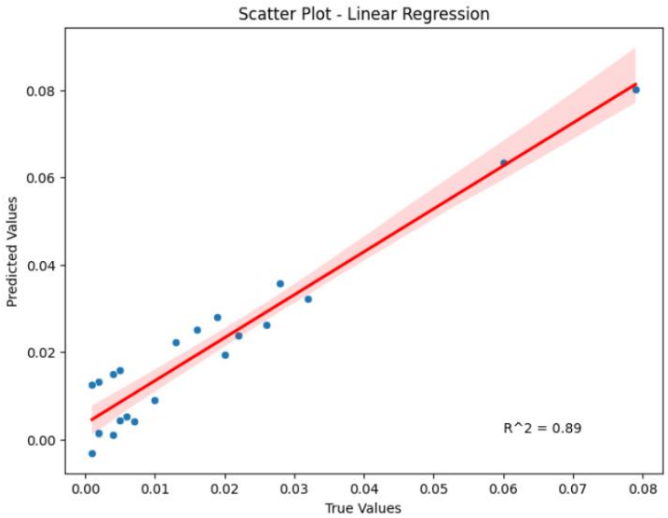

Figure 2. Plot for the linear regression model

Source: authors calculations, based on carburant consumption

Given the large volume of data points within the database, to enhance efficiency and scale the process, a linear regression model was developed and applied to the broader dataset. The regression model's scatter plot, with its distinct red regression line (Figure 2), produced using Python, and Matplotlib library, indicated that 89% of the variance in the dependent

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

variable (CO_2 emissions) could be explained by the independent variables incorporated into the model. The pink confidence interval on the plot provided a predictive range with a certain level of probability for the real values. This strategic combination of manual calculations for a targeted sample and automated regression for the extensive dataset ensured efficient analysis of vehicle emissions.

In the segmentation phase of the study, a range of variables were utilized to categorize vehicles and drivers according to their CO_2 emissions. The following variables were considered monthly: distance traveled, total CO_2 emitted, additional CO_2 emitted, unique locations visited, total locations visited, and duration of travel.

In the context of our data, the "additional CO_2 quantity" refers to emissions that fall outside the typical driving activity, specifically identified as the emissions resulting from prolonged idling. Our observations indicated that certain drivers were idling their engines for periods extending beyond five minutes. Such behavior, which is not associated with regular driving or traffic conditions, was categorized as supplementary fuel consumption and, consequently, additional CO_2 emissions. This categorization is critical as it helps isolate and quantify inefficient fuel usage that contributes to unnecessary CO_2 emissions, offering a clearer target for reducing the environmental impact of the fleet's operation. This aspect of the analysis underscores the importance of behavioral interventions that could lead to a reduction in idle times and, thereby, enhance overall fuel efficiency and sustainability.

Through this segmentation, two distinct clusters emerged, differentiated by their emission patterns and travel behaviors. The silhouette score, peaking at 0.35 for the two clusters, suggests a moderate separation between them. This is further substantiated by a significant F-statistic and a p-value of 0.0075, indicating that the variability in the number of unique locations visited is significantly greater between the clusters than within them. This statistical outcome implies that the drivers in one cluster are visiting more unique locations, which could be contributing to their higher CO_2 emission levels.

Moreover, the greater F-statistic value reinforces the distinction between these two clusters, suggesting substantial differences in their travel and emission profiles. This could be reflective of varying driving habits, operational purposes of the vehicles, or other underlying factors such as the frequency of trips to certain locations that are farther away or require more fuel-intensive travel routes. The segmentation thus provides a detailed characterization of the fleet, informing targeted interventions for emission reduction and more efficient fleet management.

Following the segmentation into clusters, an exploratory analysis was to visually discern and confirm the differences between them. For each variable, including the monthly unique locations visited, boxplots were generated for each cluster. These boxplots serve as a powerful visualization tool, laying out the distribution spread and central tendencies, and highlighting any potential outliers. By comparing these boxplots across clusters, we could visually validate the significant differences that the statistical tests indicated. The boxplot for

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

the variable 'unique locations visited' (Figure 3) particularly illustrated a clear disparity between the clusters, reinforcing the findings from our statistical analysis that drivers in different clusters exhibited distinct patterns in their travel behaviors. This visual representation complements our statistical findings, providing an intuitive understanding of the data's structure and the behavioral segmentation within the fleet.

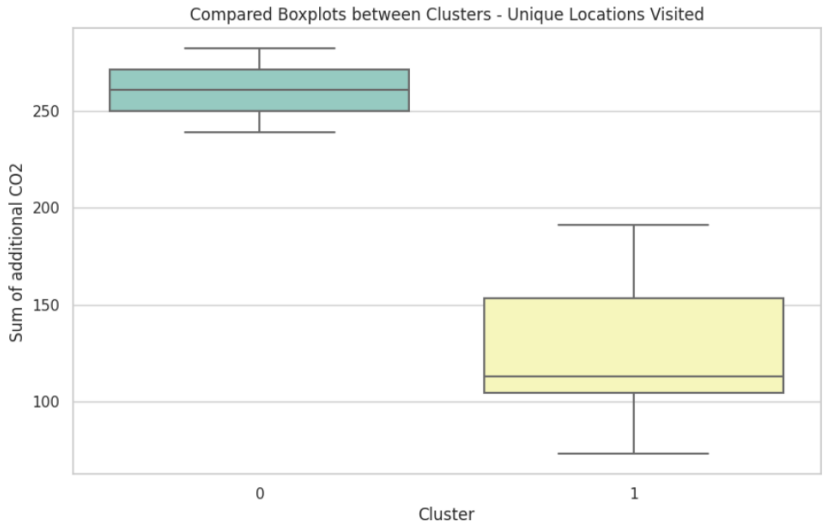

Figure 3. Boxplot for variable 'Unique locations Visited' for the two distinct clusters

Source: authors calculations

The objective of the CO_2 emissions forecasting segment of the study was to predict the future CO_2 consumption of the fleet weekly. To achieve this, we employed a Bidirectional Long Short-Term Memory (LSTM) network at the individual vehicle level. This advanced form of a recurrent neural network (RNN) is uniquely suited for this task due to its ability to process sequences of data while considering both past and future context, which is crucial for making accurate predictions.

Bidirectional LSTMs extend the capabilities of standard LSTMs by analyzing the data in both forward and backward directions, thereby capturing patterns that may be overlooked when considering only past context. This dual-direction analysis is particularly beneficial when dealing with complex time-dependent data like CO_2 emissions, which can be influenced by a myriad of interdependent factors.

Despite the sophisticated nature of the model, the results obtained were suboptimal, with an R^2 score of less than 0.2. This suggests that the model was unable to capture the complexity of the emissions patterns within the fleet adequately. A potential reason for the model's underperformance could be the size of the dataset. With smaller datasets, the model has limited data from which to learn, reducing its ability to generalize and identify complex patterns. This limitation was further investigated by conducting a Dickey-Fuller test for each

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

time series to check for stationarity. The results confirmed that the time series for the vehicle in question was stationary, suggesting that the issue lies not in the nature of the data but possibly in the amount available for the model to learn from.

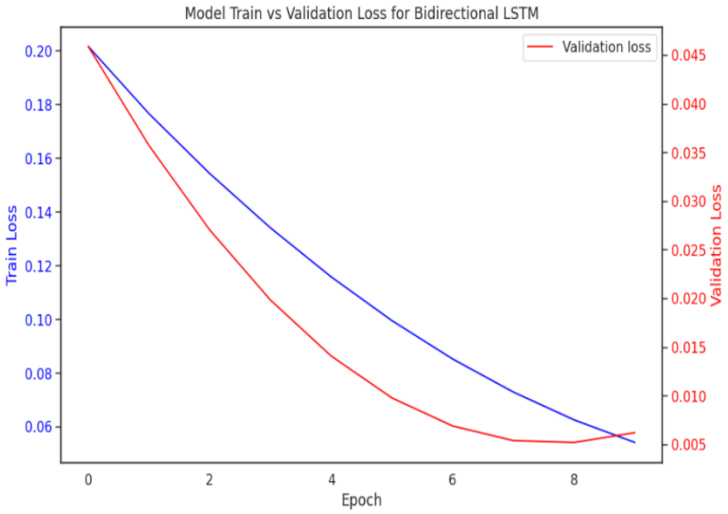

Figure 4. Loss comparison between training and validation set

Source: authors calculations

The observed pattern in the loss graph (Figure 4), where both training and validation loss decrease initially, with validation loss reducing at a faster rate before beginning to rise in the later epochs and eventually intersecting with the training loss, suggests a phenomenon known as overfitting.

At the start, as the model learns from the training data, it is expected that the loss metrics for both the training and validation sets decrease. This indicates that the model is improving its predictive accuracy. However, as training progresses, the model may start to learn not only the underlying patterns in the data but also the noise and random fluctuations specific to the training set.

When the validation loss starts to increase while the training loss continues to decrease, it implies that the model's adjustments are too specific to the training data, and its generalizability is suffering. This is where the model has essentially memorized the training data rather than learning to generalize from it. As a result, its performance on the validation set—representing new, unseen data—begins to worsen.

The intersection of the training and validation loss curves towards the later epochs further confirms that the model's performance on the training set is no longer a reliable indicator of its performance on unseen data. This calls for early stopping, regularization techniques, or adjustments to the model's complexity to prevent overfitting and to ensure that the model remains robust and performs well on both the training and validation sets.

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

Figure 5 is presenting the hystory of CO_2 emissions for all 9 cars analyzed and the estimation for the future three weeks. The estimation is an aggregate result off all the LSTM forecasts at the car level.

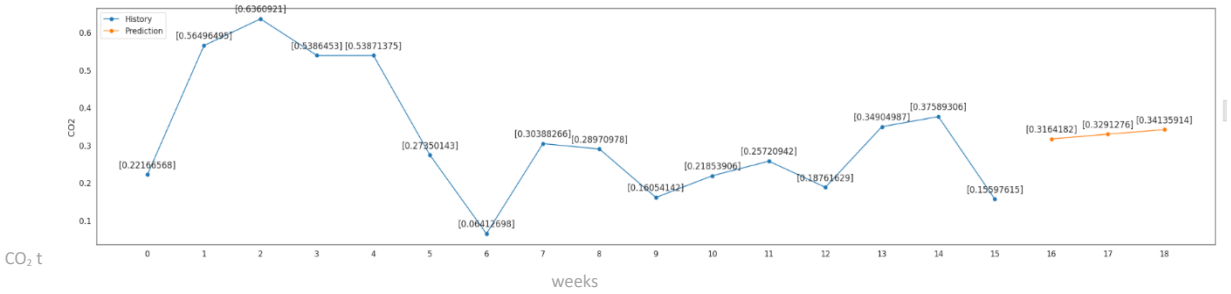

Figure 5. Time series for CO_2 History at the entire fleet analyzed plus future estimation for the next 3 weeks

Source: authors calculations, based on carburant consumption

The segmentation analysis at the route level involved a detailed examination of CO_2 emissions. These variables included the quantity of CO_2 emitted per route and per vehicle, the additional quantity of CO_2 emitted in the same context, the number of vehicles utilizing each route, and the frequency of route usage on a weekly basis.

With a high silhouette score of 0.72 for three clusters, the segmentation process suggested a robust distinction between the different groups. This was further corroborated by ANOVA results, where a p-value significantly less than 0.05 indicated substantial differences among the clusters.

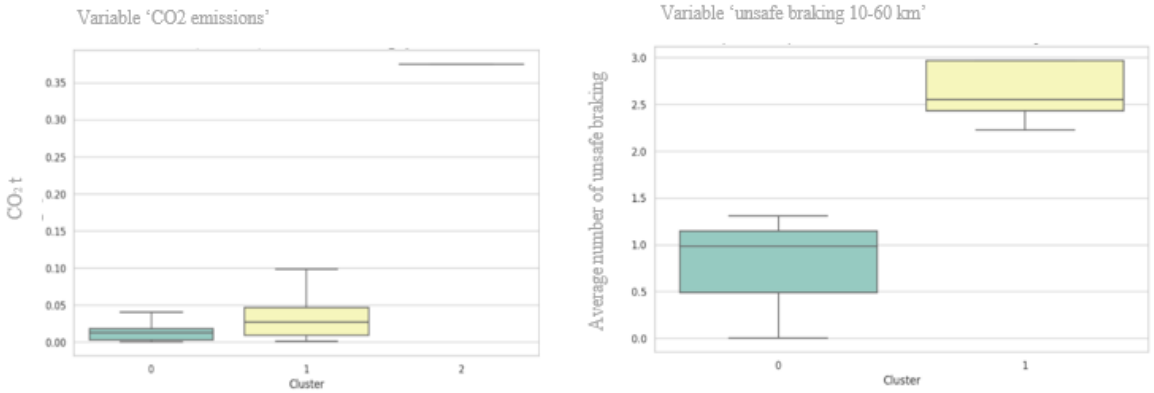

Figure 6. Boxplots for the distinct clusters

Source: authors calculations

To visually confirm and explore these differences, boxplot analyses were conducted for each cluster. These boxplots (Figure 6) were instrumental in highlighting the variations in CO_2

Eastern European Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

emissions among the clusters. For example, the analysis of CO_2 emissions across clusters using boxplots revealed significant disparities, indicating that some routes or vehicles were associated with higher emissions.

In the segmentation analysis focused on driver behavior, a comprehensive set of variables (Table 1) was employed to categorize drivers based on their driving safety metrics. The variables used for segmentation included the average, sum, and count of unsafe braking events at different speed intervals (10-60 km/h, 60-110 km/h, and over 110 km/h), as well as the average, sum, and count of corners taken dangerously within the same speed brackets. Additionally, metrics for unsmooth trips, which include sudden movements or jolts within medium and high-speed ranges, were also considered.

Table 1. Some variables used for driver behavior segmentation (number/car)

Car ID	AVG Unsafe braking 60 - 110 km/h	AVG Unsafe braking > 110 km/h	AVG Corner taken dangerously 60 - 110 km/h	AVG Unsmooth trip > 110 km
0001	0.14	0.00	0.13	0.07
0002	0.16	0.00	0.08	0.21
0003	1.46	0.26	1.42	1.62
0004	0.97	1.77	0.70	1.48
0005	0.00	0.00	0.00	0.00
0006	0.56	0.26	0.88	1.59
0007	1.39	0.45	0.43	2.33
0008	0.53	0.30	0.25	1.34

Source: authors calculations, based on IoT data

With a silhouette score of 0.36 for two clusters, the analysis effectively divided the drivers into groups with distinct driving behaviors. Particularly, it highlighted a segment of drivers who demonstrated unsafe driving practices that could negatively impact vehicle maintenance. Notably, all rented and service vehicles fell into this category, suggesting a trend that these vehicles tend to be driven less carefully.

Boxplots were also used to illustrate the differences between clusters (Figure 6), offering a visual representation of the data distribution for each variable. For example, the variable 'average unsafe braking between 10 and 60 km/h' was depicted in a boxplot, revealing significant variations between the two clusters. This visualization underscores the clusters' behavioral differences, with one showing a higher propensity for unsafe braking, thus requiring targeted interventions to improve safety and reduce potential maintenance issues.

Eastern European
Journal of Regional Studies
ISSN: 1857-436X / ISSN: 2537-6179
Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

Conclusions

The study's conclusions are diverse, spanning various models and analyses. They encapsulate the nuanced outcomes of different machine learning applications in the study, ranging from the descriptive analysis of driver behavior clusters to the intricate details of CO_2 emission segmentation. The findings also delve into the specific challenges and limitations encountered, such as the limited performance of certain predictive models on smaller datasets. Overall, these conclusions offer a multifaceted view of the study's impact and its contributions to understanding vehicle and driver behavior dynamics.

The cluster-level descriptive analysis, utilizing boxplots and ANOVA, played a crucial role in our study. It enabled to discern of significant behavioral differences among various driver groups and conducted to the conclusion that the clusterisation algorithms and the chosen number of distinct clusters was significant. By analyzing these clusters, we identified distinct driving behaviors and their corresponding carbon emission variabilities. This analysis not only highlighted the diversity in driving patterns but also provided valuable insights into the environmental impact of these behaviors, emphasizing the importance of tailored strategies for emissions reduction and efficient driver management.

The study's use of machine learning algorithms to address unwarranted fuel consumption was particularly revealing. It identified a specific segment of drivers who habitually leave their engines running when idle, contributing significantly to fuel siphoning. This discovery highlights the urgent need for comprehensive strategies aimed at curtailing such harmful practices. Effective measures are required not only to prevent fuel wastage and misappropriation but also to promote more environmentally responsible behavior among drivers, thereby enhancing overall sustainability in vehicle usage.

The study's CO_2 emissions segmentation through algorithms revealed a pivotal aspect: a particular vehicle cluster demonstrated significantly higher CO_2 emissions, despite comparable mileage to other groups. This cluster, characterized by more unique location visits, stood out for its disproportionately high additional CO_2 emissions. This finding underscores the complexity of emission patterns, highlighting that factors beyond mileage, like location frequency, play a critical role in environmental impact. It brings to light the need for more nuanced approaches to addressing vehicle emissions.

The bidirectional LSTM model used in our study, while demonstrating high effectiveness in analyzing larger datasets, encountered limitations when applied to smaller datasets. Its performance, as indicated by an R^2 value not exceeding 20% at the individual vehicle level, suggests that its predictive accuracy diminishes with the reduction in data size. This outcome highlights the need for model adjustments or alternative approaches when dealing with less voluminous data, ensuring that the predictive quality is maintained across different dataset sizes.

Volume 10 / Issue 1 / June 2024
DOI: https://doi.org/10.53486/2537-6179.10-1.05

The route segmentation aspect of our study unveiled a significant pattern: certain routes were consistently used more frequently and were associated with markedly higher CO_2 consumption. This pattern points to unnecessary fuel wastage or inefficient fuel use along these routes, raising important questions about the causes of such excessive consumption and the potential need for targeted interventions to optimize fuel efficiency. This insight not only highlights specific areas requiring immediate attention but also underscores the need for more targeted measures to address and mitigate such irregularities in CO_2 emissions.

The study's driver segmentation revealed a distinct group characterized by potentially aggressive driving behaviors. This particular cluster, predominantly comprising individuals using rented or company vehicles, showed tendencies that could lead to adverse effects on vehicle maintenance. This finding is crucial for understanding the implications of driving behaviors on vehicle wear and tear, highlighting the need for targeted interventions or education programs, especially for drivers of rental and company-owned vehicles.

The research faced several limitations which could impact the breadth and depth of its conclusions. The small sample size due to IoT devices installed in a limited number of vehicles may affect the generalizability of the results across the entire fleet. Relying on an external CO_2 calculator and the absence of IoT measurements for pollutants like NOx may introduce discrepancies in pollution emission assessments. Additionally, the lack of driver identification data, such as gender and age, restricts the analysis of how these factors may influence driving behavior and emissions.

For future developments, this study envisages a holistic expansion across the entire fleet by incorporating IoT devices into all vehicles, thereby enhancing data representativeness and result generalization. Direct measurements of pollutants through specialized IoT sensors will offer accurate pollution emission data. The integration of maintenance records will augment telematic predictors with technical vehicle health data. Additionally, introducing sensors to monitor driver behavior, like driving style and habits, will yield further insights to elevate efficiency and reduce emissions.

As a proposed further development of this study, incorporating a meteorological dimension could significantly enhance the analysis. Recognizing a potential correlation between weather conditions and the incidence of external impacts on vehicles, this extension would involve overlaying weather data with the timing and characteristics of recorded impacts. The aim would be to uncover patterns that might indicate weather-related vulnerabilities, thus providing a more complete understanding of the factors influencing driving behavior and vehicle safety.

This proposed inclusion of meteorological data aspires to lead to more informed decisions regarding fleet management, driver training programs, and vehicle maintenance schedules. This exploration of the interplay between weather conditions and driving incidents

could offer valuable insights, potentially guiding strategic adjustments in fleet operations to mitigate risks associated with adverse weather.

Acknowledgments

This work was supported by the project "Societal and Economic Resilience within multi-hazards environment in Romania" funded by European Union – Nextgeneration EU and Romanian Government, under National Recovery and Resilience Plan for Romania, contract no.760050/ 23.05.2023.

This paper was financed by the Bucharest University of Economic Studies during the PhD program.

References

Hasan, G. M. S., Timothy, A. B., Surawski, N., Md Komol, M. R., Sajjad, M., Thuy Chu-Van, Ristovski, Z., & Brown R. J. (2023). Real-driving CO_2, NOx and fuel consumption estimation using machine learning approaches. *Next Energy*, *1*(4). https://doi.org/10.1016/j.nxener.2023.100060

Roussou, S., Garefalakis, T., Michelaraki, E., Brijs, T., & Yannis, G. (2024). Machine Learning Insights on Driving Behaviour Dynamics among Germany, Belgium, and UK Drivers. *Sustainability*, *16*(2), 518. https://doi.org/10.3390/su16020518

Jain, N., & Mittal, S. (2022). Bayesian Nash Equilibrium based Gaming Model for Eco-safe Driving. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7482-7493. https://doi.org/10.1016/j.jksuci.2021.07.004

Brunheroto, P. H., Pepino, A. L. G., Deschamps, F., & Loures, E. F. R. (2022). Data analytics in fleet operations: A systematic literature review and workflow proposal. *Procedia CIRP*, *107*, 1192-1197. https://doi.org/10.1016/j.procir.2022.05.130

Yuan, Y., Wang, X., Calvert, S., Happee, R., & Wang, M. (2022). A risk-based driver behaviour model. *IET Intell. Transp. Syst*, *18*(1), 88-100. https://doi.org/10.1049/itr2.12435

Kumar, R., & Jain, A. (2022). Driving Behaviour Analysis and Classification by Vehicle OBD Data Using Machine Learning. *PREPRINT (Version 1).* https://doi.org/10.21203/rs.3.rs-2353524/v1

Delussu, F., Imran, F., Mattia, C., & Meo, R. (2021). Fuel Prediction and Reduction in Public Transportation by Sensor Monitoring and Bayesian Networks. *Sensors*, *21*(14). https://www.mdpi.com/1424-8220/21/14/4733

Islam, M., Ahmed, M., & Begum, S. (2023). Interpretable Machine Learning for Modelling and Explaining Car Drivers' Behaviour: An Exploratory Analysis on Heterogeneous Data. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence.* (Vol. 2, pp. 392-404). ICAART https://doi.org/10.5220/0011801000003393

Kmail, E., Lail, Y., Sawalha, R., Mousa, M., & Eid, D. (2023). *Driving Behavior Classification at Highways Using Vehicle Kinematics: Application of Unsupervised Machine Learning.* https://doi.org/10.13140/RG.2.2.31961.70243.

Khanfar, N., Elhenawy, M., Ashqar, H., Hussain, Q., & Alhajyaseen, W. (2022). Driving behavior classification at signalized intersections using vehicle kinematics: Application of unsupervised machine learning. *International Journal of Injury Control and Safety Promotion, 30*. 1-11. https://doi.org/10.1080/17457300.2022.2103573.

Ziryawulawo, A., Kirabo, M., Mwikirize, C., Serugunda, J., Mugume, E., & Miyingo, S. P. (2023). Machine learning based driver monitoring system: A case study for the Kayoola EVS. *SAIEE Africa Research Journal*, *114*(2), 40-48. https://doi.org/10.23919/SAIEE.2023.10071976.

Daoudi, N., Smail, Z., & Aboussaleh, M. (2023). Machine Learning Based Predictive Maintenance: Review, Challenges and Workflow. *Artificial Intelligence and Industrial Applications.* https://doi.org/10.1007/978-3-031-43524-9_6.

Mohammd, T., Almsre, A., & Ashqar, H. (2022). *Effect of roundabout design on the behavior of road users: A case study of roundabouts with application of Unsupervised Machine Learning*. https://doi.org/10.13140/RG.2.2.19955.50723.

Aljohani, A. (2023). Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility. *Sustainability*, *15*(20). https://www.mdpi.com/2071-1050/15/20/15088

Almallah, M., Hussain, Q., Alhajyaseen, W., & Brijs, T. (2020). *Improved Road Safety at Work Zones using Advanced Traveler Information Systems.* https://doi.org/266. 10.29117/quarfe.2020.0243.

Arumugam, S., Bhargavi, R. (2023). Road Rage and Aggressive Driving Behaviour Detection in Usage-Based Insurance Using Machine Learning. *International Journal of Software Innovation (IJSI)*, 11(1), 1-29. http://doi.org/10.4018/IJSI.319314

Khanfar, N., Ashqar, H., Elhenawy, M., Hussain, Q., Hasasneh, A., & Alhajyaseen, W. (2022). Application of Unsupervised Machine Learning Classification for the Analysis of Driver Behavior in Work Zones in the State of Qatar. *Sustainability*, *14*(22). https://doi.org/10.3390/su142215184.

Tawanda, T., Shamuyarira, S. B., & Munapo, E. (2023). Truck Fuel Consumption Prediction Using Logistic Regression and Artificial Neural Networks. *Int. J. Operat. Res. Inf. Syst*, *14*(1), 1–17. https://doi.org/10.4018/IJORIS.329240

Musril, H., Saludin, F., Winci, U., Kundori, K., & Rahim, R. (2023). Using k-NN Artificial Intelligence for Predictive Maintenance in Facility Management. *International Journal of Electrical and Electronics Engineering*, *10*. 1-8. https://doi.org/10.14445/23488379/IJEEE-V10I6P101.