Journal of Computer Science

Original Research Paper

Fuel Prediction Model for Driving Patterns Using Machine Learning Techniques

[1]Manjunath TK and [2]Ashok Kumar PS

[1]Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bangalore, India
[2]Department of Computer Science and Engineering, ACS College of Engineering, Bangalore, India

Article history
Received: 15-08-2023
Revised: 18-10-2023
Accepted: 27-12-2023

Corresponding Author:
Manjunath TK
Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bangalore, India
Email: itsme.tkm@gmail.com

Abstract: In recent years, the demand for fuel efficiency has become a crucial aspect of the automotive industry. Predicting fuel economy accurately is essential for optimizing vehicle performance and reducing environmental impact. This proposed model presents a machine learning-based approach for developing a fuel prediction model using the real dataset. The model aims to predict fuel efficiency for different drivers by considering input features related to driving conditions and driver behavior. The study explores the application of linear regression and support vector regression, to achieve accurate and reliable predictions. The dataset is pre-processed to handle missing values, normalize numerical features, and encode categorical variables. Feature engineering techniques are employed to select the most relevant features and enhance the model's performance. A thorough assessment is carried out utilizing diverse performance measures, including mean squared error and R-squared score, to evaluate the forecasting aptitude of the created models. The linear regression model exhibits exceptional performance, as evidenced by its high R-squared values (0.9963%, approaching 1) and low values for MAE, MSE, and RMSE. The outcomes illustrate the efficacy of the suggested method in precisely forecasting fuel economy for varying drivers for even real-time values. The findings provide valuable insights for vehicle manufacturers, policymakers, and individuals interested in optimizing fuel consumption and reducing greenhouse gas emissions. Overall, this study contributes to the growing body of knowledge in machine learning techniques and reinforces the significance of machine learning techniques in addressing fuel economy challenges with driver's behaviors.

Keywords: Greenhouse Gas Emissions, Fuel Prediction, Fuel Economy, Machine Learning, Performance Evaluation, Linear Regression, Support Vector Regression, R-Square, Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE)

Introduction

The introduction provides an overview of the importance of fuel economy prediction in the automotive industry. It highlights the motivation behind the research, the significance of accurate fuel prediction models, and the potential benefits of optimizing fuel efficiency.

The topic of fuel consumption deviations sparks intense discussions. Automakers provide conservative benchmarks, drivers may exhibit inefficient driving habits, traffic congestion exacerbates fuel usage, and various strategies like traffic optimization and driver guidance are proposed to achieve significant fuel savings. To unravel these complexities and establish a baseline for individual driving styles' impact on fuel consumption, diverse methods, and techniques have been employed to monitor individual driving behaviors.

A vehicle's fuel economy during the trips depends on several parameters, important and main parameters are listed below:

- Route/road conditions (traffic lights, roundabouts, lane width, road surface, etc.)
- Congestion and other road users
- Individual driving style

- Vehicle type and condition (mass, engine size, number of gears, tire pressure, etc.)
- Engine specification
- Fuel type
- Ambient conditions (weather)

Most of these aspects require a change in technology, infrastructure, or transport demand to result in lower fuel consumption and therefore lower CO_2 emissions. Partly because it is the only parameter with an immediate impact on CO_2 emission reduction, Individual driving style is a popular fuel-saving measure. However, the potential effect of changes in driving behavior is hard to determine, mainly because a good description of the distribution of average driving behavior is lacking. The current paper reports results from recent studies on European driving behavior and its relation to fuel consumption, such as the eco-driver analysis (Heijne *et al*., 2017) and a TNO study on driving behavior in several European countries (Heijne *et al*., 2016).

Central to the distinction of Individual driving styles is the separation of driving behavior enforced by the environment on the one hand and the remaining personal choices and anticipatory behavior on the other hand. To assess the fuel consumption reduction potential, it is crucial to separate Individual driving style from the other factors mentioned above, of which the infrastructure, speed limits, and degree of congestion are the main influences.

A traffic light or a car ahead may require a driver to brake, but in some cases, braking is unnecessary or can be avoided by keeping more distance or through better anticipation. Within the Udrive project, an effort was made to untwine the different aspects that may lead to higher fuel consumption. If one can successfully do this, the remaining bandwidth of personal driving styles, between the worst and the best-performing driver, determines the potential room for fuel saving related to more economic driving behavior. The three main aspects that are evaluated are braking gear-shifting behavior and speed choice.

The primary objectives of the eco-driver study, the primary objectives were to enhance comprehension regarding the diversity in driving patterns and the roles played by various driving styles in shaping the "average" driving behavior concerning eco-driving. Additionally, the aim was to evaluate the potential reduction in fuel consumption and CO_2 emissions linked to the adoption of an eco-driving approach. This should improve insight into the overall net potential for eco-driving at the national and EU level, by studying different parts of the driver population, different road types and traffic situations, and different vehicle applications.

Eco-driving in the context of this study denotes a driving style associated with low fuel consumption. Some practical examples of an eco-driving style are:

- Shift gear up as soon as possible, between rpm 2000-2500
- Anticipate traffic flow (to minimize dynamics and limit braking)
- Maintain a steady speed
- Decelerate smoothly by coasting

Integrating the findings from the examination of driving styles with the fuel consumption data gathered in other research endeavors enables a meticulous examination of how diverse driving styles influence fuel usage and emissions (Kaggle, 2023). This scrutiny yields a valuable understanding of the mechanisms contributing to the existing environmental impact (CO_2) of the transportation sector within the EU. The paper is divided into the following sections; literature review, data processing, prediction modeling, result and discussion, and conclusion.

Various researchers, while appraising driver fuel consumption, have also delved into the examination of driving conditions, which can be deduced from or directly obtained through real-world driving data by Lee *et al.* (2011), resulting in more intricate evaluations. The influence of external environmental factors on vehicle fuel usage was extensively explored by Ehsani *et al.* (2016), providing a comprehensive understanding of the matter. However, they did not thoroughly analyze the impact of driving behavior, simply citing speed and acceleration as the two most crucial elements. By analyzing real-world data, Rios-Torres *et al*. (2019) divided driving styles into three groups and then they explained how each driving style affected fuel usage (Rios-Torres *et al*., 2019). Depending on the driving circumstance and the driver's driving style, the study's findings demonstrate that vehicle fuel usage can vary significantly from normal US Environmental Protection Agency (EPA) driving cycles.

Although the studies examining the link between driving behavior and fuel consumption stated above have produced positive findings, there are still many open-ended issues. To examine the correlation between driving habits and fuel usage, most of these studies have employed statistical or rule-based techniques. However, these approaches demand extensive volumes of extended driving data and a solid grasp of the statistical attributes of the data. Traditional techniques for extracting insights from raw data usually necessitate specialized expertise. While machine learning approaches also mandate substantial data, they possess the ability to autonomously uncover knowledge and inherent patterns from the raw data. In 2013, a single study estimated that approximately 32% of reported fatalities in China's 74 major cities were linked to ambient PM 2.5 air pollution, accounting for around 1.03

million deaths (Fang *et al*., 2016). As a result, a significant portion of research endeavors has been aimed at reducing automobile emissions. Multiple studies (Liimatainen, 2011; Meseguer *et al*., 2015; D'agostino *et al*., 2014) have highlighted the inherent risks associated with driving.

Several investigations Liimatainen, 2011; Meseguer *et al*., 2015; D'agostino *et al*., 2014) have demonstrated that irrespective of the vehicle type, driving habits such as speed management, favored acceleration rate, and vehicle stability significantly influence fuel consumption. Enhanced Driving Assistant Systems (ADAS) can be designed to offer more accurate and intelligent eco-driving guidance by effectively identifying connections between driving conduct and fuel consumption (Bengler *et al*., 2004; Barkenbus, 2010).

By analyzing the influence of driving styles on fuel consumption, we can identify drivers who consume more energy compared to others. This insight can facilitate the adoption of fuel-efficient driving techniques among high-energy users. Moreover, the driving behavior-energy consumption model, once established, can serve as a fundamental technology within Advanced Driving Assistance Systems (ADAS) or eco-driving coaching systems. This has the potential to decrease fuel expenses for commercial vehicles (Sullman *et al*., 2015), optimize the placement of charging stations (Lee and Wu, 2015b), curtail transportation-related greenhouse gas emissions (Nègre and Delhomme, 2017), and more. To curb vehicle emissions and enhance fuel efficiency, comprehending the precise correlation between driving behavior and fuel consumption is of utmost importance. The motivation for our study stems from this very need. However, research on the influence of driving behavior on fuel consumption is limited. We evaluated the factors affecting driving behavior by analyzing the distributions of these factors derived from prior driving instances. Mathematical modeling was employed to investigate the impact of diverse driving behaviors on fuel efficiency (Javanmardi *et al*., 2017). The study initially categorized driving behavior into two levels: Maneuvering level and control level behavior. Subsequently, the authors simulated three distinct driving behaviors aggressive driving, eco-driving, and typical driving by adjusting various parameters of the model. Their results illustrated that the proposed model effectively replicated actual driving behavior and the corresponding fuel consumption patterns similarly, Lv *et al*. (2018).

In an effort to enhance energy efficiency, a study utilized an unsupervised machine learning approach based on Gaussian mixture models to outline optimal control strategies for each of the three prevalent driving behaviors (Lv *et al*., 2018). While the mentioned studies successfully identified fuel-efficient driving habits, their practicality is limited due to the influence of various static or dynamic environmental factors on driving behaviors (af Wåhlberg, 2007; Ericsson, 2020). To advance the understanding of fuel-efficient driving conduct, our paper introduces a method that employs two machine learning techniques. In the initial phase, we evaluate the driver's fuel economy by recording timestamps using the OBD Tool, building on insights from previous research endeavors (Lee and Wu, 2015a; Qi *et al*., 2015; Constantinescu *et al*., 2010).

Unlike prior research that employed machine learning for driving behavior analysis, our study takes a different approach by utilizing basic clustering techniques to classify the collected driving signal dataset from different drivers. We analyze the drivers driving cycle with respect to their fuel consumption while completing the trip. Nearly 100 driver’s data has been collected in a fixed route; the source destination is fixed, same vehicle, same weather conditions. We conducted an experiment to collect real driving data and used proposed algorithms to predict the fuel efficiency of individual drivers.

The contribution of the work is to select suitable algorithms and evaluate those algorithms to fit into the real world of vehicle automation. In this proposed work we selected linear regression and super vector regression algorithms for predictions.

Therefore, this study presents a novel approach that leverages two machine-learning techniques to advance the understanding of fuel-efficient driving behavior. In the initial phase, we utilize an unsupervised machine learning method to assess driver behavior's fuel efficiency, achieved by recording data through the OBD tool. Building upon the utilization of machine learning for driving behavior analysis in previous studies (Lee and Wu, 2015a; Qi *et al*., 2015; Constantinescu *et al*., 2010), our study employs a straightforward clustering approach to categorize the collected driving signal dataset from multiple drivers. The main objectives of this study are defined below and same way paper is organized into sections:

- Real-driving data collections using the OBD interface
- Filtering the data, preprocessing, and cleaning the data
- Features extractions
- Model building and evaluation
- Results comparisons

Dataset Description and Data Processing

The fuel economy dataset is introduced, outlining its key features and variables. The pre-processing steps, including handling missing values, normalizing numerical features, and encoding categorical variables, are described in detail.

This section discusses the process of feature selection and engineering, where relevant features are identified and engineered to enhance the model's performance.

Various techniques, such as correlation analysis and domain knowledge, are utilized to select the most influential features.

OBD II Standard

The On-Board Diagnostic (OBD) standard originated in the United States with the main purpose of aiding in the detection of engine malfunctions. Its core objective revolves around identifying any rise in harmful gas emissions that surpass acceptable thresholds. This system functions by continually monitoring a range of sensors designed to transmit electrical signals as feedback to the vehicle's central Electronic Control Unit (ECU). These sensors oversee various aspects of engine management, specifically focusing on air/fuel volume detection, which allows the ECU to precisely determine the optimal mixture in real time. Additional sensors, like the oxygen sensor and Mass Air Flow (MAF) sensor, also contribute to the air/fuel mixture regulation. To communicate with a vehicle's ECU, an OBD scanner is employed. This scanner serves as a diagnostic tool for identifying issues within the vehicle's electrical and emission systems. When a malfunction is detected, the ECU stores an error code in its memory, which can then be read and interpreted by the scanner.

The initial OBD standard, OBD-I, was developed to monitor a smaller number of parameters in comparison to OBD-II. With the emergence of fuel-injection systems in the automotive industry, OBD-I primarily focused on detecting faults within engines' ignition, emission, and injection systems. The diagnostic techniques during this period were rudimentary and OBD-I did not establish a benchmark for acceptable emission levels in vehicles. Consequently, conditions such as running too rich or too lean, which contribute to increased fuel consumption, went undetected. Ignition systems at that time were not as advanced and sophisticated as today's standards. Many other non-engine electrical error codes were not encompassed within the standard. Failures were simply indicated through visual warnings to the driver, with the error stored in the ECU's memory.

The subsequent generation of OBD referred to as OBD-II, introduced comprehensive standards for various components, including the diagnostic plug, connector specifications, Diagnostic Trouble Codes (DTCs), and signaling protocols on the Controller Area Network (CAN) bus. Additionally, the standard defined an extensive list of Diagnostic Trouble Codes (DTCs). OBD-II also established parameters that could be monitored and assigned unique codes (Identification IDs) to each Parameter (PID).

The Elm327 scanner is a popular On-Board Diagnostics (OBD-II) device used to interface with a vehicle's onboard computer system. It allows users to access and retrieve diagnostic information, such as trouble codes, sensor data, and various vehicle parameters.

The specific sensor data that can be scanned by an Elm327 scanner on a Suzuki Swift petrol car model from 2018 may vary depending on the car's OBD-II implementation and the software or app being used with the scanner. However, in general, the following sensor data can typically be accessed with an OBD-II scanner:

- Engine Revolutions Per Minute (RPM)
- Vehicle speed
- Throttle position
- Coolant temperature
- Intake air temperature
- Mass Air Flow (MAF) sensor reading
- Oxygen sensor (O_2) readings
- Fuel Pressure
- Fuel trim values
- Engine load
- Timing advance
- Battery voltage
- Emission-related Trouble Codes (DTCs)

Figure 1 shows an example of ELM 237 OBD-II connectors. In this particular device, all the above sensor data has been recorded with the knowledge of the OBD-II (Autocom, 2020) device that offers a connection between the vehicle’s internal bus and a personal computer/laptop using a Bluetooth connection.

Table 2 shows a list of some sensor data recorded during live driving, the data set may have null values, outliers, and a few unwanted values.

Dataset Description

Actual driving data is collected using ELM 327 OBD-2 Tool, which is connected to a car and through a laptop and collected nearly 13 Sensor data. The details of data fields and values are described in Tables 1-2. The Raw data may have null values, Nan Values, and missing values, due to continued recording from sensors for the car. The default frequency of recording the data through the tool is taken default.

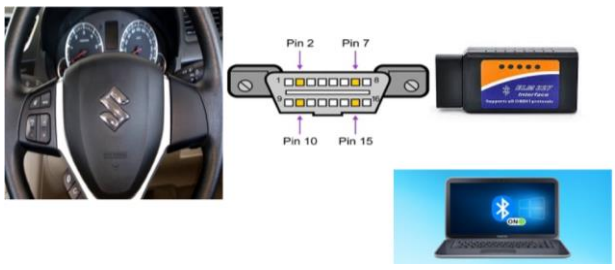


Fig. 1: ELM 237 is interfaced with the Swift car and connected to the laptop via Bluetooth

Table 1: Sensor sample data values

ENG_C -TEMP	FUEL _LEV	ENG _LOAD	AMB_AIR _TEMP	ENG _RPM	IN _MP	MAF	AIR_In _TEMP	SPEED	STFT _BANK	TPS	TIMING _ADV	MPG
95.00	33.30	39.60	17.00	784.00	34.00	4.35	31.00	0.00	-7.80	20.40	56.10	0.35
96.00	32.20	51.00	17.00	762.00	27.00	1.69	31.00	6.00	-3.10	31.80	53.30	0.86
96.00	32.30	34.90	17.00	795.00	38.00	3.55	31.00	0.00	-1.60	19.20	62.00	0.29
96.00	33.70	39.60	17.00	797.00	32.00	453.00	31.00	0.00	0.00	20.80	5220.00	0.36
97.00	34.50	0.00	17.00	0.00	100.00	0.00	32.00	0.00	0.00	23.90	63.10	0.00

Table 2: List of field names and their types

#	Column	Non-null count	Dtype
0	V_ID	1743 non-null	object
1	ENG_C_TEMP	1741 non-null	float64
2	FUEL_LEV	1741 non-null	float64
3	ENG_LOAD	1738 non-null	float64
4	AMB AIR_TEMP	1740 non-null	float64
5	ENG_RPM	1741 non-null	float64
6	IN MP	1743 non-null	int64
7	MAF	1735 non-null	float64
8	AIR_IN_TEMP	1716 non-null	float64
9	SPEED	1655 non-null	float64
10	STFT_BANK	1736 non-null	float64
11	TPS	1732 non-null	float64
12	TIMING ADV	827 non-null	float64
13	MPG	1743 non-null	float64

dtypes: Float64 (12), int64 (1), object (1)

Data cleaning is a pivotal phase in the machine learning workflow, encompassing the identification and rectification of errors, disparities, and inaccuracies within a dataset. This process aims to enhance the overall quality and dependability of the data, as clean data forms the foundation for constructing precise and resilient machine learning models. Here are several widely used data-cleaning techniques (Analytics, 2023). Table 3 describes the complete Information about all the fields and their counts. The list of all field individual values of min, max mean, etc., are shown.

Handling missing values: Missing data can lead to biased or inaccurate results. You can deal with missing values by either removing the rows or columns with missing data or imputing the missing values using techniques like mean, median, and mode, or more sophisticated methods like interpolation or machine learning-based imputation.

Outlier detection and treatment: Outliers are data points that deviate significantly from the rest of the data. They can negatively impact model performance. Detecting outliers and either removing them or applying transformations to mitigate their impact is essential.

Data formatting: Ensure that the data is in the correct format and data types (e.g., numerical, categorical, dates) to match the requirements of the machine learning algorithm.

Standardization and normalization: Scaling the data can be necessary, especially when using algorithms sensitive to the scale of the features. Standardization (z-score scaling) and normalization (scaling between 0 and 1) are common techniques for this purpose.

Handling duplicates: Identify and remove any duplicate records in the dataset, as they can skew the results and cause issues during modeling.

Removing irrelevant features: If some features do not contribute meaningful information to the model, consider removing them to reduce noise and improve the model's efficiency.

Encoding categorical variables: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding to make them compatible with machine learning algorithms.

Data partitioning: Split the dataset into training, validation, and test sets to evaluate the model's performance accurately.

Data cleaning is an iterative process and it's essential to examine the data thoroughly, understand its characteristics, and make informed decisions based on the domain knowledge. Properly cleaned data can lead to more accurate and robust machine learning models, ultimately enhancing their performance and usefulness in real-world applications.

After applying all the methods data is cleaned and ready for building models. The Processed data set will be the samples for training and testing the models. Using processed data set features may be selected for further process.

Data Features Selection

Data feature selection is an essential step in the process of machine learning. It involves choosing a subset of relevant and significant features from the original set of variables in our dataset. The goal of feature selection is to improve the model's performance by reducing the number of input features while retaining the most informative and discriminative ones. Here are some common methods for data feature selection in machine learning.

Heat maps are graphical representations of data where individual values are represented as colors on a 2D grid. The intensity of the color represents the magnitude of the values. Heat maps are commonly used to visualize correlation matrices, where each cell in the heat map represents the correlation coefficient between two variables.

In the context of feature selection, heat maps can help identify strong correlations between different features. If two features are highly correlated, it might indicate redundancy and one of them could potentially be removed during feature selection to avoid multico linearity in the model.

DOI: 10.3844/jcssp.2024.291.302

Table 3: List of each field and its insights

ENG_C_ TEMP ----------		_LEV	FUEL _LOAD	ENG _TEMP	AMB_AIR _RPM	ENG _MP	IN MAF	_TEMP	Air_in Speed	_Bank	STFT TPS	_ADV	Timing MPG
Count	1741.00	1741.00	1738.00	1740.00	1741.00	1743.00	1735.00	1716.00	1655.00	1736.00	1732.00	827.00	1743.00
Mean	85.90	51.53	48.37	5.90	1756.31	43.16	15.97	31.66	41.12	1.02	30.89	63.50	1.28
STD	4.30	5.28	27.38	3.57	1010.74	23.19	20.90	1.81	29.31	4.61	16.11	5.70	1.68
Min	79.00	31.80	0.00	3.00	0.00	16.00	3.00	28.00	0.00	-16.40	17.30	46.30	0.00
25%	83.00	49.00	23.50	4.00	1046.00	24.00	4.34	30.00	23.00	-1.60	20.80	60.30	0.35
50%	8.00	52.90	40.80	4.00	1518.00	35.00	795.00	31.00	37.00	0.00	26.70	63.10	0.65
75%	88.00	54.90	62.40	7.00	2066.00	52.00	15.00	33.00	56.00	3.90	31.40	67.10	1.21
Max	100.00	60.80	100.00	17.00	5994.00	102.00	110.39	38.00	159.00	23.40	83.50	82.40	8.89

Table 4: Raw data set having null values

V_ID	0
ENG-C-TEMP	2
FUEL-LEV	2
ENG_LOAD	5
AMB_AIR_TEMP	3
ENG_RPM	2
IN_MP	0
MAF	8
AIR_IN_TEMP	27
SPEED	88
STFT_BANK	7
TPS	11
TIMING_ADV	916
MPG	0
type: int64	

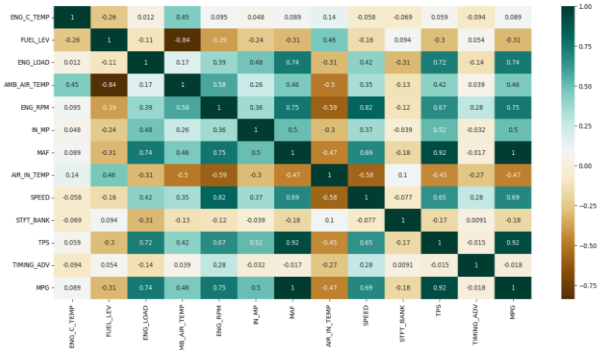


Fig. 2: Heat maps show relationships between the fields

Table 4 shows the number of recorded features, while driving vehicle the features values may be null values. This shows how many null values are recorded in each field shown clearly.

Table 5 and Fig. 2, show the heat map, we can find that the MPG feature depends mainly on the ENG_LOAD, ENG_RPM, MAF, SPEED, and TPS.

Histograms are graphical representations of the distribution of a single numerical variable. The data is divided into bins or intervals and the height of each bar in the histogram represents the frequency or count of data points falling within that bin.

Histograms provide a clear visualization of the data's central tendency, spread, and shape of the distribution. They help identify patterns such as whether the data is symmetric, skewed to the left or right, or multimodal.

In feature selection, histograms can assist in understanding the distribution of each feature. Features with similar distributions or those that are heavily skewed might be candidates for further investigation or transformation as shown in Fig. 3.

Based on the above information, the following features are selected for building the model. these features are closely related to each other by targeting the MPG values.

There are five important features are selected, the distribution between the features are shown in Fig. 4 for these features ENG_LOAD, ENG_RPM, SPEED, MAF, and TPS.

Prediction Modeling

As we discussed in the above sections, we have proposed two models, the implementations with respect to our preprocessed data sets and selected features model have been designed. The models are working as per the expectations.

Support Vector Regression (SVR)

It is a type of supervised machine-learning algorithm used for regression tasks. SVR is an extension of Support Vector Machines (SVM), which are primarily used for classification tasks. The key idea behind SVR is to find a hyperplane in a high-dimensional feature space that best fits the training data, while still allowing a specified amount of error or tolerance.

The goal of SVR is to build a model that can predict a continuous target variable (i.e., a real-valued output) based on input features. Unlike traditional regression methods, SVR aims to find a function that best fits the training data while controlling the error within a certain range, margin, and tolerance.

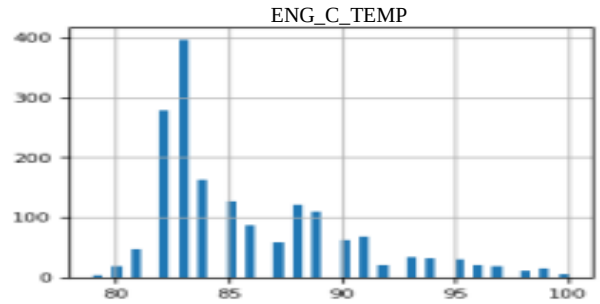
In SVR, there are two hyperparameters that control the margin and tolerance around the regression line. The margin is the region between the two hyperplanes and data points within this region are considered support vectors. The tolerance is the maximum allowed deviation from the target values for the support vectors.

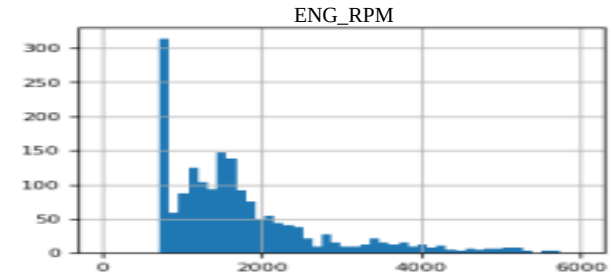
SVR uses the kernel trick to transform the input features into a higher-dimensional space, which allows the algorithm to find a non-linear decision boundary. Common kernel functions include linear, polynomial, Radial Basis Function (RBF), and sigmoid.

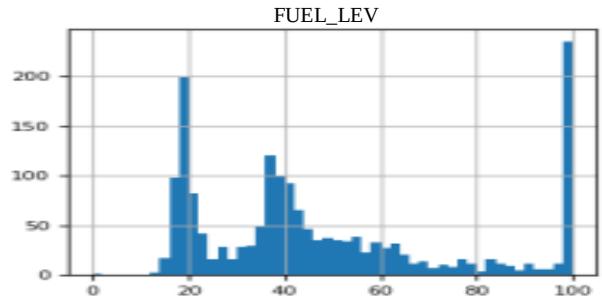
DOI: 10.3844/jcssp.2024.291.302

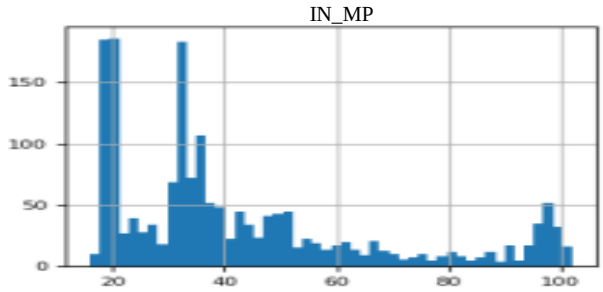
Table 5: Shows the dependent variables

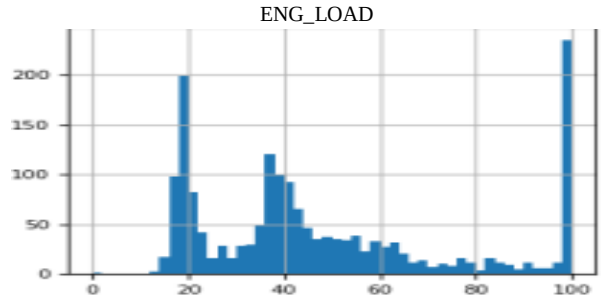
ENG_C_ TEMP ----------		FUEL _LEV	ENG _LOAD	AMB_AIR _TEMP	ENG _RPM	IN _MP	MAF	Air_in _TEMP	Speed	STFT _Bank	TPS	Timing _ADV	MPG
ENG_C TEMP	1.00	-0.26	0.01	0.45	0.10	0.05	0.09	0.14	-0.06	-0.07	0.06	-0.09	0.09
FUEL _LEV	-0.26	1.00	-0.11	-0.84	-0.39	-0.24	-0.31	0.46	-0.16	0.09	-0.30	0.05	-0.31
ENG _LOAD	0.01	-0.11	1.00	0.17	0.39	0.48	0.74	-0.31	0.42	-0.31	0.72	-0.14	0.74
AMB _AIR _TEMP	0.45	-0.84	0.17	1.00	0.58	0.26	0.46	-0.50	0.35	-0.13	0.42	0.04	0.46
ENG _RPM	0.10	-0.39	0.39	0.58	1.00	0.36	0.75	-0.59	0.82	-0.12	0.67	0.28	0.75
IN_MP	0.05	-0.24	0.48	0.26	0.36	1.00	0.50	-0.30	0.37	-0.04	0.52	-0.03	0.50
MAF	0.09	-0.31	0.74	0.46	0.75	0.50	1.00	-0.47	0.69	-0.18	0.92	-0.02	1.00
Air_in _TEMP	0.14	0.46	0.31	-0.50	-0.59	-0.30	-0.47	1.00	-0.58	0.10	-0.45	-0.27	-0.47
Speed	-0.06	-0.16	0.42	0.35	0.82	0.37	0.69	-0.58	1.00	-0.08	0.65	0.28	0.69
STFT _Bank	-0.07	-0.09	0.31	-0.13	0.12	-0.04	-0.18	0.10	-0.08	1.00	-0.17	0.01	-0.18
TPS	0.06	-0.30	0.72	0.42	0.67	0.52	0.92	-0.45	0.65	-0.17	0.01	-0.01	0.92
Timing _ADV	-0.09	0.05	-0.14	0.04	0.28	-0.03	-0.02	-0.27	0.28	0.01	-0.01	1.00	-0.02
MPG	0.09	-0.31	0.74	0.46	0.75	1.00	1.00	-0.47	0.69	-0.18	0.92	-0.02	1.00

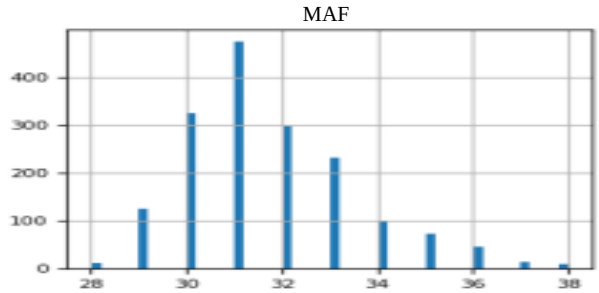


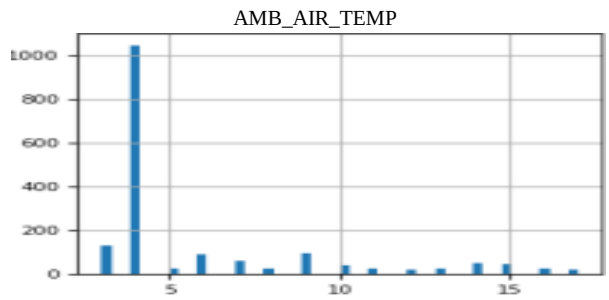


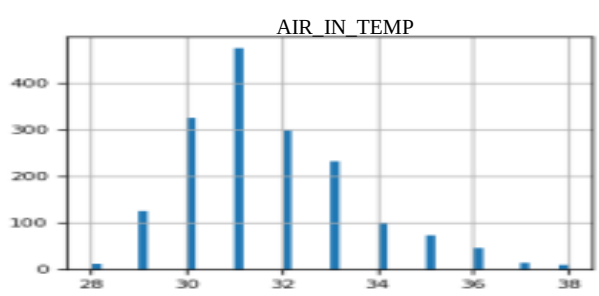












DOI: 10.3844/jcssp.2024.291.302

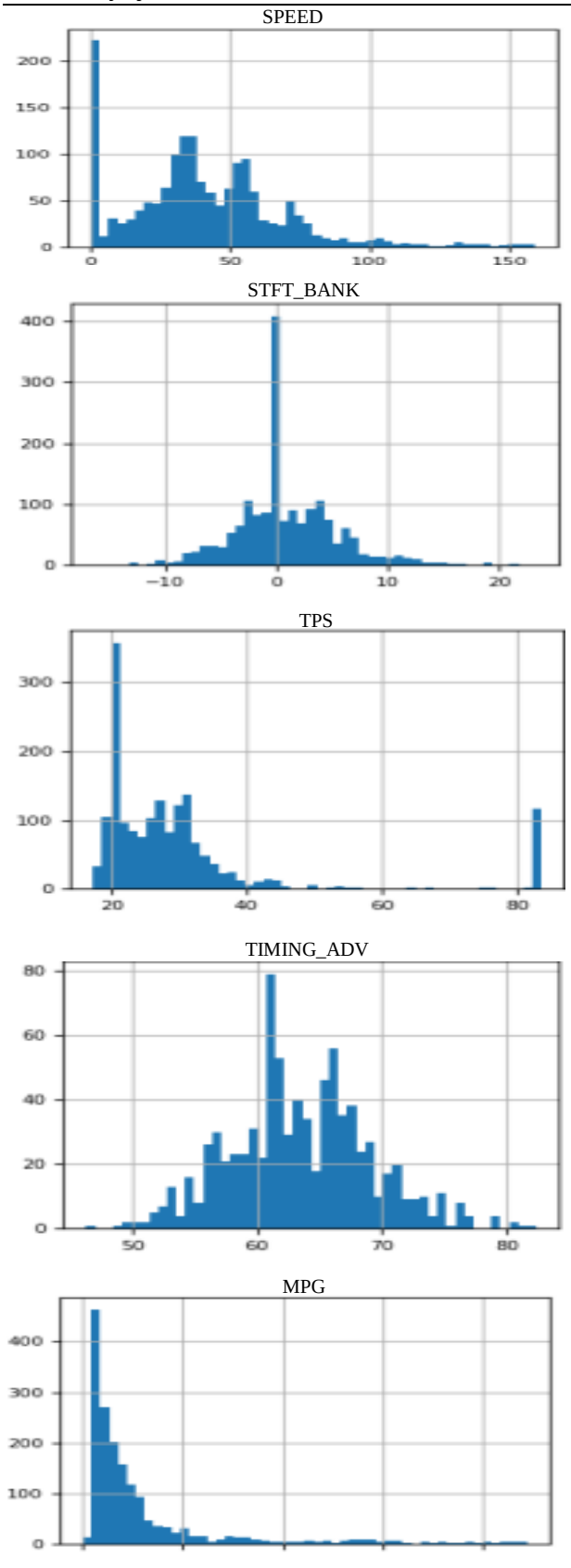


Fig. 3: Histogram shows density between the fields

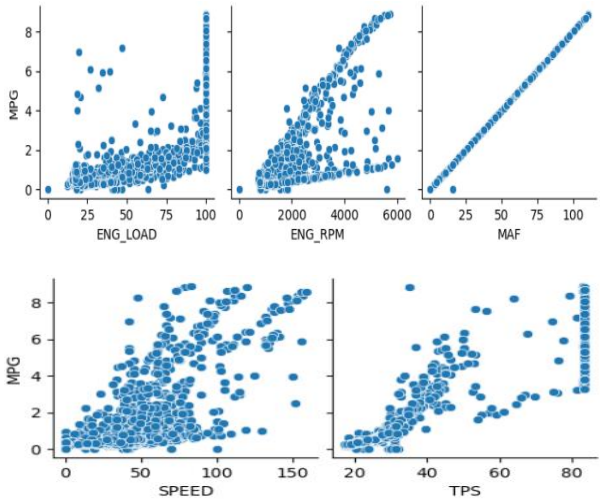


Fig. 4: Histogram shows the distribution of selected features

Train the SVR model on the preprocessed training data. The training process involves finding the optimal hyperplane in the transformed feature space that best fits the training data while considering the specified error tolerance (epsilon) and regularization parameter.

Evaluating the SVR Model for all Kernel Types

Evaluate the SVR model on the given data sets, the data sets contain 13 variables and 17000 records. The SVR model has been trained and tested using appropriate evaluation metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2) score, or explained variance score.

Linear Kernel: The linear Kernel simply computes the dot product between two feature vectors:

$$K\left(x_i, x_j\right) = x_i * x_j$$

The results are from SVR Kernel, performance evaluation metrics given below:

Kernel type	:	Linear
Mean absolute error	:	0.36664713011773303
Mean square error	:	0.2767185521332285
Root mean squared error	:	0.5260404472407312
R-squared	:	0.8859718788916411

Polynomial Kernel: The polynomial kernel maps the data into a higher-dimensional space using a polynomial function:

$$K(x_i, x_j) = (\gamma * x_i * x_j + r)\hat{}\, d$$

where:

γ = The kernel coefficient
r = An optional term (usually set to 1)
d = The degree of the polynomial

The results are from SVR Polynomial, performance evaluation metrics given:

Kernel type	: Poly
Mean absolute error	: 0.3260513303770413
Mean square error	: 0.438114129671142
Root mean squared error	: 0.661901903359661
R-squared	: 0.8194651907062166

Radial Basis Function (RBF) Kernel (Gaussian Kernel): The RBF kernel maps the data into an infinite-dimensional space using a Gaussian function:

$$K(x_i, x_j) = exp(-\gamma * \| x_i - x_j \|^ 2)$$

where:
γ = The kernel coefficient (0.1, 1, 10)

The results are from SVR RBF, performance evaluation metrics given:

Kernel type	: rbf
Mean absolute error	: 0.3996687652748071
Mean square error	: 0.7230742749434017
Root mean squared error	: 0.8503377416905601
R-squared	: 0.702040935246073

Sigmoid Kernel: The sigmoid kernel maps the data into a higher-dimensional space using a sigmoid function:

$$K\left(x_i, x_j\right) = tanh\left(\gamma * x_i * x_j + r\right)$$

where:
γ = The kernel coefficient.
r = An optional term (usually set to 0).

The results are from SVR Sigmoid, performance evaluation metrics given:

Kernel type	: Sigmoid
Mean absolute error	: 17.340058867897437
Mean square error	: 1025.5996086515252
Root mean squared error	: 32.02498413194806
R-squared	: -421.6214523112149

Linear Regression

It is a widely used statistical technique for modeling the relationship between a dependent variable and one or more independent variables. In the context of driving patterns and fuel consumption, Linear Regression can help us understand how different driving patterns (e.g., speed, distance, etc.,) affect fuel consumption.

The Linear equation for multiple Linear Regressions is:

$$y = b_0 + b_1 \times 1 + b_2 \times 2 + ... + b_n * x_n$$

Y	=	The target variable (dependent variable)
$x_1, x_2, ..., x$	=	The input features (independent variables)
b_0	=	The y-intercept
$b_1, b_2, ..., b_n$	=	The coefficients (slopes) corresponding to each independent variable

The Linear Regression model is trained using a learning algorithm that adjusts the values of coefficients (weights) to minimize the difference between the predicted values and the actual target values (MPG). The common method for fitting the model is the least squares method, where the sum of the squared residuals (the difference between the actual and predicted values) is minimized.

The linear regression model is trained with the preprocess data sets which are segmented into training data sets and testing data tests.

The results from the linear regression model are shown:

Mean absolute error	: 0.013297705574118742
Mean square error	: 0.009434814423425876
Root mean squared error	: 0.0971329728950261
R-squared	: 0.9962825551703455

Findings

Support Vector Regression (SVR) and Linear Regression (LR) are both regression algorithms used to predict continuous target variables. However, they have distinct differences in their approach, assumptions, and performance characteristics. Let's compare SVR and LR.

Linear Regression is widely used due to its simplicity, interpretability, and efficiency. However, it assumes a linear relationship between the features and the target, which might not hold true for all datasets. In cases where the relationship is non-linear, other regression techniques like polynomial regression and support Vector Regression (SVR) can be more suitable.

Based on the evaluation metrics for each model (linear regression, SVR with different kernels-RBF, Polynomial, and Sigmoid), we can have the following discussions with evaluation metrics shown in Table 6.

Linear regression model: The linear regression model performs exceptionally well, with high values for R-squared (close to 1) and low values for MAE, MSE, and RMSE. These indicate that the model fits the data very well and has low prediction errors shown in Table 6 and the equivalent performance of evaluation is shown in Fig. 9.

The R-squared value of approximately 0.9963 indicates that around 99.63% of the variance in the target variable is explained by the linear relationship with the features.

SVR with Kernel: The SVR with kernel performs reasonably well, with an R-squared value of

approximately 0.8860. This means that around 88.60% of the variance in the target variable is explained by the model's predictions shown in Table 6 and the equivalent performance of evaluation is shown in Fig. 5. The Mean Absolute Error (MAE) of 0.3666 indicates that, on average, the model's predictions deviate from the actual target values by approximately 0.3666 units. The MSE (Mean Squared Error) of 0.2767 and Root Mean Squared Error (RMSE) of 0.5260 indicate that the model's predictions have relatively low errors compared to the target variable. The lower the MSE and RMSE, the better the model's accuracy.

Table 6: Evaluation metrics of both the models

Evaluation metrics	Linear regression	SVR-Kernel	SVR-poly	SVR-rbf	SVR-sigmoid
MAE	0.0133	0.3666	0.3261	0.3997	17.3401
MSE	0.0094	0.2767	0.4338	0.7231	1025.5996
RMSE	0.0971	0.5260	0.6619	0.8503	32.0250
R-square	0.9963	0.8860	0.8195	0.7020	-412.6215

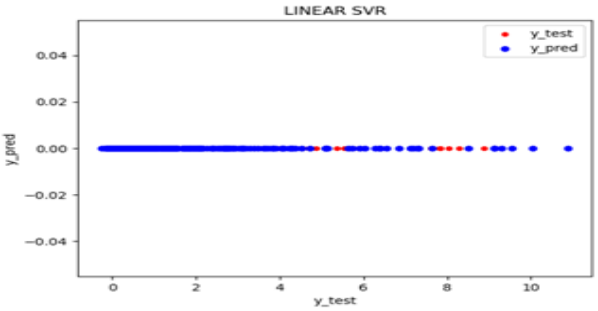


Fig. 5: Performance of SVR-linear model

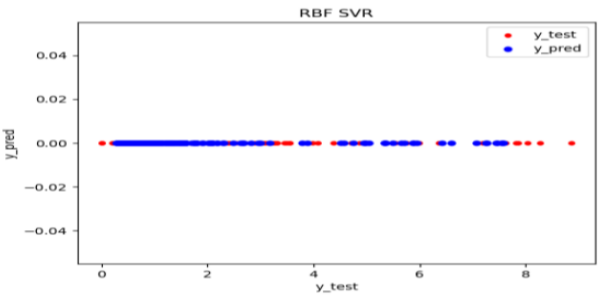


Fig. 6: Performance of SVR-polynomial model

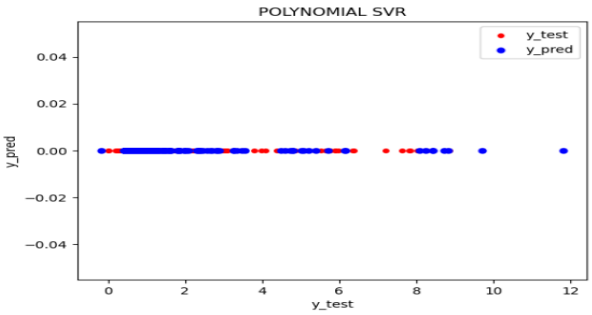


Fig. 7: Performance of SVR-RBF model

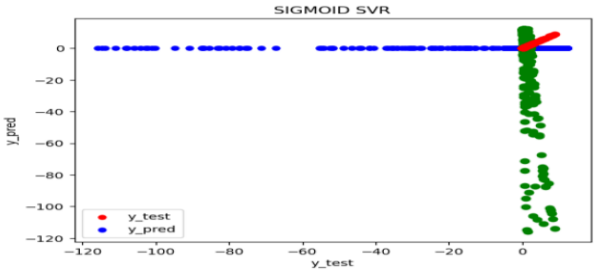


Fig. 8: Performance of SVR-sigmoid model

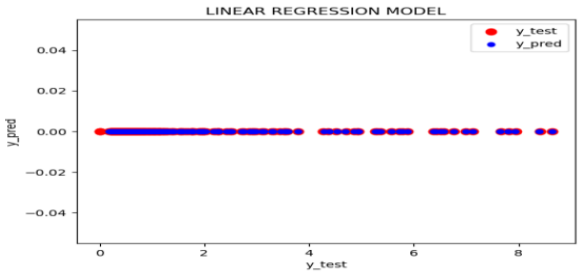


Fig. 9: Performance of linear-regression model

Overall, the SVR with kernel is performing well and providing a good fit to the data. However, it is still slightly performing well and providing a good fit to the data. However, it is still slightly outperformed by the linear regression model.

SVR with RBF Kernel: The SVR with RBF kernel performs relatively well but has higher errors compared to Linear Regression shown in Table 6 and the equivalent performance of evaluation is shown in Fig. 7. The R-squared value of 0.7020 suggests that the model explains around 70.20% of the variance in the target variable. The RMSE value of 0.8503 indicates that the average prediction error is higher compared to linear regression.

SVR with polynomial Kernel: The SVR with Polynomial kernel shows moderate performance shown in Table 6 and the equivalent performance of evaluation is shown in Fig. 6. The R-squared value of 0.8195 suggests that the model explains around 81.95% of the variance in the target variable.

The RMSE value of 0.6619 is lower than the SVR with RBF kernel, but it is still higher than the Linear Regression model.

SVR with the sigmoid kernel: The SVR with Sigmoid kernel performs the worst among all models as shown in Table 6 and the equivalent performance of evaluation is shown in Fig. 8. The R-squared value is negative (-421.6215), which indicates a poor fit of the model to the data.

The extremely high RMSE value of 32.0250 suggests that the predictions are way off from the actual target values, indicating poor model performance.

Comparisons Between LR and SVM Models

The Regression models were evaluated in terms of R-squared, MAE, MSE, and RMSE. It provides the best fit to the data and has the lowest prediction errors.

Among the SVR models, the SVR with RBF kernel performs better than the SVR with polynomial and sigmoid kernels, but it still lags behind linear regression. The SVR with a sigmoid kernel performs the worst among all models, with negative R-squared and very high RMSE, indicating that this SVR with a Sigmoid model is not suitable for the given dataset.

Based on this study, the linear regression model appears to be the most appropriate and effective choice for the given dataset, providing high-quality fit and accurate predictions.

The driving cycles are not classified as city driving, highway driving, or stop and go drive. Fuel consumptions rely on the type of driving cycles also weather conditions and road conditions. In this study only driving inputs are targeted on the targeted car for a duration of 20 min data has been collected for experimental purposes, 13 features are selected for the experiment, and other parameters that influence fuel consumption are assumed to be ideal for all drivers on a fixed route, trip, timings, and minimum traffic.

Conclusion

The Linear Regression model is the most appropriate and effective choice for the real-time driving style dataset. It outperforms all other SVR models, providing high-quality fit and accurate predictions with very low prediction errors. The SVR with kernel and RBF is the second-best performer among the SVR models, but it still lags behind the linear regression model.

The SVR models with Polynomial and Sigmoid kernels are not recommended for this dataset due to their inferior performance for the given data sets.

The linear regression model is effective and accurately predicts the fuel consumption of individual driving styles over the SVR models. As per the experiments conducted, the Linear Regression model is built and evaluated for the real driving data sets of the individual driving patterns.

Acknowledgment

The authors acknowledge the support from the Don Bosco Institute of Technology for the facilities provided to carry out the research.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Manjunath TK: Literature survey, found research gap, data analysis and interpretation of data, model designed, simulated, and drafted the manuscript.

Ashok Kumar PS: Data collection, data analysis and interpretation of data, model testing, and reviewed and edited.

Ethics

It is ensured that all the authors mentioned in the manuscript have agreed to authorship, read and approved the manuscript, and given consent for submission and subsequent publication of the manuscript.

Future Work

Duration of the trip to be increased till the trip completion and driving cycles to be classified with respect to real-time traffic. Drivers are to be selected from different ages and gender for the experiment, to ensure the model efficiency. Maybe features can be reduced by applying PCA techniques to the collected without affecting the results.

References

af Wåhlberg, A. E. (2007). Long-term effects of training in economical driving: Fuel consumption, accidents, driver acceleration behavior, and technical feedback. *International Journal of Industrial Ergonomics*, *37*(4), 333-343. https://doi.org/10.1016/j.ergon.2006.12.003

Analytics, V. (2023). Data Cleaning Using Pandas in Python Complete Guide for Beginners. https://realpython.com/python-data-cleaning-numpy-pandas

Autocom. (2020). CARS highlights for release. https://autocom.se/en/products/cars

Barkenbus, J. N. (2010). Eco-driving: An overlooked climate change initiative. *Energy Policy*, *38*(2), 762-769. https://doi.org/10.1016/j.enpol.2009.10.021

Bengler, K., Dietmayer, K., Färber, B., Maurer, M., Stiller, C., & Winner, H. (2004). Three decades of driver assistance systems. *IEEE Intelligent Transportation Systems*. https://doi.org/10.1109/MITS.2014.2336271

Constantinescu, Z., Marinoiu, C., & Vladoiu, M. (2010). Driving style analysis using data mining techniques. *International Journal of Computers Communications and Control*, *5*(5), 654-663. http://www.unde.ro/mvladoiu/papers/16-driving%20style-IJCCCopen.pdf

D'agostino, C., Saidi, A., Scouarnec, G., & Chen, L. (2014, June). Rational truck driving and its correlated driving features in extra-urban areas. In *2014 IEEE Intelligent Vehicles Symposium Proceedings* (pp. 1199-1204). IEEE. https://doi.org/10.1109/IVS.2014.6856440

Ehsani, M., Ahmadi, A., & Fadai, D. (2016). Modeling of vehicle fuel consumption and carbon dioxide emission in road transport. *Renewable and Sustainable Energy Reviews*, *53*, 1638-1648. https://doi.org/10.1016/j.rser.2015.08.062

Ericsson, E. (2000). Variability in urban driving patterns. *Transportation Research Part D: Transport and Environment*, *5*(5), 337-354. https://doi.org/10.1016/S1361-9209(00)00003-1

Fang, D., Wang, Q. G., Li, H., Yu, Y., Lu, Y., & Qian, X. (2016). Mortality effects assessment of ambient PM2. 5 pollution in the 74 leading cities of China. *Science of the Total Environment*, *569*, 1545-1552. https://doi.org/10.1016/j.scitotenv.2016.06.248

Heijne, V. A. M., Kadijk, G., Ligterink, N., van der Mark, P., Spreen, J., & Stelwagen, U. (2016). *NOx emissions of fifteen Euro 6 diesel cars: Results of the Dutch LD road vehicle emissions testing programme 2016*. Delft: TNO.

Heijne, V., Ligterink, N., & Stelwagen, U. (2017). Potential of eco-driving. *UDRIVE Deliverable 45.1. EU FP7 Project UDRIVE Consortium*. https://doi.org/10.26323/UDRIVE_D45.1

Javanmardi, S., Bideaux, E., Trégouët, J. F., Trigui, R., Tattegrain, H., & Bourles, E. N. (2017). Driving style modelling for eco-driving applications. *Ifac-Papersonline*, *50*(1), 13866-13871. https://doi.org/10.1016/j.ifacol.2017.08.2233

Kaggle. (2023) Data Cleaning Course: Kaggle offers a free course on data cleaning, which covers various techniques and best practices. https://www.kaggle.com/learn/data-cleaning https://doi.org/10.1109/TVT.2005.844685

Lee, C. H., & Wu, C. H. (2015a). A novel big data modeling method for improving driving range estimation of EVs. *IEEE Access*, *3*, 1980-1993. https://doi.org/10.1109/ACCESS.2015.2492923

Lee, C. H., & Wu, C. H. (2015b). An incremental learning technique for detecting driving behaviors using collected EV big data. In *Proceedings of the ASE Big Data and Social Informatics 2015* (pp. 1-5). https://doi.org/10.1145/2818869.2818935

Lee, M. G., Park, Y. K., Jung, K. K., & Yoo, J. J. (2011). Estimation of fuel consumption using in-vehicle parameters. *International Journal of u-and e-Service, Science and Technology*, *4*(4), 37-46. https://www.earticle.net/Article/A167085

Liimatainen, H. (2011). Utilization of fuel consumption data in an ecodriving incentive system for heavy-duty vehicle drivers. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1087-1095. https://doi.org/10.1109/TITS.2011.2142182

Lv, C., Hu, X., Sangiovanni-Vincentelli, A., Li, Y., Martinez, C. M., & Cao, D. (2018). Driving-style-based codesign optimization of an automated electric vehicle: A cyber-physical system approach. *IEEE Transactions on Industrial Electronics*, *66*(4), 2965-2975. https://doi.org/10.1109/TIE.2018.2850031

Meseguer, J. E., Calafate, C. T., Cano, J. C., & Manzoni, P. (2015, January). Assessing the impact of driving behavior on instantaneous fuel consumption. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)* (pp. 443-448). IEEE. https://doi.org/10.1109/CCNC.2015.7158016

Nègre, J., & Delhomme, P. (2017). Drivers' self-perceptions about being an eco-driver according to their concern for the environment, beliefs on eco-driving and driving behavior. *Transportation Research Part A: Policy and Practice*, *105*, 95-105. https://doi.org/10.1016/j.tra.2017.08.014

Qi, G., Du, Y., Wu, J., & Xu, M. (2015). Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis. *IET Intelligent Transport Systems*, *9*(8), 792-801. https://doi.org/10.1049/iet-its.2014.0139

Rios-Torres, J., Liu, J., & Khattak, A. (2019). Fuel consumption for various driving styles in conventional and hybrid electric vehicles: Integrating driving cycle predictions with fuel consumption optimization. *International Journal of Sustainable Transportation*, *13*(2), 123-137. https://doi.org/10.1080/15568318.2018.1445321

Sullman, M. J., Dorn, L., & Niemi, P. (2015). Eco-driving training of professional bus drivers–Does it work?. *Transportation Research Part C: Emerging Technologies*, *58*, 749-759. https://doi.org/10.1016/j.trc.2015.04.010