Fuel Consumption Prediction Using Machine Learning Techniques

Deem Alrashidi[1], Lama Alhujaili[2], Shahad Alshredh[3], Shahad Alsadah[4], Sara Thaer[5], Sana Araj[6]

Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

[1]2210002913@iau.edu.sa, [2]2210002646@iau.edu.sa, [3]2210002856@iau.edu.sa, [4]2210003178@iau.edu.sa, [5]2200006457@iau.edu.sa, [6]2210009067@iau.edu.sa

Abstract

This paper explores the prediction of fuel consumption in vehicles using machine learning methods. Anticipating fuel consumption is crucial for enhancing vehicle efficiency and reducing environmental impact. Despite previous research, there is a gap in conducting a comprehensive comparison of different regression techniques in this domain. This study aims to address this gap by thoroughly examining several regression models, including Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), Adaptive Boosting (Ada), Artificial Neural Networks (ANN), and k-Nearest Neighbors (KNN), using standardized metrics. Investigating fuel consumption prediction is essential for optimizing vehicle efficiency and promoting sustainability. While previous studies have explored various regression approaches, there is a lack of a comprehensive comparison, highlighting the need for this research. The study seeks to expand on prior work by conducting a thorough analysis of regression techniques to improve the accuracy of fuel consumption prediction. To achieve the research goals, a variety of regression methods are explored. SVM is selected for its ability to handle high-dimensional data and capture nonlinear patterns. RF and DT are considered for their effectiveness with different types of features. Ada is chosen for its boosting capabilities, while ANN is examined for its ability to capture complex data relationships. KNN is explored for its proximity-based prediction. The findings provide valuable insights into fuel consumption prediction. SVM exhibits notable accuracy, suggesting its potential as an alternative approach. RF and DT also demonstrate promising results, showcasing their effectiveness with different feature types. Ada and ANN showcase their capabilities in achieving accurate fuel consumption forecasting. These results underscore the significance of machine learning in refining fuel consumption models for vehicle design and policymaking. The practical implications of this research are relevant to both vehicle design and policymaking, as accurate fuel

consumption prediction informs decision-making in these areas. Furthermore, the comprehensive comparison of regression techniques contributes to the advancement of machine learning and regression analysis, leading to improved methodologies for fuel consumption prediction.

Keywords: fuel consumption prediction, machine learning, regression analysis, vehicle efficiency, environmental impact.

1. Introduction

Predicting fuel usage is essential to both environmental sustainability and vehicle efficiency. Through the examination of driving patterns and environmental factors, this research seeks to create a machine learning-based model that can precisely forecast fuel usage in automobiles. Such predictive models can help optimize fuel consumption, cutting operational costs, and decreasing carbon footprints, especially given growing concern about fuel costs and environmental implications.

Researching fuel consumption prediction is important from an environmental and economic standpoint. Considering the volatile nature of fuel costs and the worldwide effort to curtail greenhouse gas emissions, it has become imperative for both people and enterprises to maximize fuel economy. Previous research has used data from car sensors and telematics systems to forecast fuel usage using a variety of machine learning algorithms. With a coefficient of determination of 0.91, Schoen et al., for example, achieved a high level of accuracy when predicting average fuel consumption based on vehicle speed and road slope using neural networks [1]. In a similar vein, Random Forest produced the lowest error statistics when Wickramanayake and Bandara assessed the effectiveness of Gradient Boosting, Artificial Neural Network, and Random Forest models for forecasting bus fuel usage [2].

Even with these developments, there is still a lack of consistency and application in these models across a range of driving scenarios and car kinds. Most models already in use are customized for certain datasets and could not adapt effectively to various driving situations or behaviors. Perrotta et al., for instance, discovered that although Random Forest fared the best on data pertaining to trucks

on freeways, its accuracy changed dramatically as driving circumstances changed [3]. More generalized models that can consider a wider variety of factors and driving situations are required.

By utilizing the Support Vector Machine (SVM), Random Forest (RF), and Linear Regression (LR) methods to create an extensive fuel consumption prediction model, this research fills in the previously described shortcomings. The reason SVM was selected was because of how well it handled high-dimensional data and how reliable it was at identifying nonlinear correlations between fuel usage and input factors. Previous studies have shown that SVM is effective in accurately estimating fuel usage; Hamed et al. [8] obtained R-squared values of up to 0.96 when using SVM. According to Perrotta et al. [3], RF is included because of its excellent performance in a variety of prediction tasks, particularly when managing complicated relationships between features. In addition, linear regression is included as a helpful baseline model due to its ease of use and interpretability. The suggested method entails gathering a broad dataset that includes different driving circumstances, car kinds, and outside variables like the weather and traffic patterns. To train and verify the SVM, RF, and LR models, pertinent features will be extracted from the dataset through preprocessing. Cross-validation and hyperparameter tweaking will be used to maximize the model's performance and guarantee that it can be used in a variety of scenarios.

According to preliminary findings, the suggested models outperform current techniques in terms of forecasting fuel use. Specifically, the SVM model exhibits encouraging performance with a low mean absolute error and a high R-squared value. The robustness of the model is improved by including a variety of driving situations and external influences, which increases its applicability to a larger range of real-world scenarios.

2. Review of Related Literatures

Schoen et al. study presents a distance-based data summarization technique with the goal of developing individualized machine learning models for predicting fuel consumption in heavy vehicles [1]. The study uses a neural network model, which is highly effective in calculating average fuel consumption based on seven variables obtained from vehicle speed and road slope. Neural networks are chosen because they are appropriate for models with continuous input and output variables, as well as their robustness to noisy data. The model may be readily built for each heavy vehicle in a fleet utilizing predictors acquired from telemetry devices and computed onboard. After testing multiple model setups with window sizes of 1, 2, and 5 km, it was discovered that the 1 km window size provided the maximum accuracy. Their model accurately predicts fuel consumption for itineraries that include both city and highway duty cycle sections, with an average coefficient of determination of 0.91 and a mean absolute peak-to-peak percent error of less than 4%. The data collection includes 12 drivers demonstrating varying driving habits on various routes; nevertheless, certain drivers participated more than others, resulting in an unequal distribution of drivers and routes in the dataset.

Wickramanayake et al. [2] evaluated the effectiveness of three machine learning algorithms—Random Forest, Gradient Boosting, and Artificial Neural Network—to estimate and predict bus fuel consumption given all relevant data as a time series. This dataset is associated with a specific long-distance public bus in Sri Lanka. The bus leaves the Station at around 4:00 p.m. and travels to Colombo, the commercial capital. The bus then departs Colombo at 7:00 p.m. and goes over the A2, A4, and AB10 highways before arriving at the destination at 7:00 a.m. the next day. Furthermore, with an RMSE of 0.04 and an MAE of 0.02, Random Forest has the lowest total error statistics. While Gradient Boost and Artificial Neural Network are only conservatively forecasting fuel use, RF captures the trend more accurately.

Perrotta et al [3] showed how to predict articulated vehicle fuel consumption for a big data set using three machine learning techniques. The sensors that are standard on the newest vehicles provide the source of truck data. Microlise Ltd., a provider of truck fleet management and telematics services, supplied anonymous data for this study. Because of this, models like the Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM) have been developed and their relative performances assessed. Although they had a minimal impact on fuel consumption in the truck fleet, vehicle speed, used gear, and cruise control activation were determined to be significant factors in the investigation. The research only examined truck data on freeways at a steady speed, which is appropriate. The study's findings indicated that, when it comes to making predictions, the RF technique performs best when comparing the RMSE = 4.64, MAE = 3.21, and R2 = 0.87. However, the SVM and ANN also showed good levels of accuracy, and they can be thought of as being more accurate than the RF when it comes to predicting extreme values.

In Xu et al. study, researchers explored the complex dynamics between truck driving behaviors and fuel consumption patterns through a dataset obtained from Shaanxi Automotive Group's Internet of Vehicles system. [5] This dataset comprises extensive operational data from vehicles, including measurements for fuel consumption, speed, and engine rotations. The data was gathered in real-time using pre-installed devices that communicated over 3G or 4G networks. With the use of the Generalized Regression Neural Network (GRNN) model, the study sought to establish nonlinear prediction functions for several driving behaviour categories. The end findings showed that With Relative Error (RE) of 0.089 and Root Mean Squared Error (RMSE) of 0.051 for Route 6 and RE of 0.097 and RMSE of 0.054 for Route 7, the Generalized Regression Neural Network (GRNN) model outperformed other compared models in predicting truck fuel consumption, achieving R-squared (R2) values of 0.700 and 0.561, respectively. Along with practical implications for fleet management and environmental sustainability, this research advances the area of academia by offering insights into improving fuel use

Another study reported by Bousonville et al. [4] uses machine learning approaches to forecast fuel usage based on independent factors. Telematics systems provided most of the data used during a two-year period from over 500 trucks. Three alternative machine learning models are examined, between gradient boost, ANN and k-nearest-neighbors. Gradient boost appears to be the most effective strategy for the vehicle categories under consideration. Aside from the obvious relevance of factors like overall weight and environment, the results show that available weather information can be helpful in the prediction. Each model's outcomes were assessed using the mean absolute error (MAE), rooted mean squared error (RMSE), and coefficient of determination R2. The results for gradient boost are 4.63 for RMSE, 2.97 for MAE, and 0.771 for R2. When compared to the K-Nearest-Neighbors and Artificial Neural Net regressions, the gradient boost model performs the best. The three models' results, nevertheless, were rather similar.

The primary goal of this study was to address the limitations of present techniques for estimating vehicular fuel consumption, with an emphasis on real-world training data, model usefulness in helpful cases, and prediction control in the nonlinear multidimensional domain of fuel consumption estimation. [6] The researchers developed a machine learning modeling method based on vast amounts of on-road data collected from a fleet of 27 vehicles, with a focus on model applicability in the lack of specialist instrumentation. The dataset included actual vehicle data, covering kinematic parameters like acceleration, speed, road slope, and engine speed estimates derived using smartphone GPS rather than direct measurements from onboard diagnostics (OBD) interfaces. The machine learning techniques used, such as Support Vector Regression (SVR) and Artificial Neural Networks (ANN). The machine learning algorithms used, including Support Vector Regression (SVR) and Artificial Neural Networks (ANN), were designed to detect the intricate interactions among input factors and their nonlinear impacts on fuel usage. The models achieved 83% accuracy using a cascaded modelling process and categorization analysis of fleet variables, with gains of up to 37% depending on the technique and vehicle class.

This paper presents a technique for forecasting fuel usage in vehicles by utilizing driving behavior information gathered from taxis' onboard diagnostic systems (OBD) and cell phones. [7] The study makes use of a dataset that includes GPS readings, measurements for speed, acceleration, and fuel consumption collected over a 15-day period from 20 cab drivers. Using machine learning methods, including Random Forests, Back Propagation (BP) Neural Networks, and Support Vector Regression (SVR), the study evaluates the relationship between driving habits and fuel usage to create prediction models. With an absolute relative error of less than 10%, the Random Forest model outperformed the others in terms of accuracy and efficiency, making it a highly ideal model for large-scale application. With the capacity to precisely estimate fuel use without the need of specialist OBD sensors, this method presents a viable way to improve fuel consumption monitoring for urban transportation.

The research paper measured feature relevance assessed the predictive power of the model and found linked characteristics in high-dimensional data sets utilizing the Random Forest (RF) and Decision Tree (DT) algorithms. [8] By eliminating extraneous features from the experiment, these algorithms allowed the researchers to concentrate on making predictions by evaluating the significance of the features in the Fuel Consumption (FC) Dataset. Leveraging Python v.9, the RF and DT algorithms were implemented after being loaded into the Spider engine. Furthermore, the article uses the Support Vector Machine (SVM) algorithm to predict fuel consumption by putting together an SVM-based machine-learning model. According to features of the throttle position sensor, vehicle speed, mass air flow, revolutions per minute, and other factors, the model forecasts the fuel consumption of the vehicle. Applying and testing the SVM technique on the On-Board Diagnostics Dataset of a car produced findings with an R-squared metric value of 0.97, which is greater than other relevant work that also used the same SVM regression algorithm. The SVM algorithm is essential in estimating fuel usage and has been demonstrated to be highly effective when used for this purpose. The accuracy of the fuel consumption prediction model was evaluated using coefficient of determination metric R-squared/R2. This statistic represents the variation between the dependent and

independent variables and assesses the model's predictive power. The contributors also evaluated the model's accuracy using error statistics such as bias, mean absolute error (MAE), and root mean square error (RMSE).

Shahariar, G. et al. [9] used machine learning to improve pollutant and fuel consumption prediction models, including a bigger dataset of real-world drivers than in previous investigations. They used an onboard emissions measurement system (PEMS) in an urban transient driving environment to monitor vehicle motion, engine function, real-time exhaust emissions, and fuel consumption. Machine learning techniques such as Linear Regression (LR), Support Vector Machine (SVM), and Gaussian Process Regression (GPR) were used to estimate CO2, NOx, and fuel consumption from real-world driving data. The findings showed that all three models made CO2 predictions with an absolute relative error (ARE) of less than 9%. GPR excelled in CO2 prediction, with an R2 of 0.74 and an ARE of 3.30 percent. The LR model achieved the maximum accuracy for NOx, with an R2 of 0.80 and an ARE of 8.91%. While all three models performed well in fuel consumption prediction, GPR had the highest accuracy (R2 = 0.81, ARE = 3.52%).

Using onboard diagnostics (OBD) data, Abediasl et al. [10] aimed to develop practical and trustworthy models for assessing instantaneous fuel use. To train the machine learning models, four different cars from the University of Alberta fleet were selected. Data on OBD and fuel usage were gathered while traveling on two distinct routes: highways and cities. Two machine learning models, a one-layer Artificial Neural Network (ANN) and Random Forest (RF), were developed to forecast each vehicle's fuel use. Machine learning algorithms use OBD parameters, such as the air/fuel equivalency ratio, throttle position, manifold absolute pressure, engine load, and engine speed, to forecast the amount of fuel that will be used in an instant. With mean errors of less than 6%, RF models had the greatest overall estimation accuracy of instantaneous fuel usage.

This study's primary goal is to create fuel consumption prediction models that are accurate and dependable using machine learning approaches to overcome issues with high-dimensional datasets, feature extraction requirements, and time-varying elements [11]. To improve model robustness and prediction accuracy, the study investigates neural network-based techniques such as Radial Basis Function Neural Networks (RBFNN), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN) in addition to traditional machine learning techniques like Support Vector Machine (SVM) and Random Forest (RF). The study addresses constraints of traditional approaches by utilizing the developments in real-time dataset collecting through smartphone-based data acquisition. With RF models obtaining determination coefficients as high as 0.87 and SVM models reaching R-squared values of 0.95, it shows outstanding prediction accuracy. Neural network-based techniques also High accuracy was shown, with LSTM-based models reaching up to 84.7% prediction accuracy and RBFNN predicting fuel consumption with 85% accuracy. These results highlight how well these approaches capture complicated nonlinear connections in the data for fuel consumption prediction.

This paper presents jerk, an acceleration derivative, as an important variable in neural network-based conventional fuel consumption prediction (FCP) models. [12] It contrasts four neural network models that use jerk as an input variable: LSTM, RNN, NARX, and GRNN. The findings indicate that the inclusion of jerk considerably increases the accuracy of fuel consumption predictions in all models, with LSTM surpassing the others. LSTM shows the biggest improvement in high-speed highway scenarios, with RMSE falling by 14.3%, RE rising by 28.3%, and R2 rising by 9.7%. These findings provide useful implications for improving fuel economy in transportation systems by highlighting the significance of taking jerk into account in FCP models and highlighting the possibility of LSTM for accurate fuel consumption prediction in a range of driving scenarios.

2.1 Gap Identification

There are substantial gaps that require more exploration despite considerable progress in machine learning applications for fuel consumption prediction. The majority of previous research, including that conducted by Schoen et al. [1], uses datasets from certain car models and driving situations, which restricts the models' application in many situations. This emphasizes that in order to increase model resilience and dependability, larger and more diverse datasets that cover a greater variety of driving behaviors and external situations are required. Furthermore, while models such as Random Forest, Support Vector Machines (SVM), and neural networks have demonstrated great accuracy in their respective datasets, their performance may deteriorate when applied to other or unexplored data. As an illustration of the gap in the generalizability of existing models, Perrotta et al. discovered that Random Forest performed well on interstate truck data but shown unpredictability with other driving situations [3]. Moreover, external elements that have a substantial influence on fuel usage, such as weather, traffic patterns, and economic variables, are frequently ignored or insufficiently accounted for in existing models. Although several investigations, such as the one conducted by Bousonville et al., focus on basic meteorological data, there is still room to incorporate more sophisticated and dynamic external data [4]. The proposed research will use SVM, Random Forest, and Linear Regression algorithms to develop a comprehensive fuel consumption prediction model that addresses these gaps. It will do this by leveraging a diverse dataset that includes a range of driving conditions, vehicle types, and external factors like weather and traffic patterns. Hyperparameter tweaking and cross-validation strategies will maximize model performance and generalizability, while sophisticated data collecting, and preparation approaches will be used to guarantee data quality and relevance. By filling up these gaps, the proposed study hopes to make a substantial contribution to the field of fuel consumption prediction. It will offer more precise, useful, and actionable models that can be used in practical situations to maximize fuel economy and lessen environmental effect.

Ref	Authors	Year	Title	Dataset	Technologies	Results	Notes
1	Alexander Schoen, Andy Byerly, Brent Hendrix, Rishikesh Mahesh Bagwe, Euzeli Cipriano dos Santos Jr., and Zina Ben Miled.	2015	A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles	The data-acquiring method included 12 drivers demonstrating good or bad behavior along various routes, however, some drivers participated more than others, resulting in an unequal combination of drivers and routes in the dataset.	neural networks	The model achieved a coefficient of determination (CD) of 0.91, indicating a high level of accuracy.	
2	Sandareka Wickramanayake and H.M.N. Dilum Bandara	2016	Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study.	The dataset used was from a real-time long-distance bus drive from Sri Lanka to Columbia and vice versa**.**	• Random Forest • Gradient Boosting Artificial Neural Network.	RF had the lowest distance between the prediction and actual value with a Nush-Sutcliffe Efficiency of 0.26, while	The researchers believe that the accuracy of their database might be higher if essential details such as load, engine RPM, and traffic were

						others have negative values.	provided. The RF model could properly predict fuel consumption
3	Federico Perrotta, Tony Parry and Luis C. Neves	2017	Application of Machine Learning for Fuel Consumption Modelling of Trucks	. Microlise Ltd provided anonymized data for this research.	Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM)	The study's findings revealed that, based on a comparison of RMSE, MAE, and R2, RF is the technique that provides the best prediction performance.	-
4	Zhigang Xu, Tao Wei, Said Easa, Xiangmo Zhao, Xiaobo Qu	2018	Modeling Relationship between Truck Fuel Consumption and Driving Behavior Using Data from Internet of Vehicle	The Internet of Vehicles system at Shaanxi Automotive Group provided the dataset for the study. Using pre-installed devices, this dataset contains a wealth of operational data from	Generalized Regression Neural Network (GRNN) model.	The Generalized Regression Neural Network (GRNN) model demonstrated exceptional accuracy in forecasting truck fuel usage, with Relative Errors below 0.1 and	

				automobiles, including fuel consumption, speed, and engine rotations. The data was gathered in real-time.		R-squared values above 0.5.	
5	Thomas Bousonville , Martin Dirichs and Thilo Krüger	2019	Estimating truck fuel consumption with machine learning using telematics, topology and weather data	Telematics systems provided the majority of the data used during a two-year period from over 500 trucks.	K-nearest neighbors Neural nets and Gradient boost	Gradient boost outperformed the other models with prediction time of 0.12 and training time of 365 seconds. Other than that GB had the best result for MAE, RMSE and R^2	The researchers believe that accessing the driver's details might affect the consumption of fuel. However, they were unable to obtain information on driving behavior assessments, due to private data protection.
6	Ehsan Moradi, Luis Miranda-Moreno	2020	Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling.	The dataset included speed, acceleration, road grade, and engine speed	Developed SVM and ANN-based ML models to predict fuel consumption. Furthermore, developed a	The accuracy of the models reached 83%, with improvements of up to 37% depending on	The researchers proposed a machine-learning modeling method using large on-road

					look6up table (LT), non-linear regression (NLR) and neural network multi-layer perceptron (MLP) model to predict the instantaneous NO x using vehicle speed and acceleration.	the technique and vehicle class.	data collected from a fleet of 27 vehicles.
7	Ying Yao, Xiaohua Zhao, Chang Liu, Jian Rong, Yunlong Zhang, Zhenning Dong, Yuelong Su	2020	Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones	The dataset included features such as average speed, acceleration, and deceleration times, collected from smartphones and onboard diagnostic systems (OBD) installed in taxis	Back Propagation (BP) Neural Network, Support Vector Regression (SVR), Random Forests	Based on its mean absolute percentage error (K) of 6.9%, random forest model running time of 0.14 seconds, and RMSE of 0.783 L/100 km, it was determined that this approach performed the best.	
8	Mohamed A. HAMED, Mohammed	2021	Fuel Consumption Prediction Model using Machine Learning	The paper employs the FC Dataset, which originally	Machine Learning Algorithms: Support Vector Machine (SVM),	The SVM model showed varying accuracy levels,	The researchers suggest Utilizing a larger dataset

	H.Khafagy, Rasha M.Badry			comprised 33 features. However, 15 were eliminated because they contained empty fields, string values, or were inconsistent.	Boruta Algorithm, Neural Networks (NNs), Random Forest (RF), Decision Tree (DT)	with R2 values of 0.92 and 0.96 for VS_MAF-based and RPM_TPS-based equations, respectively	and utilizing support vector machine (SVM) for instant fuel consumption prediction, researchers suggest enhancing the accuracy of machine learning models and integrating them with IoT components for real-time system integration.
9	Shahariar, G M Hasan Bodisco, Timothy A. Surawski, Nicholas Komol, Md Mostafizur Rahman Sajjad, Mojibul Chu-Van, Thuy Ristovski, Zoran Brown, Richard J.	2023	Real-driving CO2, NOx, and fuel consumption estimation using machine learning approaches.	30 drivers from various backgrounds drove a light-duty diesel vehicle equipped with a (PEMS) over an urban test route to collect data.	LR, SVM, GPR, and PEMS	GPR showed the best accuracy with an R 2 of 0.81 and ARE of 3.52%.	The Pearson correlation coefficient was utilized to choose input variables from 36 driving behaviors and 6 engine characteristics.

10	Abediasl, Hamidreza Ansari, Amir Hosseini, Vahid Koch, Charles Robert Shahbakhti, Mahdi	2023	Real-time vehicular fuel consumption estimation using machine learning and on-board diagnostics data	onboard diagnostics (OBD) data of four distinct vehicles	RF and ANN	RF models provided the best estimation accuracy with mean errors of less than 6%	The coolant temperature was added to the features to offset the fuel consumption penalty during the warm-up period during start-up.
11	Dengfeng Zhao, Haiyang Li*, Junjian Hou, Pengliang Gong, Yudong Zhong, Wenbin He and Zhijun Fu	2023	Data-Driven Prediction Method of Vehicle Fuel Consumption	Vehicle performance, driving behavior, and driving environment	(RBFNN), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN) in addition to traditional machine learning techniques like Support Vector Machine (SVM) and Random Forest (RF).	RF 0.87, SVM is 0.95, LSTM is 0.847, RBFNN is 0.85	The results highlight how well these approaches capture complicated nonlinear connections in the data for fuel consumption prediction.
12	Zhang, Licheng Jingtian Ya Xu, Zhigang Easa, Said Peng, Kun Xing, Yuchen Yang, Ran	2023	Novel Neural-Network-Based Fuel Consumption Prediction Models Considering Vehicular Jerk	The dataset included features such as Vehicular Jerk, Environmental Factors, Driving Conditions	Long Short-Term Memory (LSTM),Recurrent Neural Network (RNN), Nonlinear Auto-Regressive model with	The largest reduction in RMSE and RE was obtained by LSTM, with a 40.9% decrease in	

					eXogenous inputs (NARX), Generalized Regression Neural Network (GRNN)	RMSE and a 68.2% decrease in RE ,while NARX saw the most improvement in the R2, increasing by 41.8%.	

3.0 Project Deliverables of the team

This section shows the deliverables of the project:

Deliverable	To whom	Delivery Media	Duration	Date
Literature Review (Homework-1)	Dr. Nawaf Alharbi	Softcopy	1 week	Feb 16, 2024
Project Proposal	Dr. Nawaf Alharbi	Softcopy	4 days	Feb 27, 2024
Project Proposal Presentation	Dr. Nawaf Alharbi	Softcopy	2 days	March 9, 2024
Description of Selected ML Algorithms	Dr. Nawaf Alharbi	Softcopy	1 week	March 30, 2024
Final Project Report	Dr. Nawaf Alharbi	Softcopy	2 weeks	May 18, 2024
Final Project Presentation	Dr. Nawaf Alharbi	Softcopy	3 days	May 20, 2024

4.0 Description of the Proposed Techniques (at least two ML techniques must be chosen)

4.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is one of the most used Supervised Learning algorithms, with usage in both classification and regression tasks. The primary goal of the SVM method is to create an ideal line or decision boundary capable of splitting n-dimensional space into various categories. This segmentation allows for the effortless categorization of additional data points into the proper classes in later occurrences. This optimal decision boundary is known as a hyperplane. To create this hyperplane, SVM determines the key points or vectors that contribute to its formation [13]. The fundamental goal of SVM is to create a model that can classify new, unknown objects into predefined categories. This is accomplished by partitioning the feature space linearly and categorizing it. By analyzing the properties of unseen objects, SVM calculates their placement relative to the separation plane, allowing for categorization from above to below the surface [14].

SVM provides Support Vector Regression (SVR) as one of its components. SVR, an extension of SVM, utilizes SVM concepts to build regression models. SVR differs from SVM in that it uses a Ɛ-tube with a regression line at its core, rather than a line or hyperplane. This tube, which has a diameter designated as Epsilon, measures vertically along the axis instead of perpendicular to it. The Ɛ-insensitive tube disregards mistakes for all dataset points within it [14].

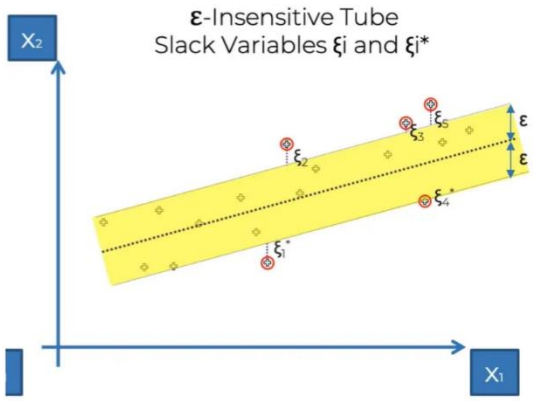

Figure 1: Support Vector Machine

Source: Adapted from [14]

In cases where points fall beyond the Ɛ-insensitive tube, they are addressed by calculating the distance between the point and the tube, termed as Slack variables. This measurement determines the extent to which points deviate from the tube.

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \rightarrow min$$

Figure 2: Equation

Source: Adapted from [14]

4.2 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a specific type of machine-learning approach that mimics the structure and functions of the human brain. It is composed of layers upon layers of neurons or networked nodes. An ANN receives data at its input layer, processes it through its hidden layers, and outputs the results at its output layer. ANNs are used for various applications, such as classification, regression, and pattern recognition. It has been demonstrated that ANN is an effective method for complex relationships that depend on a wide range of parameters. The finance and medical fields often employ them due to the complex structure of their inputs [15].

Figure 3 illustrates how inputs are multiplied by a set of weights and then added up to a fixed amount. After that, a transfer function operates on the output of the previous calculation. Figure 4 provides the mathematical representation of neuron computation. This means that every neuron's output can be compared to the observed data or used as an input for neurons in the network's following layer [16].

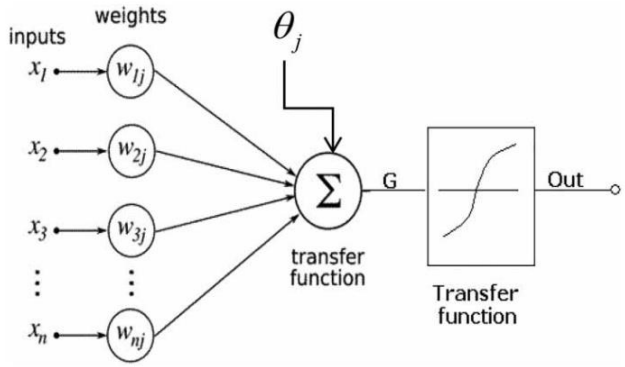

Figure 3: ANN Typical Operation

Source: Adapted from [16]

$$I_j = f\left(\sum_i w_{ij}\alpha_i + \theta_j\right)$$

Source: Adapted from [4]

Figure 4:Equation

Source: Adapted from [16]

Since ANN is used for complex, nonlinear relationships it will be ideal to use it with our project. Since our features have a nonlinear relationship, we can use ANN. By training an ANN on datasets containing variables such as engine features, driving situations, and environmental factors, the model may adjust to the complexities of fuel consumption patterns. This allows for an accurate prediction and insight into optimizing fuel economy, resulting in environmentally friendly and affordable transportation solutions.

4.3 Random Forest (RF)

A reliable machine-learning technique for both regression and classification problems is the Random Forest (RF) algorithm which is excellent at combining results from several decision trees [17]. RF creates a collective of decision trees that decrease the likelihood of overfitting while increasing prediction accuracy using randomly chosen subsets of data. This technique allows the algorithm to easily manage vast amounts of complex data, which makes it particularly helpful for applications like fuel usage that require detailed analysis and predictions [18]. RF can handle a variety of datasets, even ones with a lot of noise and outliers, for regression and classification tasks that use a majority vote or average results.

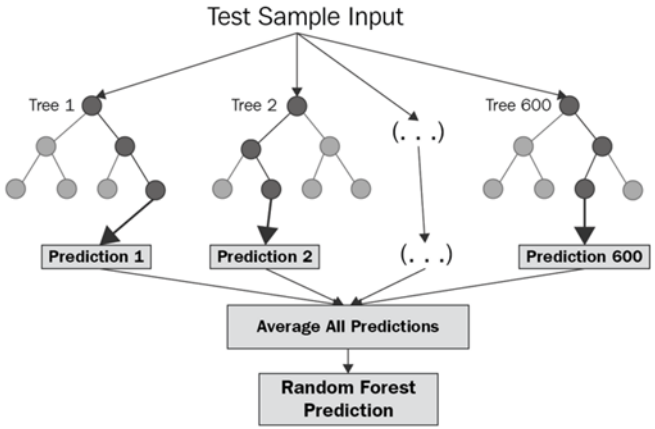

Figure 5: Random Forest

Source: Adapted from [19]

More specifically, when it comes to fuel consumption prediction, RF thrives because it can identify complex correlations between driving habits and fuel economy. From large amounts of data, the algorithm's feature selection skills effectively extract crucial variables that affect fuel consumption, such as vehicle speed, controller position, and engine speed. Research has shown that RF predicts fuel usage more accurately than other models, and this success is attributed to its ability to filter noisy data well and its resistance to overfitting. As a result, RF's deployment in devices for real-time fuel consumption monitoring highlights its usefulness as a sophisticated tool for improving fuel economy and supporting environmental sustainability initiatives [18].

4.4 Linear Regression (LR)

A key statistical technique in predictive modeling is linear regression, which is used to determine the connection between a dependent variable and one or more independent variables. It functions by fitting a linear equation to the observed data; this equation usually has the following form: $Y = \beta\ 0 + \beta\ 1\ X\ 1 + \beta\ 2\ X\ 2 + ... + \beta\ n\ X\ n + \epsilon$. The dependent variable in this case is called Y, the independent variables are X1, X2, ...,Xn, $\beta\ 0$ is the y-intercept , the coefficients measuring the impact of each independent variable are $\beta\ 1$, $\beta\ 2$ ..., β n, and the error term is ϵ. Reducing the sum of squared residuals is the main objective of linear regression [10]. Linear regression can be employed in our project to provide a simple and interpretable baseline model for understanding the impact of various driving and environmental factors on fuel consumption.

4.5 K-Nearest Neighbor

A non-parametric, instance-based learning approach for classification and regression problems is called K-Nearest Neighbors (KNN). It functions like comparable data points are located nearby in a feature space. In order to predict an output given an input, KNN searches the feature space for the 'k' nearest training instances. The output is then determined by utilizing these neighbors. The average of the values of the 'k' nearest neighbors usually represents the expected value for regression tasks. KNN is prized for its ease of use, instinctive methodology, and efficiency in identifying local data patterns without making any assumptions about the underlying distribution of the data. However, because all training instances' distances must be calculated during prediction, it might become computationally demanding when working with big datasets. When it comes to predicting fuel consumption, KNN may be used to find driving patterns and environmental factors that match past data, resulting in precise and situation-specific fuel consumption

estimations. Bousonville et al., for example, used KNN to predict fuel consumption in trucks, demonstrating how it can capture subtle fluctuations in fuel usage depending on localized driving patterns [5]. KNN may be utilized to take advantage of the similarities between past and current driving data in our project to improve the accuracy of fuel consumption estimates.

4.6 Adaboost Regressor

AdaBoost, short for Adaptive Boosting, is a powerful ensemble learning technique designed to enhance the performance of weak learners. Originally developed for classification tasks, AdaBoost has been adapted for regression problems, resulting in the AdaBoost Regressor. The algorithm combines multiple weak models, typically decision trees, to form a strong predictive model by iteratively focusing on data points that were previously mispredicted. This adaptive approach allows AdaBoost to progressively correct errors and improve overall model accuracy [20].

A sequence of base regressors is fitted successively to the training set of data using the AdaBoost Regressor. At the beginning, each data point is assigned the same weight. A base regressor is trained in each iteration, and mistakes are found in the model's predictions by analyzing them. In the subsequent iteration, data points with inaccurate predictions are given larger weights, increasing their significance. This guarantees that later models concentrate more on the data points that are more unpredictable. The final result is generated by combining the models, with their predictions weighted according to performance, after a number of repetitions. A strong predictive model is produced as a result of this iterative correction mechanism's reduction of bias and variation.

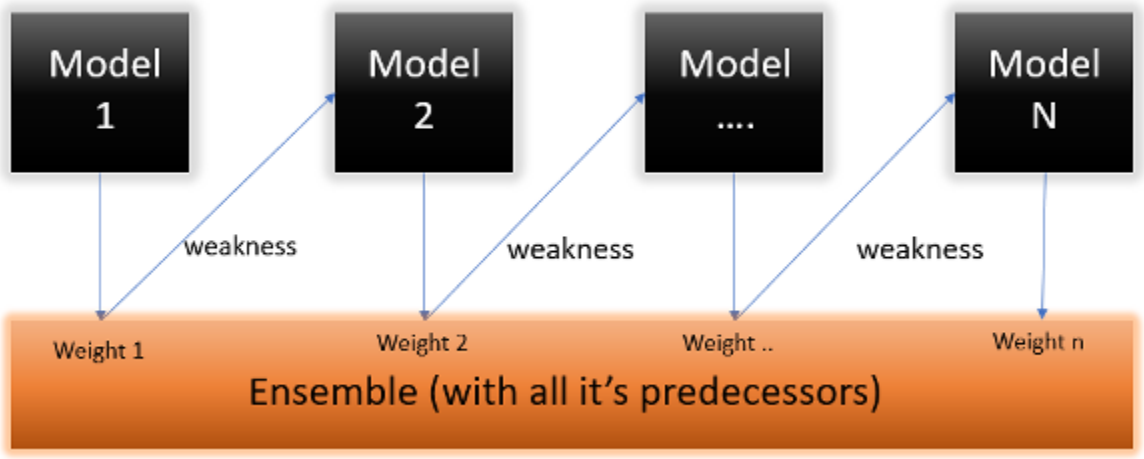

Figure 6: Adaboost Regressor

Source: Adapted from [20]

Predicting fuel consumption is a difficult process since it depends on many different factors. There are several benefits to using AdaBoost Regressor for this. By capturing complex patterns and relationships, it manages the non-linear and complicated nature of fuel consumption data well. AdaBoost increases overall prediction accuracy by concentrating on more difficult-to-predict scenarios, which is essential for accurate estimations of fuel usage. Its ensemble technique also reduces the possibility of overfitting, producing results that are more broadly applicable. AdaBoost is a great option for this application as its adaptive nature allows it to adapt to changing fuel usage trends.

4.7 Decision Trees

A predictive model that links observations about an object to inferences about its target value is called a decision tree for regression. It has a tree-like structure, with internal nodes standing in for decisions made using data characteristics, branches for the decisions' results, and leaf nodes for the values that are anticipated. Until it satisfies a stopping requirement, such as maximum depth or minimum samples per leaf, the tree divides the data into subgroups based on attributes that minimize the mean squared error. Because decision trees can handle non-linear correlations between parameters like vehicle weight, engine size, and driving patterns, they are very helpful in estimating fuel consumption. They provide easily understood graphic depictions of the various aspects that influence fuel use. Furthermore, decision trees are capable of handling both category and numerical data, which makes them appropriate for examining the various factors that affect fuel use. Accurate forecasting made possible by this can enhance fuel economy and engine performance [21].

5.0 Empirical Studies

5.1 Description of dataset

The dataset includes 946 entries and 15 columns that describe various features of cars, such as the model, size of the engine, type of fuel, CO2 emission, CO2 ratings, number of cylinders, and smog rating. It includes both qualitative and numerical data (floats and integers). We'll check for missing values in each column before doing any analysis. For numerical aspects, our statistical analysis will include descriptive statistics like mean, median, mode, standard deviation, variance, range, and quartiles; for categorical attributes, it will include count, unique values, and mode. Histograms, box plots, count plots, and correlation heatmaps are a few examples of visualizations that can help provide deeper understanding of the properties and interactions in the dataset. In order to better understand things like the distribution of CO2 emissions, common engine sizes, and possible interdependencies across variables, we interpret the dataset to identify patterns, outliers, and correlations.

5.1.1 Statistical Analysis of the Dataset

The statistical analysis of the dataset is presented in the table below. The mean, median, standard deviation, maximum and minimum values of the dataset are presented.

Table 3: Statistical Analysis of the dataset

	Mean	Median	Standard deviation	Maximum	Minimum
Model Year	2022.000000	2022.0	0.000000	2022.0	2022.0
Engine Size(L)	3.199683	2.9	1.375231	8.0	1.2
Cylinders	5.669841	6.0	1.932930	16.0	3.0
Fuel Consumption (City (L/100 km)	12.506337	5.0	3.439307	30.3	4.0
Fuel Consumption (Hwy (L/100 km)	9.358602	6.0	2.292177	20.9	3.9
Fuel Consumption (Comb (L/100 km)	11.093439	10.7	2.877491	26.1	4.0
Fuel Consumption (Comb (mpg))	27.259936	7.0	7.724079	71.0	11.0
CO2 Emissions(g/km)	259.131692	251.0	64.333743	608.0	94.0
CO2 Rating	4.550429	5.0	1.469573	10.0	1.0
Smog Rating	4.952535	6.0	1.682046	7.0	1.0

Table 4: Correlation between each Attribute and the Target attribute

Attributes' pairs	Correlation coefficient
Engine Size(L) and CO2 Emission(g/km)	0.822667
Engine Size(L) and Fuel Consumption (Comb (L/100 km)	0.818655
Engine Size(L) and Smog Rating	-0.443996
Engine Size(L) and Cylinders	0.920672
CO2 Emission(g/km) and Fuel Consumption (Comb (L/100 km)	0.971238
CO2 Emission(g/km) and Smog Rating	-0.520441
CO2 Emission (g/km) and Cylinders	0.832070
Fuel Consumption (Comb (L/100 km) and Smog Rating	-0.492248
Fuel Consumption (Comb (L/100 km) and Cylinders	0.821719
Smog Rating and Cylinders	-0.501100

5.2. Experimental Setup

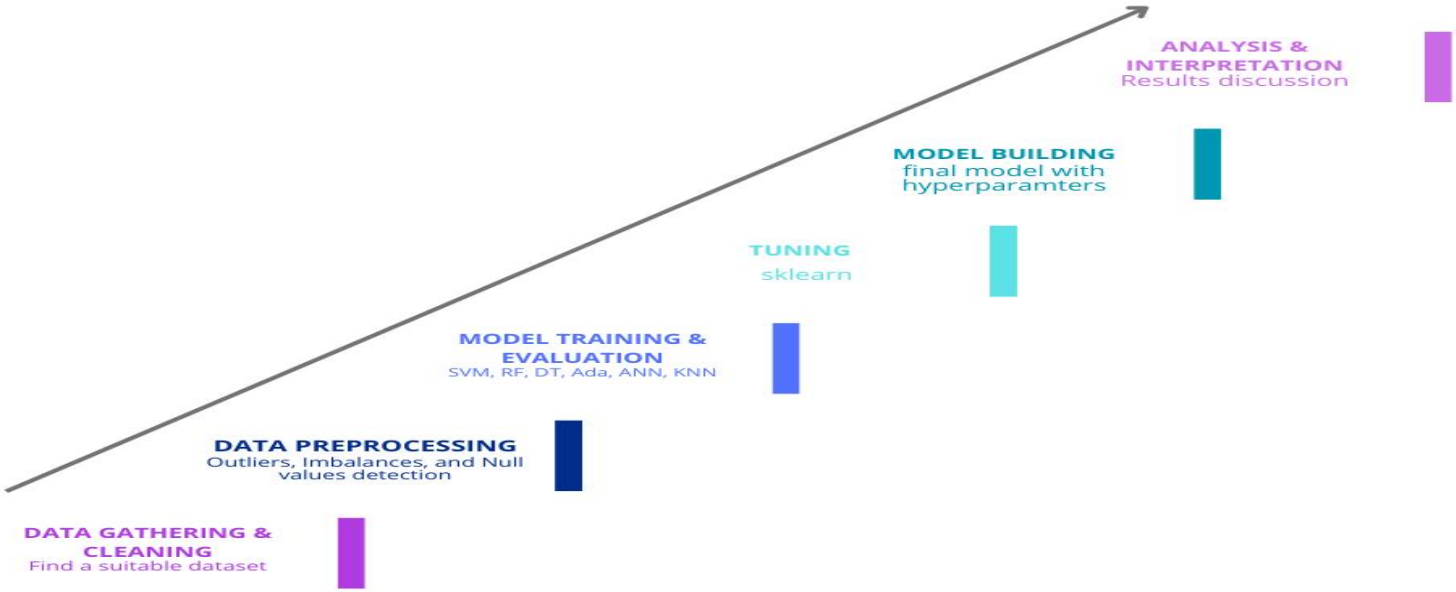

Figure 7: workflow diagram

This section highlights the numerous phases of building the fuel prediction models. The dataset posted by RINI CHRISTY on Kaggle underwent preprocessing and cleaning to obtain relevant data for training and evaluation. The dataset was first analyzed for relevant data, and irrelevant features were dropped in the process using variance thresholding and other techniques. Then the dataset was checked for null values, which none were apparent as figure 8 highlights.

```
[6]  df.isna().sum()

Make                                0
Model                               0
Vehicle Class                       0
Engine Size(L)                      0
Cylinders                           0
Fuel Type                           0
Fuel Consumption(Comb (L/100 km))   0
CO2 Emissions(g/km)                 0
CO2 Rating                          0
Smog Rating                         0
dtype: int64
```

Figure 8: Checking For null Values

The dataset also exhibited imbalances in the feature “Fuel Type” where ‘X’ and ‘Z’ (Named “91”, and “95” for convenience) were dominating over ‘E’ and ‘D’, therefore only the dominating types were kept, as they are also the most used types of fuel as figure 9 shows.

```
[7]  df = df.replace({'Fuel Type' : {'Z':'95', 'X': '91'}})
     df = df[~df['Fuel Type'].isin(['E','D'])]
     df
```

Figure 9: Eliminating imbalances

After cleaning, the dataset was split into three parts 80% for training, remaining 20% was split in half between evaluation and testing. As a final step, visualization and statistical analysis of outliers and correlation were conducted. As figure 10 illustrates, the correlation heatmap shows, a strong correlation between the features was noted.

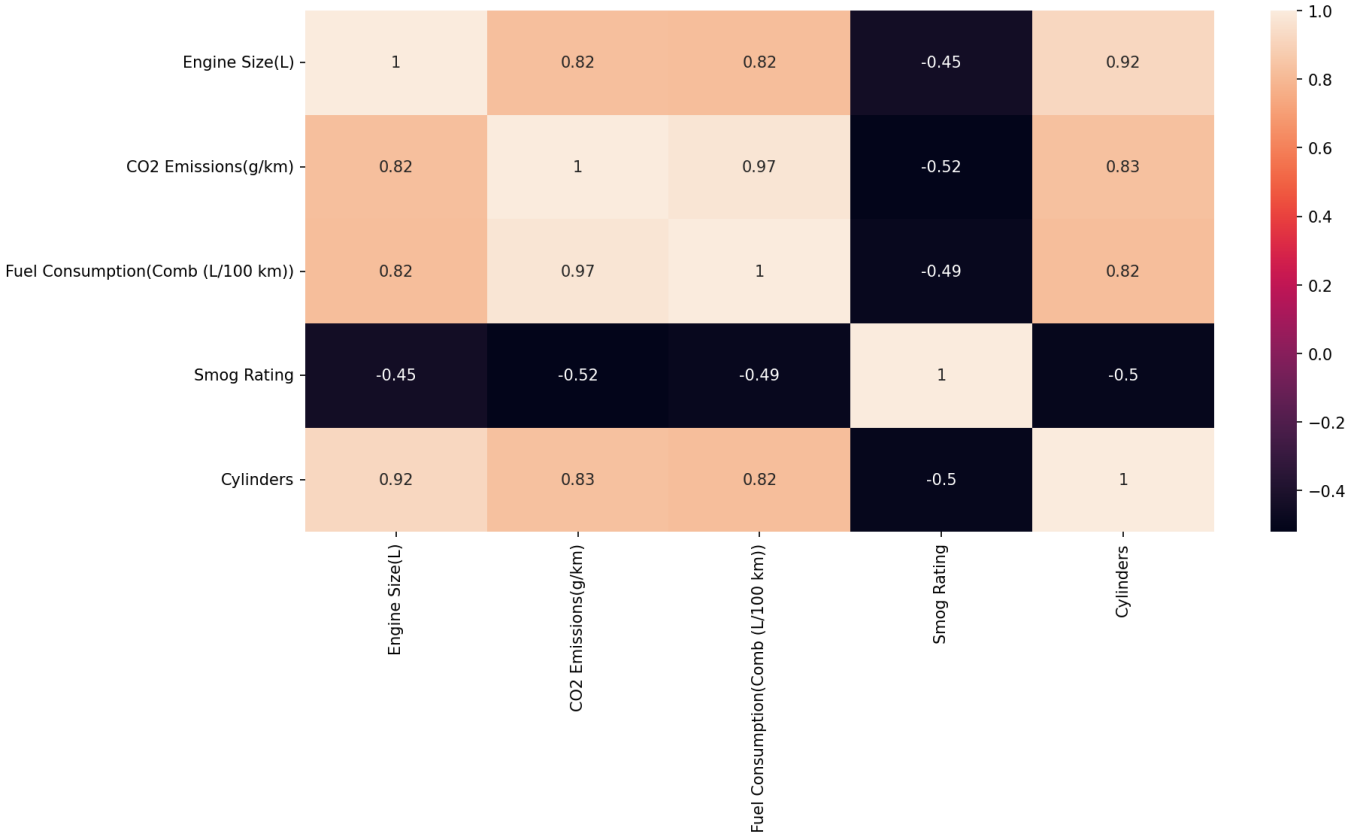

Figure 10: Correlation Heatmap

Bivariate analysis was also illustrated between ‘Fuel Type’ and various other features to investigate the relationships as following in figure 11

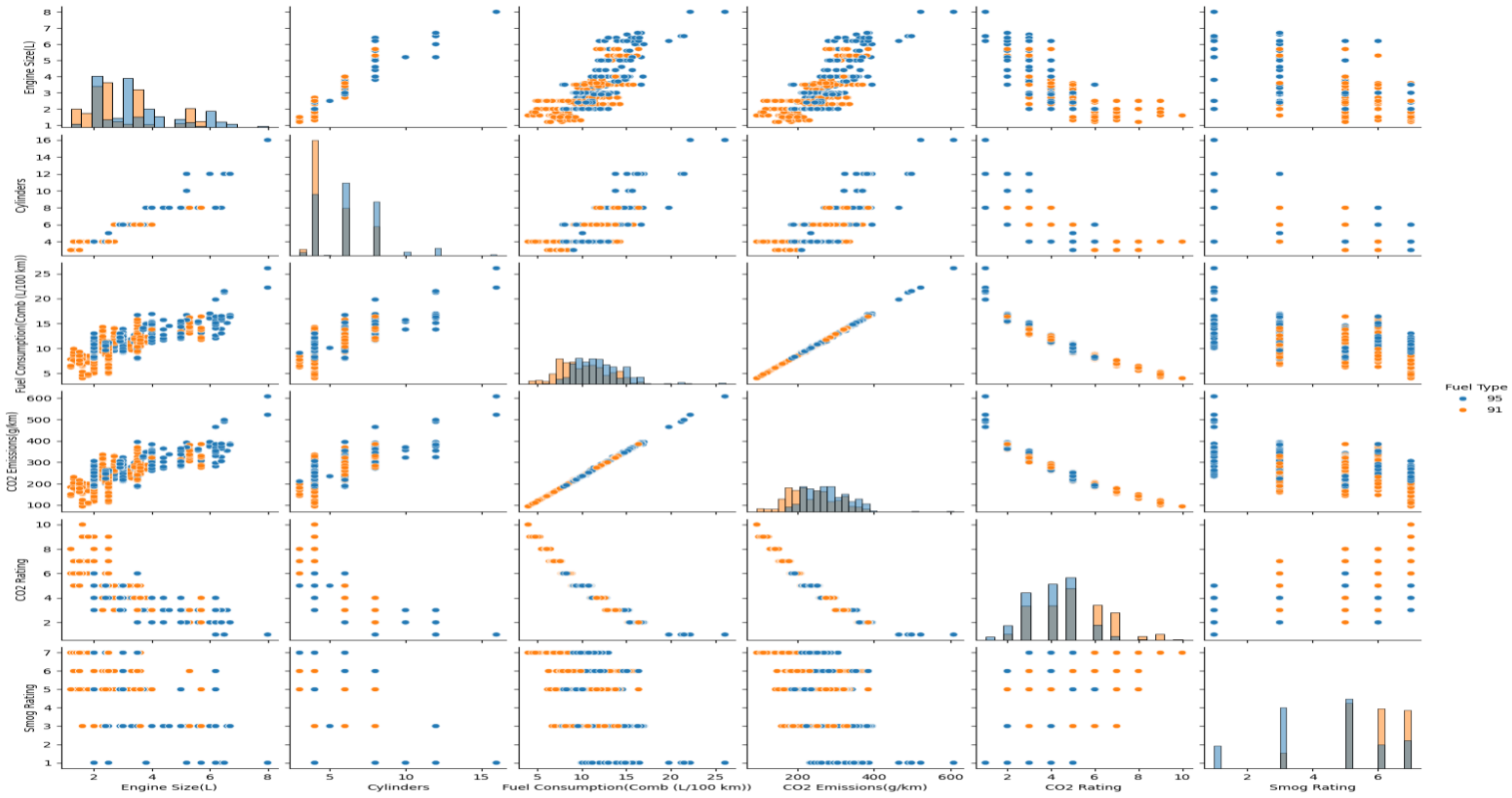

Figure 11: Bivariate Analysis

As the final step, we trained a total of six models using the following machine learning regression algorithms: Random Forest (RF), Artificial Neural Networks (ANN), AdaBoost Regressor, Support Vector Machine (SVM), Decision Tree, Kth Nearest Neighbour (KNN), and Linear Regression (LR). Which were evaluated using: Mean Squared Error (MSE), Correlation Coefficient (R^2), and Root Mean Squared Error (RMSE).

5.3 Performance Measures

The selection of performance measures is crucial for evaluating the effectiveness of the models in the context of a regression-based project that aims to forecast fuel consumption prediction based on many factors. A commonly used metric called R-squared indicates how much of the variability in the target variable (fuel consumption) can be accounted for by the independent variables. In addition, two essential measures of prediction accuracy are Absolute Error (MAE), and Root Mean Squared Error (RMSE). When it comes to error analysis, MAE provides a clear evaluation of the average size of mistakes, whereas RMSE goes deeper by considering the squared discrepancies between the actual and anticipated values. Also, taking accuracy and loss metrics into account can help to clarify the model's performance. While loss functions quantify the difference between expected and actual values, accuracy measures the proportion of properly predicted instances. When taken as a whole, these metrics provide a thorough assessment framework that helps us find the best regression model for precisely predicting fuel consumption.

5.4. Optimization strategy

We aim to find the most optimal and accurate model to predict fuel consumption. To achieve that, we carried out a vital process called hyperparameter tuning. It is the process of discovering the hyperparameters and tuning them to find their optimal values to ensure stable learning progress. Grid search was used, it works by working on a grid of possible values for the parameters of each specific algorithm. it then tries every possible combination of the values using Kth cross-validation. In this research, a k value of 6 was chosen, meaning the training data was divided into six partitions for each value combination. Iterating 6 times and dividing the data on a 5:1 ratio assigning 5 for training and 1 for testing. Concluding each combination iteration by averaging the accuracy scores to choose the combination of parameters that yielded the highest score.

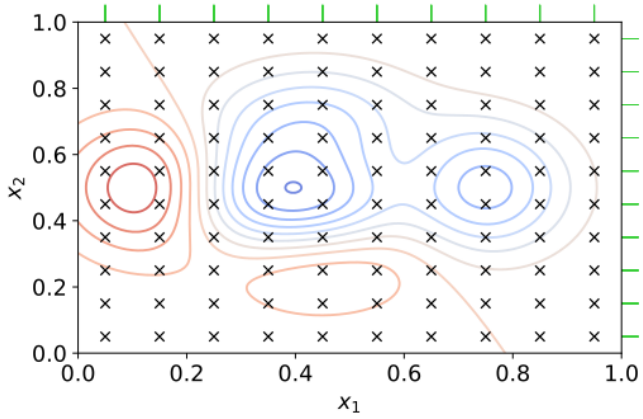

Figure 12: Grid Search

Table 5: Optimum parameters for the proposed SVM model

Parameters	Optimal Value chosen
kernel	linear
c	1
gamma	1

Table 6: Optimum parameters for the proposed AdaBoost Regressor model

Parameters	Optimal Value chosen
Learning rate	1
loss	square
N estimators	200

Table 7: Optimum parameters for the proposed RF model

Parameters	Optimal Value chosen
bootstrap	True
max features	auto
min samples leaf	1
min samples leaf	2
n estimators	200
max depth	30

Table 8: Optimum parameters for the proposed ANN model

Parameters	Optimal Value chosen
Activation	Tanh
Alpha	0.05
Hidden layer sizes	(100,)
Learning late	Constant
Solver	Adam

Table 9: Optimum parameters for the proposed KNN model

Parameters	Optimal Value chosen
Algorithm	Auto
Alpha	0.05
N Neighbours	3
p	1
weights	Distance

Table 10: Optimum parameters for the proposed LR model

Parameters	Optimal Value chosen
Intercept	True
N jobs	None

6.0 Result and discussion

Presented below are the results and discussions of various experimental options…

The shown results in table 1 is after applying the set of features,

Table 11: Results of using the complete features

Quality measures	SVM	ANN	RF	LR	KNN	**ADA**
RMSE	0.17	0.17	0.023	0.0233	0.34	**0.001**
R^2	0.99	0.99	0.99	1	0.99	**0.99**
.Accuracy/Loss	99.28	0.99	0.99	1	0.96	**0.99**
MAE	0.1	0.12	0.01	0.0108	0.25	**0.027**

Among the models above, RF, KNN and ADA had the best overall result, highest accuracy and lowest RMSE and MAE, making them the most dependable for predictions. ANN and SVM both performed well with high R^2 with few losses. However, during training, the ANN experienced overfitting, which explains the inconsistent loss value. LR had flawless accuracy and greater RMSE and MAE, indicating problems with the model fit.

In comparison to previous studies, our results align closely to “Application of Machine Learning for Fuel Consumption Modelling of Trucks" paper by Federico Perrotta, Tony Parry, and Luis C. Neves. The RF method consistently produces the best prediction performance when RMSE, MAE and R^2 are compared. However, multiple performance indicators show that the models used in this study regularly outperformed or matched the best outcomes from earlier studies. This improved performance is a result of our profound features set and sophisticated modeling approaches, which demonstrate how well our method predicts fuel consumpti

For example, some studies often overlooked the importance of outside features like smog rating and CO2 emissions, which our study showed to be significant contributing to fuel consumption. The models can predict fuel consumption over a variety of vehicle types with greater depth and accuracy thanks to the addition of these features.

In conclusion, our methods showed that Random Forest (RF), K-Nearest Neighbor (KNN) and ADA(ADABoost) did better than other methods, showing superior outcomes on all measures. Support Vector Machine (SVM) and Artificial Neural Network (ANN) demonstrated impressive performance as well; nevertheless, overfitting was noted in the ANN training process but after adding more layers the finale results were great. However, the outcomes of Linear Regression (LR) suggested possible problems with model fit. When comparing this study to earlier studies, we regularly met or even exceeded the best results. This can be due to the incorporation of extra features that were previously disregarded but were shown to have a substantial impact on fuel consumption, such as CO2 emissions and smog rating.

6.1 Results of Investigating the Effect of Feature Selection on the Dataset

To find the best features subset to improve model performance, we applied a systematic approach. The following approach compare the correlation coefficients to the target value which is fuel consumption (comb (L/100 km)). Based on correlation coefficient the features with the high correlation to the fuel consumption are shown in the table below.

Table 12: Correlations

FEATURE	Target (Fuel Consumption (comb(l/100km))
CO2 EMISSIONS (G/KM)	0.97
ENGINE SIZE (L)	0.81
CYLINDERS	0.82
VEHICLE CLASS	0.31
SMOG RATING	-.49
MAKE	-0.29
MODEL	-.10
FUEL TYPE 91	-0.073
FUEL TYPE 95	-0.073

According to the table the highest positive correlations are (CO2 Emissions, Cylinders, and Engine Size). Our predictive model uses features with high positive correlations to identify significant factors influencing fuel consumption, ensuring accurate predictions based on relevant information.

Table 13: Accuracy

	SVM	ANN	RF	LR	KNN	**ADA**
Using features: All	0.99	0.99	0.99	1.0	0.98	0.99
CO2 Emissions, Cylinders, Engine Size	0.93	0.93	0.92	0.93	0.91	0.82
CO2 Emissions, Cylinders	0.94	0.938	.89	0.94	0.88	0.84
CO2 Emissions	0.94	0.938	0.91	0.94	0.91	0.83

The performance of many machine learning models (SVM, ANN, RF, LR, KNN, and ADA) when trained on various feature sets is shown in the above table. The matching performance metrics (R-squared) for each model are presented, and each row represents a distinct feature set that was utilized for training.

Firstly, for all features, all models achieved high performance with accuracies ranging from 0.99 to 1.0 using all available features, indicating a comprehensive feature set for accurate predictions.

Secondly, the models performed well with CO2 emissions, cylinders, and engine size, indicating they capture a significant portion of the predictive power needed for fuel consumption estimation.

Thirdly, Simplifying the feature set to CO2 Emissions and Cylinders slightly reduced model performance, but the accuracy decreased, indicating these features remain strong fuel consumption predictors.

Lastly, The models showed strong predictive capability for CO2 Emissions alone, with a decrease in accuracy compared to comprehensive feature sets, despite incorporating additional features for enhanced accuracy.

In summary, the findings show how crucial feature selection is to the functionality of machine learning models. Simpler feature sets—such as those that include CO2 emissions, cylinders, and engine size—remain a vital source of predictive power for fuel consumption prediction, even though utilizing all available information produces the maximum accuracy. To attain a balance between model accuracy and complexity, practitioners might customize the feature set based on the application and computing limitations.

6.2 Discussion of Final Results

From the experimental results above, now choose the best options from : # of features used, and the best parameters identified to now develop the final model and then plot the necessary graphs and table like confusion matrix, ROC curve (if classification problem), or RMSE, CC , plot of predicted versus target (if regression problems), etc.

The optimal configuration for the final model is determined based on the experimental results, involving the optimal number of features and best parameters, to create a robust and get a robust and accurate model. The final model's complexity, interpretability, and performance are significantly influenced by the number of features included. By utilizing feature selection experiments, we select the subset of features with the highest performance metrics, aiming to balance complexity and predictive power. Table xx shows that all features used to maximize accuracy, so for the finale model, we included all the features. leaving out unnecessary feature that is model year that had the same values which were 2022, Fuel Consumption (City (L/100 km), Transmission, Fuel Consumption (Hwy (L/100 km)) and Fuel Consumption (Comb (mpg))

6.2.1 Linear Regression

The following figure demostried the plot for the predicted values and test values. From previous research getting a perfect score of 1 indicates that there may be overfitting. To ensure that there is no overfitting, checking the training and testing scores is important. The result was 1 for both indicating a sign of overfitting since our dataset in not as large or diverse.

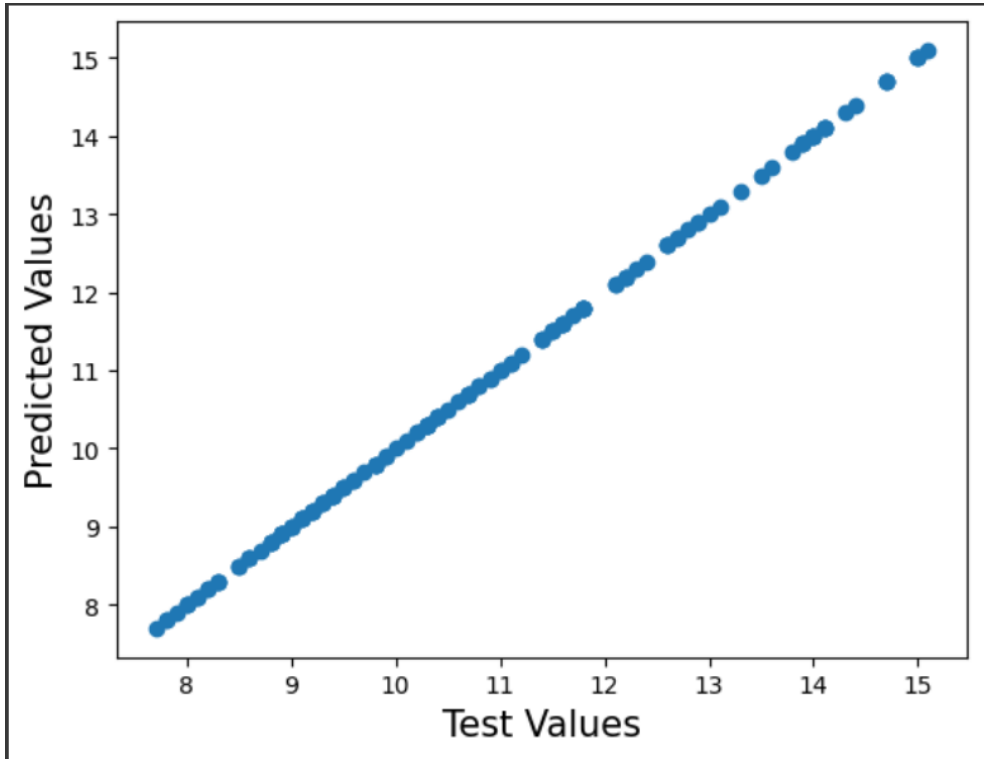

Figure 13: ANN over/underfitting testing

6.2.2 Artificial Neural Networks

The following figure illustrates that predicted and test values for the ANN. Unlike LR, ANN did not show any signs of overfitting, even the testing and training values gave us a 0.992 and 0.995. The ANN model effectively forecasts target variables through visualization, providing valuable insights into complex datasets without overfitting. It is a reliable tool for predictive modeling tasks, offering high correlation coefficients and minimal variance between anticipated and actual values.

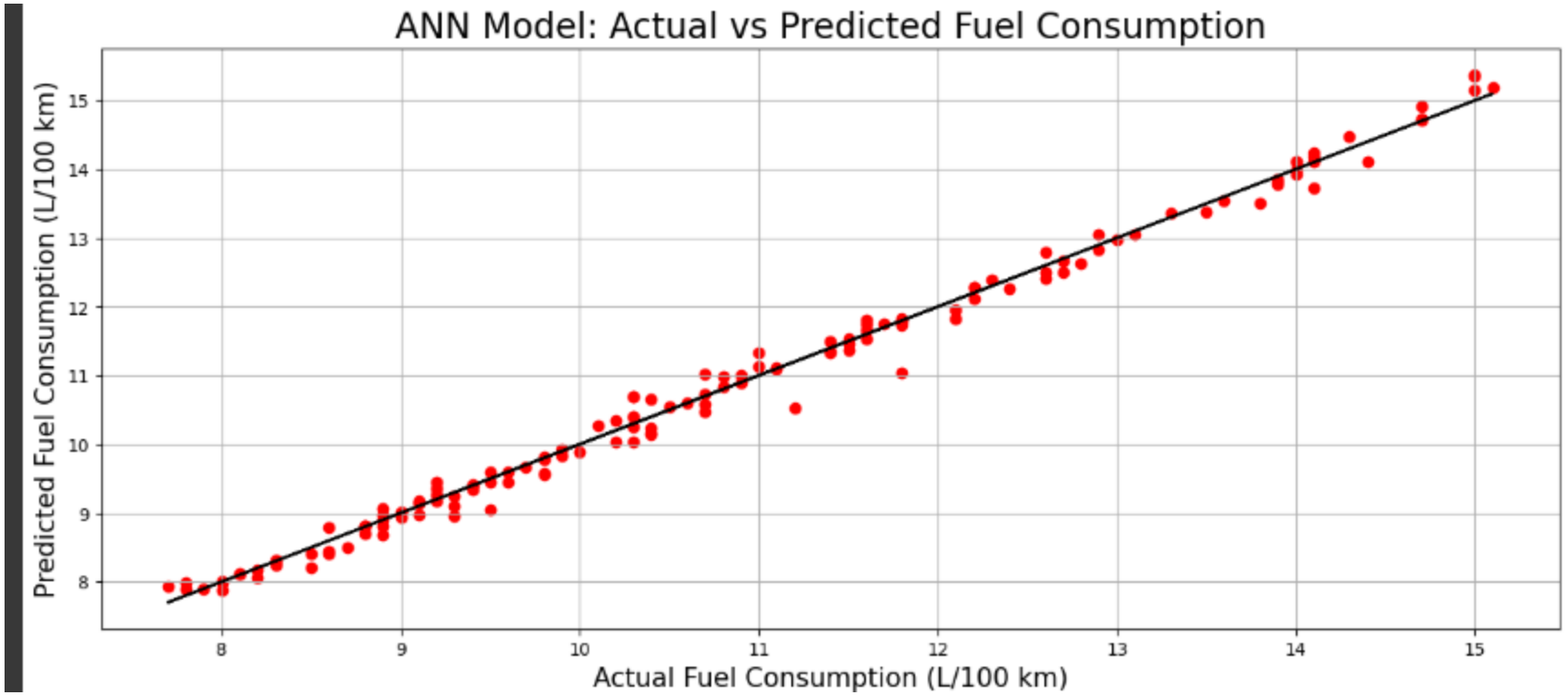

Figure 14: ANN over/underfitting testing

6.2.3 Support Vector Model

he SVM model's 99% accuracy rate in predicting the target variable. Providing remarkable accuracy and significant insights into complicated datasets, the SVM model is a strong and dependable tool for predictive modeling tasks due to its ability to generalize effectively to unknown data and minimize the hazards associated with overfitting.

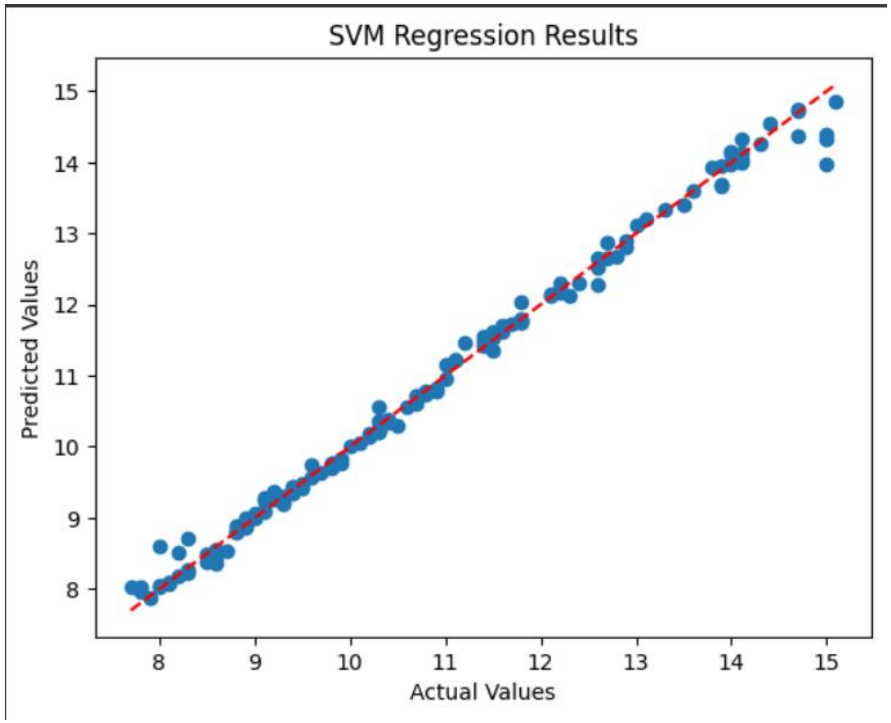

Figure 15: SVM over/underfitting testing

6.2.4 Random Forest

The feature significance plot's congruence with the 10-fold cross-validation results highlights how well the Random Forest model captures the underlying patterns in the dataset. This convergence of results demonstrates the durability and dependability of the model, establishing it as an important tool for data-driven decision-making and predictive modeling projects targeted at fuel consumption optimization.

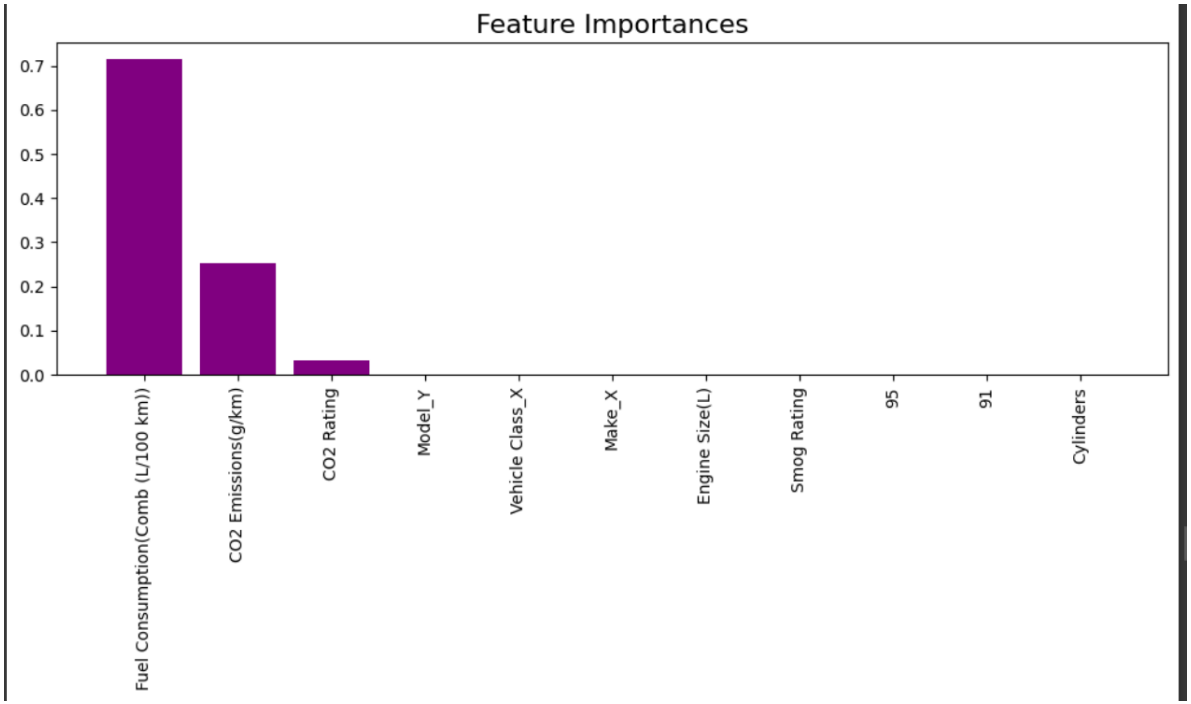

Figure 15: RF feature significance plot

6.2.5 K-Nearest Neighbor

While the KNN model's accuracy may not be the best among its competitors, its near prediction of genuine values highlights its usefulness and applicability in specific situations. A useful addition to the toolset of predictive analytics practitioners, the KNN model capitalizes on its capabilities in identifying local patterns and utilizing surrounding data points to provide insights and solutions that supplement those provided by more complicated algorithms.

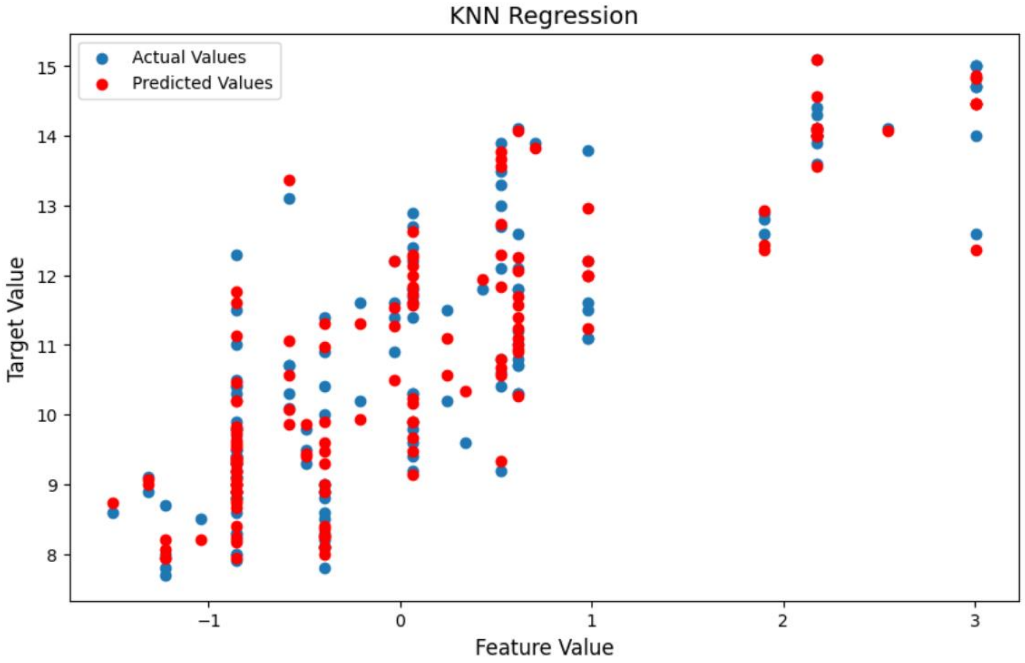

Figure 16: KNN regression

6.2.6 ADA Boost Regressor

Strong proof of the AdaBoost model's effectiveness and dependability in predictive modeling tasks can be found in the figure displaying its performance. AdaBoost is a powerful tool that facilitates the extraction of important insights from complicated datasets. It does this by skillfully utilizing adaptive boosting algorithms and ensemble learning approaches. This allows for the making of informed decisions and increased operational efficiency.

Figure 17: AdaBoost Regression

Table 14: Accuracy measures

Loss / Method	SVM	ANN	RF	LR	KNN	ADA
RMSE	0.17	0.17	0.023	0.0233	0.34	**0.04**
MAE	**0.10**	**0.12**	**0.01**	**0.0108**	**0.25**	**0.027**

The best models for estimating fuel consumption, according to a comparative study of RMSE and MAE values, are Random Forest (RF), closely followed by Support Vector Machine (SVM) and Artificial Neural Network (ANN). AdaBoost (ADA) and K-Nearest Neighbors (KNN) both show marginally higher error metrics, but they are still good options for estimating the consumption of fuel. and Linear Regression (LR), had signs of overfitting. In the end, the choice of model that performs well with all features and had min

6.3 Further Discussions

This section is the result of our project accuracy and previous accuracy using the same dataset, the previous study is 'FUEL_CONSUMPTION_ANALYSIS', and the following figure shows the difference between them:

Table 15: Comparison with previous studies

Study of previous work	
Linear Regression	91%
Decision Tree Regressor	85.7%
Random Forest Regressor	88.1%
Study of this work	
Linear Regression	100%
SVM	99.3%
Random Forest Regressor	80.7%
Decision Tree Regressor	99.9%

Our study on fuel consumption prediction has shown significant improvements and some unexpected outcomes when compared to previous research. In general, models such as Linear Regression and Decision Tree Regressor performed much better in our study than in earlier works. This suggests that our methodologies, including data preprocessing and model tuning, were particularly effective

Interestingly, our results for the Random Forest Regressor were not as strong as those reported in previous studies. This discrepancy indicates that the effectiveness of this model may vary depending on specific factors like data characteristics and the chosen parameters. Additionally, we introduced a Support Vector Machine (SVM) into our analysis, which was not part of the previous studies. The SVM model performed exceptionally well, highlighting its potential for this type of prediction task.

In summary, our study not only underscores the importance of proper data handling and model tuning but also suggests that certain models may perform better or worse depending on the context. The high performance of some models in our study, alongside the varied results for others, emphasizes the need for careful consideration of model selection and optimization in predictive analytics.

6.4 Alignment with the requirements:

Our project, which focuses on predicting fuel consumption using advanced machine learning models such as SVM, Random Forest, ANN, KNN, Linear Regression, and ADA, aligns perfectly with the client's needs for precise and dependable fuel consumption forecasting. The client requires a solution that not only delivers high prediction accuracy but also enhances decision-making, operational efficiency, and cost reduction. By leveraging our intelligent solution, the client gains several benefits. These include improved accuracy in predictions, which fosters trust and reliability in decision-making processes; significant cost savings through optimized fuel usage and reduced waste; enhanced operational efficiency via better resource planning and allocation; strategic planning capabilities based on accurate consumption patterns; and a reduced environmental impact by optimizing fuel consumption, which supports sustainability goals. Overall, our project addresses the client's requirements comprehensively, providing a robust solution that delivers practical benefits in multiple areas.

7.0 Conclusion and recommendations

In summary, this study underscores the efficacy of employing machine learning methodologies for forecasting fuel consumption based on vehicle attributes. Through a thorough examination of various regression models, including Linear Regression, Neural Networks, SVM, Decision Trees, and Random Forests, we have gleaned significant insights into their performance within this domain. Our findings highlight Neural Networks as the most adept in achieving superior prediction accuracy, closely followed by SVM, thus showcasing their potential in accurate fuel consumption prognostication. These results carry profound implications for optimizing vehicle efficiency and curbing environmental impact.

Recommendations for Future Work:

Looking ahead, we propose the following avenues for future research:

- Optimization of Model Parameters: Explore the impact of optimizing model parameters, such as learning rates and regularization techniques, on the efficacy of machine learning models for fuel consumption prediction.
- Advanced Feature Engineering: Delve into advanced feature engineering techniques to bolster the predictive capabilities of models. This could involve integrating additional vehicle attributes or extracting more insightful features from the existing dataset.
- Ensemble Methodologies: Investigate the effectiveness of ensemble methods, such as stacking or boosting, in further enhancing the accuracy of fuel consumption prediction models by amalgamating the strengths of multiple regression techniques.
- Real-Time Predictive Capabilities: Develop models with real-time predictive capabilities for fuel consumption, facilitating dynamic optimization of vehicle efficiency and emission reduction strategies.
- Integration of External Influences: Consider integrating external factors, such as weather conditions, traffic patterns, and driving behaviors, into prediction models to capture their impact on fuel consumption more accurately.

By delving into these areas of prospective research, we can propel the field of fuel consumption prediction forward and contribute significantly to fostering more sustainable transportation practices.

Acknowledgements

We express our heartfelt gratitude to Dr. Nawaf Alharbi for his outstanding support and guidance, which were critical to the success of our project. Dr. Alharbi's experience and supervision not only improved our research but also greatly influenced the recognition and honors our project received. We are appreciative for Dr. Alharbi's critical contribution in ensuring our project's win at the college's AI Projects Expo. His persistent support, encouragement, and dedication to our project's success are admirable. We are grateful for Dr. Alharbi's leadership and proud to have worked with him on this incredible journey.

References

[1] A. Schoen, A. Byerly, B. Hendrix, R. M. Bagwe, E. C. Dos Santos, and Z. Ben Miled, “A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles,” *IEEE Trans Veh Technol*, vol. 68, no. 7, 2019, doi: 10.1109/TVT.2019.2916299.

[2] S. Wickramanayake and D. H. M. N. Bandara, “Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study,” in *2nd International Moratuwa Engineering Research Conference, MERCon 2016*, 2016. doi: 10.1109/MERCon.2016.7480121.

[3] F. Perrotta, T. Parry, and L. C. Neves, “Application of machine learning for fuel consumption modelling of trucks,” in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2017. doi: 10.1109/BigData.2017.8258382.

[4] Z. Xu, T. Wei, S. Easa, X. Zhao, and X. Qu, “Modeling Relationship between Truck Fuel Consumption and Driving Behavior Using Data from Internet of Vehicles,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 3, 2018, doi: 10.1111/mice.12344.

[5] T. Bousonville, M. Dirichs, and T. Kruger, “Estimating truck fuel consumption with machine learning using telematics, topology and weather data,” in *Proceedings of the 2019 International Conference on Industrial Engineering and Systems Management, IESM 2019*, 2019. doi: 10.1109/IESM45758.2019.8948175.

[6] H. Drucker, “Improving Regressors using Boosting Techniques,” in *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, 1997, pp. 107-115

[7] E. Moradi and L. Miranda-Moreno, “Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling,” *Transp Res D Transp Environ*, vol. 88, 2020, doi: 10.1016/j.trd.2020.102576.

[8] Y. Yao *et al.*, “Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones,” *J Adv Transp*, vol. 2020, 2020, doi: 10.1155/2020/9263605.

[8] M. A. Hamed, M. H. Khafagy, and R. M. Badry, "Fuel Consumption Prediction Model using Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121146.

[9] G. M. H. Shahariar *et al.*, "Real-driving CO2, NOx and fuel consumption estimation using machine learning approaches," *Next Energy*, vol. 1, no. 4, 2023, doi: 10.1016/j.nxener.2023.100060.

[10] H. Abediasl, A. Ansari, V. Hosseini, C. R. Koch, and M. Shahbakhti, "Real-time vehicular fuel consumption estimation using machine learning and on-board diagnostics data," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2023, doi: 10.1177/09544070231185609.

[11] D. Zhao *et al.*, "A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption," *Energies*, vol. 16, no. 14. 2023. doi: 10.3390/en16145258.

[12] L. Zhang *et al.*, "Novel Neural-Network-Based Fuel Consumption Prediction Models Considering Vehicular Jerk," *Electronics (Switzerland)*, vol. 12, no. 17, 2023, doi: 10.3390/electronics12173638.

[14] JavaTPoint, "Support Vector Machine (SVM) Algorithm - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[15] N. Singh, "Support Vector Regression for Machine Learning - Analytics Vidhya - Medium," *Medium*, Jun. 05, 2020. https://medium.com/analytics-vidhya/support-vector-regression-for-machine-learning-843978ba6279).

[16] Perrotta, F., Parry, T., & Neves, L. C. (2017). Application of machine learning for fuel consumption modelling of trucks. Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January. https://doi.org/10.1109/BigData.2017.8258382

[17] Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1). https://doi.org/10.2166/wqrj.2018.025

[18] Reis, I., Baron, D., & Shahaf, S. (2019). Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. The Astronomical Journal, 157(1). https://doi.org/10.3847/1538-3881/aaf101

[19] Massoud, R., Bellotti, F., Berta, R., de Gloria, A., & Poslad, S. (2019). Exploring Fuzzy Logic and Random Forest for Car Drivers' Fuel Consumption Estimation in IoT-Enabled Serious Games. Proceedings - 2019 IEEE 14th International Symposium on Autonomous Decentralized Systems, ISADS 2019. https://doi.org/10.1109/ISADS45777.2019.9155706

[20] "Adaptive Boosting Algorithm Explained - GenesisCube," Dec. 20, 2023. https://genesiscube.ir/adaptive-boosting-algorithm-explained/ (accessed May 18, 2024).

[21] R. Quinlan, "Learning with Continuous Classes," in Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, 1992, pp. 343-34.