VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's Thesis

Final Energy Consumption in Road Transport by Means of Transport

Galutinis energijos suvartojimas kelių transporte pagal transporto rūšį

Dainius Tamuliūnas

Supervisor : Prof. habil. dr. Marijus Radavičius

Reviewer : Dr. Rimantas Eidukevičius

Vilnius
2025

Acknowledgements

I sincerely thank the social partner, the State Data Agency, for suggesting this master's thesis topic and Aušra Jablonskienė, Head of the Green Deal Statistics Division at Statistics Lithuania, for facilitating a two-month internship that provided access to Lithuanian car fleet data. Her support, guidance, and assistance with all my questions and ideas are sincerely appreciated.

Summary

This master thesis develops a comprehensive methodology for estimating annual fuel consumption in Lithuania's road transport sector. It addresses data gaps in primary datasets, integrates supplementary data, harmonizes diverse sources, and uses predictive machine learning models. The study defines the steps of data cleaning, preprocessing, modeling, and hyperparameter tuning while proposing a unified framework adapted to different vehicle types, such as passenger cars and vans. Unlike similar studies, this study incorporates regional and demographic characteristics, such as urban and rural driving patterns and transport vehicle owner age, into the modeling process.

Moreover, the study examines discrepancies between real-world fuel consumption and the official values provided by manufacturers, proposing adjustments based on real-world data. Gradient Boosting and Random Forest Regression are identified as the most effective methods, demonstrating high predictive power, fast performance while maintaining interpretability. Key vehicle features, such as engine size, weight, and power, have a significant impact on fuel consumption predictions. Nevertheless, the model could be further enhanced by incorporating more comprehensive real-world data from underrepresented vehicle categories and exploring additional factors such as seasonal variations and other vehicle characteristics. Additionally, including older vehicles, given that the current training dataset primarily includes 2021 and 2022 models, would likely improve the generalization and robustness of the model.

Keywords: Fuel Consumption, Lithuania's Vehicle Fleet, Real-World Data, Machine Learning, Regression, Linear Regressions, Gradient Boosting, CatBoost, XGBoost, LightGBM, Random Forest Regression, SVM, K-Means Clustering, Cross-Validation, Outlier Detection, Data Cleaning

Santrauka

Šiame magistro rašto darbe sukurta išsami metinių degalų sąnaudų Lietuvos kelių transporto priemonių sektoriuje vertinimo metodika. Darbe sprendžiant spragų pagrindiniuose duomenų rinkiniuose keliamas problemas integruojami papildomi duomenys, sujungiami įvairūs šaltiniai bei taikomi prognozavimo mašininio mokymosi modeliai. Aprašomi duomenų valymo, paruošimo, modelių kūrimo ir hiperparametrų parinkimo etapai, taip pat ir naudojamas bendras metodologinis pagrindas, pritaikytas skirtingiems transporto priemonių tipams, tokiems kaip lengvieji automobiliai ir furgonai. Šiame darbe, skirtingai negu panašiuose tyrimuose, į modeliavimą įtraukiami regioniniai ir demografiniai rodikliai, pavyzdžiui, miesto, rajono ir kaimo važiavimo ypatumai bei transporto priemonių savininkų amžius. Be to, šiame darbe nagrinėjami realių degalų sąnaudų ir gamintojų deklaruotų verčių neatitikimai, pasiūlytos jų korekcijas remiantis realaus pasaulio duomenimis. Tyrimas parodė, kad Gradiento Augimo regresija ir Medžių regresija yra efektyviausi metodai, pasižymintys dideliu prognozavimo tikslumu, sparčiu veikimu ir rezultatų interpretuojamumu. Reikšmingiausios transporto priemonės charakteristikos, įtakojančios degalų sąnaudas, yra variklio darbinis tūris, svoris ir galia. Visgi, rezultatai galėtų būti dar labiau patobulinti, jeigu būtų įtraukti papildomi duomenys apie retesnes transporto priemonių kategorijas bei atsižvelgta į sezoniškumą. Taip pat galėtų būti išplėsta modelio taikymo sritis pridedant senesnes transporto priemones, kadangi dabartinis mokymo duomenų rinkinys apima tik 2021–2022 metų gamybos modelius. Kartu su platesne realaus pasaulio duomenų integracija tai padidintų bendrą modelio apibendrinamumą ir patikimumą.

Raktiniai žodžiai: Degalų sąnaudos, Lietuvos transporto priemonių parkas, realūs duomenys, mašininis mokymasis, regresija, tiesinės regresijos, Gradient Boosting, CatBoost, XGBoost, LightGBM, Random Forest regresija, SVM, K-Means grupavimas, kryžminė validacija, išskirčių aptikimas, duomenų valymas

List of Figures

List of Tables

Contents

List of symbols

- L_1 stands for the sum of the absolute values of coefficients (used in Lasso Regression).
- L_2 stands for the sum of the squares of coefficients (used in Ridge Regression).
- 1.5IQR refers to the 1.5 * Interquartile Range used for outlier detection.
- k is the number of clusters in the k-means clustering algorithm.
- ϵ is the margin of tolerance (insensitivity zone) in Support Vector Regression.
- Std is standard deviation.

List of abbreviations

EU	European Union
EEA	European Environment Agency
SVM	Support Vector Machine
SVR	Support Vector Regression
RF	Random Forest
RFR	Random Forest Regression
GBR	Gradient Boosting Regression
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
ARE	Absolute Relative Error

Introduction

Energy consumption in road transport is becoming increasingly important on a global scale, driven by economic development, environmental concerns and technological shifts. According to recent European Commission reports road transport accounts for a substantial share of overall final energy consumption within the European Union (EU), far exceeding other modes of transport such as rail or air travel [11]. In Lithuania, this trend is even more pronounced: the transport sector consumes more energy than households or industry [20]. The high dependence on petroleum products and the logistical challenges associated with integrating renewable energy sources highlight the complexity and urgency of this issue. Policymakers, industry stakeholders, and scientists are increasingly recognizing the importance of reliable, data-driven insights in decision-making about reducing emissions and increasing overall energy efficiency.

Taking this into account, the master's thesis aims to develop a comprehensive methodology for estimating annual fuel consumption in Lithuania's road transport sector, broken down by specific vehicle categories. The thesis was proposed by the social partner *State Data Agency*, emphasizing the importance and practical significance of the study for national-level data initiatives. To gain access to key primary datasets, including the Lithuanian vehicle fleet and technical inspection records, I completed a two-month internship at the State Data Agency. During this period, I was granted limited environmental access to confidential datasets, allowing me to perform detailed data cleaning, feature engineering, and modeling tasks that are essential to building a robust estimation framework. Agency experts helped me understand the data, obtain additional variables related to vehicle owner information, and ensure that methodological choices were consistent with data constraints and policy objectives.

This study systematically applies advanced data science methods, ranging from conventional regression techniques to ensemble learning algorithms, to predict fuel consumption across various vehicle categories. Significant efforts have been directed toward integrating the training dataset with supplementary data sources to address gaps in the primary datasets. Access to real-world Lithuanian vehicle fleet datasets during an internship at the State Data Agency enabled the validation and adaptation of the proposed methodology for estimating fuel consumption in Lithuania across different means of transport.

This master's thesis presents a robust methodology for modeling fuel consumption in the Lithuanian road transport sector, emphasizing data-driven insights and methodological rigor. The proposed method will contribute to the efforts of the State Data Agency by providing the basis for a new model for estimating transport energy consumption. The following sections provide a detailed description of the literature review, data sources, methodological basis and analytical conclusions, and recommendations on how to improve fuel consumption estimation in Lithuania.

Goal of the Study

The goal of this study is to develop a comprehensive methodology for estimating annual fuel consumption in Lithuania's road transport sector. This involves addressing data gaps in primary datasets, integrating supplementary data, harmonizing diverse sources, and utilizing predictive machine learning models. The study will also evaluate the performance of the model, provide practical insights, and provide recommendations for future improvements.

Objectives of the Study

1. Conduct a thorough literature review to provide a theoretical foundation for the study.
2. Identify missing or incomplete information in the primary datasets and seek supplementary data sources to address gaps, ensuring that training datasets are complete and reliable.
    - Find training datasets that include mileage, vehicle parameters, and actual fuel consumption.
3. Formulate data analysis tasks to address specific research questions and select the most suitable solutions.
4. Harmonize multiple datasets to ensure consistency and compatibility for analysis.
    - Standardize and correct discrepancies in manually entered vehicle models and brands to ensure accurate mapping and consistent data.
    - Develop a method for calculating annual fuel consumption based on technical inspection data.
5. Apply predictive machine learning models to analyze fuel consumption and evaluate their performance to identify the most efficient ones.
    - Present the process of preparing and preprocessing the training datasets and the Lithuanian vehicle fleet dataset.
6. Reflect on the methodology, unresolved problems, and future research directions.

1 Literature Review

1.1 Road Transport Energy Consumption in the EU and Lithuania

The transport sector is a major driver of global economic and social activity, but it is also one of the largest sources of energy consumption. In 2022, transport activities were responsible for 31.0% of overall final energy consumption in the European Union, ahead of households (26.9%), industry (25.1%), and services (13.4%) [11]. Furthermore, among all forms of transport, road transport stands out as the largest energy consumer, accounting for 73.6% of total EU transport energy consumption, significantly exceeding air (11.4%) and water transport (13.0%) [11]. For comparison, in 2022, the transport sector in Lithuania consumed 40.4% of the country's total final energy consumption and households - 28.4% [20]. Diesel consumption in road transport amounts to 1.63 million tons, which accounts for 81.5% of all road transport fuel. Lithuania is also much more dependent on petroleum products for road transport energy needs than the EU – 81.5% of road transport energy is obtained from diesel, compared to 65.4% in the EU. Although gasoline accounts for 25. 2% of the energy of road transport in the EU, in Lithuania it only accounts for 14%. Moreover, the EU relies more on renewables and biofuels (6.4%) and electricity (0.3%) than Lithuania, where renewables and biofuels account for only 1.3% and electricity about 0. 4%. This shows Lithuania's dependence on diesel and the difficulty of integrating alternative energy sources.

1.2 Discrepancies Between Real-World and Official Fuel Consumption Figures

Every vehicle has an official fuel consumption value provided by the manufacturer, which is very important when assessing the vehicle's efficiency. However, many drivers find that fuel consumption exceeds the stated values over time, highlighting significant discrepancies in fuel efficiency reports.

An analysis by the European Commission, based on data collected from almost 3 million cars and vans between 2022 and 2024 revealed that real-world fuel consumption is approximately 20% higher than manufacturer's stated [8]. This discrepancy is due to differences in test environments, driver behavior, and vehicle conditions.

Driver behavior significantly increases the differences in fuel consumption and emissions. Shahariar et al. (2022) found that aggressive driving style increases CO emissions by 88%, and particulate matter (PM) by 112%, due to sharp accelerations and turbocharger lag events, particularly during off-peak hours [39]. Similarly, Mohammadnazar et al. (2024) showed that aggressive driving in work zones and curves increases fuel consumption by 23% compared to normal driving. By Adopting a calm/passive driving style you can increase fuel efficiency by up to 59% in work zones and 50% on freeways [24]. Seasonal fluctuations also have an impact. A study conducted in Finland showed that increased rolling resistance, lower temperatures, and harsh road conditions in winter significantly increase fuel consumption [2].

Even the Worldwide Harmonized Light Vehicles Test Procedure (WLTP), designed to provide more realistic assessments, fails to fully account for real-world variability. TNO (2023) [23] reported that WLTP-based values underestimate the fuel consumption of plug-in hybrid electric ve-

hicles (PHEVs), which consume up to three times more fuel in the real world due to low electrical efficiency.

These examples demonstrate that advanced modeling techniques are required to fully integrate real-world data. Methods such as Energy-Based Micro-Trip take into account region-specific driving patterns and environmental factors, thus providing valuable insights for policy development and improved testing systems. [29].

1.3 Policy Implications and Sustainability Considerations

Achieving energy efficiency and carbon reduction in road transport requires consistent policy measures and the integration of sustainable technologies. Regulatory frameworks, such as the European Green Deal, emphasize the transition to zero emissions through cleaner fuels [11], electrification of transport fleets [23], and improved monitoring systems [16]. The European Green Deal sets ambitious targets to achieve climate neutrality by 2050, including reducing greenhouse gas emissions by at least 55% by 2030 compared to 1990 levels. It also promotes the decoupling of economic growth from resource use and ensures that no region is left behind [10].

Policies promoting low-carbon fuels have shown significant potential in reducing greenhouse gas emissions. Research into low-carbon fuels reveals the viability of options such as biofuels, hydrogen, and synthetic fuels, especially when supported by policies that drive infrastructure development projects [47]. Similarly, the European Commission's conclusions highlight the role of advanced monitoring systems, such as on-board fuel consumption meters, in ensuring compliance with this policy and improving the accuracy of real-world emissions monitoring [23].

Behavioral changes, such as passive driving and reduced idling, complement technology and policy measures to deliver significant fuel savings. [39, 45]. However, increased fuel efficiency encourages more travel, which remains a challenge for policy effectiveness [41]. Addressing this problem requires integrated strategies that combine behavioral interventions, urban planning, and policy incentives.

Policy implementation is further complicated by unexpected global events. The COVID-19 pandemic has temporarily reduced transport energy consumption due to reduced mobility but has also exposed the vulnerabilities of current systems. Research suggests that leveraging these insights can promote lasting behavioral shifts, such as increased remote working and active travel, to sustainably reduce energy demand [28].

Future policies should prioritize promoting low-carbon infrastructure [46], supporting active travel, and establishing effective land-use planning [16]. Combined with technological innovation and behavioral change [16], these measures can effectively address the challenge of reducing carbon emissions from road transport and achieving global climate goals.

1.4 Machine Learning Approaches to Fuel Consumption Estimation

Traditional models often struggle to capture complex, nonlinear relationships between driving behavior, environmental factors, and vehicle configurations. Machine learning (ML) methods are

well suited for this area because they use data-driven models to identify complex feature patterns. However, achieving accurate results requires clean, complete, and sufficiently large datasets, which remains a major challenge in the real world. Challenges such as missing data, inconsistent records, and the lack of standardized data collection methods across regions make it difficult to develop effective machine learning and deep learning models.

Recent studies have investigated various ML models for estimating fuel consumption. Shahariar et al. (2023) [38], compared Gaussian Process Regression (GPR), Support Vector Machine (SVM), and Linear Regression (LR) models to estimate real-world driving emissions and fuel consumption. The study used data collected from a light diesel vehicle driven on the urban route by 30 drivers of different backgrounds to capture different driving behaviors and traffic conditions. GPR emerged as the most effective model, achieving an R^2 of 0.81 and an Absolute Relative Error (ARE) of 3.52% [38]. The study also concluded that while GPR outperformed other models, each ML model effectively predicted fuel consumption, as the results are very similar, demonstrating the potential of ML methods.

Zhang et al. (2024) [44] reviewed machine learning approaches for modeling energy consumption in electric vehicles, demonstrating the effectiveness of models such as Gradient Boosting Models (XGBoost and LightGBM), Random Forest Regression, and Support Vector Regression (SVR). Gradient boosting models have shown to be able to handle high-dimensional datasets and capture nonlinear relationships effectively, yet they require careful hyperparameter tuning to avoid overfitting. Random forest regression is known to be reliable and easy to interpret, making it a good choice for datasets with missing or disordered data. However, when it comes to capturing more complex data relationships, it may fall short compared to gradient boosting models. SVR performs well with smaller datasets and handles non-linear relationships effectively using kernel functions. However, it becomes computationally expensive as dataset size increases, which limits its scalability compared to ensemble methods.

In freight transport, Fang et al. (2023) [12] compared Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Networks in predicting fuel consumption using a dataset of 14,281 records from 1,110 Euro 6 articulated trucks combined with road condition data. Among these models, RF had the best overall performance, achieving an R^2 of 0.87, and an RMSE of 4.64. It handles nonlinear relationships well and estimates each variable importance, making it very effective for datasets with diverse features and noise.

However, SVR also showed good performance, with an R^2 of 0.83, and an RMSE of 5.12. This shows that it effectively models nonlinear relationships using kernel functions, but its scalability is limited by the high computational cost associated with large datasets. When predicting extreme values, SVR showed better performance compared to RF and Artificial Neural Networks, making it particularly suitable for scenarios involving extreme cases. This highlights the advantages of RF and SVR in fuel consumption modeling, with RF providing reliable general predictions, while SVR is well suited to handling extreme cases. In conclusion, machine learning methods offer significant advantages in fuel consumption estimation because they effectively capture complex, nonlinear relationships that traditional models have difficulty handling. While models like Random Forest and Gradient

Boosting handle high-dimensional and noisy datasets well, Support Vector Regression demonstrates advantages in smaller datasets and extreme value predictions.

2 Data Overview

This section provides an overview of the datasets used in this thesis, the data preprocessing steps, and the rationale for including additional datasets. Initially, I started with two main datasets: the current vehicle fleet in Lithuania and Lithuania's technical inspections, which were provided by the State Data Agency as the primary data sources. Below is a detailed description of the main Lithuanian vehicle information datasets and additional datasets that were used to train the model and improve fuel consumption results in Lithuania for each vehicle type.

2.1 Primary Datasets

The current vehicle fleet in Lithuania dataset is a comprehensive collection of all vehicles currently registered and legally operated on Lithuanian roads. It consists of approximately 2.1 million records of unique vehicles. The dataset provides information about each vehicle such as vehicle types, technical specifications (e.g., engine displacement, power, emissions standards), fuel types (primary and additional fuels), ownership details (e.g., individual or legal entity), and registration data (e.g., registration dates, municipality). It also provides information on the vehicle's seating capacity, weight, and environmental classification. All vehicle and owner identification information, including VIN, license plate numbers, and owner details, is anonymized to ensure privacy and preserve the ability to distinguish individual records. Developing a model to estimate fuel consumption for each vehicle type requires specific data on average fuel consumption (L/100 km) and mileage. However, these important attributes are not available in the current vehicle fleet dataset. To address this limitation, the technical inspection dataset was used as an additional data source.

The Lithuania's technical inspections dataset provides a detailed record of all vehicle technical inspections performed. It contains approximately 7.1 million records. This dataset includes information on the type of inspections carried out, such as regular inspections, extraordinary inspections, or registration inspections. Also, the vehicle make model, fuel type, inspection, and expiration dates. The main feature of this dataset is the odometer readings, which record the vehicle's mileage in kilometers at the time of inspection. This data is necessary to estimate the annual mileage of each vehicle. Although this dataset provides valuable insights, its limitations, such as the lack of average fuel consumption data, make it insufficient to train a model to estimate fuel consumption. Therefore, additional data sources are needed to achieve the research objectives.

To address these limitations, additional datasets and alternative sources of information were considered to find data that could be used to develop a fuel consumption estimation model. One of the datasets evaluated was the Fuel Economy dataset from the USA ([9]). This dataset provides official average fuel economy figures for approximately 48,000 unique vehicle models, providing detailed information on city and highway fuel efficiency, engine specifications, and additional attributes such as drivetrain and transmission type. The dataset required extensive preprocessing to convert American units (e.g., miles per gallon) to European metrics (e.g., liters per 100 kilometers).

The conversion process involved converting features using standardized formulas. For example, fuel consumption in miles per gallon for city and highway driving was converted to liters per 100

kilometers using the formula:

$$\text{L/100km} = \frac{235.21}{\text{mpg}}$$

Similarly, engine displacement values recorded in cubic inches were converted to cubic centimeters, and vehicle weights in pounds were converted to kilograms. These adjustments ensured consistency with Lithuanian data, allowing for meaningful comparison of data sets.

Despite these efforts, the US dataset presented several challenges. Many of the vehicles listed in the dataset are not widely used in Lithuania, and the dataset does not sufficiently capture unique driving behavior or environmental factors due to regional differences. For example, differences in road types, speed limits, and fuel quality between the US and Lithuania can significantly impact fuel consumption patterns. Due to these discrepancies, training the model on this dataset would result in predictions that would inaccurately reflect actual fuel consumption in Lithuania. However, the dataset still provided valuable benchmark data for exploratory analysis and insights into fuel consumption trends.

Real-world data from the European Environment Agency (EEA) was identified as a highly suitable dataset for model training purposes. This dataset was created as part of an initiative to collect real-world fuel consumption and CO_2 emissions data, addressing the performance discrepancies between laboratory-tested and real-world vehicle performance.[1]. It contains approximately 3.7 million car records and 500,000 van records, offering a wide range of data on vehicle specifications and possible real-world driving conditions. The dataset includes features such as engine performance, weight, fuel consumption, and emissions. One of the unique features of the dataset is the detailed tracking of real-world fuel consumption using on-board fuel consumption monitoring (OBFCM) systems. The dataset includes variables such as total fuel consumption (in liters), total distance driven (in kilometers), and specific driving patterns for hybrid and electric vehicles, such as distance driven with the engine off or in charge depletion modes. Another significant advantage is the availability of country-specific information for each entry, allowing for regional comparisons to select the countries most similar to Lithuania. This dataset enables the model to be trained to estimate average fuel consumption using real-world driving data and compare it to the average fuel consumption of the manufacturer, thus ensuring more accurate and representative estimates.

However, this dataset also has limitations. It primarily consists of information on cars and vans, excluding other vehicle types such as motorcycles and buses. Supplementary data are required for these excluded vehicle types. Average fuel consumption values provided by manufacturers were used because this method provides the most accurate estimates when real data is not available.

After evaluating the strengths and limitations of these datasets, the EEA Real-World Data was chosen as the primary source for training the fuel consumption model due to its relevance to the European context. This dataset ensures accurate, representative, and context-specific results for Lithuania, making it the most reliable basis for model development.

2.2 Supplementary Datasets

Additional dataset were used to eliminate discrepancies in vehicle make and model information and to include fuel consumption data for specific vehicle types. Although the model focuses on estimating fuel consumption for cars and vans, manufacturer-provided average fuel consumption values were used for other vehicle types due to a lack of real-world data. The supplementary datasets are detailed in the following subsection.

2.2.1 Motorcycles

The Total Motorcycle Fuel Economy Guide [27] was web-scraped as it contains detailed information on each motorcycle model, including the year, manufacturer, model name, engine size (cc) and cylinder count, average MPG, and average fuel consumption in liters per 100 kilometers (L/100km). Covering models from 1934 to 2018, over 100 pages were scraped to extract data for 5,657 unique models.

This data provides average fuel consumption for motorcycles and mopeds. According to the Official Statistics Portal [19], at the end of 2023, mopeds accounted for 1% and motorcycles for 3.3% of all registered vehicles in Lithuania. These insights make this dataset a valuable resource to help address fuel consumption data gaps for these types of vehicles.

2.2.2 Cars

An additional dataset was downloaded from a publicly available Git repository to refine brand and model names after initial standardization with the ChatGPT API. While the API eliminated most of the discrepancies, this additional dataset eliminated part of the remaining discrepancies, improving data accuracy.

2.2.3 Buses

Data on the average fuel consumption of buses was collected from online sources for specific models currently used in Lithuania. For buses, the focus was on models that account for more than 0.4% of all bus data, covering 28 different models and accounting for about 47% of all buses in Lithuania.

2.3 Data Preprocessing

2.3.1 The Current Vehicle Fleet in Lithuania Unique Vehicles Models and Makers Preprocessing

This section describes the preprocessing steps taken to clean and standardize data on vehicle models and manufacturers of the current Lithuanian car fleet. This data is very important when combining with the technical inspection data set, as both datasets needs to have the same models and manufacturers for the same vehicle records, but often contain inconsistencies resulting from manual entry during the inspection. Preprocessing focused on addressing several key issues with the dataset,

as the data from models and developers was handwritten. These included discrepancies, incorrectly entered names of car manufacturers and models, as well as missing, incomplete, or additional unnecessary data. Additionally, there were cases where manufacturer and model information were combined into a single column, further complicating the data standardization process.

Initial Cleaning

Before analysis, the dataset contained 67,000 unique combinations of car manufacturers and models. In order to standardize this data, first of all, Lithuanian symbols in text fields were replaced with their English equivalents (e.g., *š* to *s*). After that, non-alphabetic characters were removed and all text was converted to lowercase to ensure uniformity.

ChatGPT API-Based Cleaning

To removes spelling errors, abbreviations, and inconsistencies between car manufacturers and models, the GPT-3.5 Turbo API [26] was used as the primary data cleaning tool. However, challenges have arisen due to tokens and speed limits, as well as the tendency for large batches of unique vehicle models and manufacturers to result in missed entries or incorrect mappings.

To address these issues, an optimized batching strategy and parallel processing were implemented. The maximum token count, including the prompt, was set to approximately 1,000 tokens per request. This threshold was carefully chosen to ensure that each batch is processed accurately while minimizing the risk of token overflow. A token in GPT-based systems represents a unit of text, roughly equivalent to a short word or a punctuation mark. For example, the word ”car” counts as one token, while ”cars” is two tokens (”car” and ”s”). Within this limit, approximately 30–100 records could be processed in a single request, depending on the length of the vehicle maker and model names. Each batch was dynamically sized based on the token count, ensuring that the content and prompt did not exceed the size limit. Parallel processing with up to five concurrent threads allowed for faster handling of large dataset, significantly reducing the overall processing time. This approach significantly improved the reliability of the cleaning process, allowing for accurate and consistent corrections without loss of efficiency.

The following fine-tuned prompt was used for each batch:

```
Follow these instructions exactly:

1. For each row, provide the car maker and car model in this format:
   Original Marke: [original value]
   Original Modelis: [original value]
   Fixed Marke: [corrected value]
   Fixed Modelis: [corrected value]
   - 'Original Marke' and 'Original Modelis' should match exactly what was provided in the input.
   - 'Fixed Marke' should contain only the corrected car maker's name.
   - 'Fixed Modelis' should contain the corrected car model, without repeating the car maker.

2. Fix any spelling or grammatical errors in both 'Fixed Marke'
and 'Fixed Modelis' to reflect real car makers and models.

3. Expand abbreviations in 'Fixed Marke' (e.g., "MB" becomes "Mercedes-Benz", "VW" becomes "Volkswagen").

4. If multiple models are listed in 'Original Modelis', keep only the first one in 'Fixed Modelis'.

5. Ensure that the car maker and model are real (e.g., BMW X5, Audi A100).

6. Separate each row with '---'.
```

To assess the effectiveness of the cleaning process, the Levenshtein distance between the original and corrected manufacturer names and models was calculated [4]. This metric measures the number of single-character changes (insertions, deletions, or substitutions) required to convert one string to another.

- Low Levenshtein distances indicated minor corrections, such as correcting spelling errors.
- Large distances denoted large discrepancies, which were reprocessed to ensure accuracy.

An additional action was included to identify mismatched strings based on large Levenshtein distances. These mismatched rows were re-cleaned by an additional GPT-based processing step, ensuring the consistency and accuracy of the final dataset.

This cleaning method reduced the number of unique combinations of car manufacturers and models from approximately 67,000 to 24,557, making the dataset significantly more standardized and usable.

Additional Cleaning Step

To further enhance the cleaning process, an additional step was implemented to validate and correct both manufacturer and model names in the dataset. This was achieved using predefined lists of known vehicle brands and supplementary datasets described in the subsection 2.2.

Comprehensive lists of well-known brands were compiled for various types of vehicles, including cars, motorcycles, buses, vans, mopeds and special vehicles. These lists were used as a reference for validating brand names in the dataset. To ensure consistency, brand names were cleaned and normalized:

- Special and accented characters were converted to their standard forms.
- Any additional information following certain symbols (e.g., /, \, :) was removed.
- For brand names with more than two words, only the first two words were retained.

To validate and match the cleaned brand names against the known brands list, fuzzy matching was applied using `fuzz.token_set_ratio` from the Python library `fuzzywuzzy` [37]. This method evaluates word-level similarities, making it robust against reordering or additional spaces. The process included:

- Using strict 100% thresholds for all brands names to ensure high-confidence matches.
- Names that did not meet these thresholds were flagged for manual review.

Supplementary datasets were used to validate vehicle models. These datasets provided comprehensive coverage of known models associated with specific brands and vehicle types. Model names were matched within their respective brands. This ensured that each model was validated against the correct brand list. The matching process included:

- Filtering supplementary datasets to retrieve models associated with the given brand.

- Applying `fuzz.token_set_ratio` to match models within the brand's list.
- A stricter threshold of 85% for longer models names (six or more characters).
- A lower threshold of 80% for shorter brand names.
- Flagging records for manual review if no match or low confidence match is found.

The results of brand and model matching with high-confidence matches were directly updated, while records flagged for manual review were retained for further review. This additional cleaning step further refined the dataset, reducing the total number of unique combinations from a previous count of 24,557 to a final count of 23,862.

2.3.2 The Current Vehicle Fleet in Lithuania Dataset Preprocessing

This section outlines the preprocessing steps for integrating and cleaning the primary datasets subsection 2.1 Lithuanian vehicle fleet and the technical inspection datasets to ensure compatibility and reliability for fuel consumption modeling.

Integration of Cleaned Manufacturer and Model Data

As described in subsection 2.3, the cleaning process for vehicle manufacturers and models resulted in a standardized dataset, with corrected values stored alongside the original entries. The cleaned dataset included two new columns, which contained the corrected names. These were mapped back to the primary datasets using the original entries of manufacturers and models as keys. This mapping ensured that identical vehicles across both primary datasets shared the same cleaned manufacturer and model names, enabling accurate data merging.

Inspection Data Filtering

The technical inspection dataset was filtered to include only inspections conducted between 2020 and the beginning of 2024. This period was chosen to approximate the mileage for 2023, as mileage records are available only through technical inspection data. According to [22], for new vehicles, including motorcycles, the first mandatory technical inspection is carried out after three years of registration, and then every two years.

In order to calculate mileage accurately, it is assumed that driving behavior will remain the same over time, so an approximate annual mileage can be calculated based on the available inspection data. For each vehicle, the latest inspection date before January 1, 2023, and the earliest inspection date after that were selected to determine the exact time intervals for mileage calculation. Invalid or incomplete records were removed based on the following criteria:

- **Duplicate Removal:** Duplicate odometer readings and inspection dates have been removed to avoid redundancy and ensure data consistency.
- **Incorrect VIN deletion**: Records with incorrect Vehicle Identification Numbers (VINs) were filtered out.

The annual mileage for each vehicle was calculated using the following formula:

$$\text{Annual Mileage (km)} = \frac{\text{Odometer Difference (km)}}{\text{Time Difference (days)}} \times 365 \quad (1)$$

Lithuania Fleet Sample with Mileage

Lithuanian vehicle fleet data were combined with technical inspection records using vehicle identification numbers, as well as cleaned manufacturer, model, and unique code information. The combined dataset had a total of 2.1 million records. Of these, approximately 671,000 records (33%) were successfully linked to available inspection data before and after 2023 to calculate mileage. Additionally, 715,000 records were partially mapped, containing either before or after the 2023 inspection data but not both, while 770,000 records were not mapped.

Vehicle Type Categorization

Vehicles were categorized into predefined types as shown in Table 1 (e.g., *K1* for Motorcycles, *K2* for Passenger Cars) based on their transport vehicle types. This categorization ensured uniform grouping for analysis.

Table 1 presents the distribution of vehicle types and their corresponding counts as of the end of 2023 [21]. This table summarizes the dataset used in this study, highlighting the proportion of vehicles for which mileage information can be calculated in 2023 based on technical inspection data.

Table 1. *Vehicle Types and Technical Inspection Data (2024-01-01)*

Vehicle Type	Description	Number of Vehicles (2024-01-01)	% Vehicles with Milleage Information (2023)
K0	Total	2,056,580	26.74%
K1	Motorcycles	67,283	10.26%
K2	Passenger Cars	1,700,524	25.81%
K4	Buses	7,573	68.02%
K5	Trolleybuses	383	39.95%
K6	Vans	118,035	32.36%
K7	Semi-Trailer Trucks	54,451	42.93%
K8	Semi-Trailers	52,423	0.25%
K9	Trailers	20,408	0.06%
K10	Special Vehicles	15,131	11.80%
K15	Mopeds	20,369	1.72%

Outlier Detection and Handling

Outliers in mileage data were identified for each vehicle type using the 1.5 IQR method and were removed. Resulting in a final sample size of approximately 538,000. This step ensured data reliability by mitigating the impact of anomalous records.

Among buses, 301 unique manufacturer and model combinations were observed, with the top 25 models accounting for 42% of the total bus records in the final sample. For trolleybuses, the

data revealed a highly concentrated distribution, with only 7 unique models representing 95% of the trolleybus dataset. For motorcycles (*K1*), web-scraped data will be used to obtain average fuel consumption for each model subsubsection 2.2.1. In addition, fuel consumption data for trolleybuses and buses will be used directly using average fuel consumption values found for each model online manually subsubsection 2.2.3.

Correlation Analysis

The correlation analysis examined the relationship between mileage and vehicle owner age, focusing on regional and demographic variations. Municipalities were categorized as rural, city, and metropolitan to capture demographic-specific patterns in vehicle usage. The overall correlation between mileage and owner age was weak, with notable regional differences. Stronger correlations were observed in metropolitan areas, suggesting higher vehicle usage and cost implications in these regions.

Table 2 provides an overview of correlation values grouped by municipality categories and age groups. In metropolitan regions, younger vehicle owners (ages 18–25) showed a slightly positive correlation with mileage 0.06, while older age groups demonstrated progressively negative correlations, reaching -0.22 for those aged 70 and above. Similar trends were observed in city and rural municipalities.

Table 2. *Correlation Between Mileage and Age by Municipality Category*

Municipality Category	Age Category	Percentage Count (%)	Correlation with Mileage (r)
Metropolitan	18 to 25	0.45	0.06
	25 to 30	1.01	0.03
	30 to 40	4.38	-0.01
	40 to 50	4.92	-0.03
	50 to 60	4.28	-0.05
	60 to 70	3.56	-0.08
	Over 70	9.10	-0.22
City	18 to 25	0.36	0.07
	25 to 30	0.79	-0.03
	30 to 40	2.94	-0.04
	40 to 50	3.41	-0.03
	50 to 60	3.86	-0.05
	60 to 70	3.40	-0.08
	Over 70	6.16	-0.17
Rural	18 to 25	1.27	0.11
	25 to 30	2.47	0.00
	30 to 40	8.08	-0.03
	40 to 50	9.05	-0.02
	50 to 60	11.07	-0.07
	60 to 70	8.60	-0.09
	Over 70	10.79	-0.19

Further correlation analysis by vehicle type and owner age revealed significant differences as correlations were segmented by vehicle types and municipalities. Passenger cars (*K2*) in metropolitan areas showed slightly positive correlations for younger owners (ages 18–25), reflecting higher mobility needs, while older age groups exhibited weaker or negative correlations.

2.3.3 Training Dataset Preprocessing

Real-world data from the European Environment Agency (EEA) (training dataset) was carefully preprocessed to ensure it accurately represents fuel consumption patterns in Lithuania. The preprocessing steps addressed issues related to data quality, geographic relevance, and fuel consumption of hybrid vehicles, resulting in a dataset suitable for modeling. Due to subsection 1.2, this work used real-world data rather than manufacturers' official fuel consumption data.

Dataset Cleaning and Feature Selection

Unnecessary columns such as identifiers and additional information data were discarded and focused on key variables, including fuel consumption, mileage, and vehicle specifications. Column names and fuel types were translated into Lithuanian to fulfill the language aspect of the work. Rows missing fuel type information were deleted. All of the steps below are performed on the vans and cars

dataset. For hybrid vehicles, fuel types were separated into primary and additional categories, with the primary type retained in the main fuel type column which is diesel or petrol, and the secondary-electric, placed in a separate column. After that invalid hybrid rows, such as those with missing or zero values for key fuel consumption metrics, were removed to improve data reliability. The consumption of electricity was calculated using the formula:

$$\text{Energy Consumption (Wh/km)} = \frac{\text{Total Grid Energy into Battery (kWh)} \times 1000}{\text{Distance Traveled with Engine Off (km)}}.$$

Fuel consumption during engine operation was calculated using the formula:

$$\text{Fuel Consumption (L/100km)} = \frac{\text{Total Fuel Consumed during Engine Operation (liters)}}{\text{Distance Traveled with Engine Running (km)}} \times 100.$$

For non-hybrid vehicles, fuel consumption (L/100km) was calculated as the ratio of total fuel consumed to total distance traveled, multiplied by 100.

Outliers in hybrid vehicle data for both cars and vans were separately removed by analyzing 2 ratios, first energy consumption (Wh/km) and then fuel consumption (L/100km), using the Interquartile Range method with a multiplier of 1.5 (1.5IQR). This ensures that the dataset reflects typical driving behavior and fuel consumption patterns. During this process, 116 vehicles (19%) were removed from the vans dataset, and 93,089 (21%) from the cars dataset. To ensure clarity and comparability, fuel types for hybrids were updated by adding suffixes *'_H'*, distinguishing hybrid vehicles from non-hybrids in the dataset.

Removal of Irrelevant Geographic Data

Countries with terrain and driving conditions dissimilar to Lithuania's flat landscape were excluded using information from [42]. This decision was intended for countries such as Austria, Italy, Norway, and others whose mountainous terrain has a significant impact on fuel consumption patterns. Only cars and vans using diesel or petrol are removed to preserve data for other fuel types as deleting them would significantly reduce other fuel types, see Table 3. In total, 19 countries were removed, reducing the dataset by 41% (approximately 1.5 million records) for cars and 39% (approximately 78 thousands records). These exclusions ensured the dataset's geographic relevance to Lithuanian driving conditions Table 13. Additionally, data counts for cars and vans across all countries in the dataset are provided in Table 14 to give a comprehensive overview of the remaining data distribution by countries.

If we filter the data only by country and not just by diesel and gasoline, the following changes are observed: the number of LPG for cars decreases to 211, biomethane to 32, and natural gas to 0. For vans, both biomethane and natural gas are reduced to 0.

Filtering Low Driving Distances

After analyzing the total distances driven, it was observed that a significant number of data contained very short distances. To improve the quality of the dataset, all vehicles with a total distance of less than 5 km were removed, as the average fuel consumption of those vehicles varies significantly. For cars, 78,532 vehicles were deleted, which represents 3.70% of the data. Similarly, for vans, 4,364 were deleted, representing 3.58% of the dataset.

Table 3. *Effect of Filtering by Country and Fuel Type on Data Counts*

Fuel Type	Initial Cars	Remaining Cars	Initial Vans	Remaining Vans
Petrol	2,135,477	1,270,884	12,660	4,941
Diesel	980,786	553,938	185,614	115,948
Petrol Hybrid	290,844	176,916	69	29
Diesel Hybrid	135,508	78,564	526	362
Ethanol	23,738	23,738	421	421
LPG	18,490	18,490	12	12
Biomethane	1,599	1,599	164	164
Natural Gas	174	174	88	88
Electric	3	3	0	0

Outlier Removal for Non-Hybrid Vehicles

Hybrids were excluded from this analysis to avoid redundancy since their outliers were removed in a prior step using energy consumption (Wh/km) and fuel consumption data (L/100km). The process focused on the remaining dataset to ensure that only non-hybrid vehicles underwent further filtering. The 1.5 IQR method was applied to the total distance traveled (km) for each combination of vehicle type and fuel type. This method ensured the proper identification and removal of each specific type of outlier, improving the quality and representativeness of the data set. The results show the number of outliers identified and removed, as well as the percentage of outliers for each category. For details refer to Table 4.

Table 4. *Outlier Removal for Non-Hybrid Vehicles by Total Distance Traveled (km)*

Vehicle Type	Fuel Type	Total Rows	Outliers Found	% Outliers Found
K2	Petrol	1,217,796	44,768	3.68%
K2	Diesel	542,648	23,143	4.26%
K2	LPG	12,513	1,358	10.85%
K2	Natural Gas	113	18	15.93%
K2	Electric	3	0	0.00%
K2	Ethanol	23,726	2,274	9.58%
K2	Biomethane	1,597	152	9.52%
K6	Petrol	4,250	343	8.07%
K6	Diesel	110,127	6,288	5.71%
K6	LPG	9	1	11.11%
K6	Natural Gas	80	0	0.00%
K6	Ethanol	421	0	0.00%
K6	Biomethane	163	18	11.04%

This step ensured that the data reflected realistic driving distances for non-hybrid vehicles while maintaining the integrity of the previously filtered hybrid fuel types.

Fuel Consumption Analysis

To analyze the fuel consumption data, graphs were created for each vehicle type and fuel type. After analyzing the results, it was noted that some outliers remained in the data. For example, the Figure 1 shows the fuel consumption distribution for petrol vehicles. It is clear that there are still some extreme, unrealistic values in the data set.

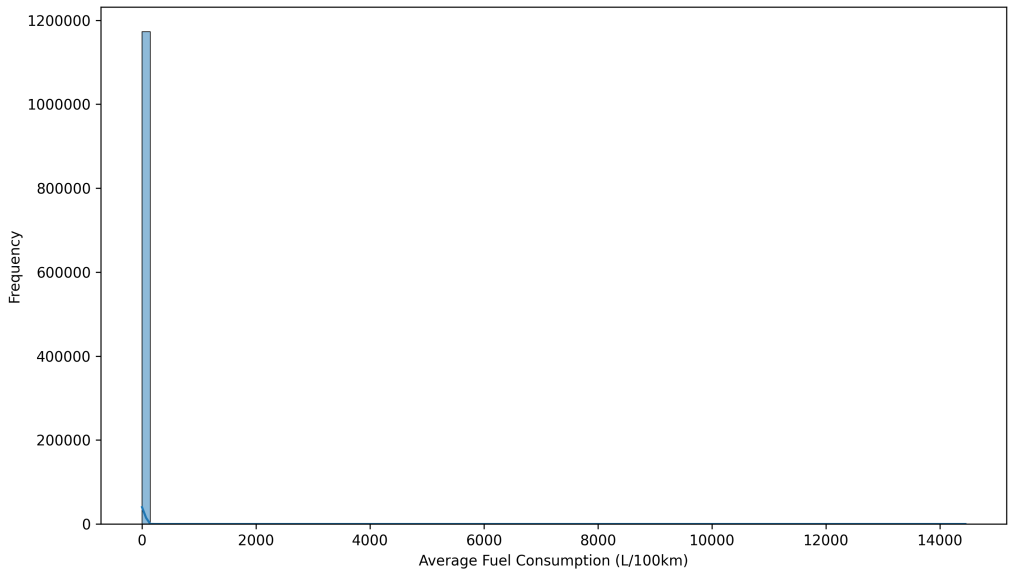

Figure 1. *K2 Petrol Fuel Consumption Distribution*

To address this issue, the 1.5 IQR method was applied to remove outliers from non-hybrid vehicles based on their fuel consumption ratios. As a result, a total of 200,462 outliers were deleted. The majority of these outliers were from K2 petrol vehicles, accounting for about 150,000 records, while around 41,000 records were removed from diesel vehicles.

Table 5. *Outlier Removal for Non-Hybrid Vehicles Based on Fuel Consumption*

Vehicle Type	Fuel Type	Total Rows	Outliers Found	% Outliers Found
K2	Petrol	1,173,028	150,266	12.81%
K2	Diesel	519,505	41,448	7.98%
K2	LPG	11,155	13	0.12%
K2	Natural Gas	95	0	0.00%
K2	Electric	3	0	0.00%
K2	Ethanol	21,452	2,308	10.76%
K2	Biomethane	1,445	147	10.17%
K6	Petrol	3,907	168	4.30%
K6	Diesel	103,839	6,070	5.85%
K6	LPG	8	1	12.50%
K6	Natural Gas	80	7	8.75%
K6	Ethanol	421	21	4.99%
K6	Biomethane	145	13	8.97%

After examining the outliers in the non-hybrid data, the fuel consumption of hybrid vehicles was further analyzed. However, outliers were removed during the hybrid preprocessing steps, but additional unrealistic fuel consumption values were identified. For example, the highest fuel consumption for petrol hybrids was 625 L/100km, and for diesel, 173 L/100km. To correct for this, all values exceeding the highest fuel consumption observed in non-hybrid cars between diesel and petrol vehicles (15.27 L/100km) were filtered out.

The fuel consumption distributions for each vehicle type and fuel type were then re-plotted to confirm that outliers had been removed and to ensure data accuracy.

Table 6. *Maximum Fuel Consumption Values for K2 Transport Vehicles*

Fuel Type	Maximum Fuel Consumption (L/100km)
Petrol	15.27
Petrol Hybrid	625.18
Biomethane	15.33
Diesel	13.07
Diesel Hybrid	173.33
Electric	9.40
Ethanol	10.21
Natural Gas	31.59
LPG	40.00

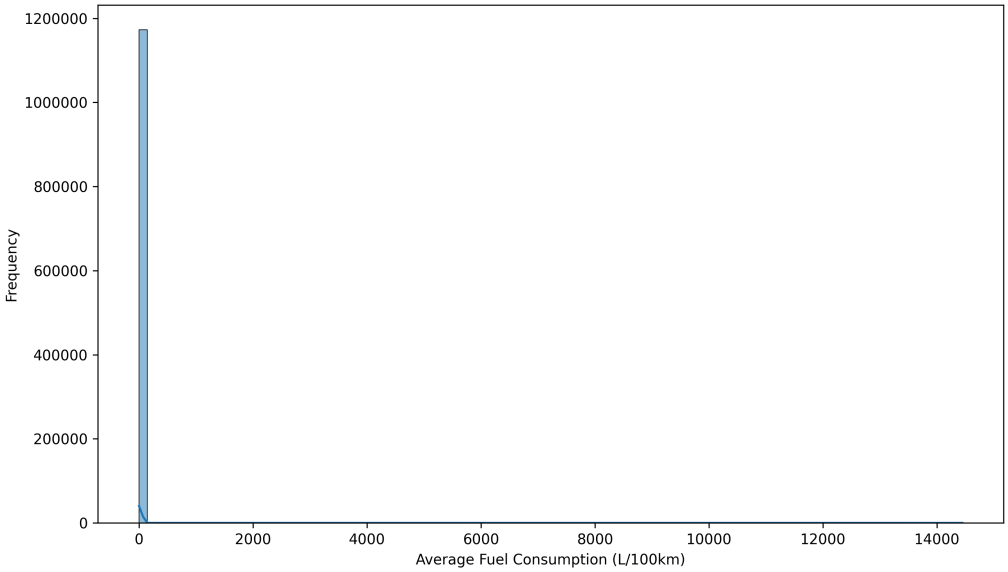

Figure 2. *K2 Petrol Fuel Consumption Distribution*

Table 7 shows the final number of records for each fuel type and vehicle type. In addition to this Figure 3 and Figure 4 present box plots of real-world fuel consumption and manufacturer-provided fuel consumption data for K2 and K6 vehicle categories, respectively. These plots highlight the variations of fuel consumption across different fuel types. It shows that the real world fuel consumption is higher than manufacturer's stated as it was discussed in the [8].

At the beginning of the analysis, we noticed significant differences between hybrid petrol and diesel fuel consumption. However, the final cleaned results show that the differences for K2 vehicles are insignificant, and for K6 diesel vehicles as well. The only exception is petrol, where hybrid petrol fuel consumption (L/100 km) is lower in both real-world and manufacturer-provided data. This difference appears to be related to the small number of records in the petrol hybrid dataset, which contains only 24 records compared to 3,739 records for petrol. The small sample size for hybrid petrol vehicles makes it difficult to draw robust conclusions about their actual fuel efficiency.

Table 7. *Final Row Counts by Vehicle Type and Fuel Type*

Vehicle Type	Fuel Type	Row Count
K2	**Petrol**	**1,022,762**
K2	Petrol Hybrid	157,569
K2	Biomethane	1,298
K2	**Diesel**	**478,057**
K2	Diesel Hybrid	74,706
K2	Electric	3
K2	Ethanol	19,144
K2	Natural Gas	95
K2	LPG	11,142
K6	**Petrol**	**3,739**
K6	Petrol Hybrid	24
K6	Biomethane	132
K2	**Diesel**	**97,769**
K6	Diesel Hybrid	344
K6	Ethanol	400
K6	Natural Gas	73
K6	LPG	7

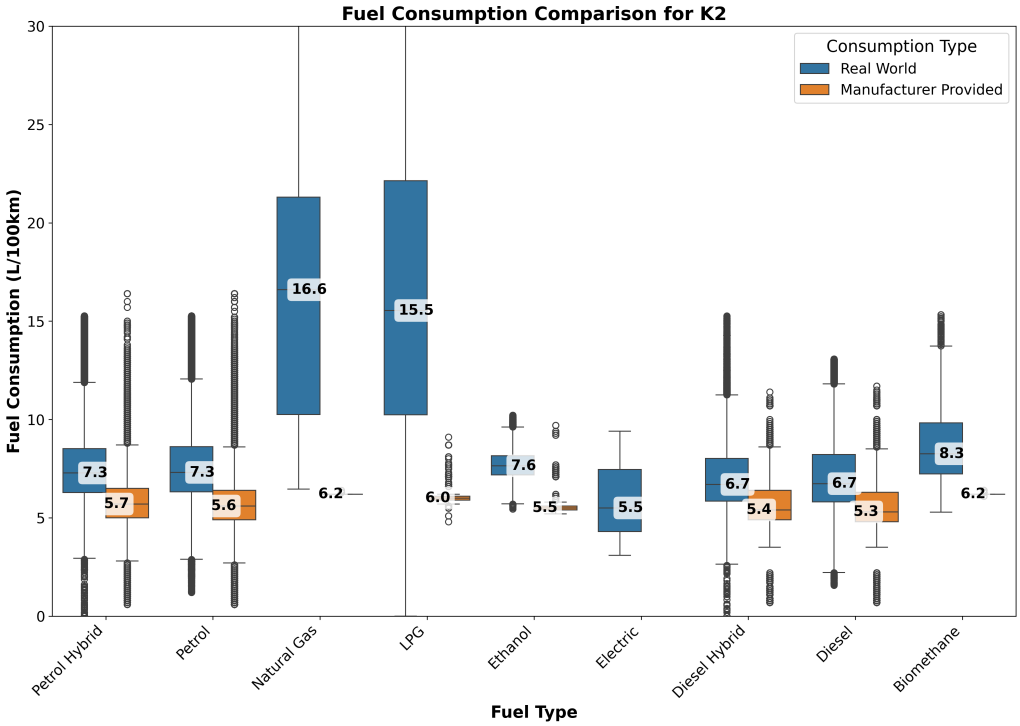

Figure 3. *Fuel Consumption Comparison for K2 Vehicles*

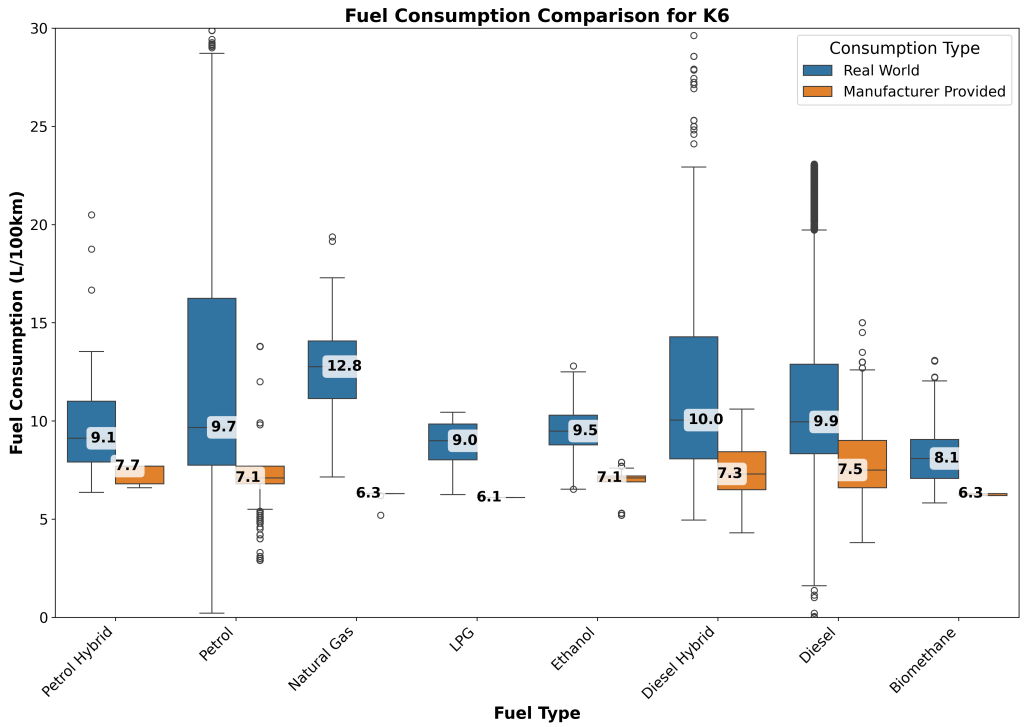

Figure 4. *Fuel Consumption Comparison for K6 Vehicles*

Comparing the number of records for each fuel type and vehicle type in Table 7 with the Lithuanian car fleet with mileage records, it is obvious that there is sufficient data for K2 category petrol, diesel and hybrid vehicle with electricity. For the K6 category, most data is concentrated on diesel vehicles, with a smaller amount for petrol vehicles.

Following this analysis, only petrol and diesel fuel types were selected for further modeling. This decision is based on the following considerations:

- **Data Availability in Training Dataset:** The Table 7 shows that the majority of the K2 and K6 vehicle dataset is concentrated in petrol and diesel types, ensuring robust statistical analysis and sufficient data. In contrast, hybrid data is limited to vehicles using electricity as one of the fuel types, and there are no records of other hybrid variations, further limiting the analysis.

- **Hybrid Complexity:** The inclusion of hybrid vehicles introduces additional variables related to electricity consumption, such as grid energy usage and battery operations. These variables are less standardized and more challenging to incorporate into a unified modeling framework effectively.

- **Data Limitations in the Lithuanian Fleet:** The Lithuanian car fleet with mileage records dataset has limited data for hybrids that use electricity as one of the fuel types, particularly for petrol hybrids in the K6 category. Due to the insufficient sample size, it is challenging to derive reliable results or make generalizations about fuel consumption for hybrids with electricity as a fuel source.

Final Training Dataset

The preprocessing steps, including outlier removal, distance filtering, and comprehensive data cleaning, an improved dataset containing 1,602,327 records was created. This dataset reflects real-world driving conditions in Lithuania and provides a reasonable data quality for fuel consumption modeling.

Several visualizations were created to gain insights into the key characteristics of the dataset. Figure 5 shows the distribution of kilometers driven by K2 diesel vehicles, illustrating typical mileage patterns for this category. The distribution of fuel consumption for K2 diesel vehicles is presented in Figure 6, highlighting the differences in actual consumption. Finally, Figure 7 presents a Spearman correlation plot for K2 diesel cars, revealing the relationship between variables such as fuel consumption, engine displacement, power, weight, and year of manufacture.

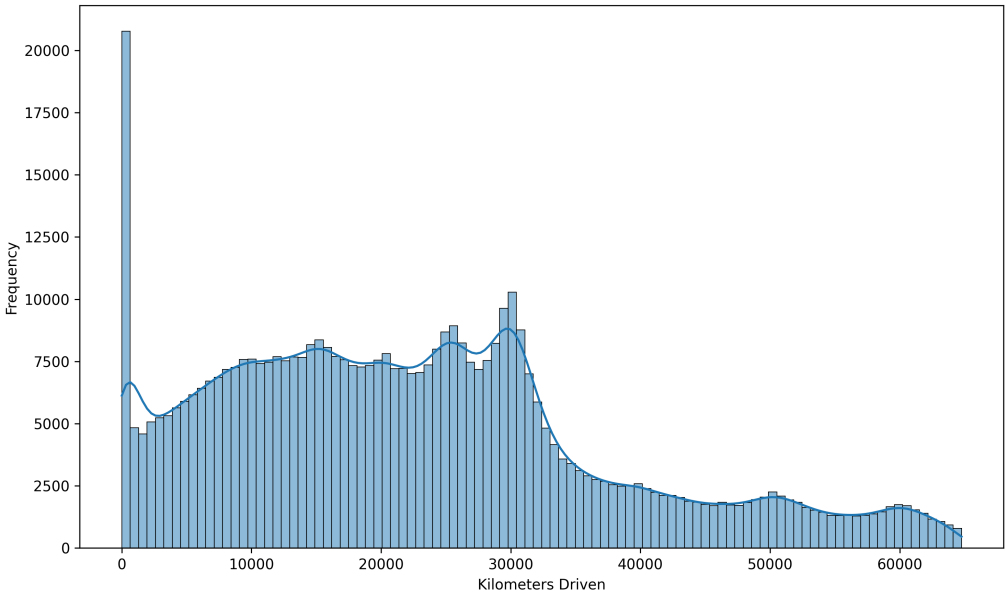

Figure 5. *Kilometers Driven Distribution for K2 Diesel Vehicles*

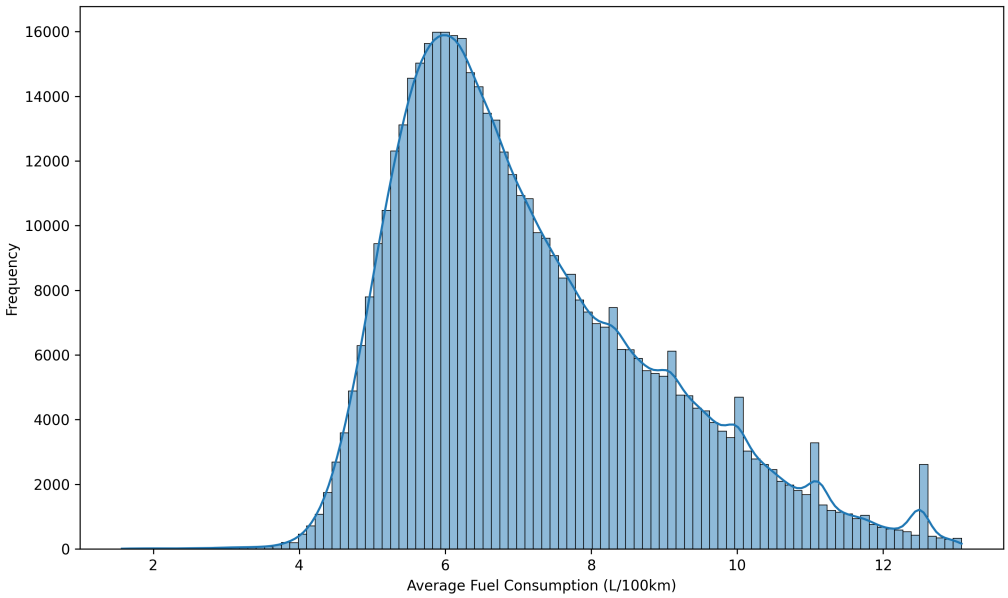

Figure 6. *Fuel Consumption Distribution for K2 Diesel Vehicles*

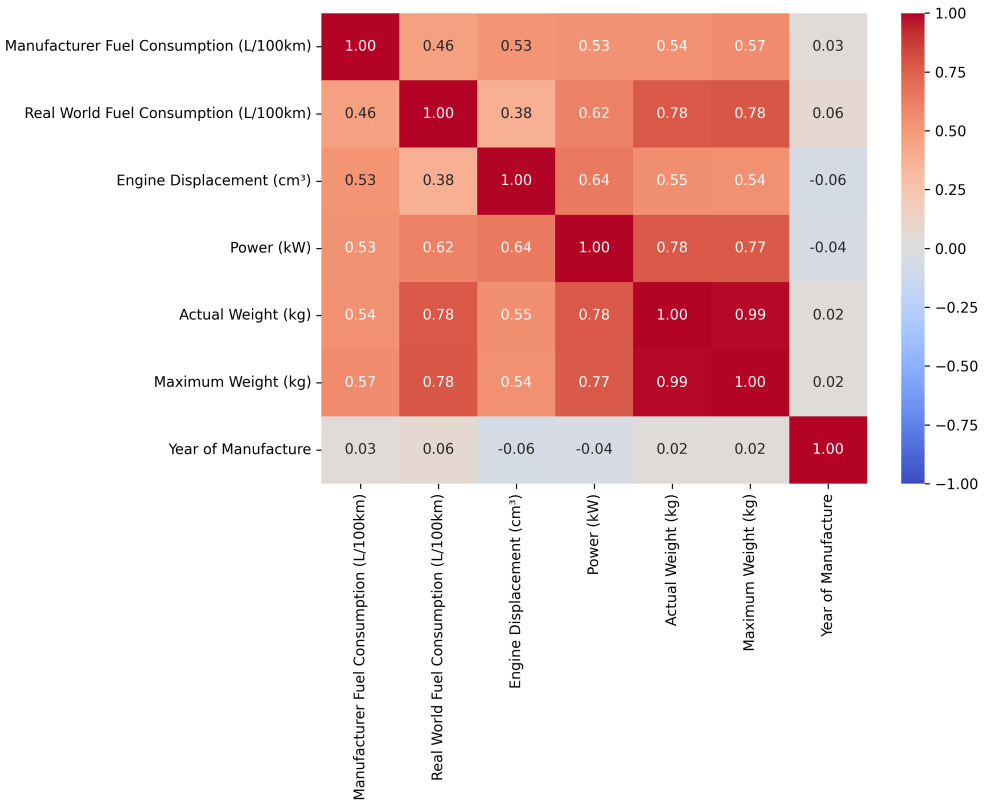

Figure 7. *Correlation Heatmap for K2 Diesel Vehicles*

Table 8. *Descriptive Statistics of Key Variables in the Final Dataset*

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Fuel Consumption (L/100km)	454,571	5.41	1.61	0.70	4.80	5.30	6.30	11.70
Fuel Consumption (Original)	478,057	7.15	1.77	1.57	5.82	6.74	8.22	13.07
Engine Displacement (cm^3)	478,057	1,961.23	382.55	999	1,950	1,968	1,993	2,997
Power (kW)	477,672	129.35	41.19	55	96	120	145	258
Actual Weight (kg)	476,683	1,773.82	322.91	1,161	1,518	1,715	1,975	2,711
Max Weight (kg)	477,994	1,937.53	366.59	1,244	1,654	1,873	2,143	3,019

The dataset ensures high data quality and geographic relevance, providing a solid foundation for accurate predictive modeling. The main observations from the descriptive statistics (Table 8) and visualizations are as follows:

- **Kilometers Driven (Figure 5):** K2 Diesel shows higher average mileage (mean: 22,612.6 km), indicating frequent use, likely for commercial or long-distance purposes. For K2 Petrol, the distribution captures both moderate and high-intensity use (mean: 11,905.3 km), reflecting diverse driving patterns in a large dataset. K6 Petrol and K6 Diesel show sharp peaks at very low mileage, likely due to fewer records, limited use, or vehicles recently added to operation.
- **Fuel Consumption (Figure 6):** The distribution shows a long tail with clear small peaks indicating subgroups of vehicles with different usage intensity or maintenance habits. The overall

shape resembles a lognormal distribution, which corresponds to the natural variability in real-world driving conditions. Despite the tail, the distribution remains well-structured and representative, providing a reliable data for robust modeling. Similarly, the K2 gasoline graph shows a similar pattern. The K6 diesel graph shows more pronounced peaks at specific values, likely due to the smaller data set and possible clustering around specific vehicle models. In contrast, the distribution of K6 gasoline has two dominant peaks, likely due to the smaller data set size and the influence of several vehicle groups.

- **Correlations (Figure 7):** The heatmaps reveal strong relationships between vehicle specifications and fuel consumption. The actual fuel consumption of K2 diesel and petrol shows a strong correlation with vehicle parameters, while the manufacturer-specified consumption correlations are weaker. In the case of K6 petrol, the lack of actual weight data limits the correlation analysis, although engine power and displacement show a slight relationship. The manufacturer-specified consumption of K6 diesel is strongly correlated with vehicle specifications, especially with maximum weight, where the correlation coefficient is 0.94, indicating that it depends on the design specifications. These patterns highlight the need for careful feature selection in model development.

The appendix contains comprehensive visualizations and descriptive statistics for K2, K6 diesel and petrol data. These include kilometers driven distributions, fuel consumption distributions, and correlation heatmaps in appendix subsections: A.1, A.2, A.3, and A.4. The refined dataset ensures both high data quality and geographic specificity, providing a strong foundation for modeling fuel consumption tailored to Lithuania's conditions.

2.4 Software and Tools

In this thesis, Python 3.10 was chosen as the main programming language because it is widely used in data science and offers an extensive library of tools and resources. The coding environment was adapted for both exploratory data analysis and machine learning model development. Additionally, different programming styles and languages were used to handle different data sets.

2.4.1 Primary Coding Environment

When working with the primary datasets – the current vehicle fleet and technical inspection datasets – the ETL (Extract, Transform, Load) method was used due to specific environmental requirements. These operations were performed using a special platform in the State Data Agency environment, which supports modular coding through a node workflow. Each node performed a specific function, ensuring clear tracking of the data pipeline. This approach ensured efficient data processing and traceability throughout the pipeline.

2.4.2 Data Preparation and Analysis Tools

Data preparation and additional analysis were performed using three different methods:

1. **PostgreSQL:** Used for SQL-based queries to efficiently extract, aggregate, filter, and transform data within a pipeline. SQL offers simplicity and fast computation and is well-suited for structured data operations.

2. **PySpark DataFrames:** Utilized for large-scale data processing within the environment, PySpark's distributed computing capabilities enabled efficient handling of extensive datasets. It was chosen over Pandas because it can perform parallel operations in batches and offers significantly faster data processing.

3. **Pandas DataFrames:** As one of the most widely used DataFrame libraries in the world, Pandas provides intuitive data manipulation and analysis. It has been used in private environments for data preparation and modeling tasks. Pandas was chosen because of its user-friendly interface and my skills in working with the library, allowing me to perform tasks efficiently and quickly.

Code Repository

The complete source code for this project is available on **GitHub (/DainiusTamuliunas)**.

The final code consisted of approximately 400 lines of SQL queries for efficient data extraction and transformation, 1,000 lines of PySpark code for data analysis and applying the final model to the Lithuanian vehicle fleet sample, and 2,650 lines of Pandas-based Python scripts for data analysis and modeling.

2.4.3 Text Refinement and Language Optimization

To ensure that the thesis text aligns with academic standards, ChatGPT 4.0 [25] was utilized for text refinement and language optimization. The tool helped increase clarity, consistency, and formal tone, which are essential for effectively presenting research findings. Additionally, Grammarly [14] was used to address grammatical accuracy and punctuation. This combination ensured that the final thesis met academic standards.

3 Methodology

3.1 Linear Models with Regularization

Linear Regression modelling is a type of supervised machine learning algorithm that models the linear relationships between independent variables X and continuous dependent variables y. A linear regression problem, depending on the number of its features, can be single or multiple. Multiple linear regression should not be confused with multivariate linear regression, in which case multiple dependent variables are predicted instead of a single scalar variable. Ordinary Least Squares (OLS) is a basic linear modeling method that forms the foundation upon which extensions such as Lasso, Ridge, and Elastic Net are built. These methods address challenges such as multicollinearity and feature selection, providing better stability and predictive performance.

3.1.1 Ordinary Least Squares (OLS) Linear Regression

The Ordinary Least Squares (OLS) minimizes the residual sum of squares between observed target values and predicted values, assuming a linear relationship between the dependent variable and independent variables. The optimization problem is expressed as:

$$\min_{w} \|Xw - y\|_2^2, \tag{2}$$

where:

- X is the matrix of input features,
- w represents the coefficients,
- y is the vector of observed target values.

The closed-form solution for the coefficients is:

$$\hat{w} = (X^T X)^{-1} X^T y. \tag{3}$$

In practice, OLS is widely used due to its interpretability and computational efficiency. Its general form is represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \tag{4}$$

where:

- y is the dependent variable,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients for independent variables $x_1, x_2, \ldots, x_n$.

This method is useful in analyzing fuel consumption data because it helps model the relationship between key vehicle characteristics (e.g. engine size, vehicle weight, and fuel type) and actual average fuel consumption, thereby creating a baseline model that can be compared with more advanced approaches.

However, OLS assumes that input features are not highly correlated. When multicollinearity exists, predictions can become unstable, which leads to sensitivity in coefficient estimates. To address this issue, regularization methods such as Ridge and Lasso regression are commonly used because they help stabilize the model by penalizing large coefficients. Even despite this, OLS remains a widely used approach for regression tasks [34].

3.1.2 Lasso Regression

The Lasso Regression, a regression method based on the Least Absolute Shrinkage and Selection Operator, is an extension of OLS by adding a penalty term, which is the sum of the absolute values of the coefficients, also known as L_1-regularization. Its objective function is defined as [33]:

$$\arg\min_{w} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |w_j| \tag{5}$$

Where λ controls the strength of the regularization. The L_1 regularization shrinks some coefficients of less significant variables to zero. As a result, features with zero coefficients are eliminated from the model, thereby performing variable selection. This is especially useful when working with large amounts of data, where there are many predictions relative to the number of observations [33]

3.1.3 Ridge Regression

Ridge Regression, also known as Tikhonov regularization, is an extension of OLS, that addresses multicollinearity among predictor variables by adding an L_2-regularization term which penalizes large coefficients and thus reduces their variance. Its objective function is defined as:

$$\arg\min_{w} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} w_j^2 \tag{6}$$

where λ controls the degree of regularization.

Unlike Lasso, Ridge does not set any coefficients to zero, ensuring that all features are retained. This method is particularly effective for data sets with correlated predictors because it stabilizes coefficient estimates and reduces variance. Ridge regression is robust to overfitting and offers a solution to multidimensionality by retaining all variables, even the less important ones, while reducing their coefficients closer to zero [36].

3.1.4 Elastic Net Regression

Elastic Net Regression was introduced by Zou and Hastie in 2005, it is a linear regression algorithm that combines L_1- and L_2-regularization with a standard least squares objective function to leverage the advantages of both methods. The objective function is defined as:

$$\arg\min_{w} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \left(\lambda \sum_{j=1}^{p} |w_j| + (1-\lambda) \sum_{j=1}^{p} w_j^2 \right), \tag{7}$$

where α controls the overall regularization strength and λ (with values between 0 and 1) balances between L_1 and L_2-penalties. Elastic Net is a regularized regression technique that can handle multicollinearity and overfitting problems. This method is advantageous in situations where neither Lasso nor Ridge individually provides optimal results [31].

3.1.5 Comparison and Applications

Each linear regression method offers unique advantages based on the characteristics of the data set and the problem being solved:

- **OLS:** Simple and interpretable, suitable when the features are uncorrelated and there is no multicollinearity. It is widely used as a baseline model for comparison in predictive tasks.
- **Lasso:** Particularly effective for large data sets with many predictors because it performs variable selection by reducing less significant coefficients to zero.
- **Ridge:** Handles multicollinearity well because it reduces the coefficients of correlated predictors to zero, thereby ensuring stability and reducing overfitting.
- **Elastic Net:** Combines the strengths of Lasso and Ridge, providing a balance between feature selection and multicollinearity handling.

These models are widely used for predictive modeling tasks due to their simplicity, computational efficiency, and ability to address a variety of data challenges. These models can be applied in a variety of fields, including finance, healthcare, environmental science, and fuel consumption estimation using real-world data, where interpretability and reliability are essential. For further details, refer to the Scikit-learn documentation [31, 33, 34, 36].

3.2 Ensemble Methods

Ensemble methods are a machine learning technique that combines multiple base models to create a single optimal predictive model. These methods are known for their robustness and ability to model complex relationships and minimize overfitting. These methods, which exploit the strengths of individual models, improve predictive accuracy, often by reducing variance and bias through generalization or sequential learning. [30]

3.2.1 Random Forest Regression

Random Forest is an ensemble learning method widely used for both regression and classification tasks. It was introduced by Breiman [5] and builds upon the principle of bagging (bootstrap aggregating). It combines the predictions of multiple decision trees to improve predictive accuracy and generalization. In regression tasks, Random Forest creates decision trees on bootstrapped subsets of the data and averages their predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(X),$$

where B is the number of trees in the forest, and $T_b(X)$ represents the prediction from the b-th tree. By aggregating the results of multiple trees, Random Forest reduces variance and minimizes the risk of overfitting.

The algorithm performs the following steps:

- **Step 1:** Determine the number of decision trees (B) to be built in the forest.
- **Step 2:** For each decision tree, create a bootstrapped dataset by sampling the original data with replacement.
- **Step 3:** At each node of the tree, select a random subset of features and determine the best split using variance reduction as the splitting criterion.
- **Step 4:** Continue splitting nodes recursively until a stopping criterion is met. In this work, the stopping criteria are as follows:
  - Nodes cannot split further if they contain fewer than 2 samples.
  - Leaf nodes must contain at least 1 sample.
  - No explicit limit on tree depth, so nodes will split until pure or until the other criteria are met.
- **Step 5:** For regression problems, the predictions of all trees are averaged to produce the final output, ensuring robust and accurate predictions.

Random Forest is robust against overfitting due to its averaging predictions. It handles both numeric and categorical data efficiently, making it very versatile. Additionally, it provides a feature importance metric that can help with feature selection and model interpretation. Despite its strengths, Random Forest can be computationally intensive for large datasets, especially when the number of trees is large. Furthermore, the interpretation of individual decision trees is complex compared to simpler models such as linear regression.

Random forest Regression is robust against overfitting due to averaging predictions from multiple trees, it also handles both numerical and categorical data effectively and provides feature importance metrics, that helps to understand feature selection. However, it has a few limitations, it is computationally intensive for large datasets with a high number of trees and difficult to interpret

individual tree decision compared to simpler models like linear regression. This study uses Random Forest Regression to model the relationship between key vehicle parameters (engine size, vehicle type, fuel type, weight) and fuel consumption. By tuning hyperparameters such as `n_estimators`, `max_depth`, and `max_features`, the model is optimized to balance bias and variance, ensuring robust predictions. Additionally, feature importance metrics provide insights into the most important features that impact fuel consumption, helping to interpret the model and potential future optimizations [35].

3.2.2 Gradient Boosting Regression

Gradient Boosting Regression (GBR) is an ensemble learning technique that builds a strong predictive model by iteratively adding weaker learners, typically decision trees, in a sequential manner. Each new learner is trained to minimize the residual errors of the ensemble model created before. GBR is highly flexible and can optimize any differentiable loss function, making it suitable for various regression tasks, including predicting energy consumption in road transport. This subsection provides a deeper look into the underlying mathematical steps of the algorithm, based on Friedman's original paper [13].

GBR is based on the principle of boosting, which sequentially combines weak learners to create a stronger model. At each iteration, the algorithm fits a new weak learner to the negative gradient of the loss function, effectively addressing the shortcomings of the current ensemble. Decision trees are commonly used as base learners due to their simplicity and ability to capture non-linear relationships.

The algorithm performs the following steps:

- **Step 1:** Start by building an initial model F_0, which is a constant prediction that minimizes the chosen loss function $L(\cdot)$ across all N training samples. The initial guess of our model helps to anchor subsequent improvements.

$$F_0 = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma).$$

- **Step 2:** Iteratively add new *weak learners* (usually decision trees) to improve the predictions of the current model. Repeat this boosting process M times.

  1. For each training sample x_i, compute the residual $r_{i,m}$ by taking the negative gradient of the loss function with respect to the model prediction from the previous iteration $F_{m-1}(x_i)$.:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}.$$

  2. Using the pairs $\{(x_i, r_{i,m})\}$, train a small regression tree that divides the input feature space (i.e., the d-dimensional space that includes all the features of the model) into J separated regions (leaf nodes). Each region $R_{j,m}$ corresponds to a leaf of the new tree. By fitting the tree to these pseudo-residuals, the model learns to correct the largest remaining errors.

3. For each region $R_{j,m}$, compute the optimal leaf value:

$$\gamma_{j,m} = \arg\min_{\gamma} \sum_{x_i \in R_{j,m}} L(y_i, F_{m-1}(x_i) + \gamma).$$

4. Finally, we combine the new tree with the existing model by adding a scaled version of the leaf predictions to $F_{m-1}(x)$.:

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J} \gamma_{j,m} \mathbf{1}(x \in R_{j,m}),$$

where $\nu \in (0,1)$ is the learning rate, controlling the contribution of each tree.

The performance of GBR depends on key hyperparameters:

- **Learning Rate (ν):** Controls the step size at each iteration. A smaller value improves generalization but increases computation time.
- **Number of Estimators (M):** Total number of boosting iterations. This is often determined by cross-validation to balance bias and variance.
- **Tree Depth (J):** Limits the maximum depth of individual trees. Simpler trees prevent overfitting and ensure computational efficiency.
- **Loss Function (L):** Common practice include the squared error for regression problems.

For this study, hyperparameter tuning was performed using grid search and cross-validation to optimize predictive performance while minimizing overfitting.

GBR is well-suited for predicting final energy consumption in road transport due to its ability to model nonliner and complex relationships between predictors and target variables. Additionally, it effectively handles heterogeneous datasets with varying feature distributions and provides feature importance metrics, enhancing model interpretability. Its iterative nature ensures that errors in initial predictions are progressively corrected, resulting in a reliable model. Despite its advantages, GBR has several challenges, one of the biggest is overfitting and computational intensity due to iterative process, cross-validation and hyperparameter tuning [32].

3.2.3 CatBoost Regression

CatBoost (Categorical Boosting) is relatively new machine learning algorithm developed in 2017 by Yandex company [15]. A gradient boosting system specifically designed for datasets with categorical variables. It uses techniques such as ordered boosting to prevent overfitting and handles categorical features on its own without requiring extensive preprocessing. Similar to Gradient Boosting Regression, CatBoost iteratively minimizes a differentiable loss function to improve predictions. CatBoost differs from other gradient boosting models in that it supports categorical features and uses ordered boosting, which reduces overfitting by sequentially training models without data leakage.

These innovations make it particularly effective on datasets dominated by categorical data. Since this work uses vehicle specifications as input features, as well as categorical variables such as fuel type and vehicle type, CatBoost is a suitable choice for this problem because it can handle categorical data naturally without extensive preprocessing, providing good model performance and interpretability [6].

3.2.4 XGBoost Regression

XGBoost (eXtreme Gradient Boosting) is an open-source library that provides an optimized and scalable implementation of gradient-boosted decision trees. It was developed by Tianqi Chen and Carlos Guestrin in 2016 as part of their research [7]. Since its introduction, XGBoost has become a widely used method for solving supervised learning problems, especially with structured datasets.

One of the main reasons for XGBoost's popularity is the inclusion of advanced software and hardware optimization techniques that allow it to efficiently process large data sets. Its innovations include clever regularization of decision trees and the use of second-order approximations, which optimize splits and improve accuracy by incorporating gradient and second-derivative information. These features allow XGBoost to reduce overfitting, provide high predictive accuracy, and ensure computational efficiency.

The algorithm starts with an initial prediction, which is usually set to 0.5. Residuals are calculated by comparing the predicted values with the actual target values. A decision tree is then created to predict these residuals, and the quality of the tree splits is evaluated using similarity scores and gain metrics. Final predictive function combines the contributions of all trees.

Despite the rise of deep learning for tabular data, XGBoost continues to outperform neural networks in many benchmarks due to its efficiency and ability to handle structured data. These advantages, combined with robust handling of missing values, make XGBoost a powerful choice for regression and classification tasks [43].

3.2.5 LightGBM Regression

LightGBM (Light Gradient Boosting Machine) [18] is an open-source gradient boosting framework based on a tree learning algorithm designed to optimize speed and performance on large datasets. Developed in 2017 as part of the Microsoft DMTK (Distributed Machine Learning Toolkit) project [17]. LightGBM offers several advantages, including high processing speed, reduced memory usage, support for GPU learning, and excellent scalability for large amounts of data . LightGBM differs from other boosting algorithms in that it uses a leaf-based tree growth strategy that grows leaves with the highest potential loss reduction, leading to faster convergence and improved accuracy. However, this strategy can lead to overfitting on smaller datasets. Parameters such as tree depth and the number of leaves can be fine-tuned to mitigate overfitting.

In addition, LightGBM includes two unique techniques:

- **Gradient-based One-Side Sampling (GOSS):** GOSS reduces the number of data points used in

training by focusing on the cases with the largest gradient, ensuring minimal loss of accuracy and reducing computational complexity.

- **Exclusive Feature Bundling (EFB):** EFB reduces the number of features by combining sparse features into dense ones, reducing memory usage and speeding up the training process.

These innovations make LightGBM particularly effective on large-scale datasets, ensuring efficient and accurate model training.

3.3 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a version of Support Vector Machines (SVM) tailored for regression tasks. Unlike traditional regression methods that aim to minimize prediction errors directly, SVR focuses on finding a hyperplane that fits the data within a specified margin of tolerance, known as the ϵ-insensitivity zone. This approach allows SVR to handle both linear and non-linear relationships effectively, making it a versatile tool for predicting continuous outcomes.

SVR solves the following optimization problem (8):

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \quad \text{subject to} \quad \begin{cases} y_i - \langle w, \phi(x_i)\rangle - b \leq \epsilon + \zeta_i, \\ \langle w, \phi(x_i)\rangle + b - y_i \leq \epsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* \geq 0, \quad i = 1, \ldots, n. \end{cases} \tag{8}$$

where w is the weight vector, b is the bias term, ϕ denotes a feature space transformation, ζ and ζ^* are slack variables measuring deviations from the ϵ-insensitivity zone, and C is a penalty parameter balancing margin width and error tolerance.

The dual formulation of this problem enhances computational efficiency by leveraging kernel functions, such as radial basis function (RBF), polynomial, and sigmoid kernels. These kernels enable SVR to model complex, non-linear relationships in data by mapping the input features to a higher-dimensional space. The prediction for a new data point x is computed as (9):

$$\hat{y}(x) = \sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b, \tag{9}$$

where α_i and α_i^* are Lagrange multipliers, $K(x_i, x)$ is the kernel function, and SV represents the support vectors, which are data points lying outside the ϵ-insensitivity zone.

SVR offers greater flexibility and robustness compared to traditional linear regression. By leveraging kernel functions, SVR can manage complex patterns in data, similar to how neural networks handle non-linear relationships. Effective hyperparameter tuning, such as selecting the appropriate kernel and setting the ϵ parameter, is crucial for maximizing SVR performance.

3.4 K-Means Clustering

K-means clustering is an unsupervised learning algorithm that partitions data into k clusters. The number of clusters is provided as an input. It forms clusters by minimizing the sum of squares within a cluster, re-assigning each data point to the nearest cluster centroid, and updating the centroids accordingly. The K-Means objective for the dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, can be written as follows:

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2,$$

where C_j denotes the j-th cluster, and $\boldsymbol{\mu}_j$ is its centroid.

In this work, the dataset is grouped by categorical features (e.g., vehicle type, municipality) to calculate aggregate metrics such as average mileage and average vehicle age. These aggregations reduce noise and highlight higher-level patterns that may be more suitable for clustering. Next, the features are standardized so that variables with different scales do not dominate the distance calculation. K-Means clustering helps reveal how annual mileage correlates with the owner's age, vehicle type, and place of residence. This provides valuable insights when analyzing fleet usage or regional differences.

Although K-Means is computationally efficient and easy to interpret, k must be chosen in advance and can be sensitive to outliers. The Silhouette score-based optimization was used to select k. This evaluates how data points fit into their assigned clusters compared to other clusters and helps reduce over-clustering and under-clustering [40].

4 Modelling and Results

Before training the models, the dataset was preprocessed to ensure compatibility with the algorithms and data quality.

The features used in the models included both numerical and categorical variables. Numerical variables included engine displacement (in cubic centimeters), engine power (in kilowatts), and vehicle weight. For K6 vehicles, the maximum weight was used as actual weight data was unavailable. When the actual mass was used for K2 vehicles and the maximum mass was excluded. A new variable called weight was created to dynamically manage these adjustments. Categorical variables included fuel type and vehicle type.

To handle the few remaining missing values in the dataset, rows with missing data were deleted, ensuring a clean and reliable input for training the model.

Feature engineering was applied to account for differences across vehicle types. A custom pipeline dynamically adjusted the inclusion of weight metrics, such as maximum or actual weight, based on vehicle type and fuel type. This approach allowed a unified modeling framework, eliminating the need for separate models and ensuring consistency in feature representation.

Scaling and encoding were applied to standardize the dataset. The numerical features were standardized to have zero mean and unit variance, which is an important preprocessing step for gradient-based algorithms such as GradientBoosting and XGBoost. Categorical features were one-hot encoded to create binary variables for each unique category, ensuring proper representation of categorical data in the models.

The dataset was then split into training (80%) and testing (20%) subsets using a fixed random seed (42) to ensure reproducibility and reliable evaluation of model performance. To maintain a proportional distribution of vehicle and fuel types, stratified splitting was used, ensuring representative training and testing subgroups.

To further improve the performance of the models and ensure reliable predictions, hyperparameter optimization methods were systematically applied during training. Grid search combined with 5-fold cross-validation was used to tune the model parameters. This iterative process evaluated various combinations of parameters on the training dataset, reducing overfitting and improving the generalization of the models. The optimal hyperparameters were selected based on their performance on the training dataset, measured by the MAE.

The final tuned models were evaluated on the held-out testing dataset, which was not exposed during training. To assess the performance of each model, key evaluation metrics were calculated, including MAE, root mean square error (RMSE), and R^2 score. A stratified split by transport type and fuel type was used to create a balanced testing dataset of 20,000 data points, ensuring that the test set maintained the same distribution as the overall dataset. This method was only used for scatter plots that show a comparison of actual fuel consumption and predicted fuel consumption. Detailed descriptions of parameter tuning are provided in each model-specific subsection.

4.1 Ordinary Least Squares (OLS) Linear Regression

The OLS linear regression model was used as a baseline to evaluate the relationship between features and real fuel consumption. The model's performance metrics are as follows:

- **MAE:** 1.9266
- **RMSE:** 2.8026
- R^2**:** 0.3572

The low R^2 (0.3572) indicates limited explanatory power, suggesting nonlinear patterns not captured by the model. Figures 8 compare predicted vs. actual fuel consumption for vehicle categories K2 (left) and K6 (right). Diesel vehicles (blue) exhibit tighter predictions, while petrol vehicles (orange) show greater spread, especially for K6. Additionally, the graph indicates that vans lack a clear linear relationship between the input variables and actual fuel consumption, suggesting potential complexities or non-linear interactions in their behavior.

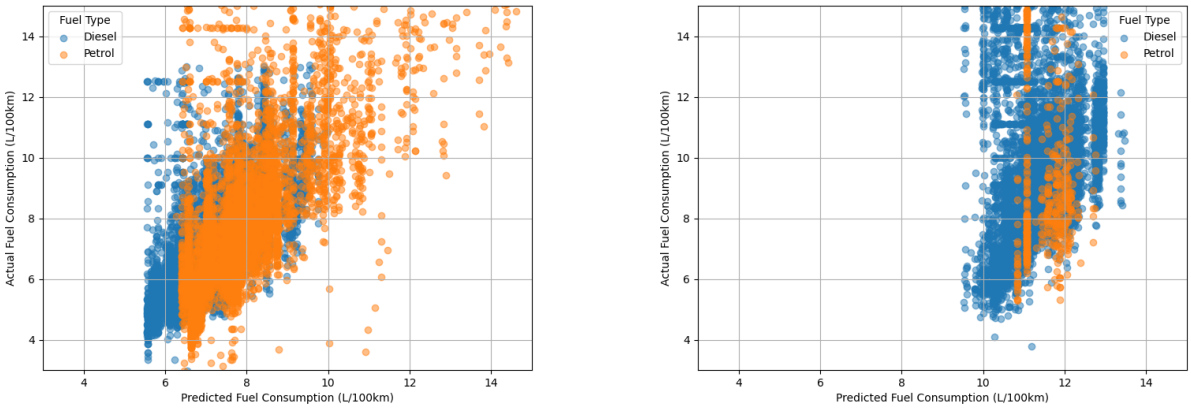

Figure 8. *Predicted vs Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right)*

Overall, the OLS model provides a simple baseline but struggles with predictive accuracy, highlighting the need for more advanced models to capture complex relationships in the data.

4.2 Lasso Regression

Lasso regression was implemented to evaluate its ability to model the relationship between features and real-world fuel consumption while applying L_1 regularization to reduce overfitting. Hyperparameter tuning was performed using grid search with 5-fold cross-validation over a range of λ values incremented from 0.05 to 1.0 in steps of 0.05. The best-performing parameters identified were:

- **Lambda (**λ**):** 0.05

The model's performance metrics are:

- **MAE:** 1.9438
- **RMSE:** 2.8129
- R^2**:** 0.3525

The R^2 value (0.3525) is slightly lower than that of the OLS model, indicating that Lasso regression do not offer improvement in explanatory power. However, Lasso's inclusion of regularization aids in feature selection by reducing the coefficients of less important variables.

Figure 9 illustrates the comparison of predicted vs. actual fuel consumption for vehicle categories K2 (left) and K6 (right). As observed, the predictions for diesel vehicles (blue) remain relatively close to the actual values, while petrol vehicles (orange) exhibit greater variability. The spread is indicative of Lasso's limitations in capturing nonlinear relationships in the data.

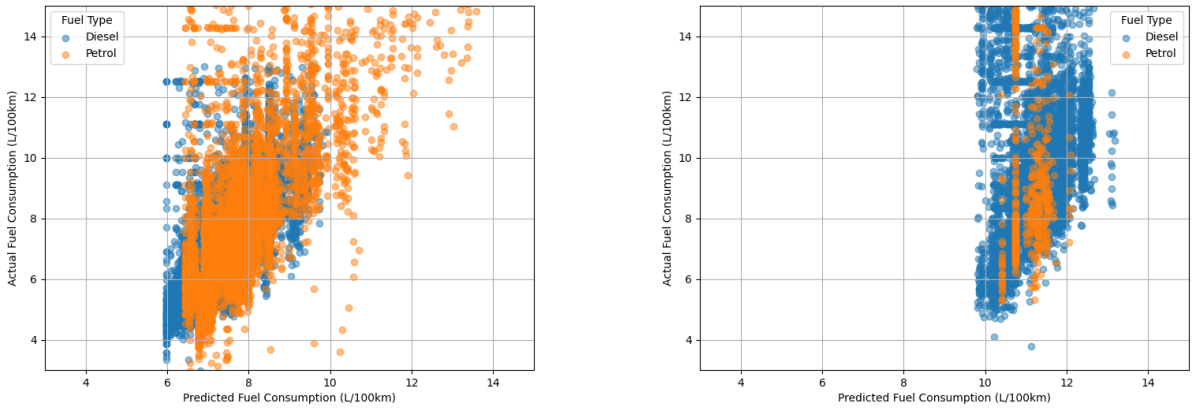

Figure 9. *Predicted vs Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right) using Lasso Regression*

While Lasso regression introduces a degree of regularization and variable selection, its performance is constrained by the limited number of features in the dataset. Furthermore, its inability to model the nonlinear relationships in fuel consumption data emphasizes the need for more advanced techniques.

4.3 Ridge Regression

The Ridge regression model was implemented to improve the baseline predictions by introducing L_2-regularization, which helps to mitigate overfitting. Hyperparameter tuning was performed using grid search with 5-fold cross-validation over a range of λ values incremented from 0.05 to 1.0 in steps of 0.05. The best-performing parameter identified was:

- **Lambda** λ**:** 0.05

The optimal λ value, being close to zero, suggests that the regularization term had minimal impact on the coefficients, and the model's behavior closely resembled of the OLS regression (subsection 4.1).

The performance metrics for this model are:

- **MAE:** 1.9266
- **RMSE:** 2.8026
- R^2**:** 0.3572

The R^2 score of 0.3572 indicates that the model still struggles to explain a significant portion of the variance in the data. The figures 10 show a comparison of predicted and actual fuel consumption for vehicles in categories K2 and K6, respectively.

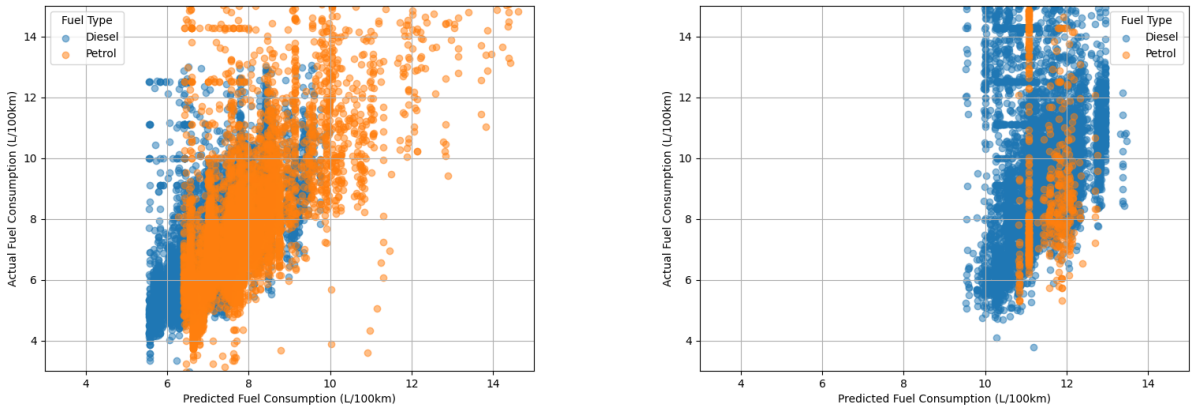

Figure 10. *Predicted vs. Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right) using Ridge Regression.*

While Ridge regression introduces regularization and slightly improves the model's stability, the similarity in performance to the OLS regression further underscores the limited utility of regularization when working with a small set of features.

4.4 Elastic Net Regression

Elastic Net regression, which combines the properties of Lasso and Ridge regression, was employed to evaluate its performance on predicting real-world fuel consumption. The model's hyperparameters were tuned using grid search with 5-fold cross-validation across the following ranges:

- **Alpha (**α**):** [0.05, 0.1, 0.15, ..., 1.0]
- **Lambda (**λ**):** [0.1, 0.2, 0.3, ..., 1.0]

The best parameters identified were:

- **Alpha (**α**):** 0.05
- **Lambda (**λ**):** 0.1

The model's performance metrics are as follows:

- **MAE:** 1.9283

- **RMSE:** 2.8137
- R^2**:** 0.3521

The R^2 value of 0.3521 indicates that while the Elastic Net model captures linear patterns better than simple OLS, Ridge, and Lasso regressions, it cannot explain complex, nonlinear relationships in the data. Figures 11 illustrate the predicted vs. actual fuel consumption for vehicle categories K2 and K6.

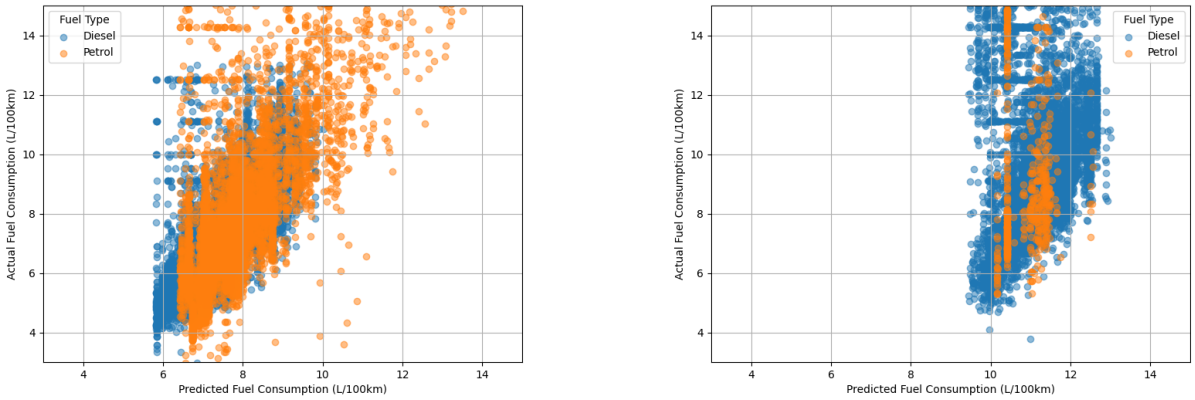

Figure 11. *Predicted vs Actual Fuel Consumption for Elastic Net Regression on Vehicle Categories K2 (left) and K6 (right).*

The hyperparameter tuning process demonstrated the value of balancing L1 and L2 regularization terms to optimize the model's predictive power while maintaining simplicity. Overall, the Elastic Net regression remains limited in predictive accuracy due to the underlying complexity and potential nonlinear patterns in the data.

4.5 Random Forest Regression

The Random Forest Regression model was utilized to predict real-world fuel consumption, leveraging its ensemble-based approach to improve prediction accuracy. Hyperparameter tuning was conducted using grid search with 5-fold cross-validation over the following parameter ranges:

- **Number of Estimators:** {200, 400}
- **Max Depth:** {None, 10, 20}
- **Minimum Samples Split:** {2, 5, 10, 20}
- **Minimum Samples Leaf:** {1, 2, 4, 10}

The best combination of hyperparameters was identified as:

- **Number of Estimators:** 400
- **Max Depth:** 20

- **Minimum Samples Split:** 20
- **Minimum Samples Leaf:** 10

The model's performance metrics are as follows:

- **MAE:** 1.4921
- **RMSE:** 2.4660
- R^2**:** 0.5153

With an R^2 value of 0.5153, the Random Forest Regression model demonstrated its ability to capture complex relationships in the data, providing robust predictive accuracy. Figures 12 illustrate the predicted versus actual fuel consumption for vehicle categories. As observed, diesel vehicles tend to have more consistent predictions, while petrol vehicles exhibit greater variance, particularly in the K2 category.

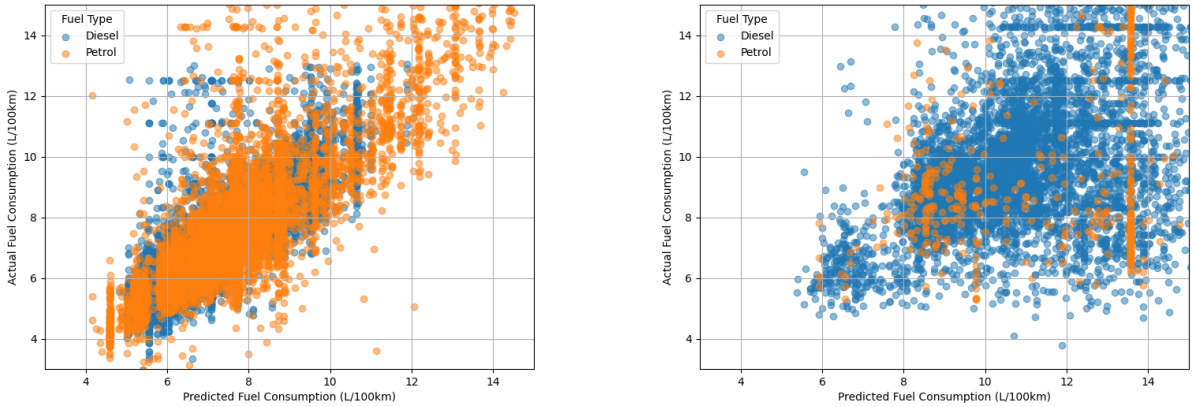

Figure 12. *Predicted vs Actual Fuel Consumption for Random Forest Regression for Vehicle Categories K2 (left) and K6 (right)*

The Random Forest model proved effective at modeling, its ensemble approach and parameter tuning enabled the model to generalize well to unseen data, making it a reliable choice for this predictive task.

4.6 Gradient Boosting Regression

The Gradient Boosting Regression model was applied to capture complex, nonlinear relationships between features and real-world fuel consumption. Hyperparameter tuning was performed using grid search with 5-fold cross-validation over the following ranges:

- **Learning Rate:** {0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2}
- **Max Depth:** {2, 3, 4}
- **Number of Estimators:** {25, 100, 200, 400}

The best combination of hyperparameters was determined to be:

- **Learning Rate:** 0.2
- **Max Depth:** 4
- **Number of Estimators:** 400

The model's performance metrics are:

- **MAE:** 1.4997
- **RMSE:** 2.4330
- R^2**:** 0.5157

The R^2 value of 0.5157 indicates a significant improvement over simpler models, suggesting that Gradient Boosting effectively captures more complex patterns in the data. Figure 13 compares predicted vs. actual fuel consumption by vehicle categories.

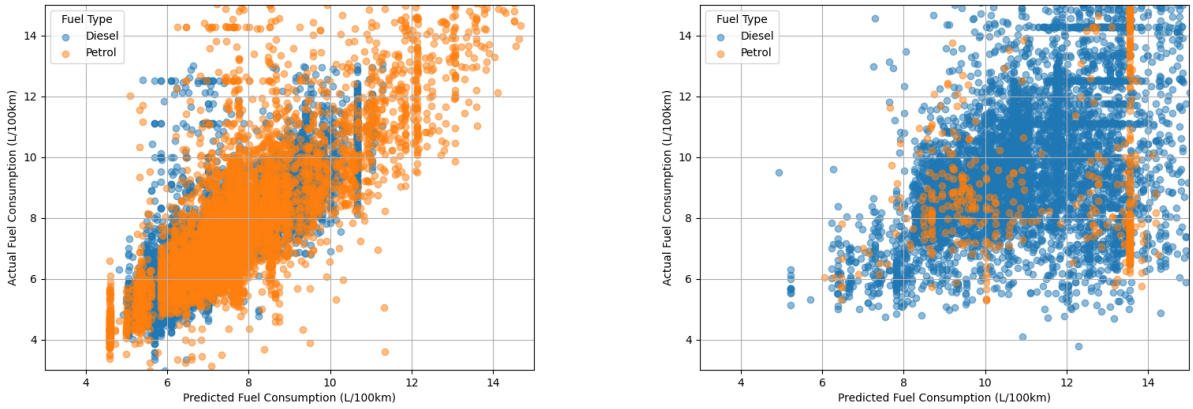

Figure 13. *Predicted vs Actual Fuel Consumption for Gradient Boosting Regression for Vehicle Categories K2 (left) and K6 (right)*

Overall, Gradient Boosting proved to be a powerful model, effectively handling nonlinearities and delivering strong predictive performance.

4.7 CatBoost Regression

The CatBoost regression model was employed to predict real-world fuel consumption using categorical and numerical features. Hyperparameter tuning was not extensively conducted for CatBoost in this instance due to its inherent ability to efficiently handle categorical variables and defaults that often provide robust results. The following parameter ranges were tested during tuning:

- **Iterations:** 50, 100, 200, 400, 1000
- **Depth:** 2, 3, 4, 6

- **Learning Rate:** 0.05, 0.1, 0.2, 0.5, 1.0

However, the following parameters were used:

- **Iterations:** 1000
- **Depth:** 6
- **Learning Rate:** 0.2

These parameters were selected based on general best practices and initial experiments, balancing model complexity and computational efficiency. The model's performance metrics are:

- **MAE:** 1.4932
- **RMSE:** 2.4397
- R^2**:** 0.5129

The R^2 score of 0.5129 demonstrates improved predictive power compared to simpler models like OLS regression, capturing more complex relationships in the data. Figures 14 compare predicted vs. actual fuel consumption for vehicle categories. Overall, CatBoost demonstrates its ability to handle both categorical and numerical variables effectively. The inclusion of categorical features without the need for explicit one-hot encoding allowed CatBoost to leverage the data structure more effectively, improving prediction accuracy and model efficiency.

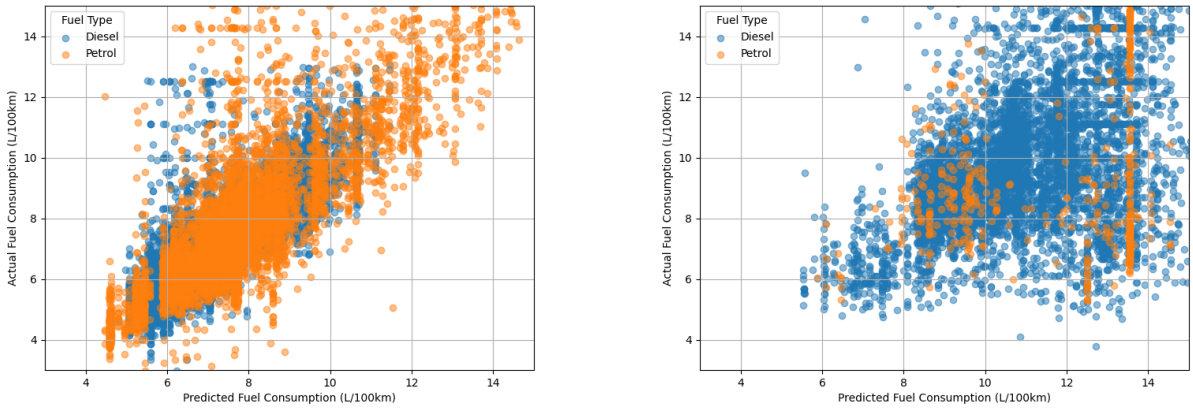

Figure 14. *Predicted vs Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right)*

4.8 XGBoost Regression

The XGBoost Regression model was employed to analyze nonlinear relationships between input features and real-world fuel consumption. Hyperparameter tuning was conducted using manual grid search, focusing on the following parameters:

- **Number of Estimators:** 25, 100, 200, 400

- **Learning Rate:** 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2
- **Max Depth:** 2, 3, 4

After testing multiple combinations, the optimal hyperparameters were identified as:

- **Number of Estimators:** 400
- **Learning Rate:** 0.2
- **Max Depth:** 4

The model's performance metrics on the test dataset are:

- **Mean Absolute Error (MAE):** 1.5003
- **Root Mean Square Error (RMSE):** 2.4363
- **Coefficient of Determination (R^2):** 0.5143

The R^2 value of 0.5143 demonstrates that XGBoost effectively captures significant nonlinear patterns in the data. The model performs comparably to Gradient Boosting Regression, highlighting its capability to balance model complexity and predictive accuracy.

Figure 15 depict the predicted vs. actual fuel consumption for K2 and K6 vehicle categories, respectively. The analysis indicates that XGBoost is a robust algorithm for modeling fuel consumption, providing a balance of flexibility, interpretability, and computational efficiency.

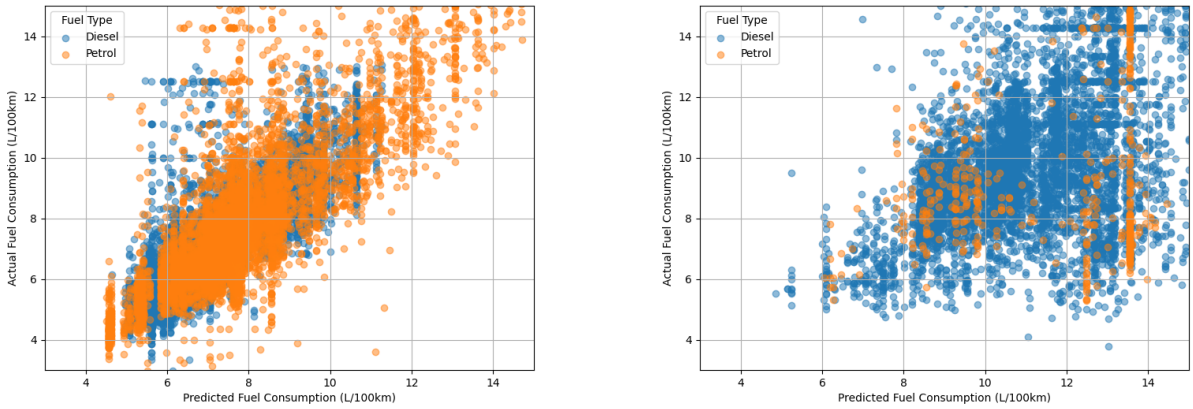

Figure 15. *Predicted vs. Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right) using XGBoost Regression.*

4.9 LightGBM Regression

The LightGBM regression model was utilized to predict real-world fuel consumption by leveraging both numerical and categorical features. Hyperparameter tuning was conducted to optimize the model's performance. The following parameter ranges were tested during the tuning process:

- **Number of Estimators:** 25, 100, 200, 400, 1000
- **Learning Rate:** 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.5, 1
- **Maximum Depth:** 2, 3, 4, 5

The best parameters identified through this process were:

- **Learning Rate:** 0.1
- **Maximum Depth:** 5
- **Number of Estimators:** 1000

These hyperparameters were selected based on iterative grid search, balancing model complexity and predictive accuracy. The model achieved the following performance metrics:

- **MAE:** 1.4927
- **RMSE:** 2.4350
- R^2**:** 0.5148

The R^2 score of 0.5148 highlights the model's ability to capture complex patterns in the data, outperforming simpler models. Figure 16 illustrates the comparison between predicted and actual fuel consumption for vehicle categories K2 and K6.

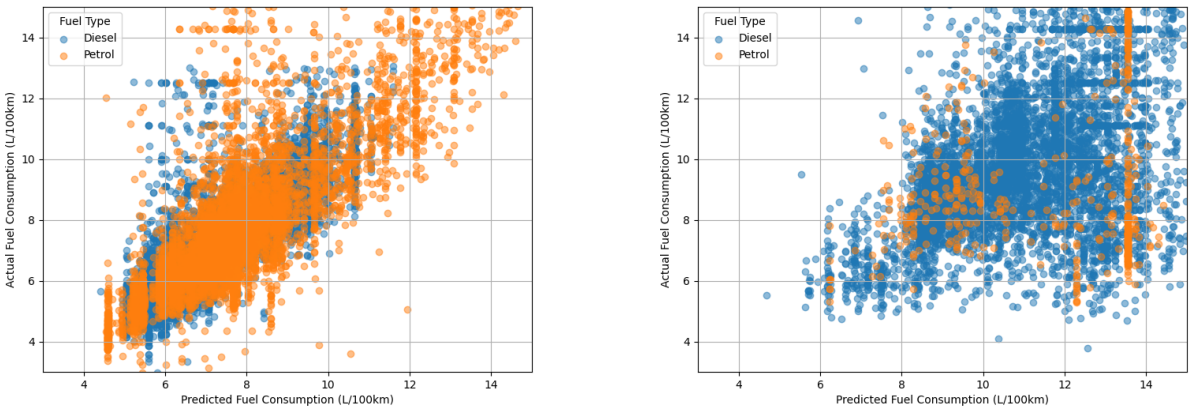

Figure 16. *Predicted vs Actual Fuel Consumption for Vehicle Categories K2 (left) and K6 (right)*

Overall, LightGBM proved to be a robust choice for fuel consumption prediction, effectively capturing relationships between vehicle characteristics and real-world consumption patterns. The use of optimized hyperparameters and integrated preprocessing contributed to the model's ability to generalize across different vehicle types and fuel categories.

4.10 Support Vector Regression (SVR)

Support Vector Regression (SVR) was employed to model real-world fuel consumption due to its effectiveness in handling nonlinear relationships. Hyperparameter tuning was conducted using grid search with 5-fold cross-validation over the following ranges:

- **Regularization Parameter (C):** {10, 100, 200}
- **Epsilon (ϵ):** {0.1, 0.2, 0.5}

The best combination of hyperparameters was determined to be:

- **Regularization Parameter (C):** 200
- **Epsilon (ϵ):** 0.2

The model's performance metrics are:

- **MAE:** 1.5017
- **RMSE:** 2.6796
- R^2**:** 0.4124

The R^2 value of 0.4124 indicates that while SVR captures some relationships between features and fuel consumption, its performance is lower compared to other advanced models such as Gradient Boosting and Random Forest 17. Overall, Support Vector Regression provides an alternative approach for modeling nonlinear relationships. However, it requires further optimization or feature engineering to improve its predictive accuracy for fuel consumption tasks. Additionally, when applied to large datasets, SVR becomes computationally intensive and less efficient compared to models like Gradient Boosting, which not only offer faster training times but also demonstrate superior predictive performance.

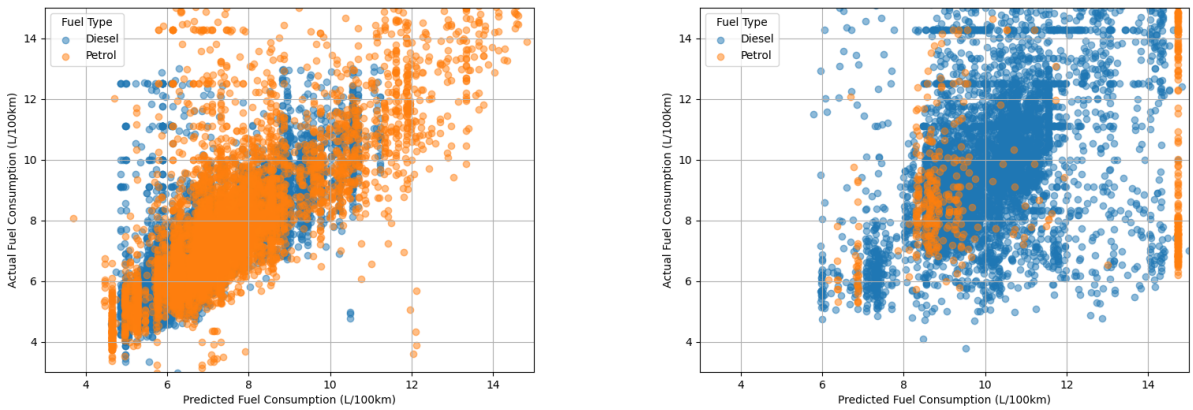

Figure 17. *Predicted vs Actual Fuel Consumption for SVR Regression for Vehicle Categories K2 (left) and K6 (right)*

4.11 Results

The performance of all models was evaluated using MAE, RMSE, and R^2 metrics on the test dataset. The results for each model are summarized in Table 9:

Table 9. *Summary of Model Performance Metrics*

Model	**MAE**	**RMSE**	R^2
OLS Linear Regression	1.9266	2.8026	0.3572
Lasso Regression	1.9438	2.8129	0.3525
Ridge Regression	1.9266	2.8026	0.3572
Elastic Net Regression	1.9283	2.8137	0.3521
Random Forest Regression	**1.4921**	**2.4660**	**0.5153**
Gradient Boosting Regression	**1.4997**	**2.4330**	**0.5157**
CatBoost Regression	1.4932	2.4397	0.5129
LightGBM Regression	1.4927	2.4350	0.5148
Support Vector Regression	1.5017	2.6796	0.4124
XGBoost Regression	1.5003	2.4363	0.5143

The table above highlights key performance metrics for each model:

- **Mean Absolute Error (MAE):** Random Forest achieved the lowest MAE (1.4956), closely followed by LightGBM (1.4927) and CatBoost (1.4932). These results indicate their superior accuracy in minimizing average prediction errors, with Random Forest slightly outperforming the others.

- **Root Mean Square Error (RMSE):** Gradient Boosting Regression had the lowest RMSE (2.4330), suggesting it performs best in handling larger prediction errors compared to other models. LightGBM (2.4350) and CatBoost (2.4397) were also competitive.

- R^2 **Score:** Gradient Boosting achieved the highest R^2 value (0.5157), indicating it explains the most variance in the data. Random Forest (0.5153) and LightGBM (0.5148) were close contenders.

Overall, **Random Forest Regression** and **Gradient Boosting Regression** emerged as the best-performing models. They effectively balance accuracy and error minimization, making them reliable choices for predicting fuel consumption. Simpler models, such as linear regressions with regularization, struggled to perform well, as indicated by their relatively high MAE and RMSE values and low R^2 scores. On the other hand, Support Vector Regression, while moderately effective, was computationally intensive and less efficient on large datasets compared to Gradient Boosting and Random Forest.

5 Lithuania Vehicle Fleet Results

This section presents the results of fuel consumption modeling and scaling for Lithuania's vehicle fleet. The analysis primarily focused on K2 (passenger cars) and K6 (vans) vehicle types, with diesel and petrol as the main fuel types. Supplementary fuel consumption data detailed in subsection 2.2 was integrated into the dataset to calculate fuel consumption for K1 (motorcycles) and K4 (buses). The Random Forest Regression model (subsection 4.5) was employed for predictions, followed by reweighting and rescaling procedures to ensure alignment with population-level distributions.

5.1 Scaling Results to Estimate Fuel Consumption of Lithuania's Vehicle Fleet

The primary goal of this analysis was to ensure that the final dataset accurately reflected the full vehicle population distribution in Lithuania. To achieve this, scaling and clustering techniques were applied.

5.1.1 Initial Reweighting Process

The initial reweighting process aimed to ensure that the sample dataset accurately reflected the actual vehicle fleet distribution in Lithuania. This was achieved by calculating the scaling factors for each type of vehicle. These scaling factors were applied to the sample dataset to align it with the actual distribution of the Lithuanian vehicle fleet. By adjusting the sample dataset using these coefficients, the resulting dataset approximated fuel consumption across the entire Lithuanian vehicle fleet. This reweighting approach corrected sampling biases and ensured that the insights and analyses derived from the sample dataset could be reliably generalized to the entire Lithuanian vehicle fleet.

5.1.2 Alternative Reweighting Process: Clustering with K-means

The alternative reweighting process used K-means clustering to improve the representativeness of the dataset by grouping vehicles based on similarities in municipality and mileage patterns. Unlike the initial reweighting, this method leveraged geographic and usage-based contexts to gain more granular insights.

Municipalities were identified as a key factor influencing vehicle usage, as the relationship between vehicle owner age and mileage is weak. Incorporating municipalities in the clustering process provided a solid basis for grouping vehicles with similar annual mileage.

The clustering process involved the following steps:

1. **Data Preparation**: Vehicles were grouped by type and municipality, calculating the average mileage of each group. The data were standardized to ensure correct distance calculations between groups.

2. **Modeling**: K-means clustering was applied with a range of cluster values ($k = 2$ to $k = 10$). The silhouette score was used to evaluate each k and determine the optimal number of clusters.

3. **Results**: Optimal $k = 3$ clusters were identified, achieving a silhouette score of 0.776, indicating well-separated and meaningful clusters (see Table 10).

The Silhouette score evaluates the quality of clustering [3]. The Silhouette coefficient $s(i)$ for each data point $\mathbf{x}_i$, is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where:

- $a(i)$ is the average distance between $\mathbf{x}_i$ and all other points in the same cluster.
- $b(i)$ is the minimum average distance from $\mathbf{x}_i$ to points in any other cluster.

A higher Silhouette score means data points stay closer to their own centroid than others, creating more coherent, distinct clusters.

Table 10. *Silhouette Score for Different Numbers of Clusters*

Number of Clusters (k)	**Silhouette Score**
2	0.7411
3	**0.7762**
4	0.7611
5	0.4914
6	0.6439
7	0.6106
8	0.5556
9	0.5160
10	0.3392

The clustering revealed distinct patterns in vehicle usage based on type and municipality. Average mileage varied significantly across clusters, highlighting differences in usage (see Table 11).

Table 11. *Cluster Information by Vehicle Type and Municipality*

Cluster	Vehicle Type	Municipality	Average Mileage (km/year)
0	K10	Metropolitan	12,412.52
0	K10	City	12,703.63
0	K10	Rural	11,079.39
0	K2	Metropolitan	10,819.41
0	K2	City	10,299.26
0	K2	Rural	11,159.34
0	K4	Metropolitan	17,455.70
0	K4	City	19,954.81
0	K4	Rural	19,182.51
1	K7	Metropolitan	66,063.81
1	K7	City	60,558.39
2	K1	Metropolitan	1,736.14
2	K1	City	1,659.50
2	K1	Rural	1,636.47

Due to the limitations in accessing Lithuanian fleet data after the internship ended, it was not possible to properly calculate the final results using the clustering method.

However, the scaling approach ensures broad representativeness by matching sample proportions to the population distribution and effectively addressing sampling biases. The clustering method, had it been applied, would have provided additional granularity by grouping vehicles based on geographic and usage patterns, highlighting differences in mileage across regions and vehicle types.

The results for the scaling approach, which estimates the fuel consumption of Lithuania's vehicle fleet, are presented below 12:

Table 12. *Fuel Consumption of Lithuania's Vehicle Fleet by Vehicle Type and Fuel Type*

Category	Vehicle Type	Fuel Type	Total Fuel Consumption (Million L/100 km)
K1	Motorcycles	Petrol	477.83
K2	Passenger Cars	Petrol	41,455.11
K2	Passenger Cars	Diesel	91,989.66
K6	Heavy Goods Vehicles	Petrol	307.01
K6	Heavy Goods Vehicles	Diesel	20,967.18
K15	Mopeds	Petrol	29.34
K4	Buses	Diesel	3,191.91

6 Conclusion

This thesis developed a comprehensive methodology for estimating annual fuel consumption in Lithuania's road transport sector, broken down by vehicle type and fuel type. The study utilized a range of datasets, including technical inspection records, European real-world data, and supplementary sources, to address data gaps and provide robust fuel consumption estimates for cars, vans, motorcycles, and other road vehicles.

Data preprocessing involved extensive cleaning, integration, and feature engineering to ensure compatibility across datasets. Machine learning models such as Random Forest and Gradient Boosting emerged as the most effective methods, achieving R^2 values exceeding 0.51 on testing data. These models incorporated features like engine size, power, and weight to predict real-world fuel consumption accurately. Random Forest was identified as the optimal model due to its balance of accuracy, interpretability, and computational efficiency.

The results revealed significant discrepancies between real-world and manufacturer-reported fuel consumption values, highlighting the need to adjust official figures to reflect actual vehicle performance. Urban-rural differences in vehicle usage patterns were also identified, emphasizing the importance of considering regional characteristics in energy policies. These findings highlight that data-driven methodologies can help inform national and regional decision-making.

The State Data Agency supported this research, providing access to primary datasets and expert guidance during a two-month internship.

While this study provides a robust framework, limitations include underrepresenting specific vehicle categories and excluding seasonal and behavioral factors. Future research could address these gaps by incorporating additional datasets, extending analysis to older vehicles, and exploring hybrid and electric vehicle consumption.

In conclusion, this thesis contributes a scalable and data-driven approach to estimating fuel consumption in Lithuania's road transport sector. The findings offer actionable insights for policymakers, transportation authorities, and other stakeholders.

7 Limitations

This study faces several limitations that impact the scope, accuracy, results, and generalizability of its findings. These are detailed below:

- **Data Availability:** The analysis relies heavily on datasets that are not entirely representative.
- **Vehicle Ages:** The European Environment Agency's real-world dataset provides data for vehicles from 2021 and 2022. However, the Lithuanian vehicle fleet consists mainly of older vehicles, which may lead to discrepancies between the dataset and the actual fleet's fuel consumption and performance.
- **Hybrid and Alternative Fuels:** The available real-world data is insufficient to comprehensively model fuel consumption for hybrid vehicles and vehicles using alternative fuels such as LPG, ethanol, or biomethane.
- **Model Simplifications:** Data preprocessing and modeling assumptions, such as excluding certain vehicle types and using mean values for missing data, may oversimplify real-world complexity.
- **Behavioral and Seasonal Factors:** The study does not fully account for behavioral differences (e.g., aggressive driving style) or seasonal differences (e.g., increased fuel consumption in winter), which significantly affect fuel consumption in the real world.
- **Limited Historical Data:** The lack of historical mileage and fuel consumption data prevents detailed trend analysis over time, which could have provided deeper insights.
- **Time Constraints:** The analysis of the Lithuanian car fleet dataset was limited by the duration of my internship at the State Data Agency. With only two months, time constraints limited the depth of data exploration and model implementation.

Addressing these limitations requires access to more comprehensive and representative datasets, improved data collection methodologies, and the development of advanced models that can more effectively capture behavioral, geographic, and seasonal variations.

A Data Preprocessing Additional Tables

Table 13. *Countries Removed and Reasons for Exclusion*

Country Code	Reason for Exclusion
AT	Significant mountainous terrain (Alps)
SI	Contains parts of the Alps and Dinaric Alps
IT	Significant mountainous terrain (Alps, Apennines)
ES	Pyrenees, Sistema Bético, and other mountain ranges
PT	Contains Sistema Central and other ranges
NO	Mostly in the Scandinavian Mountains
SE	Partially includes the Scandinavian Mountains
FI	Contains the Scandinavian Mountains in the northwest
BG	Balkan Mountains and Rila-Rhodope ranges
RO	Predominantly in the Carpathians
CZ	Includes Sudetes, Ore Mountains, and other ranges
SK	Predominantly in the Carpathians (e.g., Tatra Mountains)
GR	Pindus Mountains and other ranges
HR	Part of the Dinaric Alps
BA	Part of the Dinaric Alps
RS	Includes Dinaric Alps and Carpathians
ME	Predominantly in the Dinaric Alps
AL	Contains the Dinaric Alps and Accursed Mountains
MK	Includes the Šar and Rila-Rhodope ranges

Table 14. *Data Counts by Country for Cars and Vans after removing Irrelevant Geographic Data*

Country Code	Vans	Cars
FR	38,084	582,715
DE	34,987	860,045
BE	10,820	144,557
PL	9,251	176,589
NL	7,948	87,714
DK	5,697	55,627
IE	4,284	33,694
HU	3,837	46,859
EE	621	7,975
LU	584	15,873
LT	537	9,739
IT	300	83,923
LV	289	4,795
IS	198	821
CY	148	3,484
SE	63	12,917
MT	42	982
GR	31	3,377
ES	27	38,601
AT	18	11,550
PT	8	5,712
FI	4	3,210
CZ	3	7,236
RO	2	3,826
SI	2	2,064
BG	1	806
HR	1	1,580
NO	1	673
SK	1	2,848

A Final Training Dataset Analysis

A.1 Fuel Consumption Analysis

The following figures illustrate the distribution of kilometers driven and the correlation analysis for the final training dataset, focusing on the two selected fuel types, Petrol and Diesel, across K2 and K6 vehicle categories.

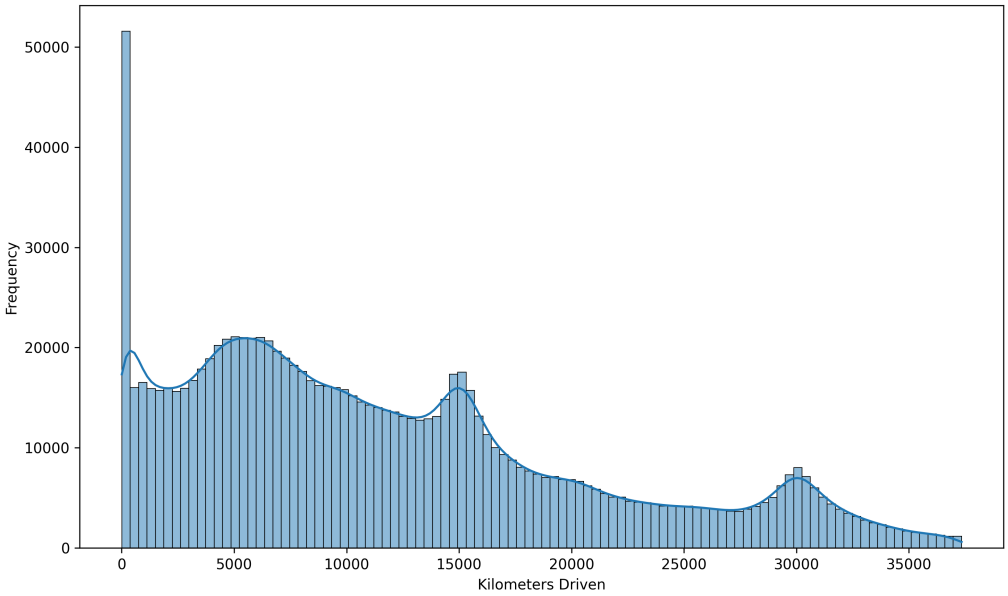

Figure 18. *Kilometers Driven Distribution for K2 Petrol Vehicles*

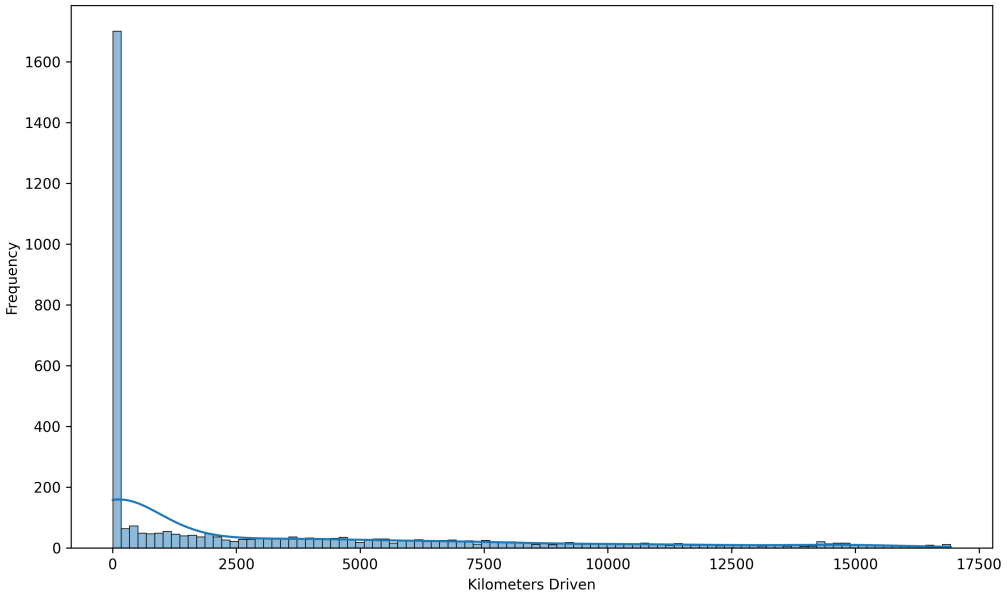

Figure 19. *Kilometers Driven Distribution for K6 Petrol Vehicles*

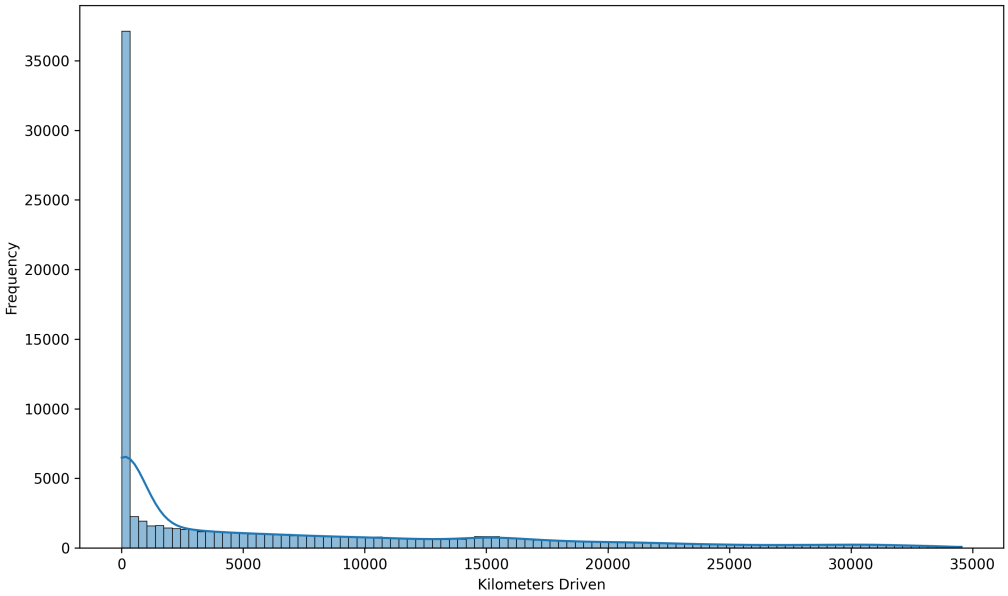

Figure 20. *Kilometers Driven Distribution for K6 Diesel Vehicles*

A.2 Correlation Analysis

The following heatmaps represent the correlation analysis between key variables such as fuel consumption, engine displacement, power, weight, and year of manufacture for the final dataset. These plots provide insights into the relationships between variables.

Figure 21. *Correlation Heatmap for K2 Petrol Vehicles*

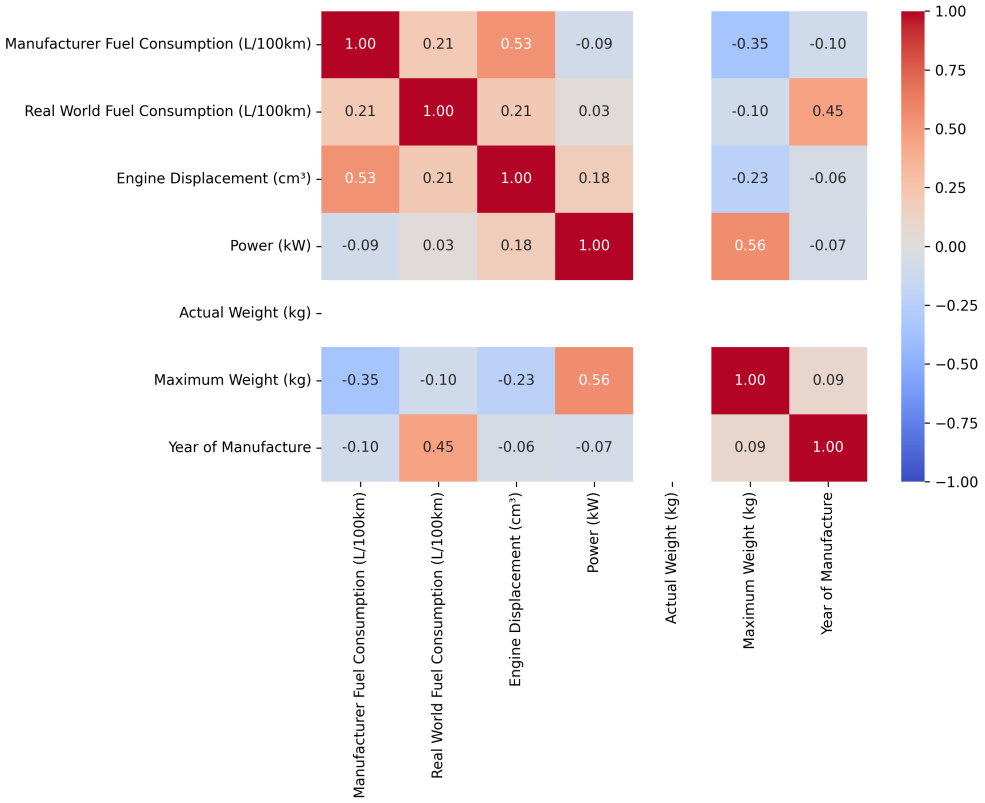

Figure 22. *Correlation Heatmap for K6 Petrol Vehicles*

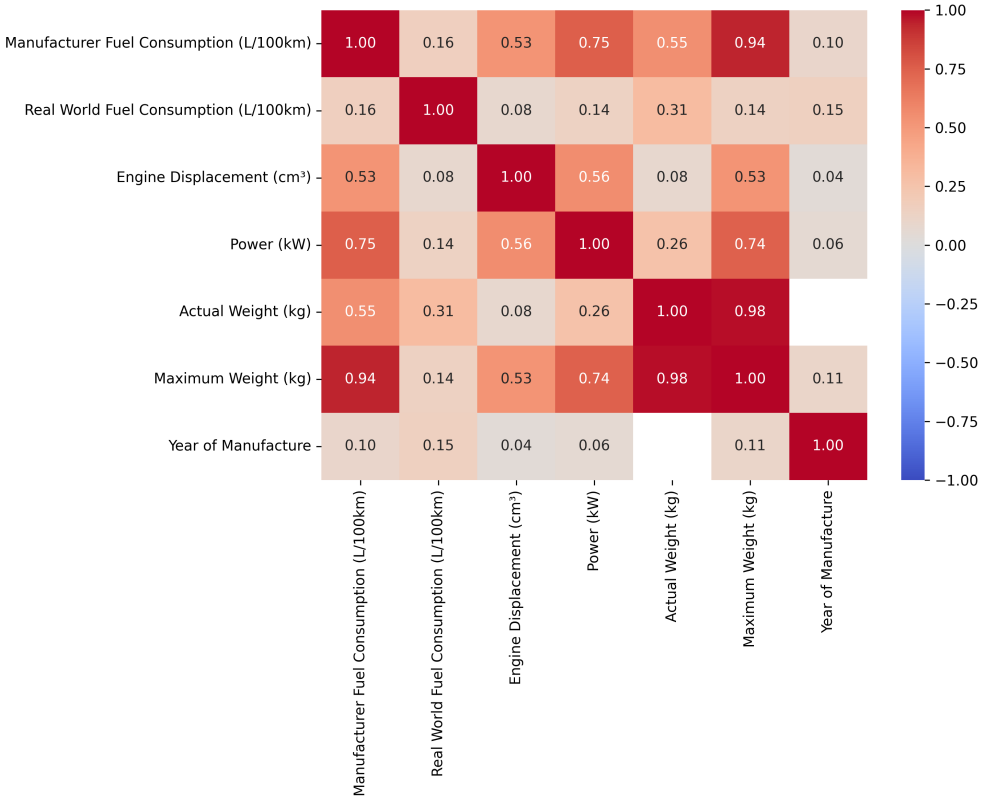

Figure 23. *Correlation Heatmap for K6 Diesel Vehicles*

A.3 Descriptive Statistics

Table 15. *Descriptive Statistics of Key Variables for K2 Petrol Vehicles*

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Fuel Consumption (L/100km)	902,033	5.43	1.99	0.60	4.90	5.60	6.40	16.40
Fuel Consumption (Original)	1,022,762	7.76	2.18	1.20	6.32	7.31	8.62	15.27
Engine Displacement (cm^3)	1,022,761	1,526.85	539.42	875	999	1,477	1,798	6,749
Power (kW)	1,016,944	109.52	53.77	44	70	100	120	537
Actual Weight (kg)	1,022,379	1,472.89	305.14	915	1,263	1,409	1,665	2,810
Max Weight (kg)	1,022,537	1,593.72	320.25	992	1,378	1,527	1,784	3,044

Table 16. *Descriptive Statistics of Key Variables for K6 Petrol Vehicles*

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Fuel Consumption (L/100km)	3,567	7.09	1.23	2.90	6.80	7.10	7.70	13.80
Fuel Consumption (Original)	3,739	12.11	5.33	0.20	7.75	9.67	16.24	29.89
Engine Displacement (cm^3)	3,737	1,368.02	275.92	996	1,199	1,462	1,462	2,956
Power (kW)	3,737	79.93	19.26	49	75	75	81	235
Actual Weight (kg)	1	1,165.00	–	1,165	1,165	1,165	1,165	1,165
Max Weight (kg)	3,736	1,472.35	295.93	1,082	1,261	1,266.50	1,631	2,921

Table 17. *Descriptive Statistics of Key Variables for K6 Diesel Vehicles*

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Fuel Consumption (L/100km)	93,369	7.65	1.50	3.80	6.60	7.50	9.00	15.00
Fuel Consumption (Original)	97,769	11.14	4.05	0.00	8.33	9.95	12.89	23.07
Engine Displacement (cm^3)	97,765	1,895.57	262.08	1,248	1,950	1,968	1,996	2,998
Power (kW)	97,769	104.74	25.79	55	88	103	125	221
Actual Weight (kg)	20	2,360.90	314.62	1,473	2,379.50	2,444	2,501	2,825
Max Weight (kg)	97,762	2,255.42	333.44	1,273	2,036	2,275	2,539	3,200

A.4 Fuel Consumption Distribution Analysis

This appendix subsection provides detailed visualizations of the fuel consumption distribution for the final training dataset.

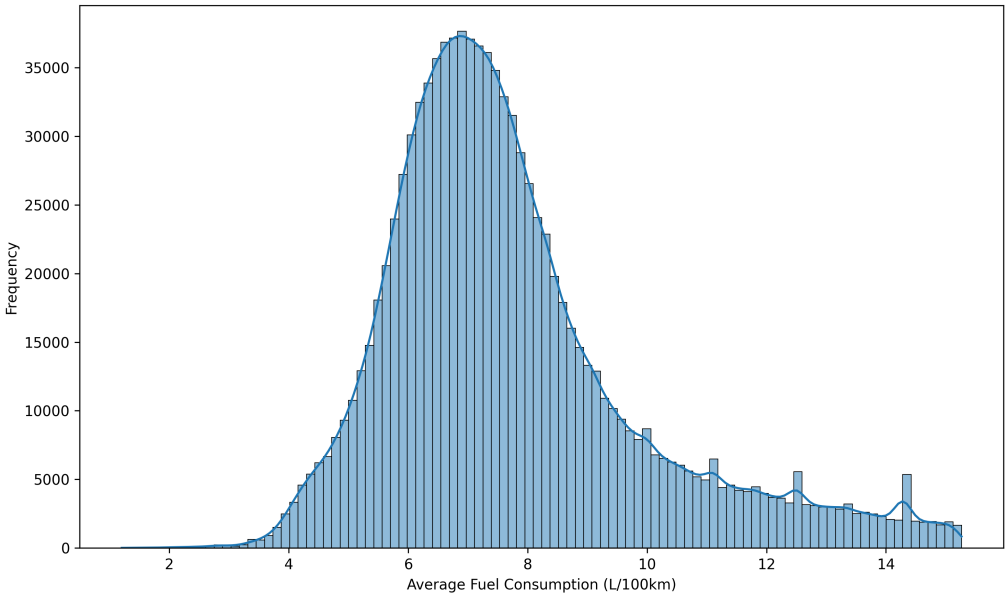

Figure 24. *Fuel Consumption Distribution for K2 Petrol Vehicles*

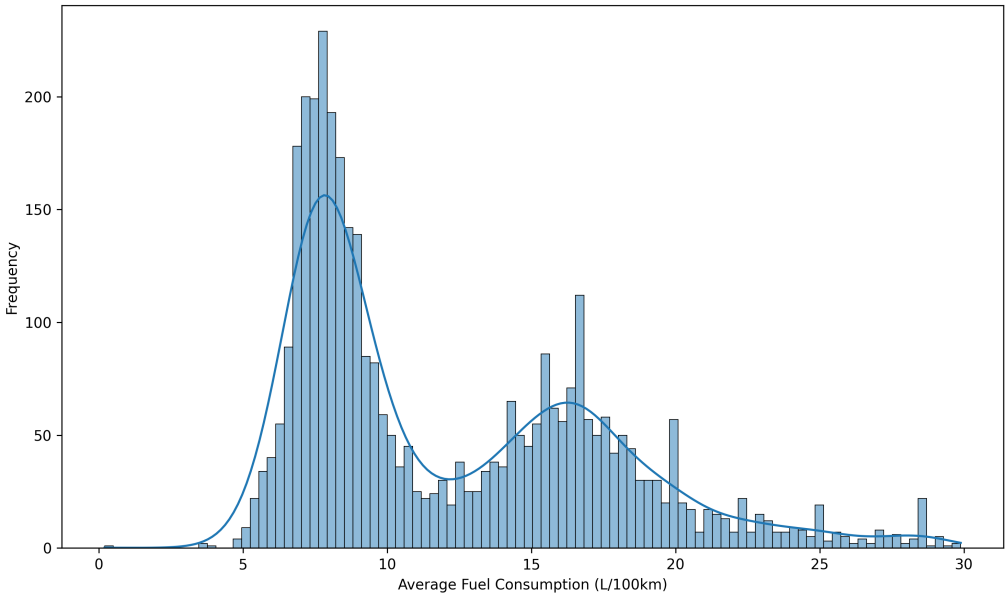

Figure 25. *Fuel Consumption Distribution for K6 Petrol Vehicles*

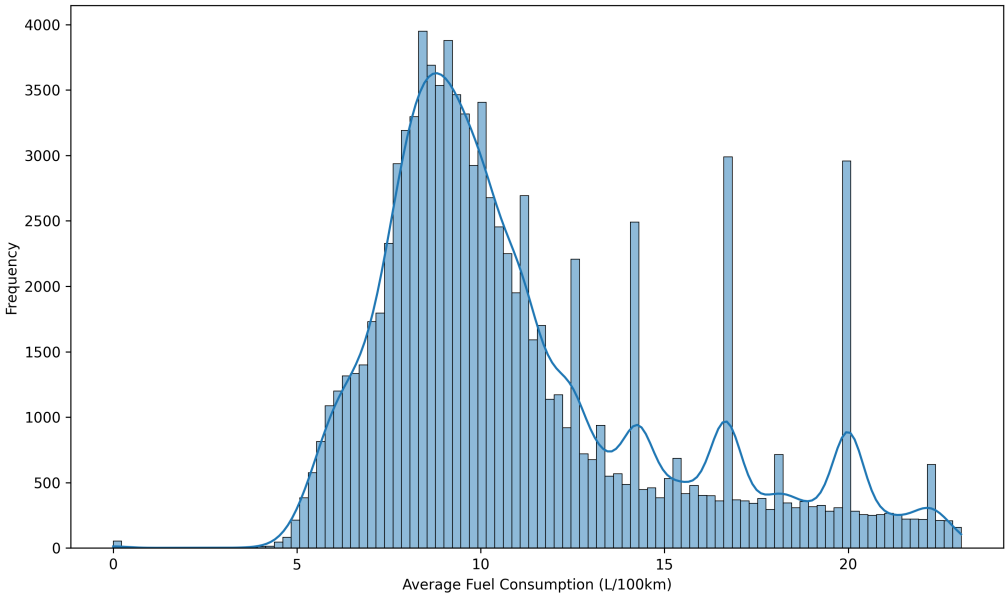

Figure 26. *Fuel Consumption Distribution for K6 Diesel Vehicles*

References and sources

[1] E. E. A. (EEA). *Collecting real-world data on the CO2 emissions of passenger cars and vans*. Accessed: 2024 December. 2024. URL: https://www.eea.europa.eu/en/datahub/datahubitem-view/1c1ffad2-34c3-471b-bd69-dd013cdd7b80#:~:text=Collecting%20real-world%20data%20on%20the%20CO2.

[2] P. Anttila, T. Nummelin, K. Väätäinen, J. Laitila, J. Ala-Ilomäki, A. Kilpeläinen. "Effect of vehicle properties and driving environment on fuel consumption and CO2 emissions of timber trucking based on data from fleet management system." In: *Transportation Research Interdisciplinary Perspectives* 15 (2022). Accessed December 2024, page 7. https://doi.org/https://doi.org/10.1016/j.trip.2022.100671.

[3] Apache Spark Community. *PySpark Overview*. Version 3.5.4, Accessed: 2024-12-17. 2024. URL: https://spark.apache.org/docs/latest/api/python/index.html.

[4] M. Bachmann. *RapidFuzz: Levenshtein Distance in Python*. Accessed: 2024 December. 2021. URL: https://rapidfuzz.github.io/Levenshtein/.

[5] L. Breiman. "Random Forests." In: *Machine Learning* 45 (2001). Accessed: 2024 December, pages 5–32.

[6] catboost. *CatBoost Regression Documentation*. Accessed: 2024 December. URL: https://catboost.ai/en/docs/concepts/python-reference_catboostregressor.

[7] T. Chen, C. Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). URL: https://api.semanticscholar.org/CorpusID:4650265.

[8] E. Commission. *Real-world CO2 emissions and fuel consumption of cars and vans collected in 2022-2024*. URL: https://climate.ec.europa.eu/news-your-voice/news/publication-real-world-co2-emissions-and-fuel-consumption-cars-and-vans-collected-2022-2024-07-26_en (viewed 2024-12-18).

[9] U. D. of Energy, U. E. P. Agency. *Fuel Economy Dataset*. Accessed: 2024 December. 2024. URL: https://www.fueleconomy.gov/feg/download.shtml.

[10] European Commission. "The European Green Deal - Priorities 2019-2024." In: *European Commission* (2024). URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en.

[11] Eurostat. *Final energy consumption in transport - detailed statistics*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Final_energy_consumption_in_transport_-_detailed_statistics.

[12] J. Fang, L. Zhou, H. Liu, Y. Zhang. "Application of machine learning for fuel consumption modelling of trucks." In: *Transportation Research Part C: Emerging Technologies* 144 (2023). https://doi.org/10.1016/j.trc.2023.103983.

[13] J. H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine.” In: *The Annals of Statistics* 29.5 (2001), pages 1189–1232. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/2699986 (viewed 2024-12-22).

[14] Grammarly. *Grammarly Writing Assistant*. Grammatical accuracy and style enhancement tool. Accessed: 2024 December. URL: https://app.grammarly.com/.

[15] Yandex. *CatBoost Documentation*. Accessed: 2024 December. URL: https://catboost.ai.

[16] S. Johnson, P. Sabharwall, Y. Ballout. “Global energy policy analysis to achieve near-term climate goals in the United States.” In: *Next Energy* 1.4 (2023). Accessed December 2024, Open Access, Under a Creative Commons license. URL: https://doi.org/10.1016/j.nxener.2023.100070.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” In: *Neural Information Processing Systems*. 2017. URL: https://api.semanticscholar.org/CorpusID:3815895.

[18] lightgbm. *LGBMRegressor*. Accessed: 2024 December. URL: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html.

[19] O. S. P. of Lithuania. *Number of registered vehicles in Lithuania by type*. 2023. URL: https://osp.stat.gov.lt/.

[20] S. Lithuania. *Kuro ir energijos suvartojimas 2023 [Fuel and Energy Consumption 2023]*. 2023. URL: https://osp.stat.gov.lt/lietuvos-aplinka-zemes-ukis-ir-energetika-2023/energetika/kuro-ir-energijos-suvartojimas.

[21] S. Lithuania. *Number of Road Vehicles at the End of the Year*. Accessed: 2024 December. 2024. URL: https://osp.stat.gov.lt/statistiniu-rodikliu-analize?indicator=S5R036#/.

[22] L. R. susisiekimo ministerija. *Motorinių transporto priemonių ir jų priekabų privalomosios techninės apžiūros atlikimo tvarkos ir Europos ekonominės erdvės šalyse atliktos privalomosios techninės apžiūros pripažinimo sąlygų ir tvarkos aprašas*. Accessed: 2024 December. 2018. URL: https://www.vta.lt/wp-content/uploads/2018/05/Periodiskumas.pdf.

[23] N. L. Misja Steinmetz Emiel van Eijk. *Real-world fuel consumption and electricity consumption of passenger cars and light commercial vehicles*. 2023.

[24] A. Mohammadnazar, Z. Khattak, A. Khattak. “Assessing driving behavior influence on fuel efficiency using machine-learning and drive-cycle simulations.” In: *Transportation Research Part D* 126 (2024), page 12. URL: https://doi.org/10.1016/j.trd.2023.104025.

[25] OpenAI. *ChatGPT 4.0*. Text refinement and language optimization tool. Accessed: 2024 December. URL: https://chatgpt.com/.

[26] OpenAI. *OpenAI API Documentation: GPT-3.5 Turbo*. Accessed: 2024 December. URL: https://platform.openai.com/docs/models#gpt-3-5-turbo (viewed 2025-01-04).

[27] M. L. Pard. *Total Motorcycle Fuel Economy Guide*. Accessed: 2024-12-21. 2024. URL: https://www.totalmotorcycle.com/MotorcycleFuelEconomyGuide.

[28] H. Patino-Artaza, L. C. King, I. Savin. "Did COVID-19 really change our lifestyles? Evidence from transport energy consumption in Europe." In: *Energy Policy* 191 (2024). URL: https://doi.org/10.1016/j.enpol.2024.114204.

[29] L. F. Quirama, M. Giraldo, J. I. Huertas, M. Jaller. "Driving cycles that reproduce driving patterns, energy consumptions and tailpipe emissions." In: *Transportation Research Part D: Transport and Environment* 82 (2020). Accessed December 2024. https://doi.org/10.1016/j.trd.2020.102294. URL: https://doi.org/10.1016/j.trd.2020.102294.

[30] scikit-learn. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/ensemble.html.

[31] scikit-learn. *ElasticNet*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.

[32] scikit-learn. *GradientBoostingRegressor*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

[33] scikit-learn. *Lasso*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.

[34] scikit-learn. *LinearRegression*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[35] scikit-learn. *RandomForestRegressor*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[36] scikit-learn. *Ridge*. Accessed: 2024 December. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html.

[37] SeatGeek. *fuzzywuzzy: Fuzzy String Matching in Python*. Accessed: 2024 December. 2023. URL: https://pypi.org/project/fuzzywuzzy/.

[38] G. M. H. Shahariar, T. A. Bodisco, A. Zare, M. Sajjad, M. I. Jahirul, T. C. Van, H. Bartlett, Z. Ristovski, R. J. Brown. "Real-driving CO2, NOx and fuel consumption estimation using machine learning approaches." In: *Next Energy* 1 (2023). https://doi.org/10.1016/j.next.2023.100060.

[39] G. H. Shahariar, T. A. Bodisco, A. Zare, M. Sajjad, M. Jahirul, T. C. Van, H. Bartlett, Z. Ristovski, R. J. Brown. "Impact of driving style and traffic condition on emissions and fuel consumption during real-world transient operation." In: *Fuel* (2022). URL: https://doi.org/10.1016/j.fuel.2022.123874.

[40] A. Spark. *KMeans*. Accessed: 2024 December. URL: https://spark.apache.org/docs/latest/ml-clustering.html#k-means.

[41] S. Tsemekidi Tzeiranaki, M. Economidou, P. Bertoldi, C. Thiel, G. Fontaras, E. L. Clementi, C. Franco De Los Rios. “The impact of energy efficiency and decarbonisation policies on the European road transport sector.” In: *Transportation Research Part A: Policy and Practice* 170 (2023). Accessed December 2024. URL: https://doi.org/10.1016/j.tra.2023.103623.

[42] WorldAtlas. *The Major Mountain Ranges in Europe*. Accessed: 2024 December. URL: https://www.worldatlas.com/articles/the-major-mountain-ranges-in-europe.html.

[43] xgboost. *XGBRegressor*. Accessed: 2024 December. URL: https://xgboost.readthedocs.io/en/stable/python/python_api.html#xgboost.XGBRegressor.

[44] X. Zhang, Y. Li, Q. Wang, Z. Chen. “A review of machine learning approaches for electric vehicle energy consumption modelling in urban transportation.” In: *Renewable Energy* 234 (2024). https://doi.org/10.1016/j.renene.2024.121243.

[45] J. Zhao, S. Heydari, M. Forrest, A. Stevens, J. Preston. “Investigating correlates of personal and freight road transport energy consumption: A case study of England.” In: *Journal of Transport Geography* 112 (2023). Accessed December 2024. URL: https://doi.org/10.1016/j.jtrangeo.2023.103693.

[46] J. Zhao, S. Heydari, M. Forrest, A. Stevens, J. Preston. “Investigating correlates of personal and freight road transport energy consumption: A case study of England.” In: *Journal of Transport Geography* 112 (2023). Accessed December 2024. URL: https://doi.org/10.1016/j.jtrangeo.2023.103693.

[47] L. Zhao, M. Wang, F. Zhang. “A review on low carbon fuels for road vehicles.” In: *Energy Reports* 9 (2023). https://doi.org/10.1016/j.egyr.2023.09.123.