Chapter 1

Methodology

1.1 Introduction

This chapter delineates the methodological framework employed to develop and evaluate an AI-based model for predicting vehicle maintenance costs, facilitating a comparative analysis of efficiency between Electric Vehicles (EVs) and Internal Combustion Engine (ICE) vehicles. The methodology encompasses data collection from diverse sources, preprocessing to ensure data quality, feature engineering to create a target variable, model selection and training using advanced machine learning techniques, and performance evaluation using robust metrics. The primary objective is to predict annual maintenance costs, which serve as a proxy for vehicle efficiency, enabling insights into the operational and economic differences between EVs and ICE vehicles under various conditions.

1.2 Research Design

The study adopts a quantitative and experimental research design, leveraging supervised machine learning algorithms to analyze historical vehicle data. The focus is on regression models to predict a continuous target variable—annual maintenance costs. This design enables a detailed assessment of factors influencing maintenance costs, providing a quantitative basis for comparing EV and ICE vehicle efficiency. The experimental approach involves training multiple machine learning models, tuning their hyperparameters, and validating their performance through cross-validation to ensure robust and generalizable results.

1.3 Data Collection

1.3.1 Data Sources

The dataset utilized in this study, vehicle_comparison_dataset_030417.csv, was compiled from multiple open-access sources, including:

- U.S. Environmental Protection Agency (EPA): Provides standardized vehicle performance metrics (https://www.epa.gov).
- Kaggle Public Datasets: Offers real-world vehicle-related datasets (https://www.kaggle.com).
- EV and ICE Manufacturer Performance Records: Includes detailed specifications and operational data from manufacturers.
- Research Publications and Academic Journals: Supplements the dataset with validated performance metrics.

The dataset comprises 2000 records, each representing a vehicle with a comprehensive set of attributes relevant to efficiency analysis.

1.3.2 Key Data Features

The dataset includes the following key attributes:

- Vehicle Type: Categorical variable indicating whether the vehicle is an EV or ICE (e.g., "EV" or "ICE").
- Energy Consumption: Numerical, measured in kWh/100 km for EVs or L/100 km for ICE vehicles (e.g., 13.80 kWh/100 km).
- CO_2 Emissions: Numerical, grams of CO_2 emitted per kilometer (g/km), ranging from 0 (post-cleaning) to values like 193.10 g/km.
- Maintenance Cost: Numerical, annual maintenance cost in monetary units (e.g., 51,768.21 units), serving as the target variable.
- Cost per km: Numerical, operational cost per kilometer (e.g., 0.14 units/km).
- Energy Storage Capacity: Numerical, battery capacity for EVs or fuel tank capacity for ICE vehicles (e.g., 80.72 kWh or 61.84 L).
- Mileage: Numerical, total kilometers driven (e.g., 138,080.30 km).
- Acceleration: Numerical, time to accelerate from 0 to 100 km/h in seconds (e.g., 5.19 sec).

- Torque: Numerical, engine torque in Newton-meters (Nm) (e.g., 328.70 Nm).
- Lifespan: Numerical, vehicle lifespan in years (e.g., 14.33 years).

These features provide a comprehensive view of vehicle performance and efficiency, with maintenance_cost_annual as the primary target for prediction.

1.4 Data Preprocessing

1.4.1 Data Cleaning

The dataset was examined for missing values using df.isna().sum(), confirming no missing entries across all columns. However, negative values in the co2_emissions_g_per_km column were identified, likely due to data anomalies or EV-specific offsets. These were addressed by clipping negative values to zero:

df["co2_emissions_g_per_km"] = df["co2_emissions_g_per_km"].clip(lower=0)

This ensures that CO_2 emissions are non-negative, aligning with physical expectations.

1.4.2 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to understand feature distributions and relationships. Key visualizations included:

- Box Plots: Compared feature distributions between EV and ICE vehicles, revealing that EVs have higher median energy consumption (13–15 vs. 6–10), higher torque (250–350 Nm vs. 150–250 Nm), lower CO_2 emissions (near 0 vs. 150–175 g/km), and lower maintenance costs (40,000–50,000 vs. 90,000–100,000 units).
- Correlation Heatmap: Identified potential multicollinearity among numerical predictors, guiding feature selection.

These insights informed preprocessing and model development strategies.

1.4.3 Normalization and Encoding

The preprocessing steps were integrated into a ColumnTransformer within a Pipeline:

- Numerical Features: Features such as energy_consumption, acceleration_0_100_kph_sec, torque_Nm, and lifespan_years were standardized using StandardScaler to ensure zero mean and unit variance, improving model convergence.

- Categorical Features: The vehicle_type column was one-hot encoded using OneHotEncoder with drop='first' to avoid multicollinearity, transforming it into binary columns (e.g., vehicle_type_ICE).

This pipeline ensured consistent preprocessing across training and testing phases.

1.5 Machine Learning Model

1.5.1 Algorithm Selection

Seven regression models were evaluated to predict annual maintenance costs:

- Linear Regression: A baseline model for capturing linear relationships.
- Decision Tree Regressor: Models non-linear patterns but is prone to overfitting.
- Random Forest Regressor: An ensemble method combining multiple decision trees for robustness.
- Gradient Boosting Regressor: Builds trees sequentially to correct errors, balancing bias and variance.
- CatBoost Regressor: Handles categorical features natively and reduces overfitting (https://catboost.ai).
- LightGBM Regressor: Optimized for speed and accuracy in gradient boosting (https://lightgbm.readthedocs.io).
- XGBoost Regressor: Selected for its high accuracy, robustness to outliers, and feature importance analysis capabilities (https://xgboost.readthedocs.io).

For validation, Random Forest and Linear Regression were emphasized due to their interpretability and performance, correcting the initial mention of Logistic Regression, which is unsuitable for regression tasks.

1.5.2 Model Training

- Pipeline Construction: Each model was integrated into a Pipeline combining preprocessing and model training, ensuring consistent data transformation.
- Data Splitting: The dataset was split into 80% training and 20% testing sets using train_test_split with random_state=42 for reproducibility.
- Cross-Validation: 5-fold cross-validation was employed using KFold with stratification on vehicle_type to maintain class balance across folds.

- Hyperparameter Tuning: For XGBoost, GridSearchCV was used to optimize hyperparameters over a grid:
  - n_estimators: [100, 200, 300]
  - max_depth: [3, 5, 7]
  - learning_rate: [0.01, 0.05, 0.1]
  - min_child_weight: [1, 3, 5]
  - subsample: [0.6, 0.8, 1.0]
  - colsample_bytree: [0.6, 0.8, 1.0]

  The optimal parameters identified were:
  - colsample_bytree: 0.8
  - learning_rate: 0.01
  - max_depth: 3
  - min_child_weight: 3
  - n_estimators: 300
  - subsample: 0.6

Feature importance was analyzed using the tuned XGBoost model to identify key predictors of maintenance costs.

1.6 Efficiency Evaluation Criteria

The efficiency of each vehicle type is indirectly assessed through the prediction of annual maintenance costs, which reflect operational and economic efficiency. Lower maintenance costs suggest higher efficiency, particularly for EVs, which may benefit from simpler mechanical designs. The model's predictive accuracy on this target enables a comparative analysis of EV and ICE vehicle efficiency.

1.7 Model Evaluation Metrics

The performance of the models was evaluated using the following metrics:

- Mean Absolute Error (MAE): Measures the average absolute difference between

predicted and actual maintenance costs:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where y_i is the actual maintenance cost, $\hat{y}_i$ is the predicted cost, and n is the number of observations.

- Root Mean Squared Error (RMSE): Quantifies the square root of the average squared differences, emphasizing larger errors.
- R^2 Score: Indicates the proportion of variance in the target variable explained by the model.

Feature importance rankings from the XGBoost model were used to interpret contributing factors to maintenance costs.