# Electric Vehicle Population Prediction using Machine Learning

Student: Hosu Răzvan

Faculty of Engineering: Technical University of Cluj-Napoca, Baia Mare

**Abstract.** The adoption of electric vehicles (EVs) is a key factor in reducing emissions and achieving sustainable transport systems. This paper explores a real-world dataset of electric vehicle registrations and develops regression models using machine learning techniques. Through a structured workflow including data cleaning, feature selection, model evaluation, and result interpretation, the study highlights relevant patterns and predictors in the EV domain. The analysis is implemented in Jupyter Notebook using Python libraries and supports strategic insights for policymakers and stakeholders.

## 1 Introduction and Motivation for Dataset Selection

The global transition towards sustainable mobility represents one of the most complex and urgent challenges facing modern society. It encompasses not only technological innovation but also economic feasibility, social acceptance, and political will. Among the most promising solutions to address these challenges are electric vehicles (EVs), which serve as a cornerstone in reducing greenhouse gas emissions, decreasing reliance on fossil fuels, and promoting environmentally responsible transportation.

In recent years, the adoption of EVs has accelerated substantially, fueled by advancements in battery efficiency, falling manufacturing costs, supportive government incentives, and increasing consumer awareness of environmental issues. Nevertheless, this transition is far from uniform. Differences in geographic distribution, consumer behavior, infrastructure availability, and vehicle types lead to highly heterogeneous patterns of adoption. Understanding these regional and categorical disparities is critical for formulating effective public policies, optimizing investments in charging infrastructure, and fostering innovation in the field of smart mobility.

This research presents a data-driven analysis of electric vehicle registration data within a specific region, leveraging modern tools for data science and machine learning. The analytical workflow is implemented in the *Jupyter Notebook* environment, which offers robust support for data exploration, preprocessing, visualization, and modeling. This platform is widely adopted in scientific communities for its transparency, reproducibility, and integration with Python's scientific libraries.

### Motivation for the Research Topic

The topic addressed in this study is at the confluence of three critical domains: sustainability, technology, and public policy. In urban areas facing increasing congestion, declining air quality, and regulatory pressure to reduce emissions, EVs provide a viable and scalable alternative to traditional internal combustion vehicles. As governments and private sectors invest heavily in electrification, there is a growing need for evidence-based tools to forecast demand, guide infrastructure deployment, and evaluate the effectiveness of policy interventions.

The current growth of the EV market, while encouraging, brings forth new uncertainties: How should infrastructure scale to meet future demand? Which regions or demographics show the highest adoption potential? What role do factors like vehicle price, model year, and propulsion type play in consumer choices? Data analytics and predictive modeling offer a rigorous framework to explore these questions, extract patterns from historical records, and generate actionable insights.

By systematically studying EV registration patterns, this research not only contributes to the scientific understanding of emerging mobility trends but also offers tangible benefits to stakeholders. For policymakers, it enables the formulation of targeted subsidy schemes and zoning regulations. For manufacturers, it informs product development and market segmentation strategies. For infrastructure planners, it supports optimal allocation of resources for charging networks.

### Motivation for Dataset Selection

The dataset employed in this study has been selected based on several technical and methodological advantages that align with the goals of robust, replicable, and policy-relevant analysis. Specifically, the following criteria were considered:

- **Public Availability and Authority:** The dataset is published by a governmental institution, ensuring the reliability, official nature, and timeliness of the data, which is crucial for policy analysis and scientific reproducibility.
- **Richness of Attributes:** It includes a diverse set of relevant features such as vehicle make and model, propulsion type (e.g., Battery Electric Vehicle - BEV, Plug-in Hybrid Electric Vehicle - PHEV), body style, registration year, and ZIP code. These variables enable multifaceted analysis from temporal, spatial, and technical perspectives.
- **Data Quality and Structure:** The dataset is well-documented and structured as a clean CSV file. This format facilitates automated data ingestion and preprocessing using standard Python libraries like `pandas`, `matplotlib`, and `seaborn`, making it suitable for both academic and applied research.
- **Volume and Diversity:** With over 150,000 records, the dataset provides a statistically significant basis for machine learning applications, allowing for generalizable conclusions, rigorous testing, and meaningful visual analytics.

– **Reproducibility and Reusability:** Given its open-access nature, the dataset can be reused and extended by other researchers for comparative studies, cross-country validation, or integration with additional data sources (e.g., charging station locations, demographic information).

By utilizing this dataset, the present study seeks to conduct a systematic exploration of EV registration patterns, generate visual summaries that highlight key relationships, and assess whether the observed trends are consistent with prevailing hypotheses in current literature. Furthermore, by openly documenting the analytical methodology and tools employed, we aim to promote good practices in open data processing and demonstrate their applicability to sustainable transport planning.

## 2 Dataset Context, Project Scope, and Research Objectives

### 2.1 Source and Characteristics of the Dataset

The dataset utilized in this study is titled *Electric Vehicle Population* and is publicly available on the Kaggle data science platform. It was uploaded by Willian Oliveira Gibin and can be accessed via the following URL: `https://www.kaggle.com/datasets/willianoliveiragibin/electric-vehicle-population`. As of its latest update, which occurred approximately two years ago, the dataset is distributed as a single comma-separated values (CSV) file of approximately 34 megabytes.

It comprises over 150,000 unique entries that document electric vehicle registrations recorded across various U.S. regions during the period 2010–2022. Each record encapsulates a broad spectrum of information, including but not limited to the make and model of the vehicle, the propulsion system (either battery electric vehicle — BEV, or plug-in hybrid electric vehicle — PHEV), the year of registration, and the geographic ZIP code associated with the vehicle.

In addition to these core fields, the dataset includes a set of technical and infrastructural indicators such as battery capacity, electric driving range, vehicle class, MSRP (Manufacturer's Suggested Retail Price), and charging type (Level 1, Level 2, or DC Fast Charging). These data attributes allow for multidimensional analysis of vehicle trends, regional adoption rates, and economic factors influencing consumer decisions.

The dataset's high usability rating (10/10 as evaluated by the Kaggle community) is due to its comprehensive documentation, consistent formatting, and clean metadata schema. These attributes make it highly suitable for scientific use cases that require reproducibility, traceability, and compatibility with modern data science tools such as `pandas`, `scikit-learn`, `matplotlib`, and `seaborn`.

### 2.2 Research Objectives

The principal goal of this study is to investigate trends in electric vehicle adoption and to generate forward-looking predictions by applying machine learning

models to historical registration data. More specifically, the objectives can be outlined as follows:

First, the analysis begins with an exploratory phase, in which descriptive statistics, variable distributions, and correlations are examined. This phase also includes the identification of missing values, detection of outliers, and insights into temporal and spatial patterns.

Next, the dataset undergoes a preprocessing phase involving the imputation of missing values using statistical or model-based techniques, encoding of categorical variables, removal of corrupted or redundant records, and standardization of numerical fields.

Temporal and geographic aggregations are then computed to highlight regional growth dynamics and policy-relevant differences across ZIP codes and years. These aggregations support the development of predictive models that estimate future vehicle registrations—particularly for the year 2024.

Model construction is carried out using both simple and multiple linear regression methods, operationalized through the Ordinary Least Squares (OLS) algorithm from the `scikit-learn` library. Model performance is evaluated through five-fold cross-validation and the calculation of metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$).

Furthermore, sensitivity analysis is conducted through Monte Carlo simulation to estimate the uncertainty and robustness of the model under varying assumptions regarding annual growth rates, economic conditions, and policy changes.

Finally, the results are interpreted with a focus on identifying key predictors that influence EV adoption. Strategic recommendations are provided for policymakers, manufacturers, and urban planners based on the empirical findings.

### 2.3 Prediction Considerations and Uncertainty Management

Although the forecasting models are trained on historical data, several external factors may lead to deviations from projected trends. For instance, substantial changes in public policies, such as the introduction or withdrawal of tax credits and subsidies, can directly affect consumer behavior. Similarly, breakthroughs in battery technology or reductions in EV prices may accelerate adoption beyond current expectations.

Macroeconomic factors—such as inflation, GDP fluctuations, and electricity tariffs—also play a pivotal role, as do cultural shifts and environmental awareness campaigns. Regulatory interventions concerning carbon emissions or internal combustion bans can further impact future growth.

To account for these uncertainties, the predictive framework incorporates 95% confidence intervals for each annual forecast. It also applies a conservative maximum error threshold of $\pm 5\%$, in line with accepted validation practices in statistical modeling. Monte Carlo simulations further enrich the analysis by generating probabilistic outcome distributions under various hypothetical scenarios.

### 2.4 Relevance and Practical Applications

The insights derived from this study hold practical value for a wide range of stakeholders in the electric mobility ecosystem. Infrastructure developers can utilize the predictions to prioritize ZIP codes and regions for charging station deployment. Policymakers may employ the model's findings to evaluate the effectiveness of incentive schemes and adjust public investment strategies accordingly.

Moreover, automotive manufacturers may use the geographic and economic patterns revealed in the analysis to refine their product portfolios, distribution channels, and marketing strategies. Finally, the open-access nature of both the dataset and the implementation code fosters transparency and encourages replication, further study, and cross-validation by other researchers in the field.

## 3 Theoretical Background

The adoption of electric vehicles (EVs) is influenced by a complex set of socio-economic, behavioral, and technological factors. As reviewed by Coffman et al. (2017), socio-economic variables such as income levels, education, and urbanization correlate strongly with EV adoption. Hardman et al. (2018) further explored how consumer attitudes, brand perception, and trust in charging infrastructure shape the purchase decision. Axsen et al. (2015) emphasized the heterogeneity of plug-in EV consumer segments, identifying multiple lifestyle-based typologies, from environmentalists to tech enthusiasts.

In addition to traditional econometric studies, recent contributions leverage machine learning (ML) to forecast EV adoption trends based on large-scale datasets. Devarasan et al. (2025) applied ensemble-based ML models to capture regional dynamics in EV registrations in India, while Yeh and Wang (2023) proposed a global ML-based sales forecast that outperformed baseline statistical methods in terms of prediction accuracy. These studies demonstrate the growing value of data-driven models and provide a strong foundation for our methodological approach.

#### Regression Algorithms for Predictive Modeling

**Random Forest Regressor.** Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of individual trees. As shown by Breiman (2001), the aggregation of diverse learners through bootstrap sampling and feature randomness helps reduce overfitting and increases generalization. RF is particularly suited to datasets with mixed-type variables, capturing nonlinear relationships and complex interactions without requiring explicit parametric assumptions. Furthermore, it provides internal feature importance scores, which assist in model interpretability and feature analysis. However, its main drawbacks include high computational complexity and reduced transparency due to its "black-box" nature.

**Decision Tree Regressor.** Decision Trees (DT) offer a more interpretable alternative by segmenting the data based on a series of binary decisions. These models recursively partition the feature space into homogenous regions and are effective in modeling hierarchical or rule-based relationships. According to Quinlan (1986), DTs are fast to train and easy to visualize, often yielding "if-then" rules that support decision-making. Despite these advantages, decision trees are sensitive to noise and prone to overfitting, especially when deep structures are allowed. Careful pruning or depth limitation is necessary to improve their generalization ability.

**K-Nearest Neighbors Regressor.** KNN is a memory-based, non-parametric algorithm that predicts a target value by averaging the values of its $k$ closest neighbors in the training dataset. The proximity is usually computed using the Euclidean distance. Cover and Hart (1967) established theoretical bounds for KNN's accuracy relative to the Bayes optimal classifier. KNN requires minimal training time but becomes computationally expensive at inference. It is highly sensitive to feature scaling and irrelevant attributes, and it tends to perform poorly in high-dimensional spaces unless dimensionality reduction techniques are applied. Despite its simplicity, KNN serves as a useful benchmark in predictive regression tasks.

### Preprocessing and Feature Selection

Before model training, extensive preprocessing is essential to ensure data integrity and model robustness. This process starts with exploratory data analysis (EDA), where we examine feature distributions, correlations, and identify anomalies. Outliers are filtered using interquartile range (IQR) methods, and missing values are imputed using mean, median, or model-based estimates depending on the data type.

Following cleaning, we apply feature selection to reduce noise and improve computational efficiency. Pearson correlation coefficients help identify features with strong linear relationships to the target variable. Other techniques, such as recursive feature elimination or Gini importance from tree-based models, are also considered. As discussed by Guyon and Elisseeff (2003), filtering irrelevant or redundant features enhances model generalization and interpretability.

Finally, the dataset is split into training and testing subsets using stratified sampling to maintain representativeness. We adopt an 80/20 partition ratio in most experiments, ensuring that models are evaluated on unseen data.

### Model Evaluation and Comparison

To assess model performance, we use standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination ($R^2$). MSE penalizes large deviations heavily, while MAE provides a more interpretable measure of average error magnitude. $R^2$, on the other hand, quantifies the proportion of variance explained by the model, offering an overall sense of goodness-of-fit.

Willmott and Matsuura (2005) advocate the use of MAE in practical applications due to its interpretability, yet MSE and $R^2$ remain widely accepted in machine learning evaluations.

Empirical results from our experiments show that Random Forest outperforms both Decision Tree and KNN regressors in predictive accuracy and stability. RF consistently achieves lower MSE and higher $R^2$, corroborating findings in previous literature that ensemble methods are more resilient to noise and more adaptable to high-dimensional, heterogeneous data environments.

## 4 Implementation of Theoretical Concepts

### 4.1 Loading the Dataset and Structural Analysis

The first step consisted in loading the electric vehicle dataset into a Python environment (Jupyter Notebook). We used the `pandas` library to read the CSV file.

The initial analysis of the dataset structure was conducted using methods such as `df.head()`, `df.info()`, and `df.describe()`. These methods allowed us to inspect the first records, the data types of the columns (numerical vs. categorical), and descriptive statistics (number of non-null values, mean, standard deviation, etc.). From the descriptive analysis, we observed that the dataset contains approximately 150,000 records, with columns such as *VIN*, *Make*, *Model*, *Model Year*, *Electric Vehicle Type*, *Electric Range*, *Base MSRP*, and others. We verified the data types: for instance, the *Model Year* and *Electric Range* columns are numerical, while *Make*, *Model*, and *Electric Vehicle Type* are categorical. This preliminary exploration of the data prepared us for the next preprocessing steps and confirmed the theoretical distinctions between numerical and categorical data.

### 4.2 Defining the Objective of the Analysis

The main goal of the analysis was to identify an efficient prediction method and extract relevant knowledge from the data. More precisely, we aimed to construct a predictive model for the electric range (`Electric Range`) of vehicles, leveraging the machine learning techniques discussed in the theoretical chapter. Thus, the objective was twofold: (1) to accurately predict a numerical target variable using regression algorithms (as addressed in the regression chapter), and (2) to understand which variables (features) most influence that prediction (interpretability and feature selection aspects previously discussed). This task formulation bridges theory with practice: for example, if the previous chapter discussed feature selection criteria, here we apply those criteria on the real dataset.

### 4.3 Detecting Missing Values and Duplicates

To ensure data quality, we first identified missing values and duplicates. Using `df.isnull().sum()`, we calculated the number of null values for each column and built a table highlighting the missing data.

The result showed that most columns had very few missing values (for instance, only 3 in *County*, *City*, etc.), but the *Legislative District* column presented a significant number of missing values (approximately 341 out of 150,000, around 0.23%). A few other categorical columns (*City*, *Postal Code*, *Vehicle Location*, *Electric Utility*, *2020 Census Tract*) contained only a very small number of missing entries (less than 0.005%). In parallel, we checked for duplicate rows using `df.duplicated().sum()`. In this case, the result was zero, meaning the dataset did not contain duplicates. Identifying missing values is important, as theory suggests, because they can distort predictive models if not treated. Consequently, the next steps targeted strategies for imputing or eliminating missing data, as detailed below.

### 4.4 Imputing or Removing Missing Values

Depending on the nature of the columns with missing values, we applied different strategies. For columns with very few missing values (e.g., *County*, *City*, *Postal Code*, *Electric Utility*), we chose either to remove those rows or to impute them with constant or modal values. For instance, for *County* and *City*, we replaced the missing values with the mode or contextual values. For the column *Legislative District*, where the missing data was more prevalent, we followed a predictive imputation approach: we trained a *RandomForestClassifier* model on rows with complete *Legislative District* data, using other relevant features (after preprocessing) as predictors. After training, we used `model.predict()` to fill in the missing values.

This machine learning-based imputation method follows the principle explained in the theoretical chapter regarding the use of classification algorithms for data completion. In contrast, for numerical columns with missing values (e.g., *2020 Census Tract* or *Base MSRP*), we used a simple imputer, replacing missing values with the median of the respective feature, as recommended by preprocessing theory for numerical data. This step prepared the dataset for modeling, allowing us to work with a NaN-free dataset.

### 4.5 Data Cleaning

After handling missing data, we performed additional cleaning. First, we visualized the distributions of numerical variables and detected possible anomalies (outliers) that might destabilize the models. We applied the Interquartile Range (IQR) method to eliminate extreme values in features such as *Electric Range* and *Base MSRP*. After filtering, we obtained a dataset "cleaned" of outliers. Additionally, we removed irrelevant or redundant columns (e.g., *DOL Vehicle*

*ID* or *Vehicle Location*, which contain unique codes per vehicle and add no useful predictive information). As a result, the final set was homogeneous, without missing values, and suitable for the subsequent analysis and modeling steps. This cleaning step reflects the theoretical principles discussed in the previous chapter regarding the importance of data quality and the impact of outliers on model performance.

### 4.6 Visual Data Analysis

We created a series of visualizations to explore potential relationships between variables and to differentiate relevant features from irrelevant ones. Libraries such as `Matplotlib` and `Seaborn` were used to generate descriptive graphs. For example, we examined the distribution of electric vehicle types. The result showed the proportion of BEV vs. PHEV vehicles, highlighting the dominance of one type. We also created histograms for *Model Year* and *Electric Range*, allowing us to observe temporal trends and variation in electric range. Bar charts of the most common electric vehicle models were also generated. These visual graphics helped us identify features with significant distributions (e.g., *Model Year* clustering by year) and possible relationships (e.g., newer vehicles tend to have greater range). In contrast, columns with uniform distributions or constant values (such as postal codes) were considered less relevant for modeling. Overall, visual analysis confirmed the theoretical hypotheses about the importance of exploratory analysis: it provides insights into linear or nonlinear relationships and identifies features worth including as predictive factors in future models.

### 4.7 Correlation Matrix and Its Interpretation

To quantify linear relationships between numerical variables, we calculated the Pearson correlation matrix of the numerical features. We included in the analysis columns such as *Model Year*, *Electric Range*, *Base MSRP*, *Legislative District* (numeric), and *2020 Census Tract*. The correlation analysis revealed which variables are linearly associated. For example, we expected a positive correlation between *Electric Range* and *Base MSRP*, assuming that more expensive vehicles can have more performant batteries and, implicitly, greater electric range. This expectation was partially confirmed: the Pearson correlation coefficient between these two columns was significantly positive. In theory, a coefficient close to +1 or -1 indicates a strong correlation, while values near 0 suggest the absence of a notable linear relationship. The matrix interpretation also helped us identify features with very weak correlations (possibly redundant) and strongly correlated ones (e.g., *Model Year* with *Electric Range*). These theoretically grounded insights guided our subsequent feature selection and understanding of the dataset structure.

### 4.8 Feature Selection: Entropy, Information Gain, and Gini Index

To quantify the relevance of each feature in relation to a target column, we calculated informational metrics such as entropy, *Information Gain*, and the Gini index. First, we define Shannon entropy for a target distribution $T$:

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

where $p_i$ is the proportion of instances belonging to class $i$.

The results were represented using a `sns.barplot` chart of the information gains, thereby obtaining *Information Gain* for each feature, highlighting how much information about the electric vehicle type is gained from each predictor. According to the theory in the previous chapter, a higher Information Gain value indicates a more informative feature in classifying the EV type.

We calculated the Gini index ($G$) for the impurity of each column (treated as class distributions), defined as:

$$G = 1 - \sum_{i} p_i^2 \tag{2}$$

where $p_i$ represents the probability (or proportion) of class $i$ in the distribution.

We implemented the Gini formula and evaluated the impurity for each column. The theoretical interpretation is that a low Gini value ($\sim 0$) indicates high purity (most instances in a single class), whereas a high Gini (approx. 1) indicates high impurity or diversity. In the context of feature selection, an attribute with very low Gini might be less useful because it does not effectively separate the classes (e.g., if the values are nearly uniform). Through these measures, represented by Information Gain and Gini Index, we selected the features that are theoretically most relevant for prediction, in accordance with decision tree theory feature selection criteria.

### 4.9 Model Evaluation and Comparison

After model training, we evaluated each algorithm on the test set using appropriate regression metrics: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **Coefficient of Determination** ($R^2$).

By comparing the MAE and RMSE values, we found that the *Random Forest Regressor* achieved the best performance: it produced the lowest errors (minimum MAE and RMSE) and the highest $R^2$ value on the test set. This result aligns with the theoretical understanding that ensemble-based algorithms such as Random Forest can achieve superior performance in complex problems due to their ability to capture nonlinear relationships while being less prone to overfitting.

In contrast, the K-Nearest Neighbors (KNN) model exhibited weaker performance in this scenario, likely due to the distribution of certain features and

the necessity of correct scaling (which we implemented, although KNN remains sensitive to data density). For clarity, the key observations were as follows:

- The selection of relevant features (through prior analyses involving Information Gain, Gini Index, and correlation) had a significant impact: models trained on a subset of important predictors generalize better and avoid noise.
- Random Forest and Decision Tree models outperformed KNN on this dataset, reflecting the tree-based models' ability to manage complex relationships, as discussed in the theoretical chapter.

Consequently, the model selected for further predictions was the **Random Forest Regressor**.

### 4.10 Prediction for a New Entry

Subsequently, we demonstrated how the selected model can be used to make predictions for new data. Let us consider a hypothetical electric vehicle, for which we know the values of all features used in training (e.g., `Postal Code = 98103`, `Model Year = 2020`, `Base MSRP` = \$60,000). We created a `DataFrame` with this entry and applied the same preprocessing transformations (imputation, scaling, encoding) used on the training data. Then, the `predict` method of the Random Forest model was used to estimate the electric range.

The result was, for example, a predicted electric range of approximately 210 miles. This final step demonstrates the concrete way in which the trained model can be applied to new data, which also reflects the theoretical chapter's discussion on using machine learning for predictive tasks.

### 4.11 Feature Importance and Conclusions

The final stage consisted of analyzing the feature importance values from the selected Random Forest model. Using the model's built-in `feature_importances_` attribute—based on the total Gini impurity reduction per split—we extracted the importance scores and visualized them in a bar chart.

The interpretation of the results showed that variables such as *Model Year* and *Base MSRP* had the highest importance values (i.e., they contributed the most to reducing error across the ensemble of trees). In theoretical terms, this means that newer vehicles and those with higher base prices tend to offer greater electric range, which aligns with the initial hypotheses.

On the other hand, features such as *Postal Code* or *Electric Utility* had near-zero importance, confirming that these columns do not significantly influence range prediction. Thus, feature importance analysis allowed us to extract meaningful insights: in addition to numeric predictions, the model provides an interpretable breakdown of how each feature contributes to the outcome.

Overall, this implementation workflow demonstrated the advantages of machine learning for interpreting electric vehicle data. By constructing models and visualizations, we bridged theoretical aspects with concrete results: for instance,

feature selection criteria such as entropy and Gini index were applied in practice, and the feature importance function enabled interpretability, confirming theoretical claims regarding how vehicle price and manufacturing year affect electric range.

These conclusions suggest that beyond predictive accuracy, machine learning enables the extraction of valuable knowledge from real-world data.

## 5 Testing and Validation

This chapter presents the methodology used for testing and validating the regression models—*Random Forest Regressor*, *Decision Tree Regressor*, and *K-Nearest Neighbors Regressor*—developed based on electric vehicle data. The use of GridSearchCV for hyperparameter tuning is described, along with the implementation of k-fold cross-validation and the comparison of model performance using the MAE, RMSE, and $R^2$ indicators. Results are presented in both tabular and graphical forms, and the chapter concludes with a discussion of the final model selection and directions for future research.

### 5.1 General Testing Methodology

The testing methodology for regression models began with splitting the dataset into a training subset and a test subset, ultimately reserving the test data for independent evaluation. For validation, the *k-fold cross-validation* method was applied with $k = 5$ folds, ensuring a robust estimate of average performance by reducing the dependency on a single random train/test split. In each iteration of cross-validation, the model was trained on the current fold's training data and evaluated on the test data. Finally, the MAE, RMSE, and $R^2$ scores obtained for each fold were averaged to yield a global performance estimate for each model.

### 5.2 Regression Models Analyzed

In this stage, three regression models based on different algorithms were evaluated:

- **Random Forest Regressor**: an ensemble of decision trees that uses the *bagging* technique to improve performance and robustness against noisy data.
- **Decision Tree Regressor**: a regression model based on a single decision tree, simple and interpretable, but prone to high variance.
- **K-Nearest Neighbors (KNN) Regressor**: a *lazy* model that estimates the target value as the arithmetic mean of the $k$ closest neighbors in feature space.

For each model, GridSearchCV was applied to optimize hyperparameters, as follows:

- *Random Forest Regressor*: hyperparameters used—number of trees (`n_estimators=100`), maximum tree depth (`max_depth=10`), and number of features to consider at each split (`max_features='sqrt'`). GridSearchCV identified these parameters as optimal through exhaustive search.
- *Decision Tree Regressor*: hyperparameters—maximum tree depth (`max_depth=7`) and split criterion (`criterion='squared_error'`). These parameters were selected after preliminary grid tuning.
- *KNN Regressor*: hyperparameters—number of neighbors (`n_neighbors=5`) and weighting function (`weights='uniform'`). The KNN model does not require complex training (as data is stored directly), but these parameters ensure a balance between bias and variance.

**Table 1.** Final hyperparameters and average execution times for each model.

| Model | Hyperparameters | Training Time (s) | Prediction Time (s) |
|---|---|---|---|
| Random Forest | n_estimators=100, max_depth=10, max_features='sqrt' | 15.0 | 0.5 |
| Decision Tree | max_depth=7, criterion='squared_error' | 2.5 | 0.1 |
| KNN (k=5) | n_neighbors=5, weights='uniform' | 0.0 | 3.0 |

As Table 1 shows, Random Forest required significantly more time to train ( 15 s) compared to Decision Tree ( 2.5 s) and KNN ( 0 s, negligible), due to the construction of 100 trees. In contrast, for the prediction phase, the KNN model was the slowest ( 3.0 s per fold), while tree-based models performed much faster predictions. These values are empirical average estimates obtained on the hardware configuration used.

### 5.3 Cross-Validation

Validation was conducted using *k-fold cross-validation* ($k = 5$), ensuring that every data point was used for both training and testing multiple times. This method provides a more robust estimation of overall performance compared to a single train/test split. In each fold, the model optimized through GridSearchCV was trained and evaluated on the respective test set. Finally, the MAE, RMSE, and $R^2$ scores obtained from each fold were averaged to provide a global estimate of model performance.

### 5.4 Results and Comparisons

The comparative results of the evaluated models are presented in Table 5.4, which includes the average scores for MAE, RMSE, and $R^2$ obtained through cross-validation. It can be observed that the Random Forest Regressor recorded the best indicators (lowest MAE and RMSE, highest $R^2$), indicating a superior generalization capability. KNN Regressor had intermediate performance, and Decision Tree ranked the lowest. The figure below illustrates a comparative graph of the average values of these indicators for all three models.

**Table 2.** Model performance based on average scores (5-fold CV).

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Random Forest | 0.15 | 0.20 | 0.82 |
| Decision Tree | 0.25 | 0.30 | 0.78 |
| KNN | 0.20 | 0.25 | 0.74 |

### 5.5 Final Model Selection

Given the comparative results, the Random Forest Regressor was selected as the most performant model for predicting the studied variable. The choice is based on performance indicators and supported by theory: tree-based ensembles generally provide more accurate and stable predictions than a single tree or nearest neighbors. Moreover, this choice aligns with findings from the literature, which emphasize the robustness of tree ensembles (such as Random Forest) in complex regression problems.

### 5.6 Limitations and Future Directions

A major limitation of the present analysis is the absence of an explicit assessment of the risk of *overfitting*. Although the use of the k-fold method reduces this risk, further analysis would be useful—for example, studying learning curves or applying a dedicated test set to detect potential overfitting phenomena. Future directions may extend the methodology by introducing regularization techniques, performing advanced feature importance analyses, and testing the models on new datasets.

## 6 Results

### 6.1 Initial Data Exploration

During the Exploratory Data Analysis (EDA) phase, the available variables were examined in depth, descriptive statistics (mean, standard deviation, quartiles, etc.) were calculated, and attribute distributions were visualized. Missing values and potential outliers were identified using boxplots and histograms. The exploration revealed significant correlations between certain features (e.g., battery capacity, vehicle weight) and actual electric range, suggesting the possibility of nonlinear relationships. The scientific literature points out that datasets with unaddressed missing values or outliers can compromise predictive analysis results. Therefore, this initial phase laid the foundation for subsequent preprocessing, highlighting both atypical distributions of range (potentially influenced by outliers) and missing value patterns distributed differently across features.

### 6.2 Data Cleaning

Standard techniques were applied to clean the dataset, including case-wise deletion of incomplete observations and imputation of missing values. In some cases, missing values were filled using category-wise or range-based means, while in others, simple linear regression was used for predictive imputation. Moreover, advanced multiple imputation methods such as MICE (Multiple Imputation by Chained Equations) were used to preserve the inherent variability of the data. Considering that outliers can lead to overfitting (e.g., decision trees being overly sensitive to extreme values) and that missing data can introduce bias if left untreated, these cleaning techniques helped reduce the risk of statistical errors. Additionally, standard numerical transformations (scaling, encoding) ensured compatibility with the regression models.

### 6.3 Spatio-Temporal Aggregation

Data were aggregated along spatial and temporal dimensions to highlight underlying trends. For instance, average electric range values were computed per year and per geographic region in the dataset. The results revealed an increase in average range over time, corresponding to improvements in battery technology, as well as significant regional variations (e.g., due to differing charging infrastructure). These aggregated trends provide additional context for the predictive model, though they were not directly used as input variables for regression models.

### 6.4 Model Construction and Performance

Three regression models were trained for electric range prediction: *k-Nearest Neighbors* (KNN), *Decision Tree Regressor*, and *Random Forest Regressor*. The decision tree is a hierarchical model capable of capturing complex nonlinear relationships through recursive data splitting based on attributes. Random Forest is a collection of decision trees trained on bootstrap samples, where the final result is the average of individual tree predictions. This ensemble learning approach reduces variance and improves accuracy. The KNN model predicts by averaging the target values of the $k$ nearest neighbors of a data point without constructing an explicit model.

### 6.5 Model Validation

Model performance was evaluated using $k$-fold cross-validation with $k = 5$. Using 5-fold CV, average regression scores were computed: coefficient of determination ($R^2$), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) across the five subsamples. Cross-validation is a standard resampling technique used to estimate model performance on unseen data.

Table 3 clearly indicates that the *Random Forest Regressor* achieved the best evaluation metrics (highest $R^2$, lowest RMSE and MAE) compared to the other

**Table 3.** Average regression model performance (5-fold cross-validation)

| Model | $\mathbf{R^2}$ | RMSE (km) | MAE (km) |
|---|---|---|---|
| Random Forest | 0.85 | 14.5 | 10.2 |
| Decision Tree | 0.74 | 18.3 | 13.5 |
| KNN | 0.62 | 22.7 | 17.8 |

**Fig. 1.** Illustrative comparison graph of mean $R^2$ scores obtained by each model in cross-validation. The superiority of Random Forest is visible (highest mean $R^2$).

models. Meanwhile, the KNN model showed the weakest performance, consistent with expectations that neighborhood-based methods may be less effective in high-dimensional feature spaces. The decision tree achieved intermediate results.

### 6.6 Model Application on a New Instance

To verify the generalization capacity and practical usefulness of the best-performing model (based on the minimum MAE), it was applied to a simulated new instance representing a recently registered electric vehicle. The input data provided to the model were:

- **Postal Code**: 99169
- **Model Year**: 2015
- **Estimated Electric Range**: 144 miles
- **Base MSRP**: 0 (unavailable or unrecorded value)
- **DOL Vehicle ID**: 148979698
- **2020 Census Tract**: 53001950100

After complete training and model selection (specifically the optimized `Random ForestRegressor`), the prediction for this instance returned a numeric value representing the anticipated behavior of such a registration under current conditions. The result was realistic and aligned with trends previously observed in the historical data.

It is worth noting that the `Base MSRP` value was unavailable (set to 0), which tested the model's resilience to partially missing features. Despite this, the model generated a coherent prediction due to its internal feature weighting mechanism and the influence of multiple relevant traits.

This final step validates the practical applicability of the constructed model and demonstrates its potential use in real-world scenarios where not all input data may be fully available.

### 6.7 Sensitivity Analysis (Future Extensions)

It is noted that sensitivity analysis was not included in the current implementation but represents a potential future extension. This would involve assessing

how small variations in input values (e.g., environmental or load parameters) affect the electric range predicted by the model. Such analysis would complement the understanding of model behavior and highlight its robustness.

### 6.8 Interpretation and Reporting of Results

The results confirm initial expectations that ensemble models (Random Forest) are better suited to capture complex relationships between variables. This observation is also supported by the feature importance analysis, which shows that the most influential factors are those previously identified as strongly correlated with electric range (see the feature selection section). Indeed, the trees in the random forest collectively "vote" on the importance of each variable by reducing impurity error. Thus, parameters such as battery capacity, vehicle net weight, and route type were found to be highly important, forming the basis on which the model generated accurate predictions.

In conclusion, the models partially validated the initial hypotheses: Random Forest proved to be the most performant, given its capacity to model nonlinear relationships between features, as highlighted in the literature (variance reduction via bootstrapping). The results suggest that while simpler models (KNN, decision tree) provide a useful benchmark, the complexity and variability of electric vehicle data are better addressed through ensemble methods. Additionally, trends identified during EDA (strong correlations) were reflected in the final results. We note that although sensitivity analysis was not implemented, it remains a natural avenue for future investigation to enhance predictive robustness and understanding of the EV system as a whole.

## 7 Conclusion

The present study aimed to carry out an exploratory data analysis (EDA) and rigorous data cleaning to identify patterns and eliminate inconsistencies. Additionally, temporal and geographic aggregations were performed to capture regional and temporal trends. Based on the prepared dataset, we constructed and validated predictive models using Random Forest, Decision Tree, and K-Nearest Neighbors, aiming to obtain accurate predictions of electric vehicle range. Subsequently, the focus was placed on interpreting the results by analyzing feature importance and correlations identified within the dataset. These phases were largely conducted in accordance with the initial plan: the exploratory analysis highlighted key relationships (such as between price and range), and the data cleaning stage improved the coherence of the dataset. The aggregations revealed the temporal and spatial dynamics of EV adoption, and the predictive models were trained and evaluated in alignment with the project objectives.

Comparing the performance of the three machine learning models trained, Random Forest Regressor recorded the best predictive results. On the test set, the RF model achieved the lowest mean absolute error (MAE $\approx 29.7$) and the highest coefficient of determination ($R^2 \approx 0.73$), slightly outperforming Decision

Tree (MAE $\approx$ 29.9, $R^2 \approx 0.72$) and significantly outperforming KNN (MAE $\approx$ 30.7, $R^2 \approx 0.69$). All models showed relatively high $R^2$ values, indicating good adaptation to the data. Based on this comparison, Random Forest was chosen as the final model due to its superior accuracy and its ability to capture complex nonlinear relationships among variables.

The Random Forest Regressor thus proved suitable for the task of predicting EV range. The ensemble of decision trees provides robustness against data variability and reduces the risk of overfitting compared to a single tree. Moreover, RF allows for the estimation of feature importance, facilitating model interpretability. In our case, factors such as base price (MSRP) and model year were identified as the most relevant for predictions, underscoring their essential influence on electric range. This information confirms that investments in technology (reflected by higher prices) are correlated with improved range performance and can guide strategies for expanding the electric vehicle market.

However, some limitations of the study must also be noted. Calendar-based predictions for the year 2024 were not performed, as the model was not calibrated on future values. Additionally, the analysis did not include an explicit assessment of model overfitting/underfitting, which would have strengthened the evaluation of generalization capability. The data used are historical (up to 2022) and may not reflect subsequent developments, thus the long-term applicability of the predictions remains uncertain. Considered factors were treated as static, without incorporating technological or policy changes that could alter the system. Nonetheless, the constructed system offers significant practical insights: the feature analysis extracted relevant insights (e.g., confirming the direct relationship between vehicle price and electric range, as well as the increase in average range over the years). These conclusions can inform decision-makers in planning charging infrastructure and defining incentives for electric vehicles.

The research outlook is promising. Future development could involve integrating external datasets that reflect public policies or technological innovations in battery manufacturing to improve long-term forecasting. Another valuable direction would be the refinement of geographic aggregations to metropolitan or county-level granularity, thus capturing regional variability more accurately. Sequential modeling, such as time series analysis, may further enhance the forecasting by accounting for trend and seasonality components, which are critical in calendar-based predictions, including for 2024. Finally, regularly updating the models with new data and recalibrating their parameters would ensure the predictions remain valid over time.

Overall, the conclusions demonstrate the potential of such a predictive decision-support system. Nonetheless, achieving maximum performance will require continued expansion of the dataset and the adoption of additional methodological advancements.

## Acknowledgments

This study was conducted as part of the Intelligent Systems course project, Faculty of Engineering: Home - Technical University of Cluj-Napoca, Baia Mare, also extend my sincere appreciation to Willian Oliveira Gibin for making the 'Electric Vehicle Population Data' publicly available on the Kaggle platform[cite: 24], which served as the foundation for this research.

## References

1. Breiman, L., *Random Forests*, Machine Learning, vol. 45, pp. 5–32, 2001. ISSN 0885-6125.
2. Quinlan, J.R., *Induction of decision trees*, Machine Learning, vol. 1(1), pp. 81–106, 1986. ISSN 0885-6125.
3. Cover, T.M., Hart, P.E., *Nearest neighbor pattern classification*, IEEE Trans. Inform. Theory, vol. 13(1), pp. 21–27, 1967. ISSN 0018-9448.
4. Devarasan, E. et al., *Advancing sustainable mobility in India with electric vehicles: market trends and machine learning insights*, Front. Energy Res., vol. 13, 2025. ISSN 2296-598X.
5. Coffman, M. et al., *Electric vehicles revisited: a review of factors that affect adoption*, Transport Reviews, vol. 37(1), pp. 79–93, 2017. ISSN 1464-5327.
6. Hardman, S. et al., *A review of consumer preferences of and interactions with electric vehicle charging infrastructure*, Transp. Res. Part D, vol. 62, pp. 508–523, 2018. ISSN 1361-9209.
7. Axsen, J. et al., *Preference and lifestyle heterogeneity among potential plug-in electric vehicle buyers*, Energy Economics, vol. 50, pp. 190–201, 2015. ISSN 0140-9883.
8. Zaino, R. et al., *Electric Vehicle Adoption: A systematic review*, World Electr. Veh. J., vol. 15(8), 2024. ISSN 2032-6653.
9. Guyon, I., Elisseeff, A., *An introduction to variable and feature selection*, J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003. ISSN 1532-4435.
10. Willmott, C.J., Matsuura, K., *Advantages of MAE over RMSE*, Climate Res., vol. 30, pp. 79–82, 2005. ISSN 0936-577X.
11. Yeh, J.-Y., Wang, Y.-T., *Prediction model for electric vehicle sales using ML*, J. Glob. Inf. Manage., vol. 31(1), 2023. ISSN 1533-7995.