**Forecasting the Population of Electric Vehicle Users in Uttarakhand Using Regression Models**

**MIHIKA**

**B.Tech student in Computer Science Engineering specializing in AI/ML at UPES Dehradun**

**Abstract**

The growing adoption of electric vehicles (EVs) plays a critical role in addressing sustainability challenges and transitioning toward cleaner transportation systems. Accurate prediction of EV usage is essential for planning and optimizing infrastructure in regions like Uttarakhand, India. This study employs machine learning techniques to estimate the population using EVs based on key parameters such as population density, altitude, income, grid availability, and EV generation. Regression algorithms, including Linear Regression, K-Nearest Neighbour Regression, Decision Tree, SVR (Support vector Regression) and Random Forest, were applied to predict EV adoption. Data preprocessing techniques were utilized to ensure model reliability, and performance was evaluated using metrics such as $R^2$ and Mean Squared Error. The results provide actionable insights into the potential EV adoption in Uttarakhand, contributing to informed decision-making for grid and infrastructure development.

**Introduction**

The rising global emphasis on sustainable transportation has positioned electric vehicles (EVs) as a pivotal solution to reduce environmental impacts and reliance on fossil fuels. In a region like Uttarakhand, India, where geographic and economic factors play a significant role in technology adoption, understanding the patterns of EV usage is critical for informed planning and policy-making. This study aims to predict the EV population in Uttarakhand using advanced machine learning techniques. By analysing key factors such as population density, altitude, income, grid availability, and EV generation, the project leverages regression models—including Linear Regression, KNN Regression, Decision Tree, SVR and Random Forest—to generate accurate predictions. Inspired by a similar study, this work seeks to provide data-driven insights to guide EV adoption strategies and infrastructure development in the region.

**Literature Review**

The transition to electric vehicles (EVs) has become a critical focus due to their potential in reducing environmental impacts and supporting sustainable transportation. Several studies have examined the effects of EV penetration on energy systems, focusing on charging behaviours, grid impacts, and the integration of EVs into the power infrastructure. For instance, some research highlights the challenges posed by uncontrolled EV charging, which can significantly increase peak demand and destabilize the grid. Studies also propose various machine learning models, such as CNN-BiLSTM and hybrid optimization techniques, to predict EV charging patterns and their impact on grid performance.

However, many of these models rely on limited historical data, often overlooking socio-economic and geographical factors like income, population density, altitude, and grid availability, which also play a crucial role in predicting EV adoption. Some studies have explored the influence of these parameters on EV penetration, but they often focus on a narrower set of variables or geographical areas.

This study differentiates itself by incorporating a wider range of socio-economic factors in predicting EV adoption, specifically focusing on Uttarakhand, India. By applying machine learning algorithms such as Random Forest, K-Nearest Neighbour,SVR and Decision Trees, this work provides a comprehensive approach to estimating EV population, considering the unique characteristics of the region.

**Methodology**

**1. Dataset Overview**

The dataset for this study contains information on various socio-economic and geographical factors influencing electric vehicle (EV) adoption in Uttarakhand, India. Key features include population density, income levels, grid availability, altitude, and existing EV generation data. This data was then processed to ensure completeness and reliability for predictive modeling.

**2. Data Processing**

- **Handling Missing Values**:
  Missing values were identified and checked using the command df.isnull().sum(). Ensuring that no data was missing helped maintain the integrity of the dataset, preventing any distortion of analysis and model performance.

- **Removing Duplicates**:
  Duplicate entries were checked using df.duplicated().sum(), and unnecessary duplicates were removed with df.drop_duplicates().reset_index(drop=True). This step was essential to avoid redundancy and bias in model training.

- **Outlier Detection and Handling**:
  Outliers were detected using the Interquartile Range (IQR) method. Extreme values were capped to the upper and lower bounds calculated by Q1 - 1.5 * IQR and Q3 + 1.5 * IQR. This ensured that outliers did not have a disproportionate impact on the model's performance.

- **Feature Scaling**:
  Standard scaling was applied to normalize the "ev_generation" feature, ensuring that the data had a mean of 0 and a standard deviation of 1. This was achieved using StandardScaler from sklearn. Scaling helped to balance the influence of features with different units or ranges.

- **Feature Engineering**:
  New interaction features were created to capture the relationships between key variables, such as:

  - population_grid_interaction = population_density * grid_availability
  - income_grid_interaction = income * grid_availability These new features were expected to provide deeper insights into the socio-economic impact on EV adoption.

- **Data Visualization**:
  Visualizations were employed to better understand the distribution of the features and their relationships. Histograms with kernel density estimation (KDE) were used to inspect distributions, and a correlation heatmap was generated to evaluate the interactions between different features, aiding in further preprocessing and feature selection.

**3. Model Selection**

Several machine learning algorithms were chosen for predicting the population of electric vehicles, including:

- **Linear Regression**
- **K-Nearest Neighbour Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Support Vector Regression**

**Model Selection**

To predict the population of electric vehicle (EV) users in Uttarakhand, four machine learning algorithms were employed: **Linear Regression**, **K-Nearest Neighbours Regression (KNN)**, **Decision Tree Regression**, **SVR** and **Random Forest Regression**. These models were selected due to their suitability for handling continuous target variables and their ability to capture the complex relationships between socio-economic and geographical factors influencing EV adoption. Here's a deeper look into each model and its role in predicting the EV population:

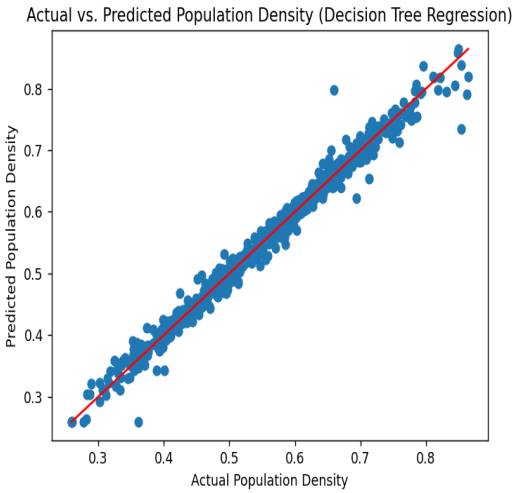
- **Linear Regression**:
  Linear regression is a simple yet effective model that assumes a linear relationship between the target variable (EV population) and the independent features (socio-economic and geographical factors). Although straightforward, it works well when the relationship between the features and target variable is approximately linear. In this case, it provided a baseline prediction model to understand the general trend in EV adoption based on the dataset.
- **K-Nearest Neighbour Regression (KNN)**:
  KNN is a non-parametric method that makes predictions based on the average value of the nearest data points (neighbour). By considering the local patterns in the data, KNN can effectively model non-linear relationships. This was particularly useful in predicting the EV population as it captured localized variations in socio-economic conditions that might not be apparent with linear models.
- **Decision Tree Regression**:
  A decision tree splits the dataset into branches based on feature values, allowing for complex, non-linear relationships between the features and target variable. It works by making decisions at each node, ultimately predicting the target by traversing from the root to the leaf. Decision trees are particularly beneficial when the relationship between the features and the target variable is hierarchical or based on thresholds. In this project, the decision tree helped to segment the data based on critical socio-economic factors like income or grid availability, thus capturing intricate patterns in EV adoption across different regions of Uttarakhand.

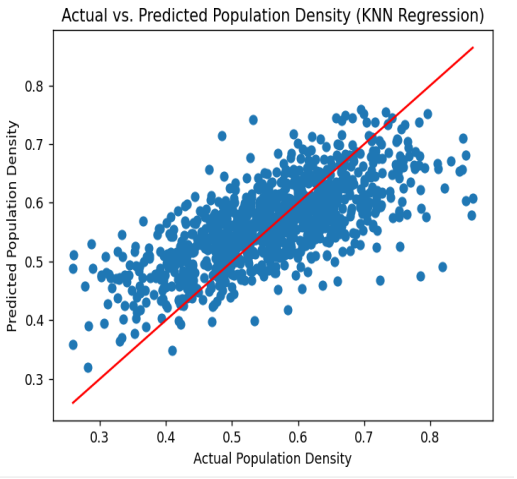
- **Random Forest Regression**:
  Random Forest is an ensemble method that constructs multiple decision trees and aggregates their predictions. It enhances prediction accuracy by reducing overfitting and handling variability in the data better than a single decision tree. Since predicting EV adoption involves multiple variables with potential interdependencies, Random Forest is especially powerful as it accounts for complex interactions and provides robust predictions even when the data is noisy or imbalanced. By averaging the predictions from many decision trees, Random Forest helped to improve the accuracy and reliability of the EV population predictions.

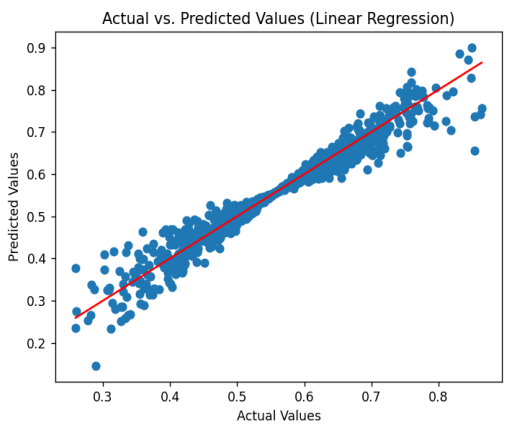
- **Support Vector Regression**

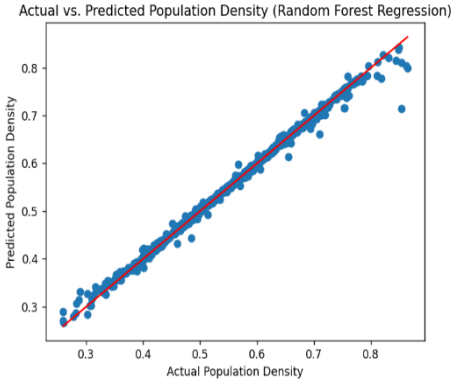
  Support Vector Regression (SVR) is a machine learning algorithm designed for regression tasks. It uses a margin of tolerance, called the ϵ\epsilonϵ-insensitive zone, to focus on significant errors while ignoring minor deviations, making it robust to noise in the data.SVR is particularly effective for capturing non-linear relationships by applying kernel functions like the Radial Basis Function (RBF). This makes it suitable for predicting EV adoption, where complex interactions exist between variables such as income, grid availability, and altitude. By tuning parameters like C, ϵ, and γ, SVR can provide accurate and reliable predictions for continuous variables like EV population density.

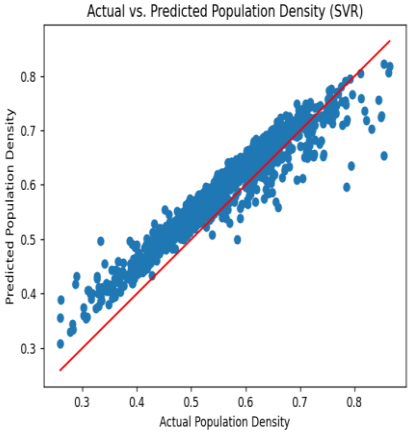
Each model was used to predict the population of EV users based on a variety of socio-economic and geographic factors. These algorithms were particularly useful in capturing the influence of features like population density, income, grid availability, and altitude on EV adoption. The combined approach of using multiple models allowed for a comparative analysis of performance, ultimately leading to the selection of the most accurate model.






---

**5. Model Evaluation**

The performance of each model was evaluated using two key metrics: **R-squared ($R^2$)** and **Mean Squared Error (MSE)**. These metrics were chosen to assess the accuracy and reliability of the models in predicting the population of electric vehicle (EV) users in Uttarakhand. $R^2$ measures the proportion of variance in the dependent variable that is explained by the independent variables, while MSE quantifies the average squared difference between the observed and predicted values.

**Evaluation Metrics:**

- **R-squared ($R^2$)**: A higher $R^2$ indicates that the model explains a large proportion of the variance in the target variable, with a value closer to 1 being preferable.
- **Formula:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- yi = Actual values (ground truth)
- y^i= Predicted values (model predictions)
- $\bar{y}$ = Mean of the actual values

- n = Number of data points

- **Mean Squared Error (MSE)**: A lower MSE indicates better performance, as it shows how close the model's predictions are to the actual values.

- **Formula:**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - yi = Actual values (ground truth)
  - y^i= Predicted values (model predictions)
  - n = Number of data points

**Evaluation Results:**

The models were compared based on their $R^2$ and MSE values. The results are as follows:

Here's the updated table with the **Mean Squared Error** and **R-squared** values for **SVR** added:

| **Model** | **R-squared** | **Mean Squared Error** |
|---|---|---|
| Random Forest Regression | 0.9943 | 7.05e-05 |
| Decision Tree Regression | 0.9846 | 0.00019 |
| Linear Regression | 0.9435 | 0.0007 |
| KNN Regression | 0.4749 | 0.0065 |
| SVR | 0.8197 | 0.0022 |

- **Random Forest Regression** achieved the highest performance, with an **$R^2$ of 0.9943** and a **very low MSE of 7.05e-05**. This suggests that the model explains 99.43% of the variance in the EV population, making it highly reliable for prediction.
- **Support Vector Regression (SVR)** demonstrated an $R^2$ of 0.8197 and a Mean Squared Error (MSE) of 0.0022. While not as high-performing as Random Forest or Decision Tree Regression, it still showed a decent ability to capture the patterns in the EV population, explaining around 82% of the variance. This makes SVR a reliable model, though it may not perform as robustly as ensemble methods in handling complex data relationships.
- **Decision Tree Regression** also performed well, with an **$R^2$ of 0.9846** and an **MSE of 0.00019**. While slightly less accurate than Random Forest, it still demonstrated strong predictive ability.

- **Linear Regression** had an **$R^2$ of 0.9435** and an **MSE of 0.0007**, indicating that it explains a significant portion of the variance but has higher prediction errors compared to the previous two models.
- **KNN Regression** showed the weakest performance with an **$R^2$ of 0.4749** and an **MSE of 0.0065**. This suggests that the KNN model struggled to accurately predict the EV population, likely due to its sensitivity to local data variations and lack of model generalization.

**Conclusion:**

Based on the $R^2$ and MSE values, the **Random Forest Regression** model was the most effective in predicting the number of EV users in Uttarakhand, offering high accuracy and low error. While not as high-performing as Random Forest or Decision Tree Regression, **SVR** still showed a decent ability to capture the patterns in the EV population, explaining around 82% of the variance. The **Decision Tree Regression** model also performed well, while the **Linear Regression** model showed reasonable but less precise results. The **KNN Regression** model was the least effective, highlighting its limitations for this type of prediction task.

These evaluation results validate the choice of Random Forest Regression as the most suitable model for estimating EV penetration in the region, providing valuable insights for energy management and infrastructure planning.

**Reference:**

**https://www.researchgate.net/publication/385353178_Performance_analysis_of_machine_learning_algorithms_for_estimation_of_EV_penetration**