

Executive Summary

- This project applies machine learning techniques to predict vehicle fuel economy using a dataset of over 40,000 vehicles from the Environmental Protection Agency (EPA). Three regression models: Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR) were implemented and evaluated. All models achieved strong predictive performance ($R^2 > 0.87$), with Random Forest performing best ($R^2 \approx 0.94$) and producing the lowest prediction errors. The results demonstrate that machine learning can effectively capture complex relationships between vehicle characteristics and fuel efficiency, providing a reliable tool for data-driven decision making in engineering and automotive applications.
- The goal of this project is to develop and compare machine learning models to accurately predict vehicle fuel economy (MPG) and support data-driven engineering decisions.

Background

- Fuel Economy is the distance a vehicle can travel per unit of fuel. Fuel Economy is a key performance indicator for consumers, manufacturers, and regulators alike. It influences consumer purchasing decisions, impacts annual fuel costs, and plays a central role in environmental policy and emissions regulations.
- The Vehicle Fuel Economy dataset, published by the **California Air Resources Board** and available through the U.S. Data.gov catalog, contains vehicle fuel economy and emissions statistics alongside engine, drivetrain, fuel type, and costing information [1]. Fuel economy figures are derived from standardized EPA testing and sourced from fueleconomy.gov, covering a broad range of vehicle makes, models, and years.
- With the increasing availability of large automotive datasets, machine learning models, including **KNN**, **RF**, and **SVR**, provide a **powerful** alternative for predicting vehicle fuel economy using measurable features such as engine size, number of cylinders, displacements, and other technical specifications. These data-driven methods can uncover complex non-linear patterns in the data that traditional linear models might miss, offering more flexible models that can adjust to new vehicle data and handle large automotive datasets efficiently.

Experimental Setup

Dataset Description:

- The dataset includes ~ **40,700** EPA vehicles dating from **1984–2019**. Source: CARB
- Key features** - engine size, fuel type, vehicle characteristics.
- Target variable** - combined fuel economy (MPG) - (comb08), to assess efficiency.

Data Cleaning:

- Cleaned data to improve quality & accuracy.
- Selected relevant predictors for efficiency.
- Ensured reliable, meaningful training data.

Preprocessing (Encoding):

- Encoded categorical variables (e.g. fuel type, transmission) to convert them numerically for machine learning models.

Preprocessing (Scaling):

- Feature scaling was applied to ensure consistent variable ranges
- This is a critical step for modeling KNN and SVR, which are both distance-based algorithms.

Modeling Approach: **Three regression models** were implemented, (Figure 1).

- Random Forest (RF):** uses decision trees to catch **complex non-linear** relationships
- KNN:** predicts outcomes based on **similarity between** the k closest **observations**.
- SVR:** uses **optimized hyperplanes** to model non-linear relationships

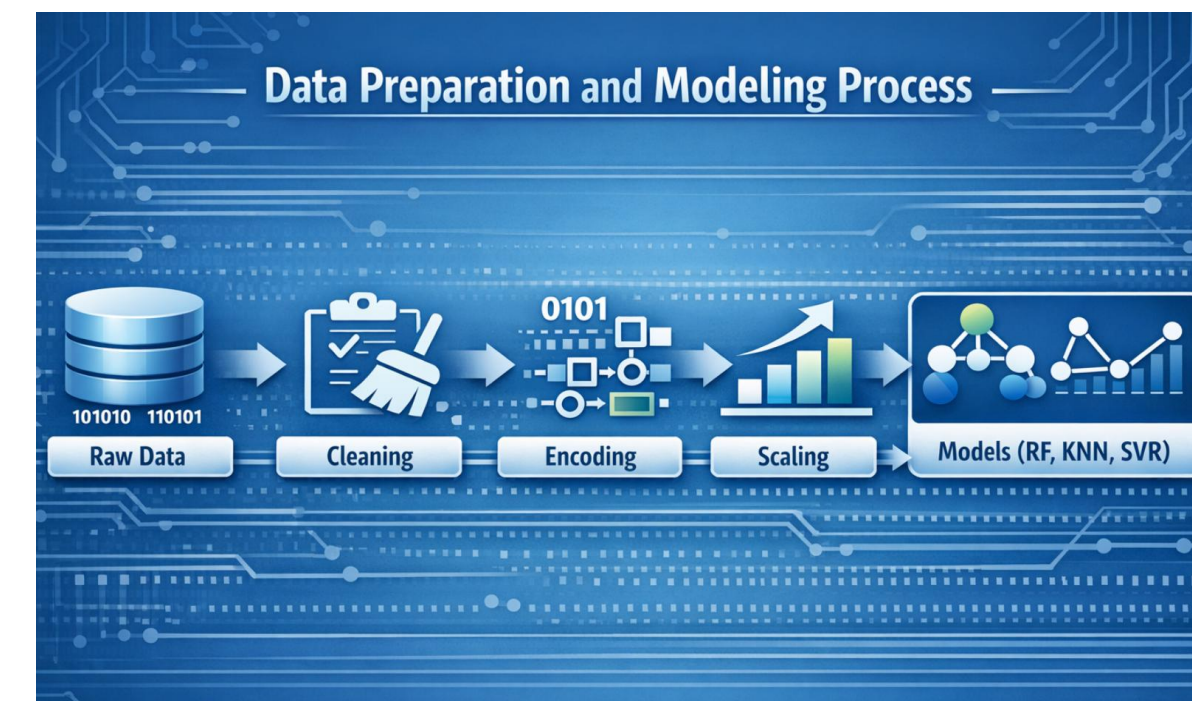


Figure 1. Data processing & modeling pipeline

Experimental Test Plan

Model Selection

The selection of **Random Forest**, **KNN**, and **SVR** was driven by the need to compare different mathematical approaches within regression analysis. By testing these three distinct algorithms, we ensure robust prediction for the continuous data set.

Random Forest Regression:

Random Forest predicts combined fuel economy by building many **decision trees** and averaging their results. This helps the model capture complex relationships between vehicle characteristics and MPG that may not follow a simple linear trend. Chosen since fuel economy depends on many factors, and RF handles better than the other models [2].

KNN:

KNN was applied to the dataset to see if a vehicle's MPG can be accurately predicted by comparing it to historically similar models. By using local interpolation, the model finds "**neighboring**" vehicles with matching specifications, such as engine size and class, to estimate fuel efficiency [4].

SVR:

SVR predicts combined fuel economy by fitting a function that captures the relationship between vehicle characteristics and MPG within an acceptable error margin. The model was tuned using "**GridSearch**" to optimize parameters such as kernel type and improve predictive performance [3].

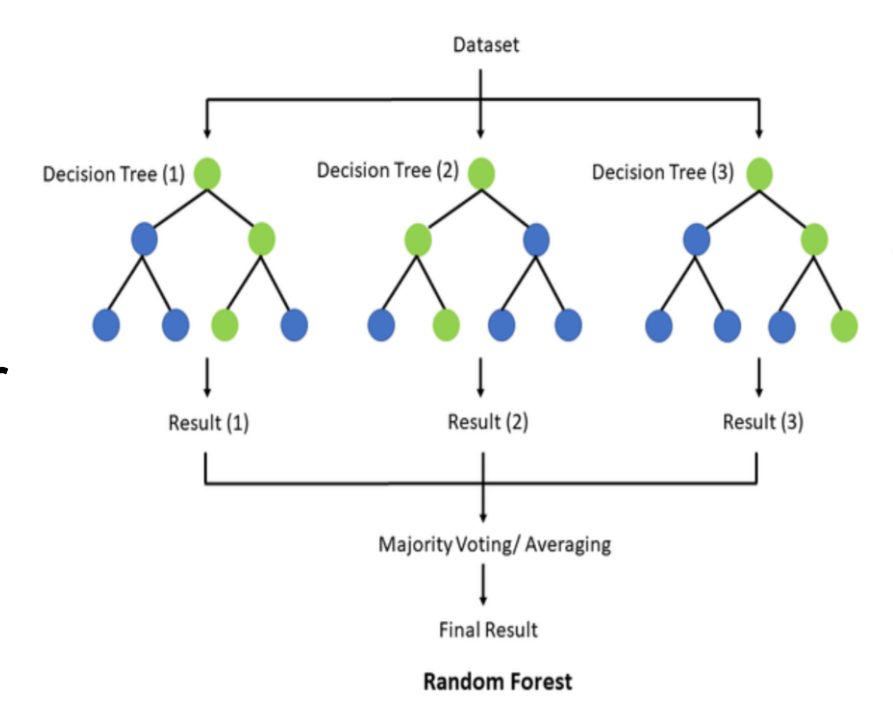


Figure 2. Random Forest model for nonlinear prediction [8].

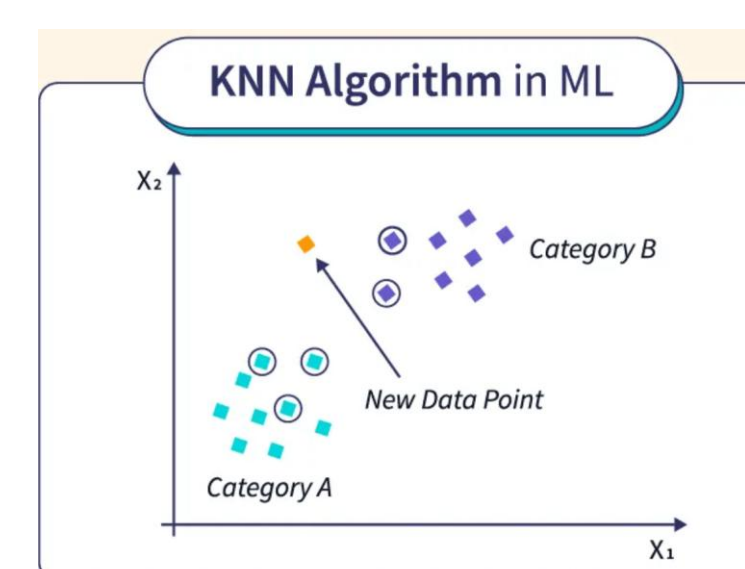


Figure 3. KNN model based on nearest neighbors [9].

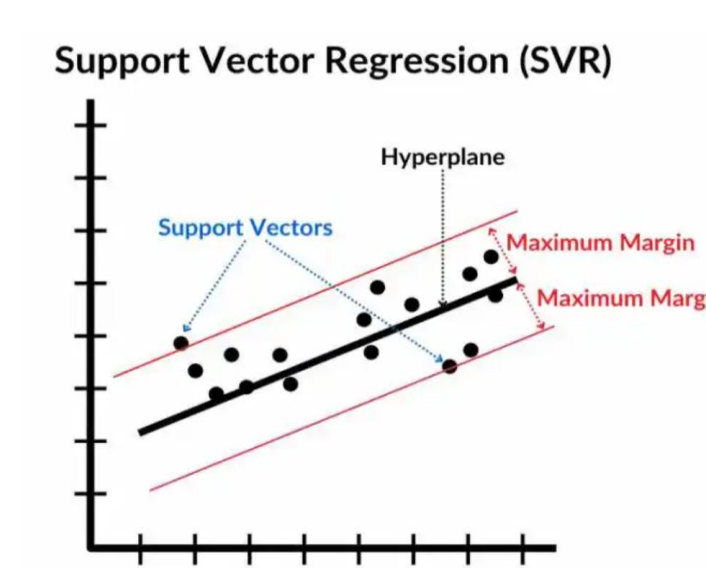


Figure 4. SVR model for nonlinear regression [10].

Experimental Results

Final Model Performance:

Table 1. Final Model Performance Comparison

Model	R^2	Adj_ R^2	RMSE	MAE
Random Forest (RF)	0.9366	0.9365	1.3385	0.7866
KNN	0.8963	0.8962	1.7111	1.1391
SVR	0.8712	0.8710	1.8733	1.1824

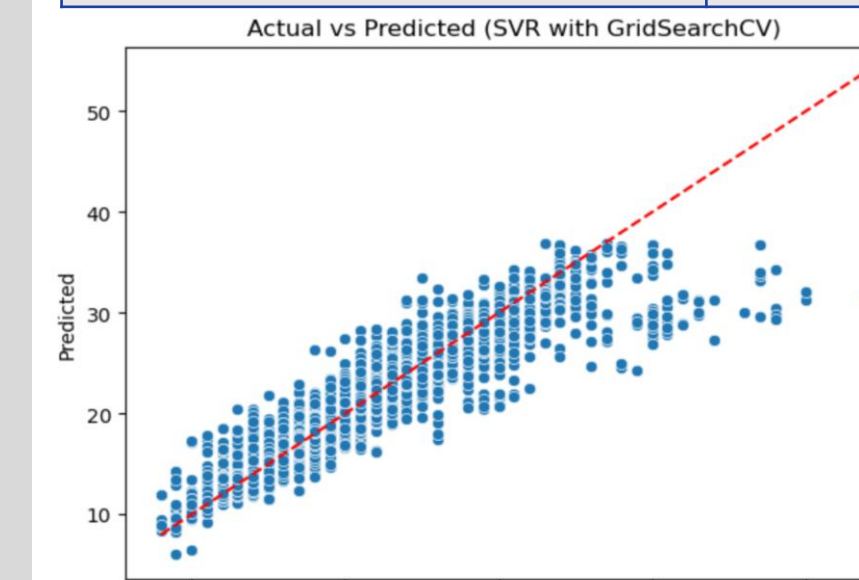


Figure 6. Predicted vs Actual (SVR)

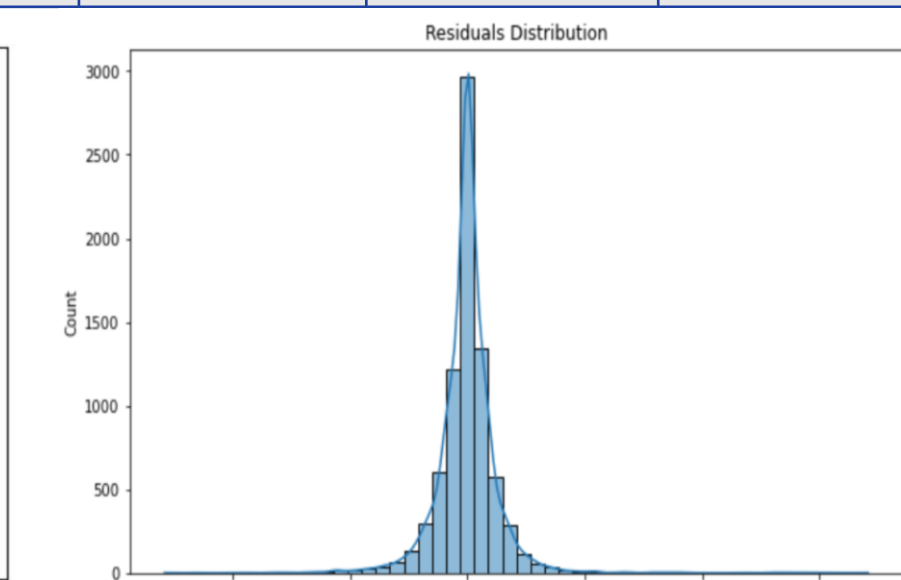


Figure 7. Residual distribution (RF)

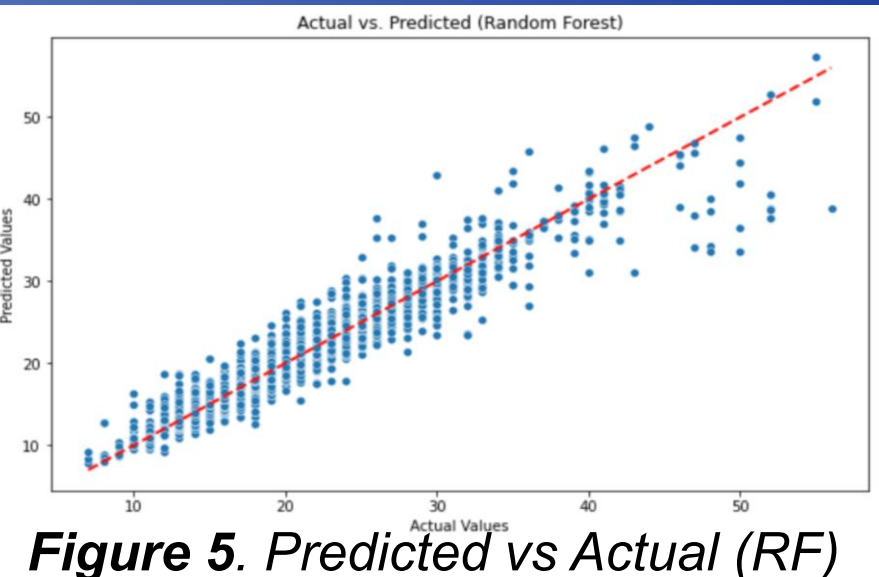


Figure 5. Predicted vs Actual (RF)

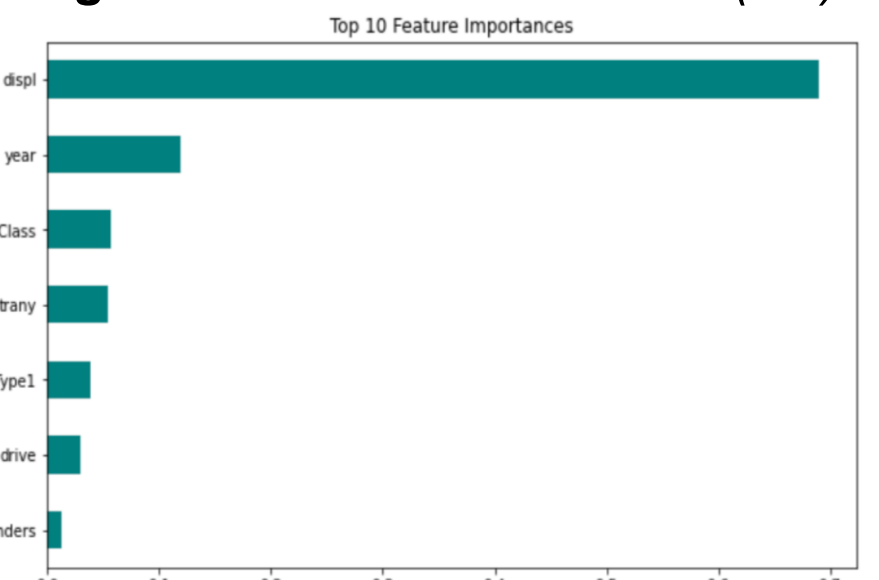


Figure 8. Feature Importance (RF)

- All models** showed **strong predictive performance** with ($R^2 > 0.87$).
- Random Forest (RF)** performed **best** ($R^2 = 0.9366$) with the **lowest errors**.
- KNN** and **SVR** were **solid** but **slightly less accurate** than RF.
- Predicted vs Actual** plots show **tight linear alignment**.
- Residuals center near zero**; **engine size (displ)** is the **top predictor**.
- Adjusted R^2** values closely **match R^2** , indicating **stable** models with minimal overfitting.

Conclusions

Highlights:

- Random Forest was the top-performing model** ($R^2 = 0.9366$, RMSE = 1.3385), outperforming both KNN and SVR in every metric, making it the most reliable choice for this regression task.
- Engine displacement (displ)** emerged as the **most influential predictor** of fuel economy, confirming that engine size plays a dominant role in determining a vehicle's combined MPG.

Future Work:

- Future research could incorporate **additional features** such as vehicle weight, aerodynamic drag, and hybrid/electric classifications to improve model accuracy beyond conventional combustion engines.
- Extending the dataset **beyond 2019** to include **more recent models** would allow the models to capture modern fuel efficiency trends driven by stricter emissions regulations and newer engine technologies.

References

- [1] California Air Resources Board, *Vehicle Fuel Economy Dataset*, 2024. Available: <https://catalog.data.gov/dataset/vehicle-fuel-economy>.
- [2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.
- [4] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [5] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
- [8] <https://interactivechaos.com/en/wiki/random-forest>
- [9] <https://www.dataexpertize.in/k-nearest-neighbor-algorithm-guide/>
- [10] <https://spotintelligence.com/2024/05/08/support-vector-regression-svr/>