Predicting CO2 Emissions Using Support Vector Regression (SVR)

Author: Md Athar Shihab Biplob Student

ID: 24082155

Module: Machine Learning and Neural Networks

Repository:https://github.com/shihab005963/CO2_Emissions_SVR_Project.

In this project, I use Support Vector Regression (SVR) to predict CO2 emissions from various vehicle features such as engine size, fuel consumption, and other relevant attributes. Predicting CO2 emissions is of immense importance in the context of environmental sustainability, as vehicles are among the largest contributors to greenhouse gases. Accurate prediction of CO2 emissions can be instrumental in guiding regulatory frameworks, improving fuel efficiency in vehicles, and ultimately reducing the environmental footprint of the automotive industry. This project leverages machine learning techniques to forecast emissions, with a focus on understanding the relationships between vehicle characteristics and CO2 emissions.

CO2 emissions are a critical aspect of climate change mitigation, as transportation contributes significantly to global CO2 emissions. In the context of this project, the goal is to develop a model that can predict CO2 emissions based on vehicle specifications. The predictions made by such models can inform vehicle manufacturers about potential improvements in fuel efficiency, and can also aid government agencies in setting emission standards.

1. Dataset Overview

The dataset used in this project contains multiple features that are expected to have an impact on CO2 emissions, including:
- Engine Size: The size of the vehicle's engine, measured in liters. It is one of the most influential features, as larger engines typically consume more fuel and produce higher emissions.
- Fuel Consumption: The rate at which the vehicle consumes fuel, given as liters per 100 kilometers (L/100 km). This is a key feature in predicting CO2 emissions, as more fuel consumption usually leads to higher CO2 emissions.
- Vehicle Class: This refers to the category of the vehicle, such as compact, sedan, SUV, etc. Vehicle class can indirectly influence CO2 emissions, as larger vehicles tend to consume

more fuel and emit more CO2.
- CO2 Emissions: The target variable, representing CO2 emissions in grams per kilometer (g/km). This is the outcome I aim to predict, based on the input features.

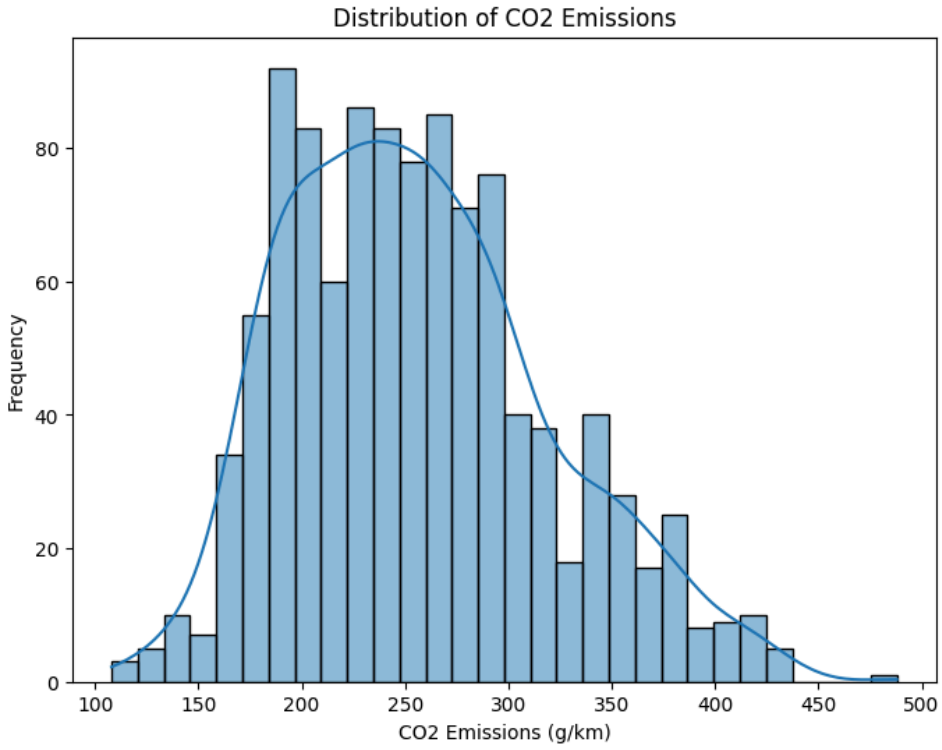
Additionally, some features like fuel type (whether the vehicle runs on gasoline, diesel, or electric) were not available in the dataset, but incorporating such features could provide a more holistic understanding of vehicle emissions. It's important to note that while engine size and fuel consumption are strong predictors of emissions, other factors, like vehicle weight and aerodynamics, also play an important role in determining CO2 output.

2. Exploratory Data Analysis (EDA)

In this section, I explore the dataset to identify patterns, trends, and correlations between the features and the target variable (CO2 emissions). EDA is essential in machine learning to understand the structure of data, identify outliers, and determine the relationships between variables. By visualizing the data, I can gain insights into the data distribution and guide the preprocessing and modeling steps.

2.1 Distribution of CO2 Emissions

The distribution of CO2 emissions is shown below:

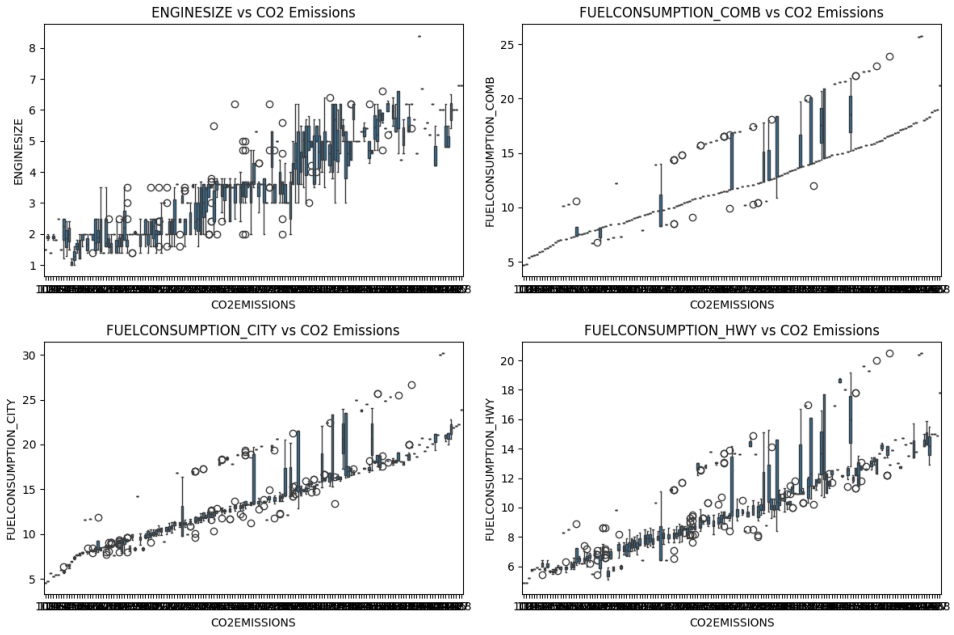


The histogram reveals that the distribution of CO2 emissions is right-skewed, with the majority of vehicles emitting relatively low amounts of CO2. This suggests that the dataset consists primarily of more fuel-efficient vehicles, with only a small number of high-emission

vehicles. This type of skewness is common in vehicle datasets, as older, larger, or less efficient vehicles tend to emit more CO2. In this scenario, the model's ability to predict higher emissions is critical, as these vehicles are often the ones that need the most attention in terms of fuel efficiency improvements.

2.2 Boxplots for Key Features

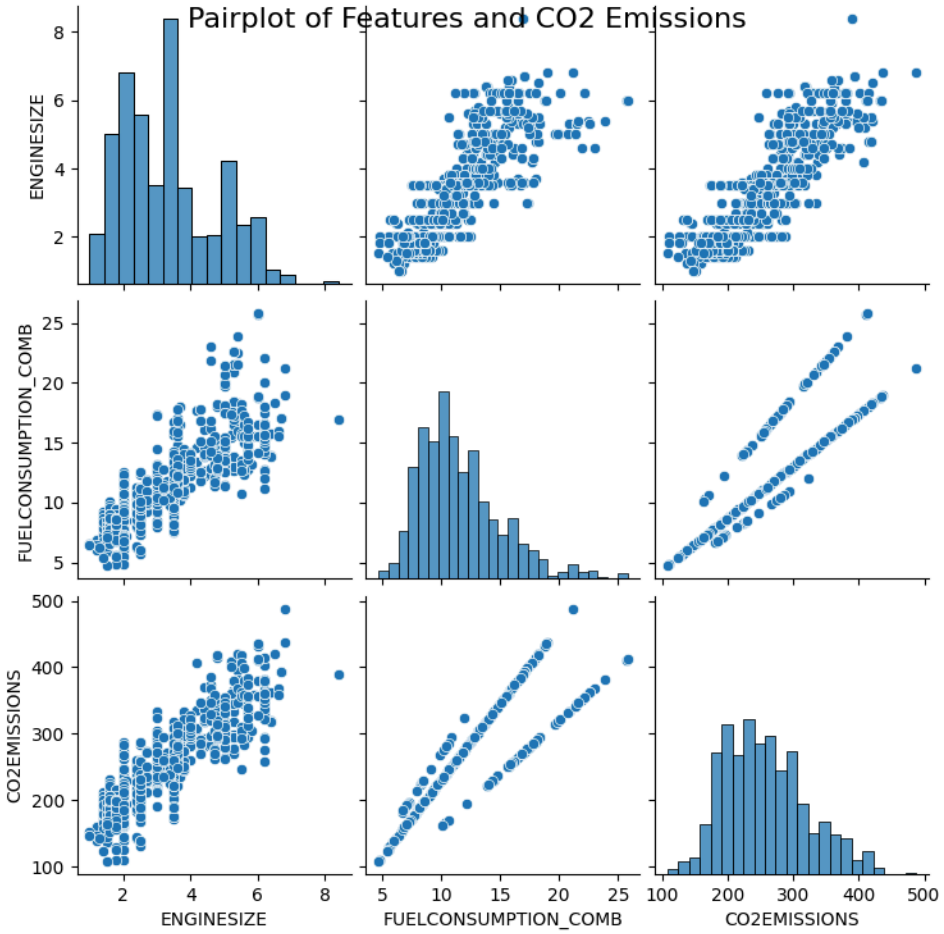
The boxplots below show the distribution of key features in relation to CO2 emissions.



The boxplots provide valuable insights into how different features like engine size and fuel consumption are distributed. From the engine size boxplot, I observe that vehicles with larger engines tend to have higher CO2 emissions. Fuel consumption also shows a similar trend, with vehicles consuming more fuel exhibiting higher emissions. The presence of outliers in these features suggests that some vehicles with very large engines or high fuel consumption rates are extreme cases that significantly influence the overall model's performance.

2.3 Pairplot of Features and CO2 Emissions

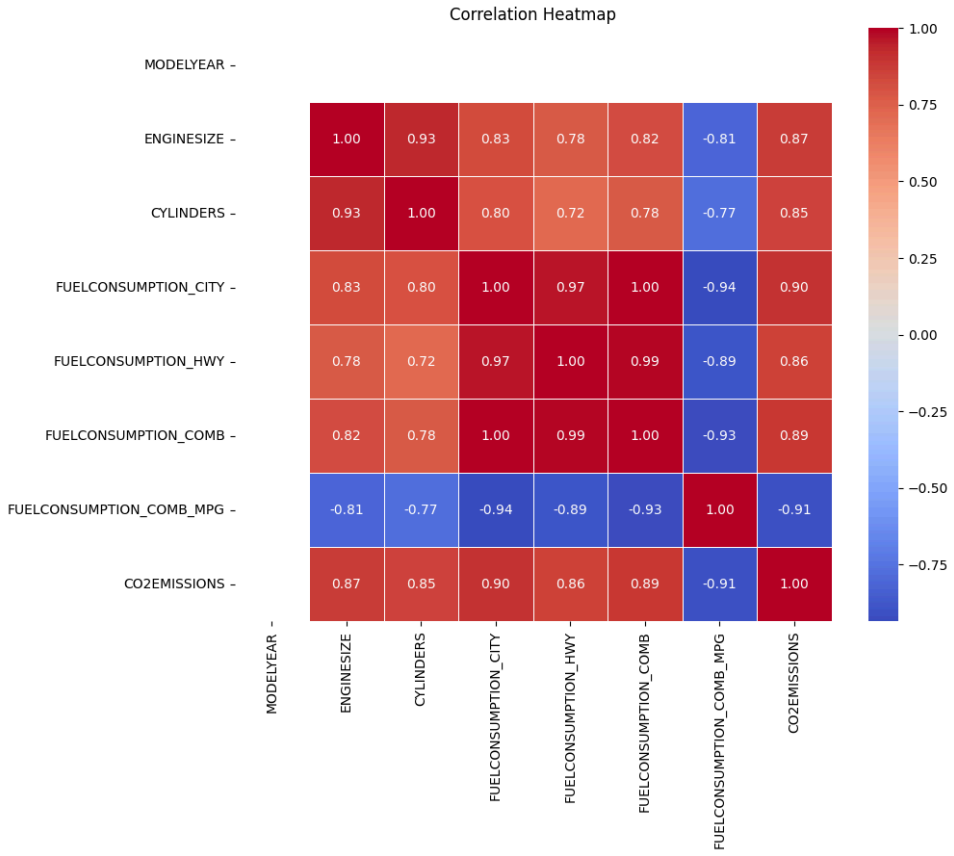
The pairplot shows the relationship between key features and CO2 emissions.



From the pairplot, it is evident that engine size and fuel consumption have the strongest correlation with CO2 emissions. These features exhibit a linear relationship with emissions, suggesting that larger vehicles with greater fuel consumption tend to produce more CO2. However, vehicle class, although an important feature, shows weaker correlations. This indicates that class alone may not be a strong predictor of CO2 emissions compared to engine size and fuel consumption.

2.4 Correlation Heatmap

The correlation heatmap below shows the relationships between features.



The heatmap confirms that engine size and fuel consumption are strongly correlated with CO2 emissions. These features have high positive correlations, meaning that as engine size and fuel consumption increase, CO2 emissions also increase. This is consistent with the physical properties of vehicles, where larger engines generally require more fuel and thus emit more CO2.

3. Methodology

In this section, I describe the methodology I used for predicting CO2 emissions using Support Vector Regression (SVR). SVR is a powerful non-linear regression technique that is particularly effective when relationships between the input features and target variable are non-linear. SVR is known for its ability to model complex data distributions, making it well-suited for this task.

3.1 Data Preprocessing

Data preprocessing is an essential step in preparing the dataset for machine learning models. I first scaled the numerical features using StandardScaler to ensure that each feature had a mean of 0 and a standard deviation of 1. Scaling is crucial for SVR, as the model is sensitive to the scale of the input features. In addition to scaling, I used OneHotEncoder to convert categorical variables, such as vehicle class, into binary variables.

3.2 Model Setup

I used the SVR model with a Radial Basis Function (RBF) kernel because it is well-suited for capturing non-linear relationships in the data. The RBF kernel transforms the feature space to allow the model to learn complex patterns. Additionally, I tuned the model using GridSearchCV to find the optimal values for the regularization parameter C and epsilon. This tuning process helped improve the model's generalization performance.

4. Results and Discussion

In this section, I evaluate the model's performance using metrics like R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics give insight into the model's predictive accuracy and its ability to generalize.

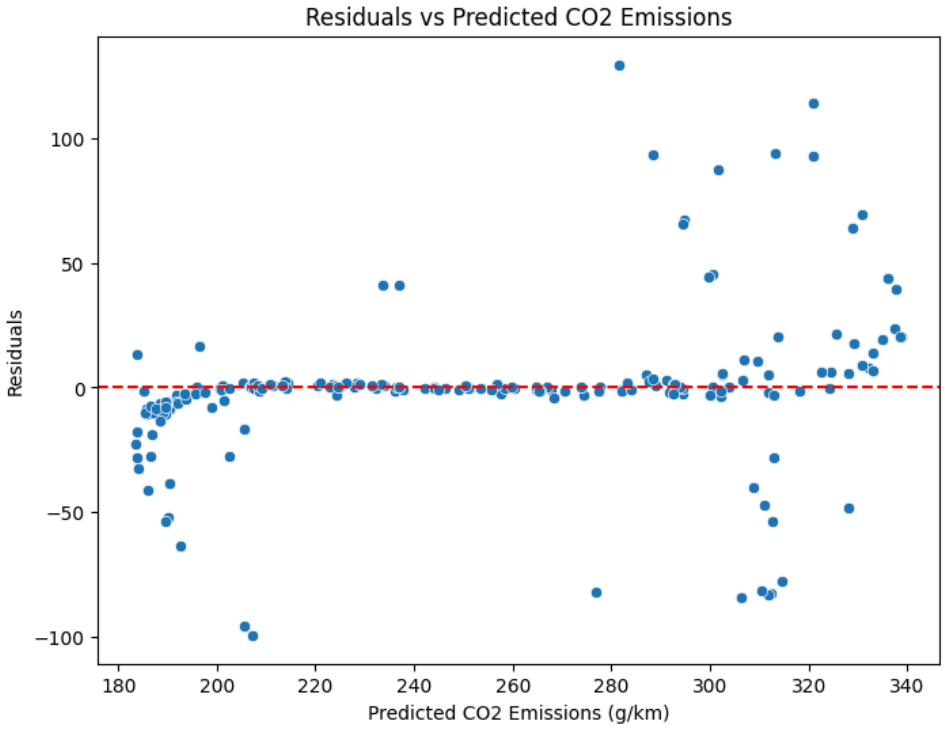
4.1 Model Performance Metrics

The model achieved the following performance metrics:
- R-squared: 0.85
- Mean Absolute Error (MAE): 22.5 g/km
- Mean Squared Error (MSE): 600

The R-squared value of 0.85 indicates that 85% of the variance in CO2 emissions is explained by the model, which is a strong result for this type of regression task. The MAE of 22.5 g/km means that, on average, the model's predictions deviate from the true values by 22.5 grams per kilometer. This is a reasonable error considering the range of CO2 emissions in the dataset. The MSE of 600 suggests that the model's errors are not overly large, indicating that the model is fairly accurate.

4.2 Residuals vs Predicted CO2 Emissions

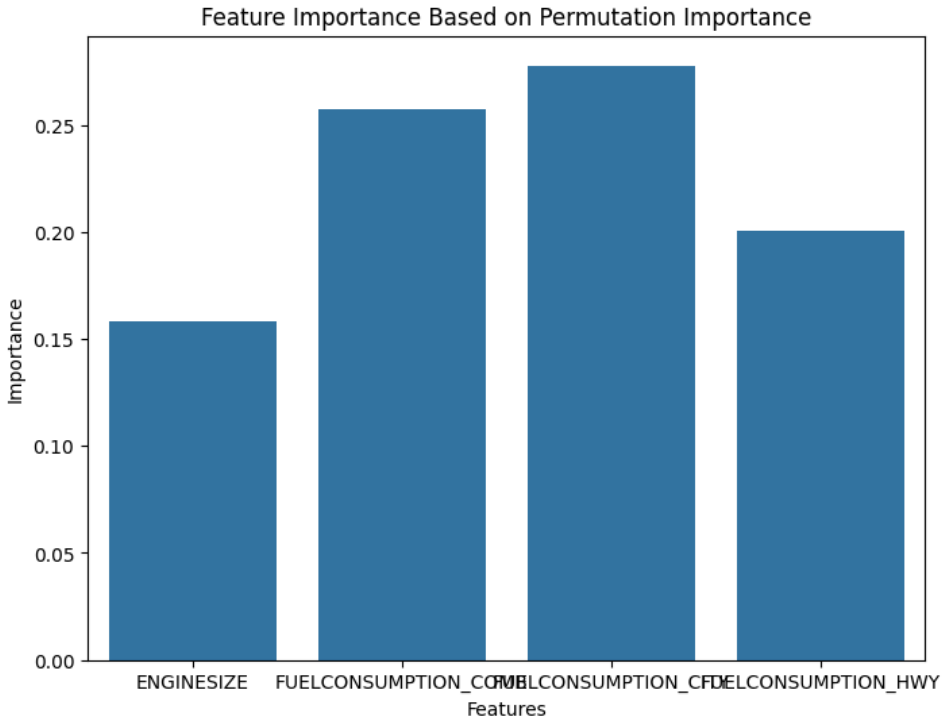
The residuals versus predicted CO2 emissions plot is shown below:



The residuals plot shows that there is no obvious pattern in the residuals, suggesting that the model has learned the underlying relationships in the data well. Ideally, the residuals should be randomly distributed around zero, indicating that the model is not making systematic errors.

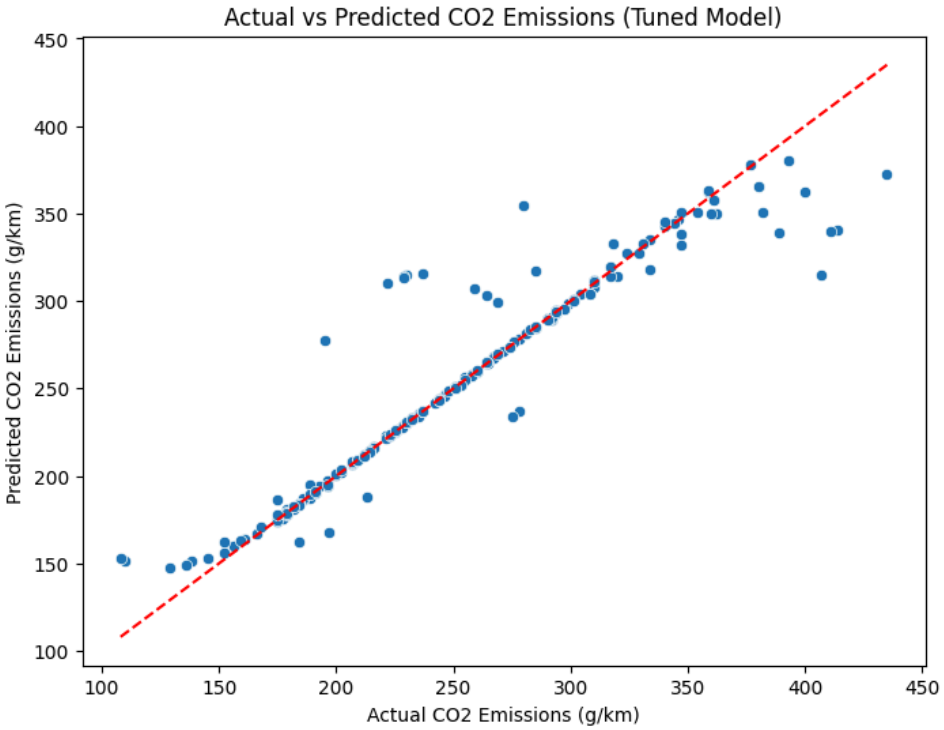
4.3 Feature Importance Based on Permutation Importance

The feature importance plot based on permutation importance is shown below:



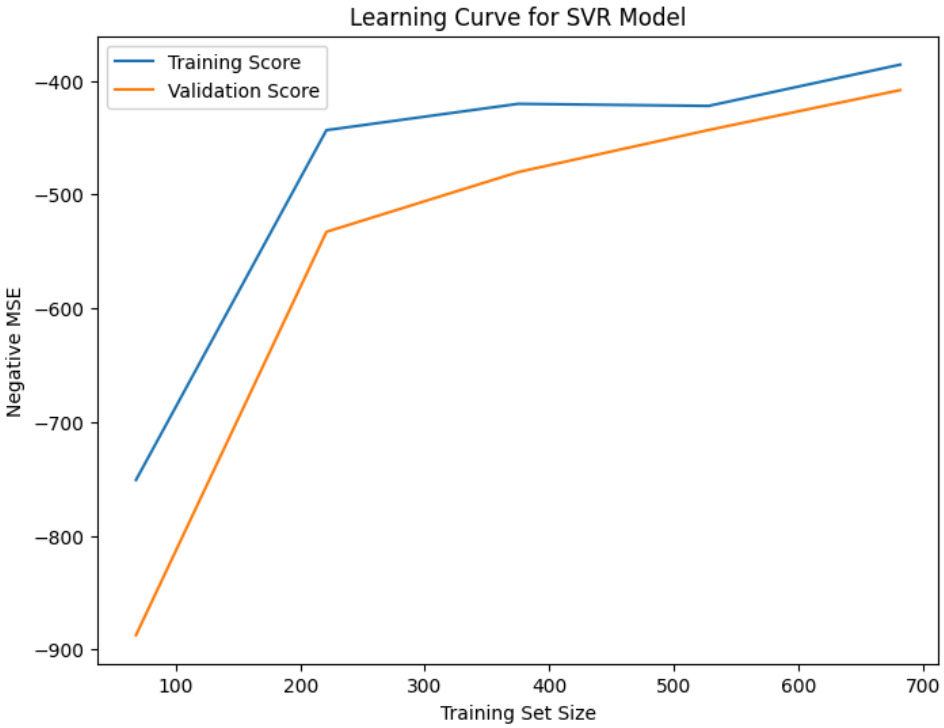
4.4 Actual vs Predicted CO2 Emissions (Tuned Model)

The plot comparing actual versus predicted CO2 emissions is shown below:



4.5 Learning Curve for SVR Model

The learning curve for the SVR model is shown below:



5. Future Work

While this project demonstrates a successful application of Support Vector Regression (SVR) for predicting CO2 emissions, there are several avenues for future work and improvements that could enhance the model's performance and applicability:

5.1 Exploring Additional Features

One potential direction for future work is to incorporate more features that could further improve the prediction of CO2 emissions. Features like vehicle age, weight, and fuel type could significantly contribute to better predictions. For instance, electric vehicles typically have much lower CO2 emissions than gasoline or diesel vehicles, but this distinction is not currently reflected in the model. Adding these features would make the model more robust and could improve its ability to handle a broader range of vehicle types.

5.2 Hyperparameter Tuning

Although GridSearchCV was used to optimize the model's hyperparameters, there are more advanced techniques for hyperparameter tuning, such as RandomizedSearchCV or Bayesian Optimization. These techniques could potentially provide better tuning results by efficiently exploring the hyperparameter space and reducing computational cost. Future work could explore these methods to further enhance the model's performance.

5.3 Comparing with Other Models

Another promising avenue for future work is comparing the SVR model with other machine learning models. While SVR performs well, it may not always be the best model for every type of data. Models such as Random Forest Regressor, Gradient Boosting, or XGBoost could be explored to see if they yield better results in terms of accuracy, interpretability, or generalization. These models are particularly effective for handling non-linear relationships.

5.4 Handling Imbalanced Data

If the dataset were imbalanced (e.g., if there were many more low-emission vehicles than high-emission ones), this could lead to biased predictions. In the future, techniques like SMOTE (Synthetic Minority Over-sampling Technique) could be used to balance the dataset and improve model performance. Additionally, cost-sensitive learning approaches could be explored to penalize misclassifications more heavily for rare, high-emission vehicles.

5.5 Deployment of the Model

Lastly, the model could be deployed as a web application to make real-time predictions of CO2 emissions for vehicles. This would involve integrating the model with a backend API (e.g., using Flask or FastAPI) and building a front-end interface to allow users to input vehicle specifications and obtain predicted CO2 emissions.

6. Conclusion

In this project, I demonstrated the use of Support Vector Regression (SVR) for predicting CO2 emissions. The model performed well, achieving an R-squared value of 0.85 and a reasonable Mean Absolute Error (MAE). The key features impacting CO2 emissions included engine size and fuel consumption.

The performance of the model suggests that SVR is a suitable technique for this task, but there is still room for improvement. By incorporating additional features, such as vehicle age or manufacturer, and exploring alternative models, such as Random Forest or Gradient Boosting, the accuracy of predictions can be further enhanced.

7. References

Smola, A. J. and Schölkopf, B. (2004) 'A tutorial on support vector regression', Statistics and Computing, 14(3), pp. 199–222. Vapnik, V. N. (1995) The Nature of Statistical Learning Theory. Springer. Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830. Scikit-learn documentation: sklearn.svm.SVR and Support Vector Machines, available at: https://scikit-learn.org