REPORT

ABSTRACT

This Project addresses the problem of regression, focusing on the estimation of vehicle CO2 emissions based on technical and fuel consumption data. Using the publicly available "2022 Fuel Consumption Ratings" dataset from Natural Resources Canada (hosted on Kaggle), the objective is to build predictive models that estimate emissions (in grams of CO2 per kilometre) based on attributes such as engine size, number of cylinders, fuel type, transmission, and combines fuel consumption. Beyond prediction, the project also explores the explainability of the results through feature importance analysis.

Several regression algorithms were implemented and compared, including **Linear Regression, Random Forest Regressor,** and **Gradient Boosting Regressor.** These models were trained and evaluated using an 80/20 train-test split. The best results were obtained using tree-based models, with Random Forest and Gradient Boosting achieving R^2 scores above 0.95 and significantly lower RMSE values compared to Linear Regression. The interpretability analysis confirmed that fuel consumption and engine characteristics are the most important predictors, aligning well with domain expectations and findings in related literature. The results demonstrate the effectiveness of ensemble methods for emission prediction tasks and highlight the value of combining predictive accuracy with model explainability.

INTRODUCTION

Carbin dioxide (CO2) emissions from vehicles represent a major contributor to climate change and environmental degradation. As urban transportation continues to grow globally, reducing vehicular emissions has become a key priority for governments, industries, and citizens. Accurate prediction of CO2 emissions based on vehicle characteristics can support the development of eco-friendly policies, inform consumers, and guide automotive manufacturers in designing cleaner technologies.

The problem addressed in this project is regression, specifically the estimation of CO2 emissions (in grams per kilometre) based on various technical features of a vehicle. These include engine size, fuel type, transmission, number of environmental data mining and predictive modelling, where the goal is not only to estimate outputs but also to understand the contribution of each input feature to the outcome.

Several studies have explored similar predictive tasks. Sahin et al. (2020) applied machine learning algorithms to transportation-related emissions data, demonstrating the effectiveness of ensemble models like Random Forest and Gradient Boosting in capturing complex non-linear relationships. Meanwhile, Chen et al. (2023) investigated hybrid models and emphasized the integration of feature importance analysis to interpret and validate predictions.

Building in these insights, the present work proposes a data mining solution that combines multiple regression algorithms – namely, Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor – to estimate CO2 emissions from vehicle attributes. Additionally, permutation-based feature importance is used to identify the most influential variables in the prediction process.