2 Literature Review

2.1 Introduction

The environmental concerns associated with greenhouse gas emissions have intensified research efforts to understand and predict vehicular CO_2 emissions. Machine learning techniques, especially ensemble models and neural networks, have emerged as powerful tools in this domain. This chapter presents a structured review of existing literature, organized into two primary segments:

- Studies related to vehicle emissions modeling
- Studies related to machine learning pipelines for structured regression problems

Finally, the chapter concludes by identifying critical research gaps that are addressed by this thesis.

2.2 Studies on Vehicle Emission Prediction

2.2.1 Traditional Statistical Approaches

Initial attempts to model vehicle emissions primarily utilized classical statistical methods, such as:

- **Linear Regression Models:** Early studies employed linear relationships between engine size, fuel consumption, and emission levels. However, these models often struggled with non-linear behaviors.
- **Multivariate Regression Analysis:** Some researchers used multivariate approaches to incorporate additional vehicle attributes, slightly improving predictive performance but still constrained by linear assumptions.

Limitations observed:

- Poor handling of complex, non-linear interactions
- High sensitivity to outliers and multicollinearity

2.2.2 Machine Learning-Based Approaches

The limitations of traditional models led to the adoption of machine learning (ML) algorithms. Key trends include:

- **Support Vector Machines (SVM):** SVM is employed by Non-linear regression tasks for solving; however, scalability to large datasets remains a challenge.
- **Random Forests (RF):** Provided significant improvements by capturing feature interactions but occasionally lacked precision for highly dynamic features.
- **Gradient Boosting Machines (GBMs):** Models such as XGBoost and LightGBM demonstrated strong predictive capabilities, especially on structured automotive datasets.

2.2.3 Domain Specific Findings

Several studies provided valuable domain-specific insights:

- **Engine Displacement and CO_2 Emissions:** Numerous papers reported a strong positive correlation.
- **Vehicle Weight Influence:** Heavier vehicles tend to exhibit higher emissions, aligning with thermodynamic principles.
- **Fuel Type Differentiation:** Emission patterns differ considerably between gasoline, diesel, hybrid, and electric vehicles.

However, these studies often emphasized model accuracy without delving into pipeline design or generalized system construction.

2.3 Studies on Machine Learning Pipelines for Structured Data

2.3.1 Data Preprocessing Techniques

Effective preprocessing is a cornerstone of successful ML applications. The literature highlights several techniques:

- **Handling Missing Values:** Mean imputation, median substitution, and more recently, K-Nearest Neighbors (KNN) imputation have been common.
- **Feature Scaling:** Standardization (z-score normalization) and min-max scaling are frequently used, particularly for tree-based models to improve convergence.
- **Categorical Encoding:** One-hot encoding, label encoding, and target encoding strategies are employed based on the algorithmic requirements.

Despite the awareness, many existing emission studies treated preprocessing as an afterthought rather than an integrated part of the pipeline.

2.3.2 Model Selection and Ensemble Techniques

Advanced pipelines often incorporate ensemble models to improve performance:

- **Bagging Techniques:** Methods like Random Forests emphasize variance reduction by training models on bootstrapped datasets.
- **Boosting Techniques:** Algorithms such as XGBoost, CatBoost, and LightGBM focus on bias reduction, yielding highly accurate models.
- **Stacking:** Some studies explored stacking multiple models to create a meta-predictor, but applications in emission prediction remain sparse.

2.3.3 Hyperparameter Tuning Methods

Hyperparameter optimization is critical for maximizing model potential. Approaches include:

- **Grid Search:** Exhaustive search across specified parameter grids; computationally intensive.
- **Random Search:** Faster but less exhaustive.
- **Bayesian Optimization:** More recent studies favor Bayesian approaches, with tools like Optuna and Hyperopt, achieving a balance between exploration and exploitation.

However, within the context of vehicle emission modeling, most published works still rely on grid or random search, underutilizing newer optimization techniques.

2.4 Research Gaps Identified

Based on the literature reviewed, several gaps persist:

- **Pipeline Integration:** A lack of comprehensive end-to-end pipelines encompassing data pre-processing, model training, feature engineering, and evaluation cohesively.
- **Ensemble and Meta-Learning:** Limited adoption of advanced ensemble techniques such as stacking regressors with neural network meta-learners for structured data.
- **Hyperparameter Optimization:** Underutilization of Bayesian hyperparameter optimization strategies in the emission prediction context.
- **Residual Analysis and Correction:** Few studies implement post-prediction residual modeling to further enhance performance.
- **Generalizability Focus:** Models are often tuned for specific datasets without ensuring broader applicability across varying vehicle types or geographies.

2.5 Summary

This subsection reviewed the evolution of CO_2 emission prediction approaches, highlighting the transition from traditional statistical techniques to sophisticated machine learning approaches. While significant progress has been made, there remains substantial scope for improvement, particularly regarding pipeline integration, ensemble learning strategies, hyperparameter optimization, and residual correction mechanisms. Addressing these gaps forms the central focus of this thesis.