****

**Assessment Report**

on

**“Vehicle Emission Classification using Machine Learning”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

By

Name Abhay Gupta

Roll Number : 202401100400004

Section: A

**Under the supervision of**

“Bikki Gupta”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

1. **Introduction**

**The goal of this project is to classify vehicles into their respective emission categories based on features such as engine size, fuel type, and CO2 emissions. This is a supervised classification problem and is crucial for environmental analysis and regulatory compliance.**

**2. Problem Statement**

Classify Vehicles Based on Engine Emissions

Predict emission category of a vehicle based on engine and fuel features.

**3. Objectives**

* Preprocess the dataset for training a machine learning model.
* Train a Logistic Regression model to classify loan defaults.
* Evaluate model performance using standard classification metrics.
* Visualize the confusion matrix using a heatmap for interpretability.

**4. Methodology**

* **Data Collection**: The user uploads a CSV file containing the dataset.
* **Data Preprocessing**:  
  + Handling missing values using mean and mode imputation.
  + One-hot encoding of categorical variables.
  + Feature scaling using StandardScaler.
* **Model Building**:  
  + Splitting the dataset into training and testing sets.
  + Training a Logistic Regression classifier.
* **Model Evaluation**:  
  + Evaluating accuracy, precision, recall, and F1-score.
  + Generating a confusion matrix and visualizing it with a heatmap.

**5. Data Preprocessing**

The dataset is cleaned and prepared as follows:

* Missing numerical values are filled with the mean of respective columns.
* Categorical values are encoded using one-hot encoding.
* Data is scaled using StandardScaler to normalize feature values.
* The dataset is split into 80% training and 20% testing.

**6. Model Implementation**

Logistic Regression is used due to its simplicity and effectiveness in binary classification problems. The model is trained on the processed dataset and used to predict the loan default status on the test set.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

* **Accuracy**: Measures overall correctness.
* **Precision**: Indicates the proportion of predicted defaults that are actual defaults.
* **Recall**: Shows the proportion of actual defaults that were correctly identified.
* **F1 Score**: Harmonic mean of precision and recall.
* **Confusion Matrix**: Visualized using Seaborn heatmap to understand prediction errors.

**8. Results and Analysis**

* The model provided reasonable performance on the test set.
* Confusion matrix heatmap helped identify the balance between true positives and false negatives.
* Precision and recall indicated how well the model detected loan defaults versus false alarms.

**9. Conclusion**

This project aimed to classify vehicle emission categories using a Random Forest Classifier based on engine size, fuel type, and CO2 emissions. The model achieved a 40% accuracy, indicating potential for improvement. Key challenges included the small dataset and limited features. Overall, the project provided hands-on experience with classification, model evaluation, and visualization in machine learning.

**10. References**

* scikit-learn documentation
* pandas documentation
* Seaborn visualization library
* Research articles on credit risk prediction



