MILES AND EMISSIONS:

How cars contribute to CO_2 levels

BINUS UNIVERSITY

3 GOOD HEALTH AND WELL-BEING | 11 SUSTAINABLE CITIES AND COMMUNITIES | 12 RESPONSIBLE CONSUMPTION AND PRODUCTION | 13 CLIMATE ACTION

ADRIAN MARC FEDIER PURNAMA - 2702247216
GERVASIUS RUSSELL - 2702247450
HANS ARDIANTA - 2702249663
JONATHAN WILLIAM GUNAWAN - 2702251794
DATA SCIENCE, BINUS UNIVERSITY

Intro: What is this about?

As we all know, air pollution has become a significant global challenge in this modern era. Vehicular carbon dioxide (CO_2) emissions has become one of the major contributor to the polluted air condition. So by this project, we are aiming to create vehicular CO_2 emission prediction based on vehicle characteristics and behaviors by implementing machine learning techniques. The dataset is obtained from Kaggle. The dataset comprises car characteristics, fuel consumption behavior, and the target variable, carbon dioxide emissions (g/km). Since the target variable in this dataset is continuous, we employ regression machine learning models for prediction. The ultimate goal of this project is to contribute to the identification of carbon dioxide emissions produced by cars, providing valuable insights that can support efforts to reduce environmental impact.

Data Understanding

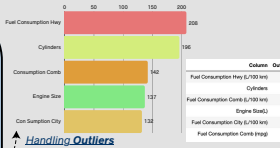
Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0 ACURA	ILX	COMPACT	2.0	4	AS5	Z	9.9	6.7	8.5	33	196

This is the **independent variable (Y)** or the **target variable**

7385 rows, **11** columns

6 numerical variables

5 categorical variables



Column	Outliers	percentage_outlier
Fuel Consumption Hwy (L/100 km)	208	2.82
Cylinders	196	2.65
Fuel Consumption Comb (L/100 km)	142	1.92
Engine Size(L)	137	1.86
Fuel Consumption City (L/100 km)	132	1.79
Fuel Consumption Comb (mpg)	114	1.54

Data Quality issues:
This data has **no null/missing values,** the only quality issue in this dataset is **outliers,** but it is also only in small quantity.

Handling ***Outliers***
To handle outliers, we use the ***Winsorizer*** *technique with caps at the* ***10th*** *and* ***90th percentiles****.*
Why Winsorizer? *Winsorization is more robust against outliers compared to other methods like IQR or z-score because* ***it limits extreme values instead of removing or fully transforming them****. Winsorizer adjusts values while preserving the overall data structure and minimizing the loss of information. This makes it particularly effective for datasets with skewed distributions or extreme outliers.*

	Feature	Correlation Score	p-value	Relation
0	Engine Size(L)	0.869299	0.0	Correlate
1	Cylinders	0.847764	0.0	Correlate
2	Fuel Consumption City (L/100 km)	0.955605	0.0	Correlate
3	Fuel Consumption Hwy (L/100 km)	0.840721	0.0	Correlate
4	Fuel Consumption Comb (L/100 km)	0.952481	0.0	Correlate
5	Fuel Consumption Comb (mpg)	-0.961673	0.0	Correlate

We can also see that every numerical feature in the dataset is **highly correlated** with the target variable

Machine Learning Model

20% test data

Data Split

80% train data

Models *that are used*

Decision Tree | Linear Regression | Random Forest | XGBRegressor

ensemble learning

	Model	R2_Score	MAE	MSE
0	Decision Tree	0.994223	2.087644	19.788222
1	Linear Regression	0.899455	10.893531	344.430175
2	Random Forest	0.992970	2.195117	24.083272
3	XGBRegressor	0.992884	2.212056	24.375210

Key takeouts

- Decision Tree - **best performer** - records highest R^2 Score and the lowest MAE and MSE values, indicating **strong accuracy**.
- Linear Regression - **weakest performer**
- Both Random Forest and XGBRegressor deliver similar results, performing well but **slightly below** the Decision Tree.

Feature Engineering

Label Encoding → **Normalization (RobustScaler)** → **Feature Selection with PCA** → Data ready to be split into **train** and **test** data

Label Encoder converts **categorical values into numerical labels,** allowing machine learning models to process and interpret categorical data.

(for example we use '**Model**' variable from this dataset)

Before	After
F-150 FFV	0
MUSTANG	1
CAMARO	2
SIERRA	3
ACCORD	4
SONIC	5
SIERRA	6
and so on..	*and so on..*

Encoding converts categorical variables into numerical formats, **preserving their information** and making them **compatible with machine learning models.**

Why do we use Robust Scaler?

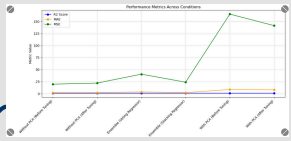
RobustScaler is chosen because it is **well-suited for skewed data** and **less sensitive to outliers.** Unlike Min-Max Scaler or Z-Scaler, which rely on the mean and standard deviation—making them prone to the influence of outliers and non-normally distributed data—RobustScaler uses the **median** and **interquartile range (IQR)**. This makes it **more effective in scaling data while maintaining robustness against extreme values,** ensuring better performance for datasets with skewed distributions or outliers.

RobustScaler formula:

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

We use **PCA** with **5 components**, determined through the **Cumulative Explained Variance (CEV)** technique with a **90%** threshold. This means **we select the smallest number of principal components that collectively explain at least 90% of the variance** in the data.

Model Performance Evaluation



In this section, we evaluate the performance of the machine learning model using metrics such as **R^2 Score**, **Mean Squared Error (MSE)**, and **Mean Absolute Error (MAE)**. The goal is to compare the model's accuracy **before** and **after** applying **PCA (Principal Component Analysis)**. By analyzing these metrics, we can determine whether PCA improves or affects the model's predictive ability.

The plot shows that **without PCA**, the model **performs well** and **improves further with tuning and ensemble methods.** In contrast, **applying PCA significantly worsens performance**, with higher errors (MSE and MAE) and near-zero R^2 scores, even after tuning, indicating **PCA is not beneficial for this dataset.**

Conclusion

By implementing various machine learning techniques, we managed to make car CO_2 emission prediction model and get the most out of it. The results show that the model's accuracy is higher without PCA, indicating that the original features are already sufficiently informative and we can use Decision Tree or Random Forest model to achieve high prediction accuracy. This project is beneficial for supporting air pollution control efforts by providing a predictive tool to identify high-emission vehicles, design strategies to reduce emissions, and support data-driven policy development. By leveraging machine learning technology, this project contributes to creating more efficient, data-based solutions to improve air quality and promote environmental sustainability.

To see the full code, you can scan this **QR code.**

The code is available and provided in **Google Colab.**