**CSE704 Semester Project: ABSTRACT**

**Title: The impact of CO2 emission on sales of new light-duty vehicles**.

Data:[**https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles**](https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles)

Project members**:** Joan Ejeta, Dequante Mckoy, Nikita Agarwal

1. **Problem Statement**

Climate change is the greatest long-term threat that we face as a global community. It also represents an enormous economic opportunity. Canada’s enhanced Nationally Determined Contribution (NDC) to the United Nations has committed Canada to cut its greenhouse gas emissions (GHG) by 40‑45 percent below 2005 levels by 2030. Canada’s NDC submission outlines a series of investments, regulations and measures that the country is taking in pursuit of its ambitious target. It includes input from provincial, territorial and Indigenous partners.

The NDC recognizes the **transport sector**as the largest source of GHG emissions in the Canada due to its high dependency on fossil fuels, with more than 90% of energy use coming from petroleum. The NDC lists a number of policies, some of which sub-national governments have already employed to reduce emissions in the sector, including: tailpipe emissions and efficiency standards; incentives for zero-emission personal vehicles; funding for charging infrastructure to support multi-unit dwellings, public charging, and long-distance travel.  The project seeks to apply data science to give a quantitative measure of CO2 emissions from light duty vehicles in densely populated area with urban traffic.

1. **What are some of the existing approaches to solve the problem?**

This is a new subject area. So far, data science, machine learning models (Ridge, Lasso, Random forest, Linear regression), deep learning models (BERT) and NLP has been used to estimate the CO2 emissions from vehicles.

1. **What is your approach**

This dataset captures the details of how CO2 emissions by a vehicle can vary with the different features. The dataset has been taken from Canada Government official open data website. This is a compiled version. This contains data over a period of 7 years.

My approach is to use Data Visualization(matplotlib) and Machine learning (linear regression/decision trees) taught in class to estimate the percentage volume of CO2 emissions from the vehicles based by its features by 2030 and estimate the pricing for carbon pollution for light-vehicles in traffic dense areas.

We will follow below step by step approach:

* Load the Dataset.
* Pre-Process the Data: Cleaning the Dataset, remove noise, anomalies, redundant data, unwanted data, Balanced Data etc.
* Analyze data – understand data with descriptive statistics.
* Data visualization.
* Create Training and Test Datasets (80% and 20%).
* Build the Model by training on the Training Dataset.
* Evaluate the Model by using the Confusion Matrix.
* Validation of the Model by 10-fold-Cross Validation Strategy.
* Prediction of the Model on Test Dataset.

1. **Expected outcome and conclusion**

If we successfully develop a machine learning model that can predict the volume/ percentage of CO2 emission from vehicles including hybrid and electric vehicles based on car features, we can be able to predict how much the transport industry will spend on carbon pollution from the current minimum price of $50 per ton which will increase by $15 each year from 2023 to 2030. Hence, vehicle dealers would be able to design more efficient cars and government can regulate the transport industry contribution to carbon pollution in order to meet its 2030 target. Following are the goals which we want to achieve with this Data Science Project.

* Which variables are significant in predicting the CO2 emission?
* How well those variables describe the CO2 emission?
* Which variable have the more carbon emission?
* which variable has least contribution to the CO2 emission?