Anim Yusufbhai Malvat
110024150
malvat@uwindsor.ca
School of Computer Science

Dhwani Gurjar
110022182
gurjar2@uwindsor.ca
School of Computer Science

Siddhi Patel
110009085
patel2e2@uwindsor.ca
School of Computer Science

Satyapalsinh Solanki
110022029
solanki@uwindsor.ca
School of Computer Science

Project Title: Data mining: Data prediction of increase in gas emissions in Ontario

Course: COMP 8157 – Advanced Database Topics

Data source: Canada's official national greenhouse gas inventory

https://www.canada.ca/en/environment-climate-change/services/climate-change/greenhouse-gas-emissions/inventory.html

GitHub: https://github.com/malvat/DataMiningGasEmissions

Milestone I

Predictive analysis means, finding trends in data using variety of statistical techniques. As mentioned in our project title and description, we are trying to predict the increase in gas emission in Ontario using historic data. We will also, try to map this with temperature and see, if we can find any relation between increase in gas emissions and temperature.

NOTE: we have used **Python Jupyter Notebook** for our project, as it provides aesthetic visuals for presentation purposes and works as an IDE at the same time.

We have formulated our project into steps, they are as mentioned below:

- Collect data
- Process the data according to the need
- Visualize the dataset to find trends
- Choose a model to map and predict
- Train parameters to improve model
- Predict results and write a report

Collecting data

After researching, we have found various data sources like, Keggel and Canada's official data sources, We found that the data was almost identical and therefore, it had not much effect but we have selected Canada's official data source (link can be found at the start of the document). After collecting our data, the data is raw and contains way more information than what we are going to use as one can see in the Figure 1. Sample of data, and therefore, it was time to process it.

:

	Year	Region	Index	Source	Sector	Sub-sector	Sub-sub-sector	Total	CO2eq	Unit
0	1990	Alberta	0	Provincial Inventory Total	NaN	NaN	NaN	y	173.0523683	Mt
1	1990	Alberta	1	Oil and Gas	NaN	NaN	NaN	y	67.57525298	Mt
2	1990	Alberta	2	Oil and Gas	Upstream Oil and Gas	NaN	NaN	y	63.93934941	Mt
3	1990	Alberta	3	Oil and Gas	Upstream Oil and Gas	Natural Gas Production and Processing	NaN	NaN	29.09710967	Mt

Figure 1 Sample of data

Process the data

To ease the process, we have divided it into 3 steps:

- Remove columns that are not necessary
- Convert the CO emission column from string to float and set the precision to two points
- Filter the data for Ontario region

After pre-processing, data looks as shown in Figure 2. Filtered Data.

	Year	CO2eq
0	1990	556.22
1	1991	549.91
2	1992	561.24
3	1993	541.90
4	1994	550.43
5	1995	563.84

Figure 2 Filtered data

Visualize the data set

To visualize the data set we have use **matplotlib** for python because, it provides various customization options and it works directly with **pandas** library of python.

```
In [7]: # plot the graph
        plt.plot(data_ontario['Year'], data_ontario['CO2eq'])
        plt.title("CO2 emissions in Ontario")
        plt.ylabel("CO2 emissions")
        plt.xlabel("Year")
        plt.show()
```

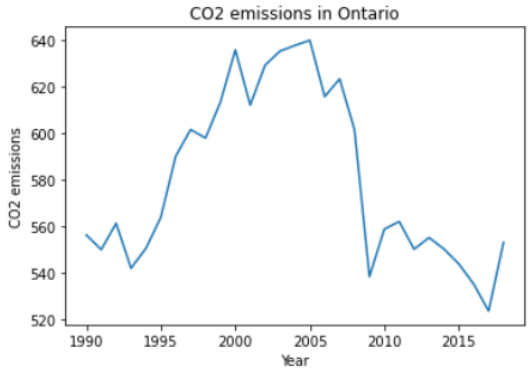

Figure 3 Co2 emissions in Ontario