Project Phase 4: Data Mining

CSI4142 - Fundamentals of Data Science

Winter 2023

uOttawa

Faculté de génie
Faculty of Engineering

School of Electrical Engineering and Computer Science
University of Ottawa

Professor Yazan Otoum

Group 10:
Mazharul Maaz - 300128179
Bill Battushig - 300109257
Xiao Meng Li - 300109886

Due Date: Apr 11th, 2023

Table of Contents

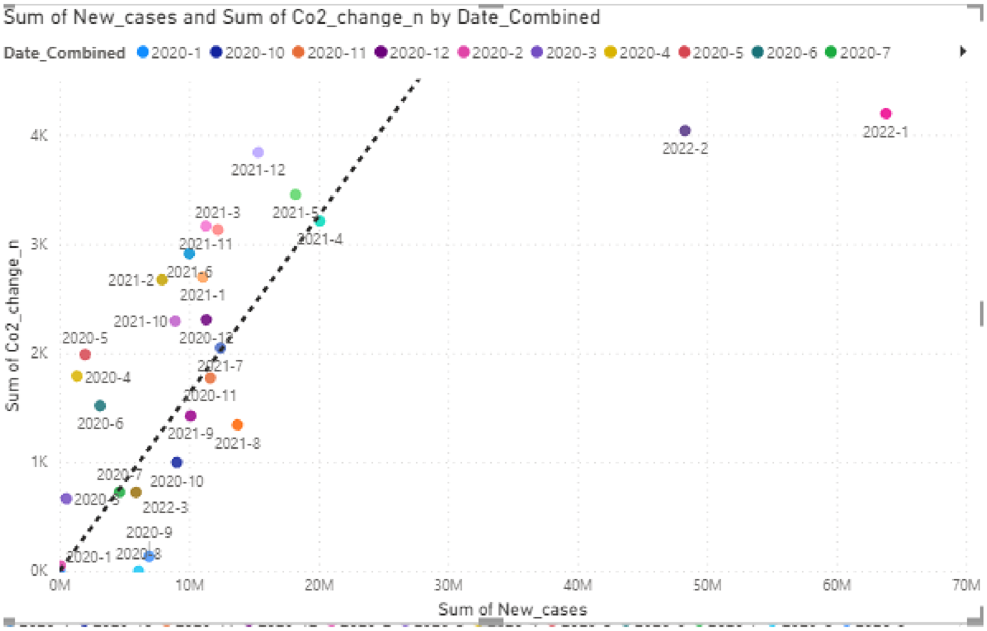
Table of Contents **1**
Data Summarization, Data Preprocessing and Feature Selection **1**
Data Summarization 2
Data Preprocessing and Feature Selection 4
Classification (Supervised Learning) **5**
Introduction 5
Table 5
Analysis 6
Detecting Outliers **6**
Preprocessing 6
Parameter Settings 6
Result 6
Conclusion 7
Distribution of Tasks **8**

Data Summarization, Data Preprocessing and Feature Selection

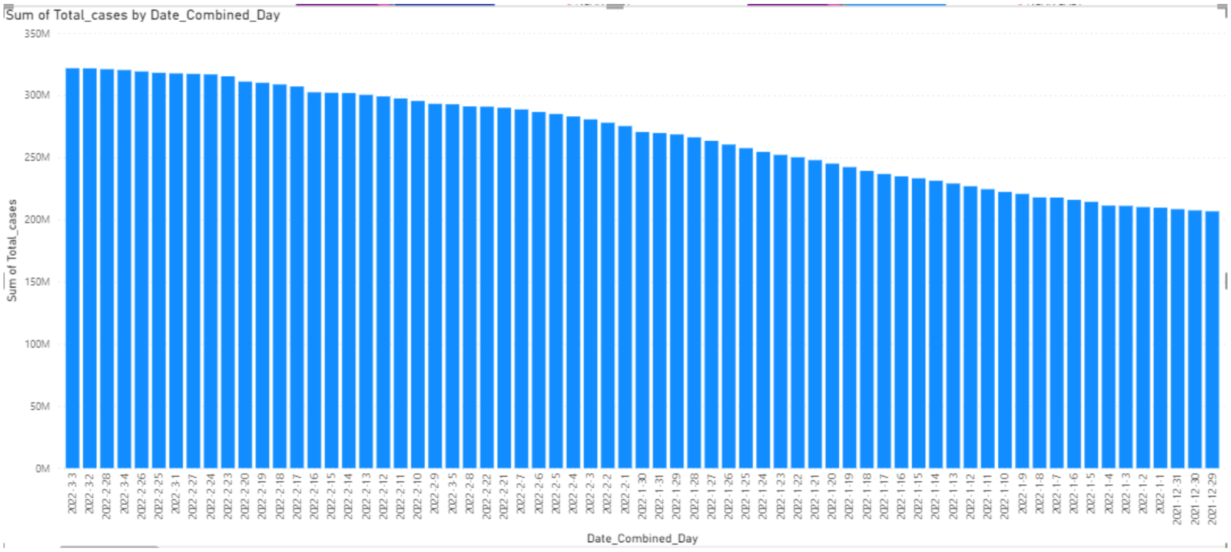
Data Summarization

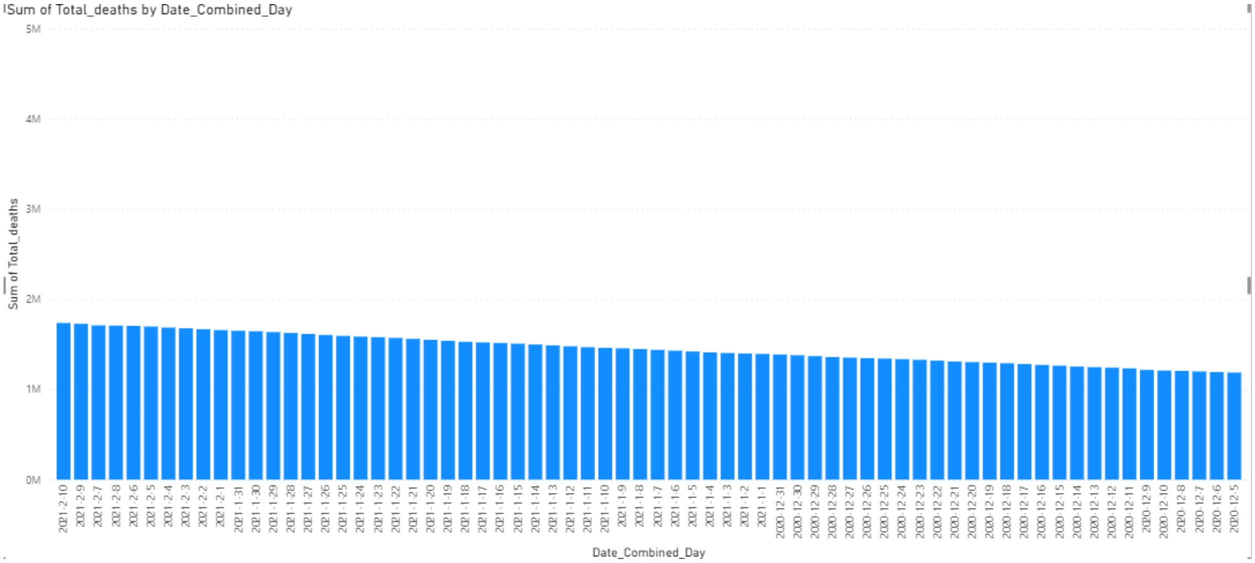
We conducted data summarization using the following techniques:

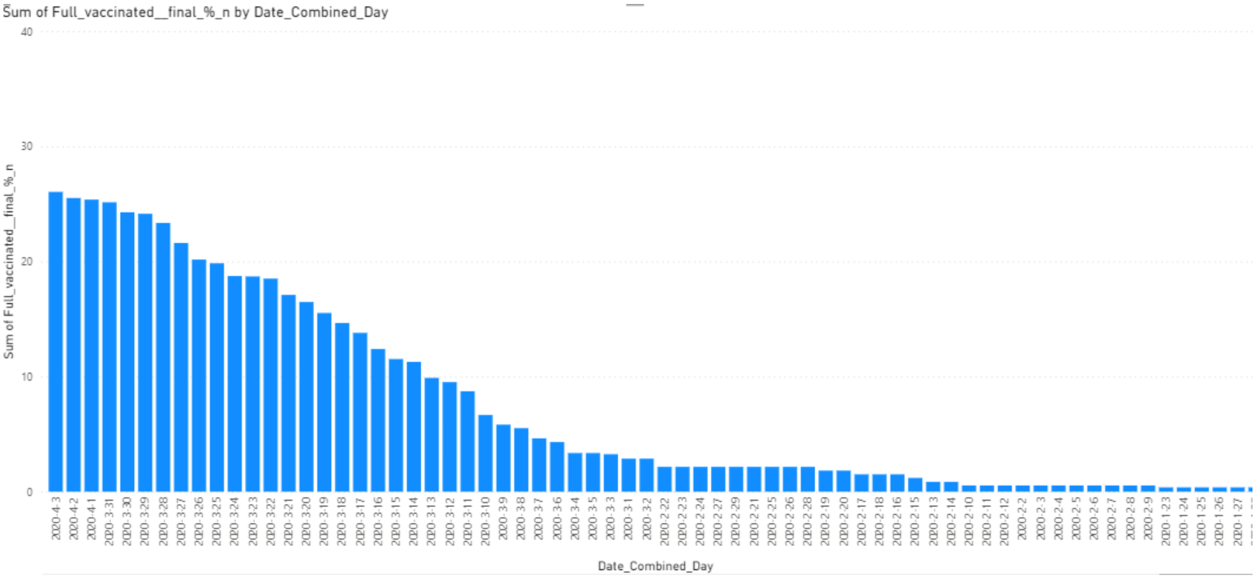
Scatter plots: To visualize relationships between different pairs of attributes, such as the Sum of new cases and CO2 change by dates. This helped us identify any correlations or trends in the data.



Bar Charts: To visualize the distribution of attributes like total cases, total deaths, and vaccination rates based on time, which helped us understand the data's overall shape.


Sum of Total_cases by Date_Combined_Day
350M
300M
250M
200M
150M
100M
50M
0M
Sum of Total_cases
2022-3-3
2022-3-2
2022-2-28
2022-3-4
2022-2-26
2022-2-25
2022-3-1
2022-2-27
2022-2-24
2022-2-23
2022-2-20
2022-2-19
2022-2-18
2022-2-17
2022-2-16
2022-2-15
2022-2-14
2022-2-13
2022-2-12
2022-2-11
2022-2-10
2022-2-9
2022-3-5
2022-2-8
2022-2-22
2022-2-21
2022-2-7
2022-2-6
2022-2-5
2022-2-4
2022-2-3
2022-2-2
2022-2-1
2022-1-30
2022-1-31
2022-1-29
2022-1-28
2022-1-27
2022-1-26
2022-1-25
2022-1-24
2022-1-23
2022-1-22
2022-1-21
2022-1-20
2022-1-19
2022-1-18
2022-1-17
2022-1-16
2022-1-15
2022-1-14
2022-1-13
2022-1-12
2022-1-11
2022-1-10
2022-1-9
2022-1-8
2022-1-7
2022-1-6
2022-1-5
2022-1-4
2022-1-3
2022-1-2
2022-1-1
2021-12-31
2021-12-30
2021-12-29
Date_Combined_Day


Sum of Total_deaths by Date_Combined_Day
5M
4M
3M
2M
1M
0M
Sum of Total_deaths
2021-2-10
2021-2-9
2021-2-7
2021-2-8
2021-2-6
2021-2-5
2021-2-4
2021-2-3
2021-2-2
2021-2-1
2021-1-31
2021-1-30
2021-1-29
2021-1-28
2021-1-27
2021-1-26
2021-1-25
2021-1-24
2021-1-23
2021-1-22
2021-1-21
2021-1-20
2021-1-19
2021-1-18
2021-1-17
2021-1-16
2021-1-15
2021-1-14
2021-1-13
2021-1-12
2021-1-11
2021-1-10
2021-1-9
2021-1-8
2021-1-7
2021-1-6
2021-1-5
2021-1-4
2021-1-3
2021-1-2
2021-1-1
2020-12-31
2020-12-30
2020-12-29
2020-12-28
2020-12-27
2020-12-26
2020-12-25
2020-12-24
2020-12-23
2020-12-22
2020-12-21
2020-12-20
2020-12-19
2020-12-18
2020-12-17
2020-12-16
2020-12-15
2020-12-14
2020-12-13
2020-12-12
2020-12-11
2020-12-9
2020-12-10
2020-12-8
2020-12-7
2020-12-6
2020-12-5
Date_Combined_Day


Sum of Full_vaccinated__final_%_n by Date_Combined_Day
40
30
20
10
0
Sum of Full_vaccinated__final_%_n
2020-4-3
2020-4-2
2020-4-1
2020-3-31
2020-3-30
2020-3-29
2020-3-28
2020-3-27
2020-3-26
2020-3-25
2020-3-24
2020-3-23
2020-3-22
2020-3-21
2020-3-20
2020-3-19
2020-3-18
2020-3-17
2020-3-16
2020-3-15
2020-3-14
2020-3-13
2020-3-12
2020-3-11
2020-3-10
2020-3-9
2020-3-8
2020-3-7
2020-3-6
2020-3-4
2020-3-5
2020-3-3
2020-3-1
2020-3-2
2020-2-22
2020-2-23
2020-2-24
2020-2-27
2020-2-29
2020-2-21
2020-2-25
2020-2-26
2020-2-28
2020-2-19
2020-2-20
2020-2-17
2020-2-18
2020-2-16
2020-2-15
2020-2-13
2020-2-14
2020-2-10
2020-2-11
2020-2-12
2020-2-2
2020-2-3
2020-2-4
2020-2-5
2020-2-6
2020-2-7
2020-2-8
2020-2-9
2020-1-23
2020-1-24
2020-1-25
2020-1-26
2020-1-27
Date_Combined_Day

Data Preprocessing and Feature Selection

We performed the following data transformations to preprocess the data:

1. **Handling Missing Values**: We analyzed the dataset to identify missing values in various attributes. To address these missing values, we replaced them with "N/A" using the fillna("N/A") function. Additionally, we dropped rows containing missing values in critical columns, such as total_deaths, new_deaths, new_cases, total_cases, average_CO2, average_CO4, and average_N2O, to maintain the quality of our dataset in phase 2.In phase 4, we used the 'dropna' function to drop rows with missing values in the selected features.

2. **Handling Categorical Attributes**: We processed the 'Country' column by removing non-alphanumeric characters using the str.replace() function. This step was crucial for ensuring the consistency of the country names and facilitating the merging of columns based on the 'Country' key in phase 2. In phase 4, we used the KBinsDiscretizer from Scikit-learn in order to perform binning to convert the output to categorical values. KBinsDiscretizer is used to divide the continuous target variable into 10 discrete bins uniformly.

3. **Normalization and Scaling**: We implemented normalization and scaling techniques to ensure that our data was represented in a consistent and comparable manner. This process included using min-max normalization to adjust emissions changes, COVID-19 cases, and vaccination information. Furthermore, we scaled the World GDP values to millions and transformed the Area measurements from square miles to square kilometers. In addition, we also used the StandardScaler class from Scikit-learn to standardize the data. It scales the data to have a mean of 0 and a standard deviation of 1. This ensures that all attributes are of equal importance during learning.

4. **Feature Selection and Engineering**: In order to enhance the information available in our dataset, We utilized the following features to predict emissions: ['Total_cases', 'Total_deaths', 'Population', 'Area(sq._km.)', 'Net_migration_rate', 'Gdp_per_capita', 'Literacy_rate', 'Unemployment_rate'] These features were used as input variables (X) in the supervised learning models (Decision Tree, Gradient Boosting,

and Random Forest) to predict the changes in greenhouse gas emissions (CO2, CO4, and N2O).

Classification (Supervised Learning)

Introduction

In this section of the project, we will be using multiple features to predict how greenhouse emissions of CO2, CO4, and N2O will change. The features we will be using to predict are 'Total_cases', 'Total_deaths', 'Population', 'Area(sq._km.)', 'Net_migration_rate', 'Gdp_per_capita', 'Literacy_rate', and 'Unemployment_rate'. We believe these features are great predictors as they help indicate the effect of Covid in a certain location, as this will increase lockdowns and reduce commute. We will be comparing the performance of three supervised learning models: Decision Tree, Gradient Boosting, and Random Forest.

Table

Target Variable	Decision Tree Performance	Gradient Boosting Performance	Random Forest Performance
CO2 Change	Accuracy: 0.95 Precision: 0.86 Recall: 0.96 Execution Time: 0.43s	Accuracy: 0.11 Precision: 0.37 Recall: 0.14 Execution Time: 10.80s	Accuracy: 0.94 Precision: 0.84 Recall: 0.95 Execution Time: 31.36s
CO4 Change	Accuracy: 0.96 Precision: 0.68 Recall: 0.97 Execution Time: 0.37s	Accuracy: 0.23 Precision: 0.26 Recall: 0.48 Execution Time: 10.78s	Accuracy: 0.96 Precision: 0.68 Recall: 0.97 Execution Time: 28.97s
N2O Change	Accuracy: 0.98 Precision: 0.97 Recall: 0.97 Execution Time: 0.33s	Accuracy: 0.35 Precision: 0.35 Recall: 0.43 Execution Time: 10.74s	Accuracy: 0.98 Precision: 0.97 Recall: 0.97 Execution Time: 25.70s
Average Change	Accuracy: 0.96 Precision: 0.83 Recall: 0.97 Execution Time: 0.33s	Accuracy: 0.23 Precision: 0.33 Recall: 0.35 Execution Time: 10.69s	Accuracy: 0.96 Precision: 0.83 Recall: 0.96 Execution Time: 28.31s
Source Code can be found in PartB_Classification.py			

Analysis

There does not seem to be a significant trend in the performance of predicting different greenhouse emissions (CO2, CO4, and N2O). Comparing the three different models, we can see that Gradient Boosting is inadequate with less than 50% accuracy, while both Decision Tree and Random Forest have great performance numbers. A decision tree seems to be the correct model with this data due to its significantly shorter execution time of just 0.33 seconds.

Detecting Outliers

The one-class SVM algorithm was used to identify global outliers in dataset. The features that are used in detecting outliers are 'Total_cases', 'Total_deaths', 'Population', 'Area(sq._km.)', 'Net_migration_rate', 'Gdp_per_capita', 'Literacy_rate', and 'Unemployment_rate'. With the help of unsupervised outlier detection algorithm one-class SVM provided by ‘sci-kit-learn’, we detected about 3952 outliers.

Preprocessing

The initial step of finding outliers was to preprocess the data in Phase 4A. We preprocessed the data by handling the ‘null’, missing values, selecting relevant features and standardizing the dataset using tools provided by ‘sci-kit-learn’. Furthermore, standardization is needed since it ensures that all features have equal importance when calculating the outliers.

Parameter Settings

The one-class SVM algorithm is versatile and the ability to handle high-dimensional data made it even more useful in our case. In our script, the kernel value was set to ‘rbf’ (Radial Basis Function), nu parameter was set to 0.05, and the gamma value to 0.1. The Radial Basis Function was used because of its ability of efficiently capturing complex patterns in data. Also, the hyperparameter that controls the proportion of outliers in the dataset like nu was set to only 0.05 because it makes the model more sensitive to outliers. Moreover, the smaller gamma value allows us to produce more flexible decision boundary.

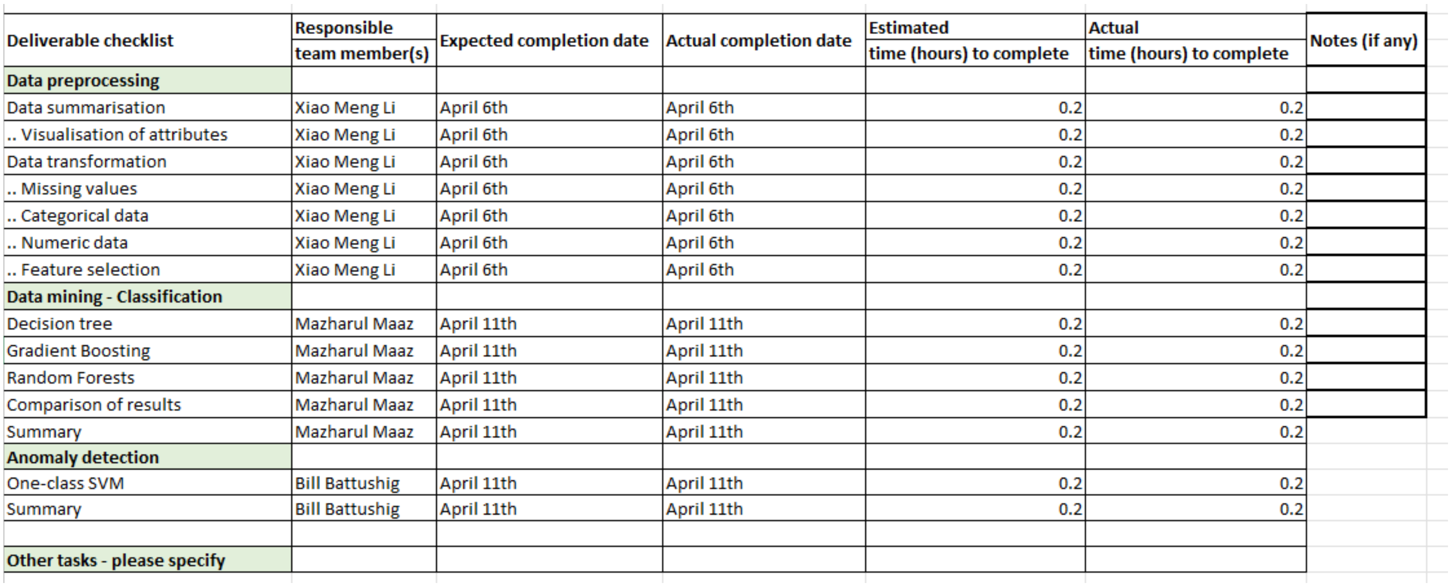
Result

The result of 3952 outliers can be due to several factors such as hyperparameter settings, data characteristics, data distribution, and feature selection. In order to reduce the high outliers, we might need to experiment more with hyperparameters of one-class SVM algorithm as it would help to find more reasonable decision boundary and a more accurate number of outliers. On the other hand, the presence of noise, extreme values can affect the decision boundary and it could lead to overestimation or underestimation of outliers. Lastly, correct features need to be selected in order to capture the characteristics of the data. Wrong features could result in a higher number of detected outliers. Simply put, more analysis and experimentation is needed to truly capture the correct outlier.

Conclusion

Finally, the detected outliers provide valuable insights into the relationships between the above selected features. For example, some outliers may represent countries or regions with unusual emission patterns or distinct socioeconomic factors. The advantage of understanding the underlying reason for these outliers can help researchers develop targeted strategies to address specific issues.

Distribution of Tasks



Deliverable checklist	Responsible	Expected completion date	Actual completion date	Estimated	Actual	Notes (if any)
	team member(s)			time (hours) to complete	time (hours) to complete	
Data preprocessing						
Data summarisation	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
.. Visualisation of attributes	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
Data transformation	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
.. Missing values	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
.. Categorical data	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
.. Numeric data	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
.. Feature selection	Xiao Meng Li	April 6th	April 6th	0.2	0.2	
Data mining - Classification						
Decision tree	Mazharul Maaz	April 11th	April 11th	0.2	0.2	
Gradient Boosting	Mazharul Maaz	April 11th	April 11th	0.2	0.2	
Random Forests	Mazharul Maaz	April 11th	April 11th	0.2	0.2	
Comparison of results	Mazharul Maaz	April 11th	April 11th	0.2	0.2	
Summary	Mazharul Maaz	April 11th	April 11th	0.2	0.2	
Anomaly detection						
One-class SVM	Bill Battushig	April 11th	April 11th	0.2	0.2	
Summary	Bill Battushig	April 11th	April 11th	0.2	0.2	
Other tasks - please specify						