 **INFOYSY.722 DATA MINING AND BIG DATA**

RESEARCH PAPER  
WEIGHTING: 20%

WORD COUNT: 4930

# ABSTRACT:

The global attention to the global warming is getting closer and closer, and carbon dioxide as one of its main effects needs to special attention.

Using a comprehensive dataset from Kaggle, this study investigated factors influencing carbon dioxide (CO2) emissions such as greenhouse gases, population and energy consumption. SPSS Modeler, Spyder and Jupyter Notebook are used for data preprocessing, feature selection and model construction.

Random Forest model investigate that energy production, population and GDP are significantly correlated with increased emissions, highlighting the environmental impact of economic activity. The logistic regression model further explores the segmentation of the energy sector, determining that oil has the highest positive correlation with CO2 levels. Visualization techniques such as pie charts and frequency plots confirm these findings by showing the distribution and trends of carbon dioxide emissions across different regions and industries.

The model results were used in discussions with real-world data to effectively reduce carbon dioxide emissions by developing a global mitigation strategy focused on transitioning to renewable energy, improving energy efficiency in fast-growing economies, and adopting cleaner industrial processes.

# Introduction:

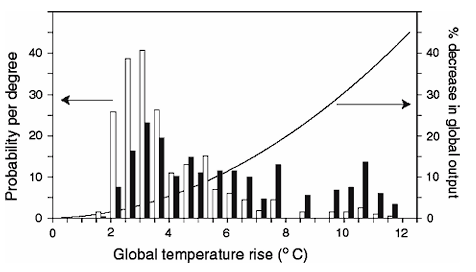
In today's world, governments are increasingly focused on mitigating the negative impacts of climate change. Although there is still some uncertainty about the extent of climate change, according to 97% of climatologists, negative impacts of climate change, which is largely driven by human activities such as burning fossil fuels, deforestation and industrial emissions (Patz, J. A et al., 2014). These activities contribute significantly to the accumulation of the greenhouse effect, resulting in an imbalance in the amount of energy absorbed and emitted by the Earth's atmospheric system, ultimately leading to a rise in temperature and thus causing serious environmental and health problems. Carbon dioxide is the only greenhouse gas that decays and displays multiple time constants. Currently, emissions of other greenhouse gases such as methane or nitrous oxide still contribute significantly to climate change, but they do not persist over time like carbon dioxide (Solomon, S et al., 2009).

Although scientists have found many signs of climate change through observations, it is still difficult to fully understand its causes and predict future trends. Today, global climate models remain a useful tool for assessing the impacts of large-scale climate change whereas the spatial resolution of the forecast needs to be improved in order to obtain more detailed regional information suitable for management purposes (López‐Moreno et al., 2007). Current climate models have limitations in explaining and predicting climate change, often failing to fully account for the complex climate system and the interactions between various influencing factors. Therefore, the research problem facing this study is how to use historical climate data to develop more accurate and reliable prediction models. These models not only need to explain past climate change, but must also be able to predict future trends in order to provide a scientific basis for making long-term planning and decisions to address climate change.

The objectives of this research aim to use historical climate data to develop predictive models that predict future conditions and contribute to climate resilience planning. Specifically, the first objective seeks to explore the relationship between carbon dioxide and multiple variables such as population, GDP, primary energy, methane to provide actionable insights at the state and national level. In addition, the second objective of this study will analyse carbon dioxide emissions from different industries to identify effective mitigation strategies. Therefore, based on study outcomes, reduce carbon dioxide emissions by 8% over the next 12 months by optimizing energy consumption, thereby supporting long-term resource sustainability and social well-being.

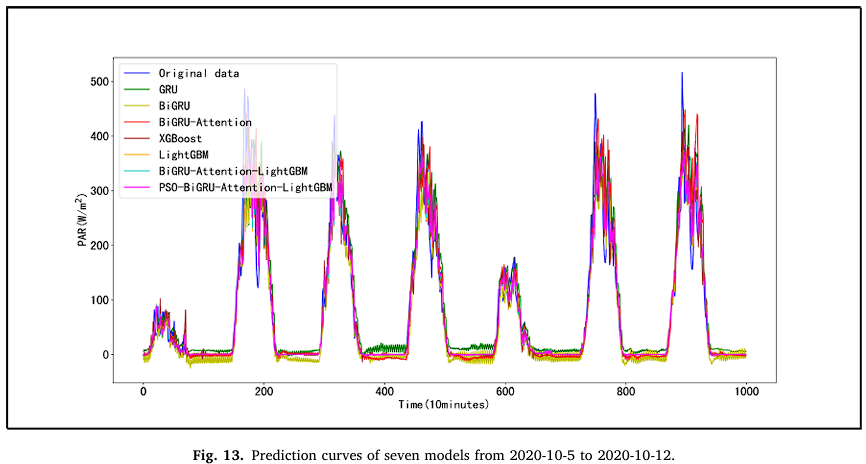
# LITERATURE REVIEW:

Whether predicting climate change is a fool's errand is still open to question. Andrew J. Watson (2007) points out that at present and in the foreseeable future, there are too many uncertain variables (such as water vapor, ice albedo, and cloud feedbacks) to accurately predict climate change in the next hundred years. However, thousands of models run through computers have been able to discover that the probability distribution function for climate sensitivity is a skewed distribution. By combining this probability distribution with the existing climate change data, a more reliable short-term climate change probability distribution image can be generated. It can be seen here that the increase in temperature directly leads to an increase in carbon dioxide through the sea surface temperature effect, which further leads to an increase in carbon dioxide emissions (See Figure1).



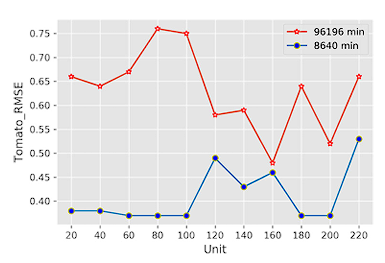
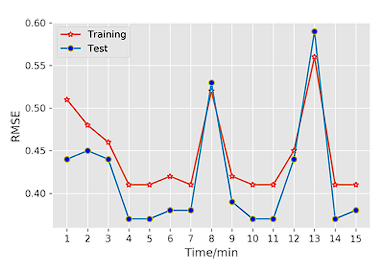
*Figure 1: Probability distribution functions for the climate change due to a doubling of CO2 and the cost of that climate change*

Mao et al. (2024) argues that the low accuracy of most prediction models is mainly due to the reliance on machine learning and deep learning techniques to continuously build and optimize single-variable prediction models. To address this problem, they try to model a combination of seven single models such as GRU, LightGBM, XGBoost and so on, constantly learning from new data, updating and optimizing the model by exploring the correlations within the data (See Figure2). This method can quickly respond to dynamic changes in the greenhouse environment, and the model accuracy does not degrade over time, which also helps to provide real-time, accurate predictions of future environmental conditions.



*Figure 2: Prediction curves of seven models from 2020-10-5 to 2020-10-12*

Liu, Yuwen et al. (2022) also emphasized the importance of using multivariate climate prediction. At the same time, the larger the number of data set units, the more details the model can capture and the better the prediction results. However, the higher the number of model units does not always mean the higher the accuracy of the model. When the number of training set and test set units is too large, the training speed of the model will slow down and it is easy to lead to overfitting (See Figure3). Therefore, selecting the appropriate number of units of data set for modelling is one of the important factors to achieve the research goal.

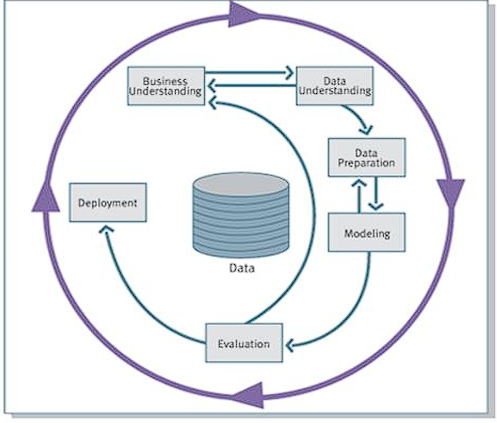


*Figure 3: RMSE value varies with the number of units in the dataset and the model training time*

Based on the literature exploring potential solutions, the approach to achieving the objectives of this study first requires generating a more reliable graph of the probability distribution of short-term climate change by combining probability functions with existing climate data. Secondly, multiple models are built and the results are combined to continuously learn and adapt to new data to maintain the accuracy of the model and provide real-time and accurate prediction of environmental conditions. Finally, choosing the appropriate data set size and variable type is crucial to achieve accurate and effective climate prediction models. Overly complex models can slow down training and lead to overfitting, and models that produce too simple can't make accurate predictions.

# RESEARCH METHODOLOGY:

For data mining research methodology, CRISP-DM (Cross-Industry Standard) process is a process model widely used in data mining and data analysis (See Figure4). Here it is used to break down the entire research project task into six different stages, each with clear tasks and objectives to help guide the implementation of the data mining project. This simple approach is designed to categorize and guide the most common steps in a data mining project. It quickly became "the standard for developing data mining and knowledge discovery projects" and remains the most widely used analytical method today (Martinez-Plumed et al., 2021).



*Figure 4: Cross-Industry Standard Process for Data Mining (CRISP-DM) process*

In this study, there are nine stages included instead of six, and the following is an explanation of each:

* Business understanding

Clearly identify the problem to be solved, such as obtaining prior knowledge of the problem through a literature review.

* Data understanding

Delve into data exploration by selecting the appropriate datasets from one or more sources based on Business understanding stage and understanding the data structure.

* Data preparing

Identify missing values, outliers, or extreme values and clean them to make sure they support the data mining methods. Finally, identify the target variables that are consistent with the objectives.

* Data transformation

Transform datasets to meet main objectives, such as normalizing data, redefining data types, or creating new variables.

* Data mining methods

Select the model to achieve the objective in Business understanding stage. Method options come from regression, classification, clustering, etc.

* Data mining algorithms

Choose the appropriate algorithm for each model selected in Data mining methods stage, such as linear or nonlinear regression, K-means clustering, hierarchical clustering, logistic regression, or decision trees.

* Data mining

Run the model with the tool and get the results. Apply the methods and algorithms in the previous steps to search for patterns, relationships, or trends in data.

* Interpreting

Evaluate output by interpreting parameters, checking model quality, visualizing results, and iterating over previous steps as needed to get the highest accuracy model and results.

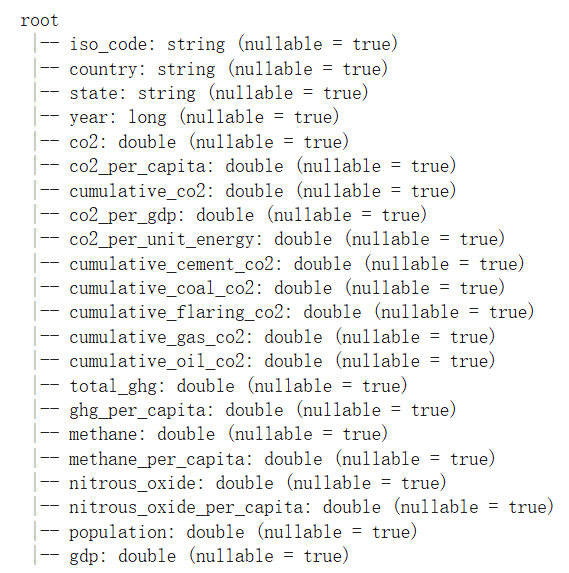
* Action

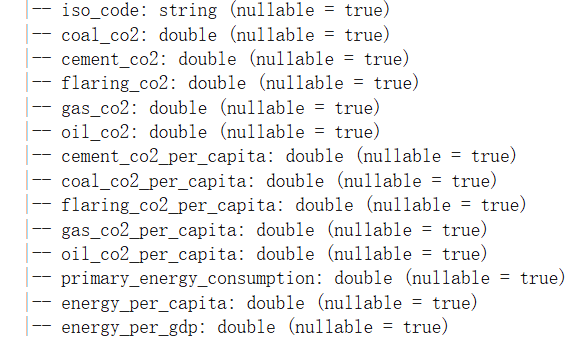
On the basis of interpretation, list and explain the meaning of the results. Provide advice to governments, industries or individuals to reduce the impact of climate change and discuss whether the conclusions are consistent with or contradict existing knowledge.

Three different tools are used in the data mining research: SPSS Modeler (ISAS), Spyder (OSAS), and Jupyter Notebook (BDAS). By using these tools to carry out steps 1-8 of the data mining process, it is found that one model accuracy rate obtained by SPSS Modeler is the highest (close to 100%) and another model obtained by Jupyter is the highest, so this study mainly researches the results of those two tooll. However, there are some differences between different tools. For example, SPSS Modeler cannot generate f1-score, precision, and recall, while it is feasible to run these metrics on Spyder using Python. As a result, ISAS are easy to use but have limited functionality. OSAS is slightly more complex to use, but as an open platform, it has a large community to contribute various features which leads to offer a wider range of choices in the data mining process. Therefore, for the parts that cannot be analysed (or hard to explore data) using SPSS Modeler and Jupyter, the Spyder will be used for support analysis.

# Design PROCESS:

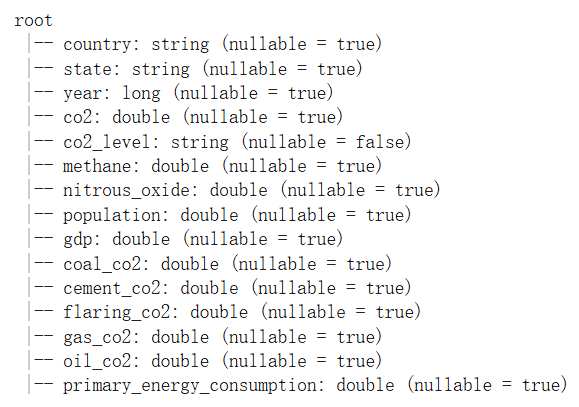
The Data set selected for this study comes from Kaggle, and the data in the data set are regularly collected through Our World in Data website, including carbon dioxide emissions (annual emissions, per capita emissions, cumulative emissions, consumption), other greenhouse gases, energy structure and other related indicators. This dataset contains 35 different types of attributes, such as country, year, carbon dioxide content, greenhouse gases, energy consumption, methane, nitric oxide, coal consumption, etc (See Figure5).





*Figure 5: All attributes contain in Dataset*

Nevertheless, data quality is critical because of the wide range and large number of numerical values for attributes, which increases the complexity of data analysis. Therefore, in order to carry out further data analysis, it is essential to carry out feature selection on multiple attributes to select the attributes that most relevant to the research objectives. According to the importance of the features, the attributes required to meet the two research objectives are selected. A new attribute named co2\_level has been created to support data mining methods of objective2 (See Figure6).

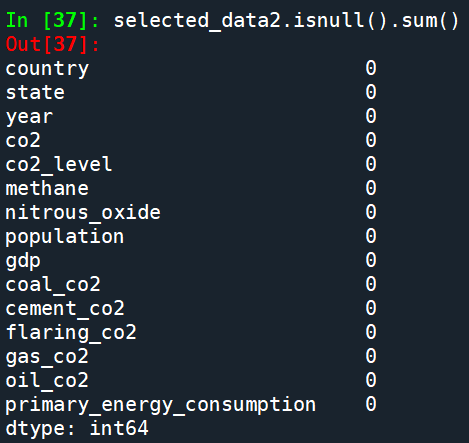
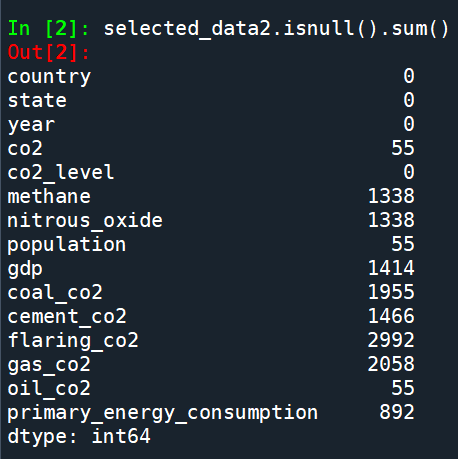


*Figure 6: All attributes satisfy data mining objectives*

The details of the selected attributes are shown below:

* Country: Country name, which identifies the country to which the data belongs.
* State: 7 state names, used to identify the region to which the data belongs.
* Year: Data were recorded from 2000 to 2022 for time series analysis.
* Co2: Total CO2 emissions, used to measure the level of greenhouse gas emissions.
* Co2\_level: Classify CO2 as low (CO2<=1000), medium (100< CO2<=3000), and high (CO2>3000) levels.
* Methane: Methane emission.
* Nitrous\_oxide: Nitrous oxide emissions.
* Population: The total population is used to analyze the impact of population on the environment.
* Gdp: Gross domestic product is a measure of economic activity in relation to environmental impact.
* Coal\_co2: Carbon dioxide emissions from coal combustion.
* Flaring\_co2: Carbon dioxide emissions from combustion processes.
* Gas\_co2: Carbon dioxide emissions from natural gas combustion.
* Oil\_co2: Carbon dioxide emissions from oil combustion.
* Primary\_energy\_consumption: Primary energy consumption.

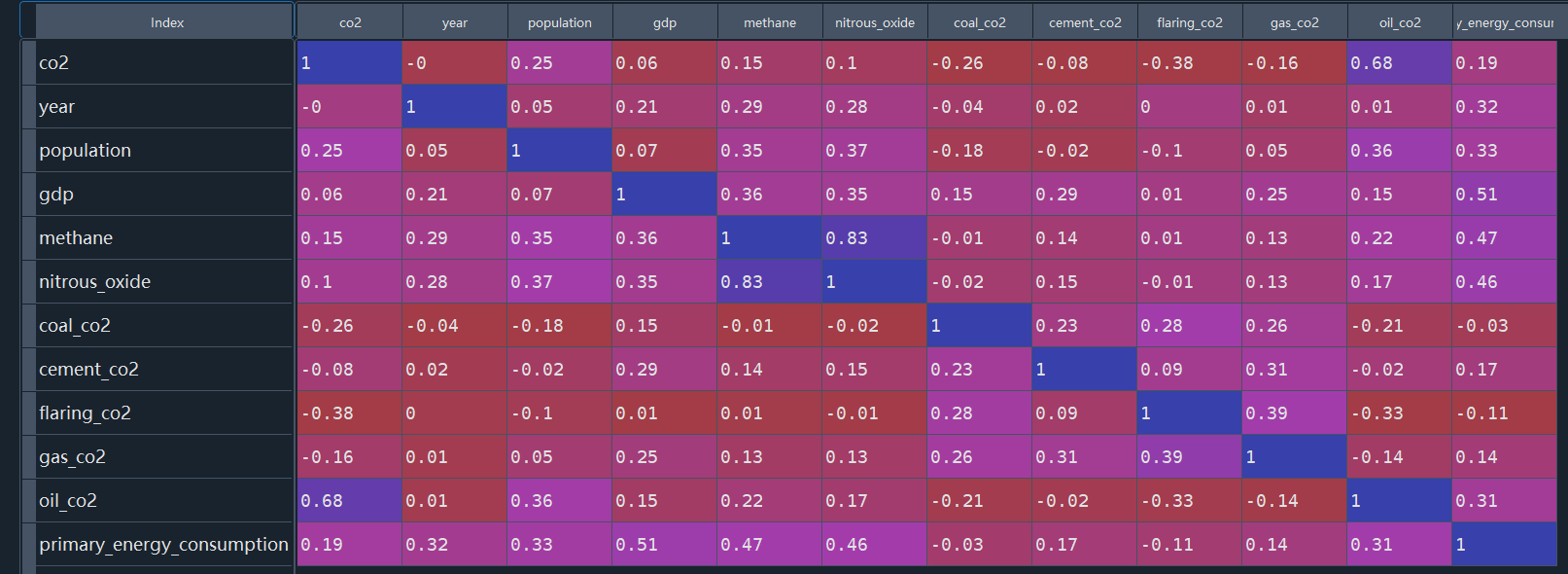
After this, it needs to ensure that the model can quickly and easily identify and resolve discrepancies or missing values between the data, and improve the efficient utilization of the data by applications and systems (Hansen, D., 2012). Since there are many missing values in the selected dataset, check and impute the selected attributes for null values, bad or inconsistencies data can help to improve the model accuracy for data mining. For null values, mean value is used to replace them. Outliers can be found by the following formula: greater than **upper=Q3+1.5 (Q3-Q1)** and less than **lower=Q1-1.5 (Q3-Q1)**, where Q3 is 75% (upper quartile) and Q1 is 25% (lower quartile). In this study, numbers that exceed these ranges are identified as data that need to be processed. For any bad data, first replaced them with null values, and then the data is reallocated using the mean value. (See Figure7).



*Figure 7: Before imputing vs After imputing null values and bad data*

# Implementation PROCESS:

For the two established objectives, the aim is to estimate future carbon dioxide content based on numerical variables, so it is necessary to make predictions for the target variables. Supervised learning is used in both objective1 and 2 as an exploratory analysis algorithm, learning mapping functions or models from training data so that the model can predict the expected output of a given input. In this study, by establishing a model, the correlation graph between the target variable and the independent variable is established according to the analysis results, so as to understand the correlation change between the actual value and the predicted value (See Figure8). The higher the degree of linear fitting, the higher the accuracy of the model.



*Figure 8: Correlation graph between target and explanatory variables*

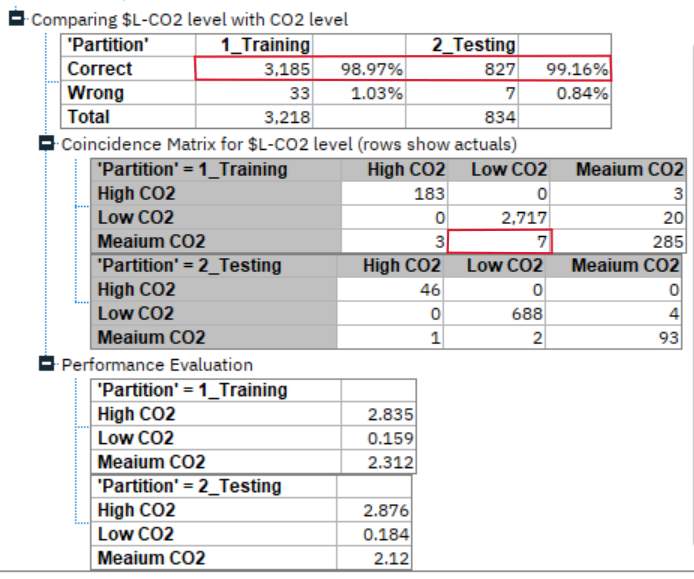
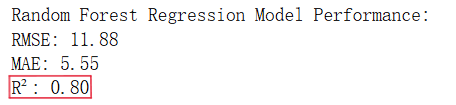
Since the experimental process is supported by three different analysis software, it is necessary to find the correlation (importance) between the target variable and the corresponding study variable in order to achieve the research purpose. For objective 1, the objective variable is continuous and the supervised learning method is used. **Regression** methods are used here as data mining method. For object 2, the data type is category rather than continuous due to the creation of the co2\_level classification target variable. Therefore, the **Classification** method is more suitable to use here.

After confirming the data mining model, this study needs to select the appropriate algorithm to obtain more accurate prediction results. Since each model contains a large number of algorithms, choosing the right data mining algorithm can get higher accuracy prediction.

* For objective1, building the model required some complexity, as it involved exploring the relationships between six independent and dependent variables. Compared to linear regression, random forest algorithms are good at learning complex and highly nonlinear relationships, often have high performance, and the results are clear and understandable. Therefore, for objective1, **Random Forest** algorithm might be more appropriate.
* For objective2, since the data type is multiple categories, multiple logistic regression can represent the probability of different CO2 levels when using fuels to produce carbon dioxide in different industries. In the previous step, null and extreme values of the required data have been imputed, so that logistic regression can better analyse the complete data structure and avoid the influence of extreme values on the model. Furthermore, due to the fact that coal, flash, cement, gas and oil variables are continuous and all belong to fuel types, there is a potential relationship between these variables. Logistic regression is good at analysing linear relationships, concise and easy to understand, and can prevent data overfitting. Therefore, it is more reasonable to choose **Logistic Regression** algorithm.

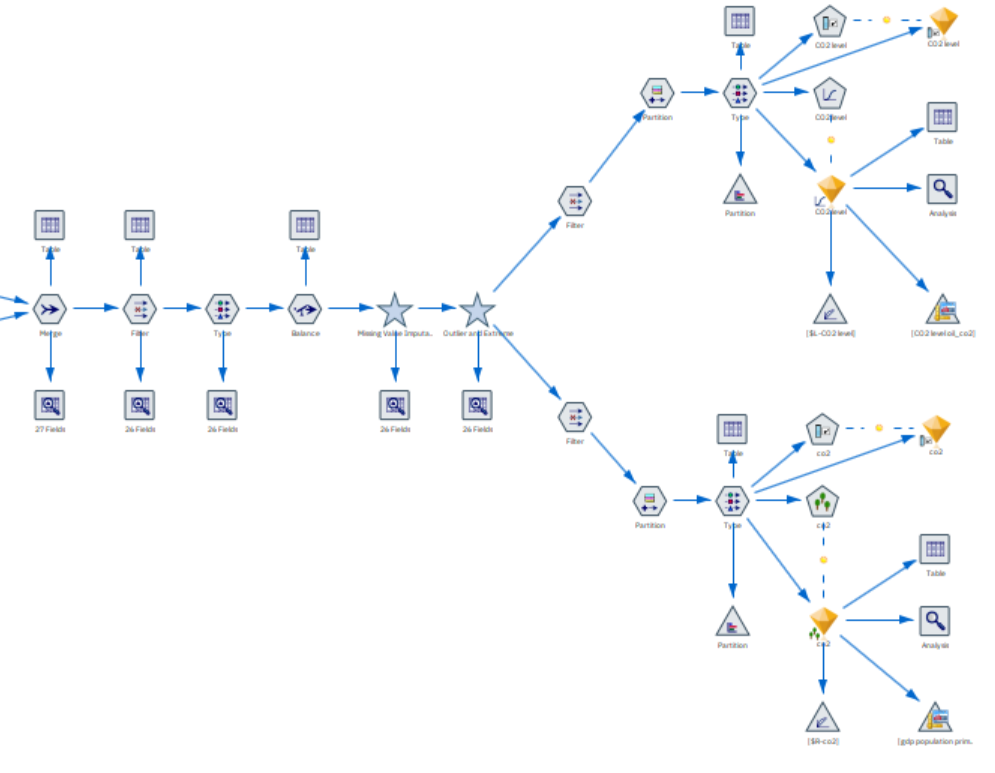
Next, Partitions were used for each model to evaluate the accuracy, sensitivity, and specificity of the model's performance between the test set and the training set. The training set is a data set used to train a model. During the training process, these parameters are constantly adjusted by the optimization algorithm, so that the model can better fit the training data to find the best model parameters. The testing set is mainly used to evaluate the generalization performance of the model. After the model has been trained and tuned using the training set and the validation set, the test set can be used to evaluate the performance of the model. The data of the test set is unknown to the model, so the test results can more accurately reflect the generalization ability and the final performance of the model. Based on this information, the study used **80%** of the samples for model training and **20%** for model testing to ensure that the total partition size is **100%**.

After construct models for each objective, model evaluation needs to conduct to test accuracy analysis. According to multiple data mining modelling iterations, the final accuracy of model 1 reached **80%**, and the accuracy of model 2 reached **99.16**% (See Figure9). These two evaluation results show that both models have good performance. The training results can be used to analyse and predict the real-world environment, and even help to predict the cross-project.



*Figure 9: Model evaluation for both models*

In order to clearly show the entire process of research from data acquisition, processing to model construction, SPSS Modeler's stream diagram is used to visually show the entire process (See Figure10). This helps to understand the relationship between the various steps and how to gradually achieve the research objectives.



*Figure 10: Full Streamline of data mining research*

# RESULT AND ANALYSIS:

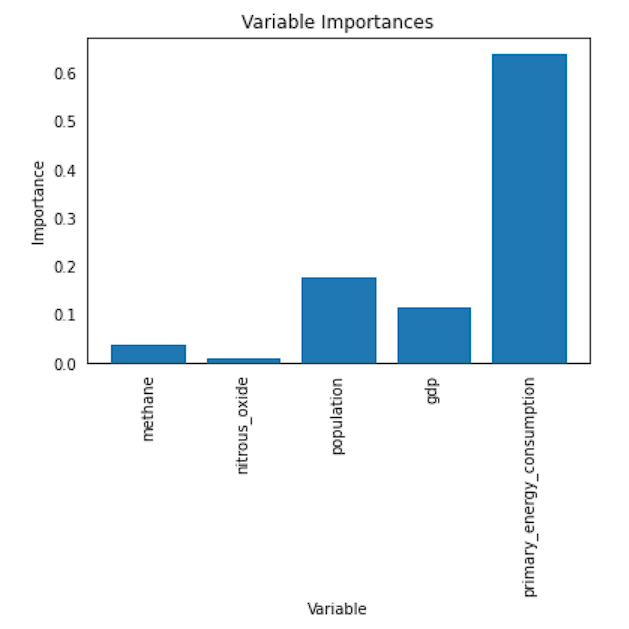
**For Objective 1**:

As a result of the random forest model, the bar chart provides an analytical operation to study the importance of variables. The following is an explanation of how to turn this visualization image into a knowledge of the effect of the carbon dioxide emission index (See Figure11).

Consistent with the estimates, energy production and consumption are the largest sources of carbon dioxide emissions. Burning fossil fuels such as coal, oil and gas is one of the main sources of greenhouse gas emissions. At the same time, industry and transportation are the main areas of energy consumption, and transportation requires a large amount of fuel, which leads to a large amount of carbon dioxide emissions.

The relationship between population and GDP and carbon dioxide content is the second. Population growth means more demand for energy, food and shelter, which leads to more energy consumption and production activities, which in turn increases carbon dioxide emissions. In addition, as the population grows, so will the demand for urbanization and industrialization. At the same time, GDP growth is usually associated with increased economic activity and progress in industrialization, which means more production and consumption demand, leading to more energy consumption and carbon dioxide emissions.

Methane is a greenhouse gas with a more potent greenhouse effect than nitrous oxide. The greenhouse effect per ton of methane is generally higher than that per ton of nitrous oxide. Thus, while nitrous oxide emissions may be higher, methane emissions produce more carbon dioxide and have a more significant impact on climate change.



*Figure 11: Random Forest result analysis for objective1*

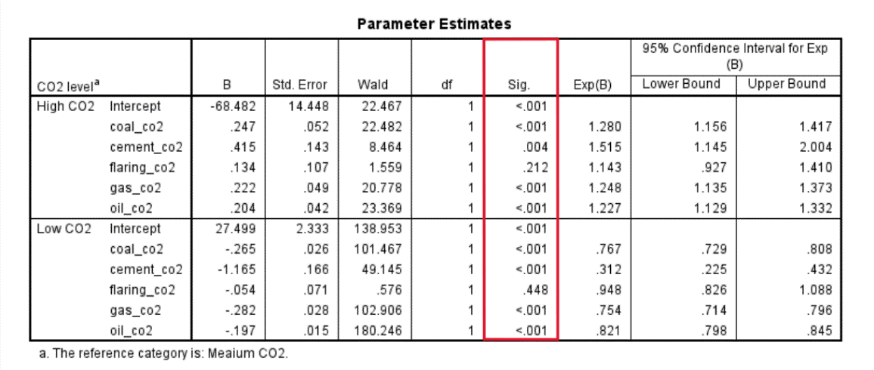
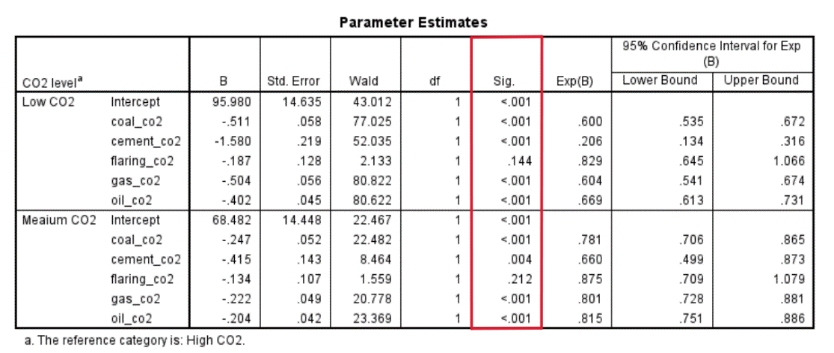
**For Objective 2**:

The parameter estimates, as the result generated by logistic regression model, provides all the necessary data analysis operations, but it is necessary to convert these data into knowledge that is easier to understand and make predictions (See Figure12).

According to the summary results, it can be seen that most industries have a significant impact on CO2 levels (P-value < 0.001). For example, the coal and oil industries burn coal and oil to produce energy, and in the process release large amounts of carbon dioxide gas into the atmosphere. In addition, industrial processes involving extraction, processing, refining, and transportation from transporters to processing plants release other greenhouse gases or pollutants that may also produce carbon dioxide when chemically reacting in the atmosphere.

For less significant segments, such as cement\_co2 (P-value < 0.05), while the industry reduces the consumption of natural gas, it still has some evidence that this industry has an impact on CO2 levels, with a higher likelihood of low CO2 levels increasing. Therefore, the implementation of mitigation measures in this industry can also effectively reduce carbon dioxide emissions but not as much as in other industries with significant impact.

The flaring industry is a special case. The use of flash is reduced, the likelihood of low CO2 levels is not significantly increased (P-value > 0.1), which is part of an interesting phenomenon that deserves further study. Therefore, as industries other than flaring reduce production and emissions of these substances, CO2 levels will gradually shift from low to high.

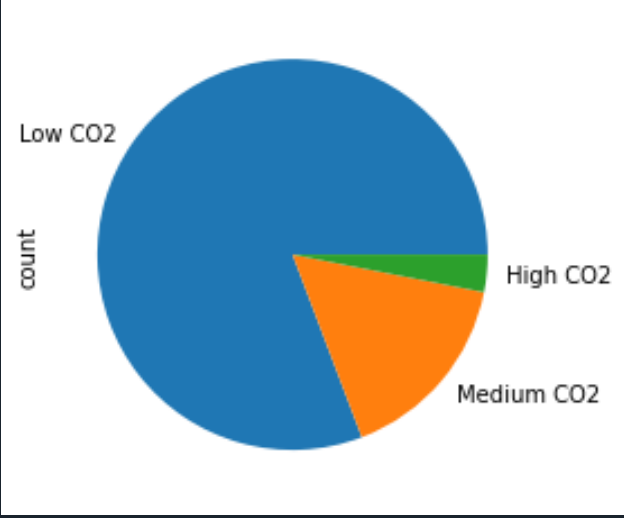


*Figure 12: Logistic Regression result analysis for objective2*

# DISCUSS the result:

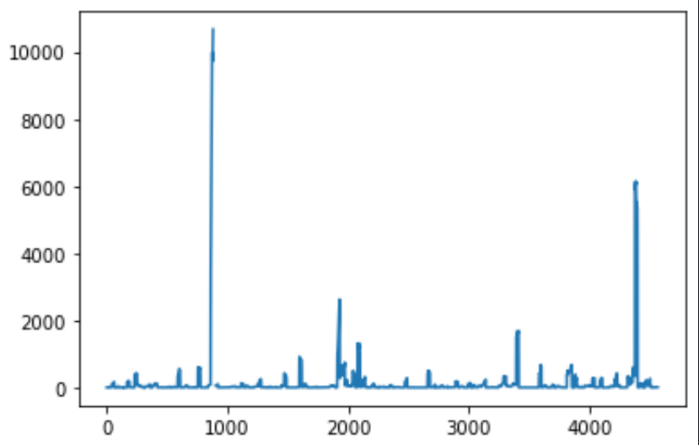
In order to discuss the research results with the actual situation. Visualizations are used here to judge the extent to which the forecast results conform to trends around the world over the past decade.

A pie chart based on the data set shows that most countries have low level of carbon dioxide, more than three-quarters of the range, while the proportion of high level is very small (See Figure13). However, this does not mean that global average carbon dioxide emissions are at a low level. Since the previously set range of levels is limited to 1000 and 3000, and carbon dioxide emissions at 1000 have already reached a higher level, overall, mitigation measures are still needed to reduce the greenhouse gas effect.



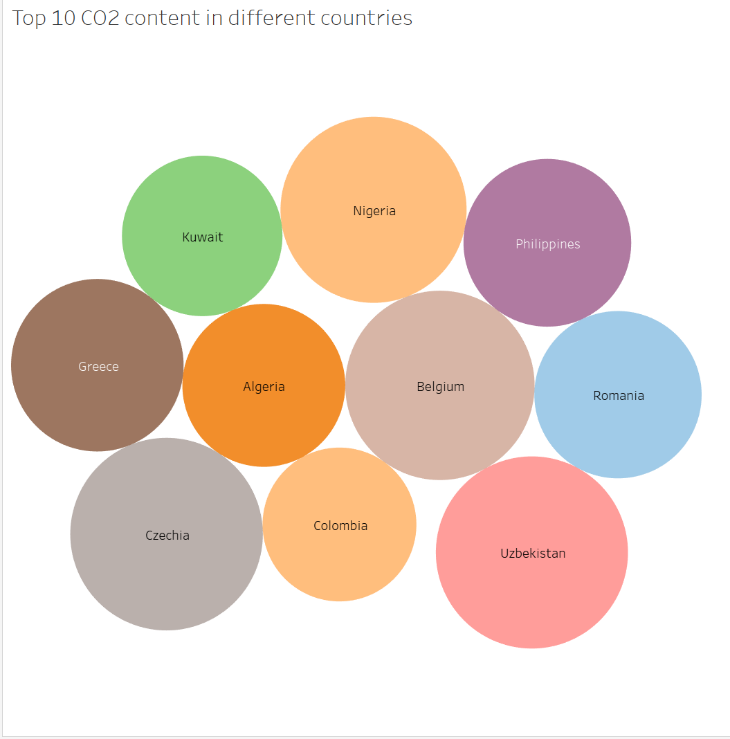
*Figure 13: Pie chart of different co2 level proportions*

The frequency line diagram overall shows a right-skew trend in carbon dioxide emissions and frequency, showing changes in carbon dioxide emissions from 0 to 5000 and frequency from 0 to over 10,000 (See Figure14). The highest distribution of carbon dioxide emissions is concentrated between 900 and 1000, with slight concentrations at 2000, 3500 and 5000 levels. Thus, this goes a long way towards explaining the global distribution of carbon dioxide emissions among countries.



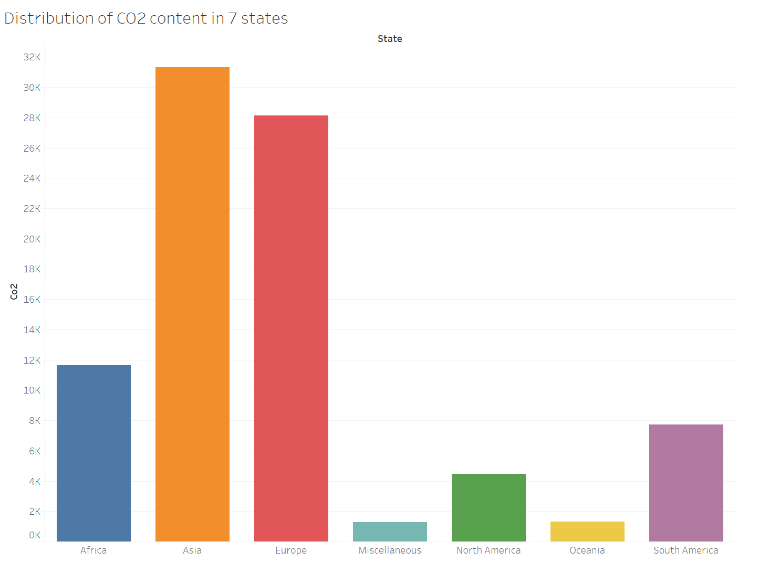
*Figure 14: Frequency line diagram of CO2 emission*

The Top 10 CO2 emissions images for different countries show the countries with the highest carbon dioxide emissions, showing that industrial and energy production activities in these countries contribute significantly to emissions. For example, Nigeria, Algeria and Kuwait generally have strong industrial sectors, so that their energy industries have the most significant impact on carbon dioxide emissions (See Figure15).



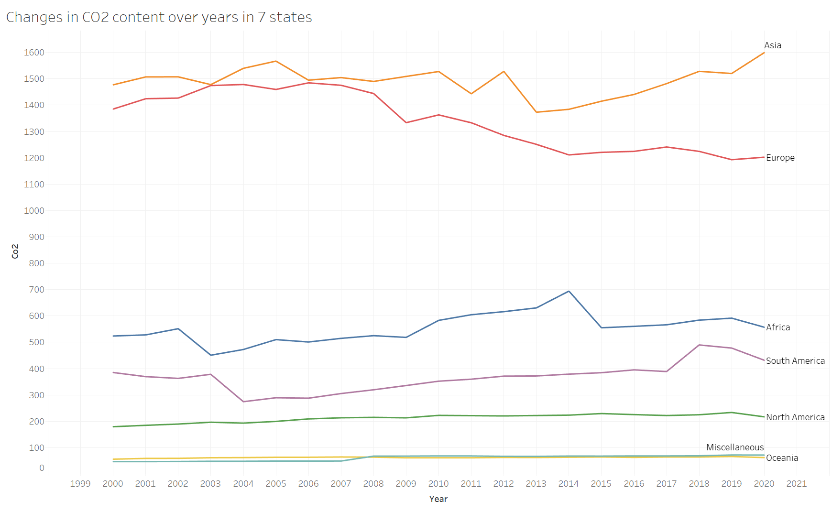
*Figure 15: Top 10 CO2 emission countries around world*

The distribution of CO2 emissions in the seven states shows the distribution of carbon dioxide emissions across different continents (See Figure16). With Asia and Europe leading the way, highlights the impact of industrial activity in these regions, including heavy industry and high energy consumption. At the same time, Asia and Europe have higher population and GDP per capita, so this is further evidence that population activity, industry and energy consumption have a significant impact on carbon dioxide emissions

**

*Figure 16: Distribution of CO2 content in 7 states*

A line chart of changes in carbon dioxide emissions for seven states over the years shows historical trends in carbon dioxide emissions (See Figure17). Notably, while trends were stable or declining in most regions, there is a slight uptick in Asia, indicating continued industrial growth and energy consumption. In other words, small increases in CO2 levels in Asia indicate that the region's industrial expansion and exploding population cannot reduce carbon dioxide emissions through mitigation strategies, whereas stable or declining trends in other regions may reflect successful mitigation strategies.



*Figure 17: Changes in CO2 content over years in 7 states*

Comparing the prediction results with current global data shows that the factors influencing carbon dioxide emissions are consistent with the explanatory variables included when the targets were originally set. Despite the increasing number of climate change mitigation policies and the gradual adoption of renewable energy sources around the world, fossil fuels will continue to play an important role in the world energy landscape for some time to come (Mac Dowell et al., 2017). Therefore, for countries dependent on the energy industry, anthropogenic carbon dioxide emissions will continue their recent growth trend. At the same time, population size is one of the main factors affecting carbon dioxide emissions, China and India's committed emissions per unit of GDP is much higher than most developed countries (e.g. the United States, Europe and Japan), indicating that in regions where industrialization is ongoing but not yet completed (Davis et al., 2010). The larger the population, the higher the GDP index, indicating the greatest inertia in infrastructure emissions.

# CONCLUSION:

Through the analysis of the data and the construction of the model, and comparing the model prediction results with the distribution of CO2 emissions in different regions of the world, it is found that Asia and Europe have the highest CO2 emissions, which is closely related to the high population density and high level of economic activity in these regions. The use of random forest models effectively shows that energy production and consumption are the main sources of CO2 emissions, especially the burning of fossil fuels. In addition, population growth and GDP growth are also closely related to CO2 emissions, reflecting the environmental impact of economic activities. Analysis of methane and nitrous oxide showed that despite their lower emissions, the impact on climate change is still present. The Logistic regression model divides the energy sector into five categories, and deeply studies the relationship between various energy production and carbon dioxide emissions. It is found that coal, oil and natural gas industries have a significant impact on carbon dioxide emissions, among which oil and carbon dioxide have the highest positive correlation.

In order to translate the findings into practical actions to reduce CO2 emissions, the following initiatives may useful:

* Energy production industries need to be focused on the coal, oil and gas sectors, and driving the transition to renewable energy sources such as wind, solar and hydropower, which will significantly reduce carbon dioxide emissions in high-emitting regions.
* Implement effective energy management strategies to promote energy efficiency and energy-efficient technologies in regions with rapidly growing populations and economies, particularly in Asia and Europe.
* Due to the need to increase people's understanding and attitudes towards climate change and climate science (Joslyn et al., 2021). It is also necessary to explain the impact of climate change with specific figures and relevant cases to call for sustainable development activities around the world to reduce the impact of population and GDP growth.
* Global efforts to reduce emissions should focus on the largest emitters. Through international cooperation and technology sharing, to develop and implement strict emission reduction policies and encourage them to adopt sustainable production and consumption patterns. Implementing these initiatives will help reduce CO2 emissions globally and mitigate the negative impacts of climate change.

Even if this study explores the relationship between carbon dioxide and human activities to a certain extent and gives appropriate mitigation measures, the factors affecting the global warming are not only from this. For example, periodic changes in solar activity affect the amount of solar radiation received by the Earth, which affects the Earth's climate and the greenhouse effect. Volcanic eruptions release large amounts of dust and gases (such as sulfuric dioxide) into the atmosphere to increase greenhouse gas concentrations. Forest fires and prairie fires not only release large amounts of carbon dioxide and other greenhouse gases, but also destroy vegetation and reduce the ability of carbon sinks. These natural factors work together with human activities to influence the intensity and variability of the greenhouse effect. Thus, exploring these factors also critical to accurately predicting and responding to climate change.

# REFERENCE:

Patz, J. A., Frumkin, H., Holloway, T., Vimont, D. J., & Haines, A. (2014). Climate change: Challenges and opportunities for global health. *JAMA - Journal of the American Medical Association*, *312*(15), 1565–1580. <https://doi.org/10.1001/jama.2014.13186>

Solomon, S., Plattner, G.-K., Knutti, R., & Friedlingstein, P. (2009). Irreversible Climate Change Due to Carbon Dioxide Emissions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 1704–1709. <http://www.jstor.org/stable/40421660>

López‐Moreno, J. I., Goyette, S., & Beniston, M. (2007). Climate change prediction over complex areas:

spatial variability of uncertainties and predictions over the Pyrenees from a set of regional climate models.

*International Journal of Climatology*, *28*(11), 1535–1550. <https://doi.org/10.1002/joc.1645>

Watson, A. J. (2007). Certainty and Uncertainty in Climate Change Predictions: What Use are Climate

Models? *Environmental & Resource Economics*, *39*(1), 37–44. <https://doi.org/10.1007/s10640-007-9173-9>

Mao, X., Ren, N., Dai, P., Jin, J., Wang, B., Kang, R., & Li, D. (2024). A variable weight combination

prediction model for climate in a greenhouse based on BiGRU-Attention and LightGBM. *Computers and*

*Electronics in Agriculture*, *219*, 108818. <https://doi.org/10.1016/j.compag.2024.108818>

Liu, Y., Li, D., Wan, S., Wang, F., Dou, W., Xu, X., Li, S., Ma, R., & Qi, L. (2022). A long short-term memory-based model for greenhouse climate prediction. *International Journal of Intelligent Systems*, *37*(1), 135–151. <https://doi.org/10.1002/int.22620>

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty years Later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. <https://doi.org/10.1109/tkde.2019.2962680>

Hansen, D. (2012, September). Big data detour ahead: new approaches in data integration. Database Trends & Applications, 26(3), 18+.

<https://link-gale-com.ezproxy.auckland.ac.nz/apps/doc/A302665032/AONE?u=learn&sid=bookmark-AONE&xid=e0f761aa>

Mac Dowell, N., Fennell, P. S., Shah, N., & Maitland, G. C. (2017). The role of CO2 capture and utilization in mitigating climate change. *Nature Climate Change*, *7*(4), 243–249. <https://doi.org/10.1038/nclimate3231>

Davis, S. J., Caldeira, K., & Matthews, H. D. (2010). Future CO 2 Emissions and Climate Change from Existing Energy Infrastructure. *Science*, *329*(5997), 1330–1333. https://doi.org/10.1126/science.1188566

Joslyn, S., & Demnitz, R. (2021). Explaining How Long CO2 Stays in the Atmosphere: Does It Change Attitudes Toward Climate Change? *Journal of Experimental Psychology: Applied*, *27*(3), 473–484. <https://doi.org/10.1037/xap0000347>