***CA2 Strategic Thinking***

**Applied machine learning to estimate CO2 adsorption in different materials**

*By*

*Rachel Mae Lademora – 2023396*

*Higher Diploma in Science in Data Analytics for Business*

*Strategic Thinking*

*James Garza*

*CCT College*

*Dublin, Ireland*

CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 1 – Capstone Project Proposal |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Rachel Mae Lademora |
| **Student Number:** | 2023396 |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | 17th December 2023 |
| **Date of Submission:** |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

[Introduction 4](#_Toc149304071)

[Objectives 5](#_Toc149304072)

[Problem Definition 5](#_Toc149304073)

[Scope 5](#_Toc149304074)

[Data Source 6](#_Toc149304075)

[Ethical Considerations 6](#_Toc149304076)

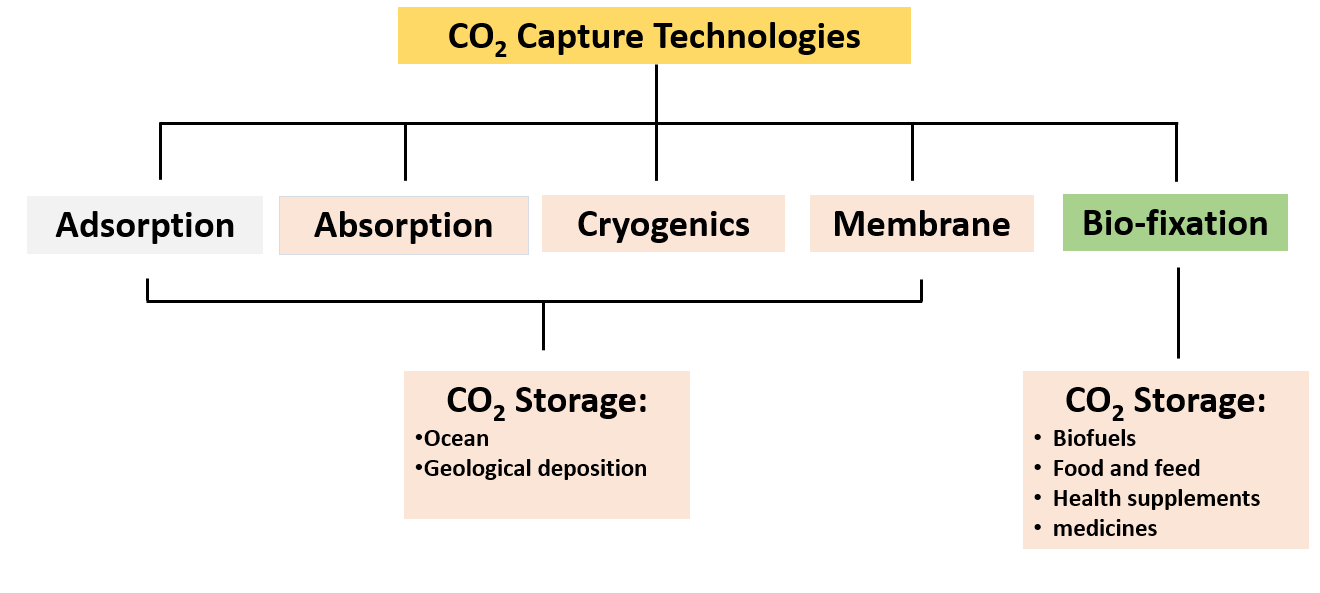
[Appendix 1 7](#_Toc149304077)

[References 10](#_Toc149304079)

# Introduction

Carbon capture and storage have been of extreme importance, as the concentration of this gas has increased, and data has shown that electricity generation will increase by 69% in 2040, with coal-based generation rising by 25% until 2040, achieving high levels of carbon dioxide (CO2) emissions. (Huetteman, Bowman and Slater-Thompson, 2016, p.81; Ren and Liu, 2023, p.1)

The mitigation of CO2 remotion from the atmosphere to reduce the greenhouse effects has become the focus of universities, government, and private institutions; furthermore, developing technologies and products to capture CO2 has gained more power. (Ren and Liu, 2023, p.1). Asif *et al*. (2018, p.4) and Daneshvar *et al*. (2022, p.6) have shown the general panorama of all techniques that we have so far (Figure 1.)



**Figure 1:** Techniques for CO2 capture.

For this capstone project proposal, we will focus on the adsorption technique. This promising method has gained notable attention due to its low operational cost, lower energy demand, ease of handling, and general reliability. (Daneshvar *et al*., 2022, p.4)

Regarding the increase in the technologies to fix CO2, we propose implementing machine learning model to evaluate which one can efficiently estimate CO2 capture. After we obtain our results, it might be possible to understand which material is more efficient and, consequently, which one should be focused on to reduce production costs and become more economically affordable.

# Objectives

Principal objective:

We propose assessing some machine learning models to predict CO2 adsorption in different materials, such as rice rusk, activated carbons, and carbon nanotubes, considering data from academic papers.

Secondary objectives:

* Identify the principal materials used in the adsorption technique.
* Use the Exploratory Data Analysis (EDA) to understand the data.
* Apply machine learning models to choose one with a high accuracy score. I would suggest that I will use the low accuracy score of materials to avoid overfitting of the model.
* Implement the chosen ML model to predict CO2 adsorption in these materials.
* I would like to suggest that I will need to change the title of our capstone project in Predicting Co2 Adsorption materials using one machine Learning Algorithms for my CA2.

# Problem Definition

Considering that the emission of CO2 is a global problem, government and private institutions are investing in research to find solutions in this field. Consequently, we expect to have a lot of data, however, we might face difficulties finding articles with comparable experimental conditions. The second problem might be in choosing the most suitable machine learning model.

Therefore, the result of this project tends to be important as it aims to provide information on which material has the best potential to be invested in and improved by public and private institutions for the removal of CO2.

# Problem Definition

Considering that the emission of CO2 is a global problem, government and private institutions are investing in research to find solutions in this field. Consequently, we expect to have a lot of data, however, we might face difficulties finding articles with comparable experimental conditions. The second problem might be in choosing the most suitable machine learning model.

Therefore, the result of this project tends to be important as it aims to provide information on which material has the best potential to be invested in and improved by public and private institutions for the removal of CO2.

# Scope

This project aims to assess some machine learning models to predict CO2 adsorption in different materials. We will give an actual situation of the research in this field, with a compilation of articles analyzed by descriptive statistics. Secondly, we will implement a ML model, such as a decision tree, k-NN, and regression.

The first boundary would be using data from exclusively adsorption technique, not others. This project will be limited by data with a maximum of five years. The third frontier is to use academic papers from reliable sources to maintain data credibility. Lastly, we will gather data with the same variables to compare the materials fairly.

At the end of this project, we expect to deliver an in-depth analysis of this technique with descriptive statistics and implement a ML model to predict CO2 adsorption in different materials. To achieve our goals, we break down the project into phases and delegate tasks as proposed in the Cross Industry Standard Process for Data Mining (CRISP-DM) method. In Appendix 1, we are presenting the timeline for this project.

Our action plan is to search for data with comparable experimental conditions, such as pressure and temperature. We will combine this data into an Excel spreadsheet. Later, we will work with Python using Jupyter Notebook to clean and organize the data and create graphs to have an overview of our data. After we have a well-developed EDA, we will test some ML models in Python to choose the best one.

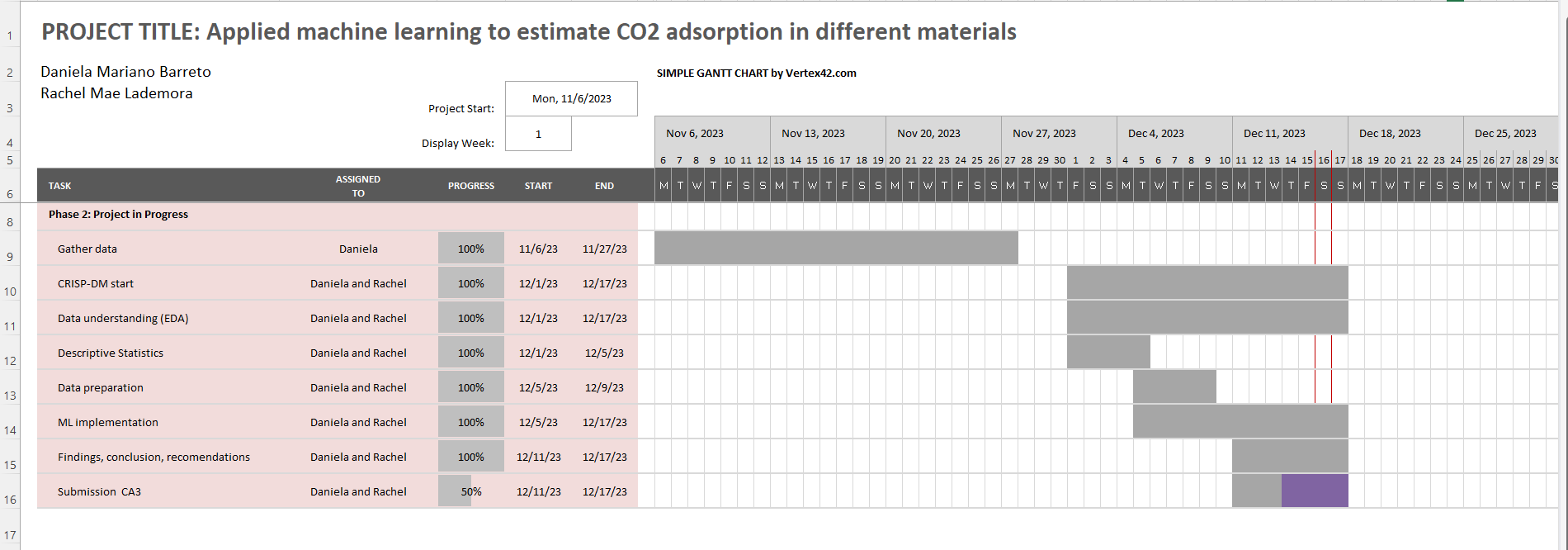
# Data Source

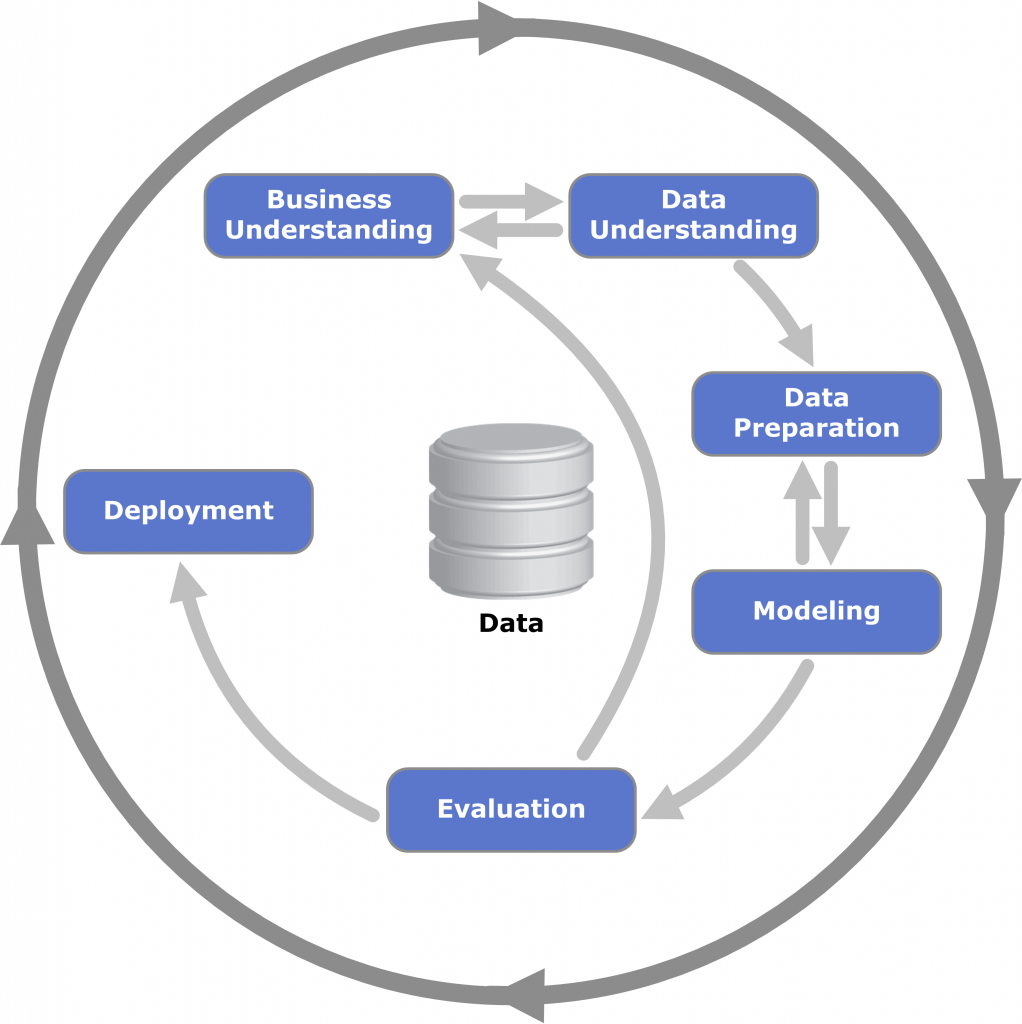
We will obtain data from academic papers in compiling at least 50 records for materials used in the adsorption CO2 capture technique. We will maintain the PDF versions of the articles from which we took data as proof of permission to access the data, and all will be appropriately referenced.

# Ethical Considerations

We will not intend to work with data that involves sensitive data, user privacy or potential social impacts; our data will be essential from laboratory research.

Appendix 1





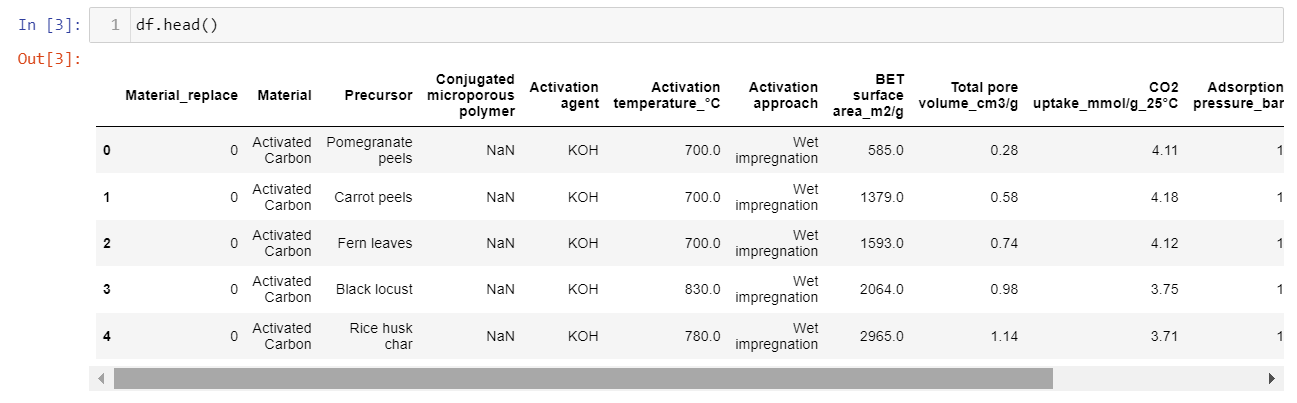
As we mention in our CA 1 enable for us to achieve our goals for the continuation of our capstone project, I will use the Cross Industry Standard Process for Data Mining (CRISP-DM) method process as a guid for the step-by-step process of data analysis of the CO2 adsorption dataset.

**Business Understanding**

The major problem that we see is the rapidly increase of the CO2 adsorption in the atmosphere that cause the greenhouse effect get worse. I will propose to used a natural material to reduce the CO2 adsorption using the accurate materials that I tested upon doing my data analysis of one model in machine learning.

In terms of business goals, it is vital to used the natural materials since it is not too expensive to use and it is more environmentally friendly, that will help to reduce the greenhouse effect. The business companies can use it first for testing’s and after they will figure it out that the materials can effectively reduce the effect of greenhouse effect and help to lessen the cost price. The business companies can offer it to other companies with a good price and after that they can sell it also to the customers in affordable price. It will be both beneficial to environment and in the business industry since it will increase their profit sales.

**Data Understanding**

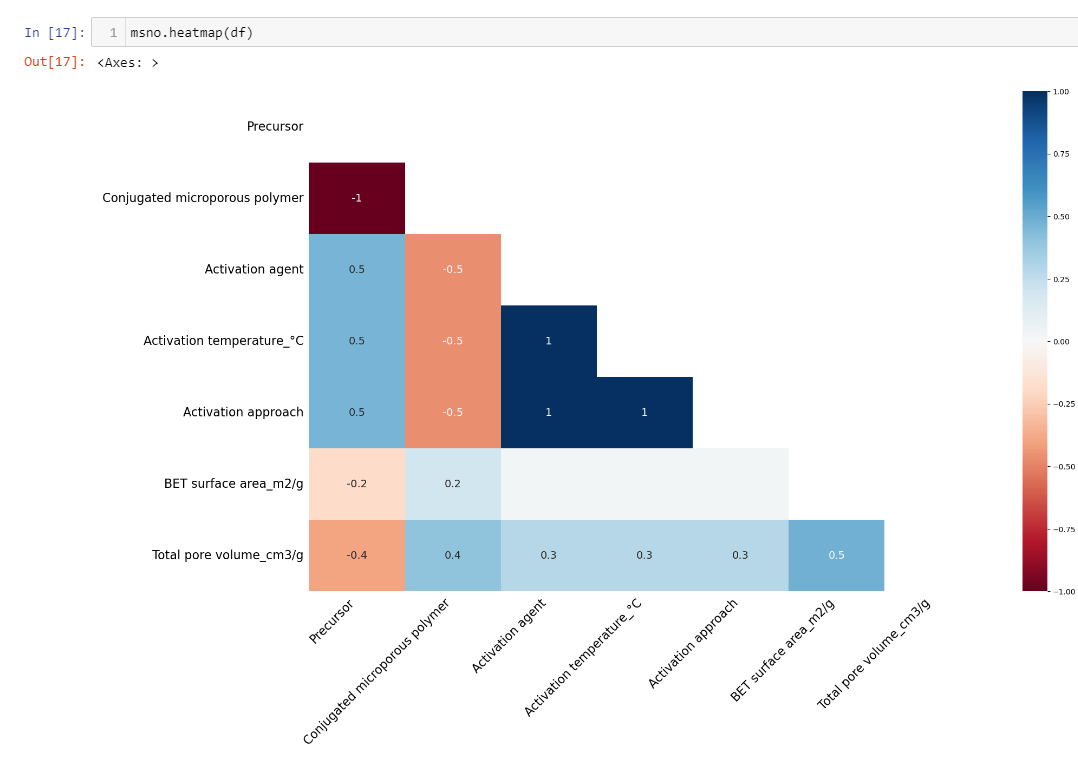
The dataset consists natural materials which is very useful to have a lower effect of adsorption to reduce the greenhouse effect. To better understand the dataset this are the dataframe head and data dictionary with definitions and the data type that the CO2 adsorption dataset. 

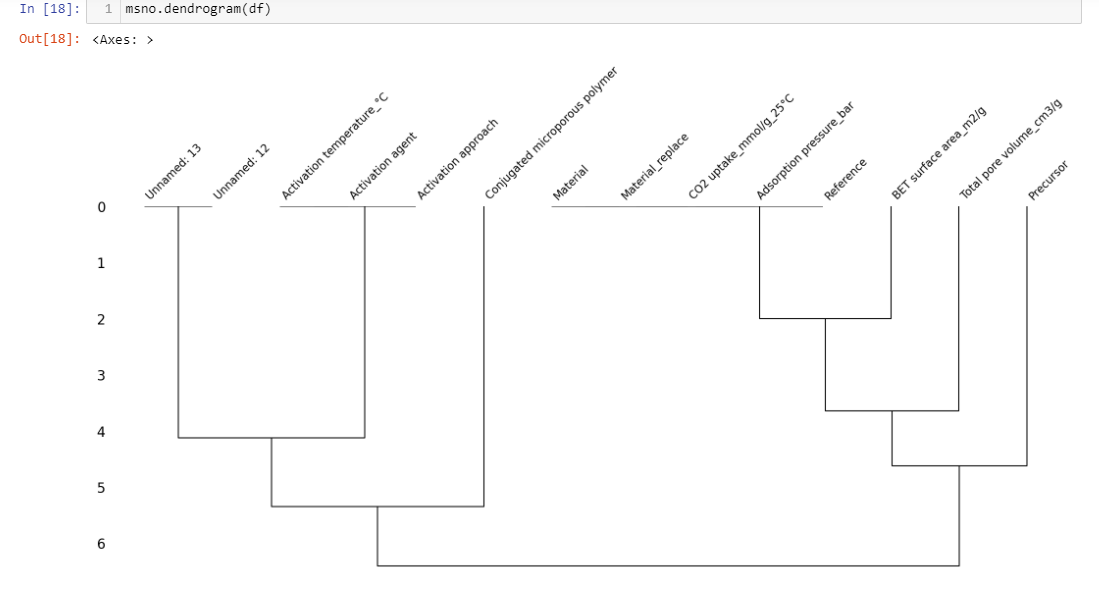
Data Dictionary

|  |  |  |
| --- | --- | --- |
| Columns name | Definition | Data Types |
| Material Replace | Code for the class material. | Int64 |
| Material | Material used in the CO2 adsorption process. | object |
| Precursor | Is the material used in the initial adsorption process. | object |
| Conjugated microporous polymer | Chemical used in polymers material in the adsorption process. | object |
| Activation Agent | Chemical compound used in the carbon- based material. | object |
| Activation\_Temperature\_oC | Temperature used to agent activation in carbon-based material. | float64 |
| Activation approach | Technique used in the activation in carbon-based material. | object |
| BET surface area\_m2/g | Surface area of the material used in the CO2 adsorption. | float64 |
| Total pore volume\_cm3/g | Volume of the pore material used in the CO2 adsorption in 25oC. | float64 |
| CO2 uptake mmol/g\_25\_ oC | CO2 adsorption in 25oC. | float64 |
| Adsorption pressure\_bar | Pressure of CO2 adsorption. | int64 |
| Reference | Reference for academic papers results (Dziejarski et al., 2023, p.69-74). | object |

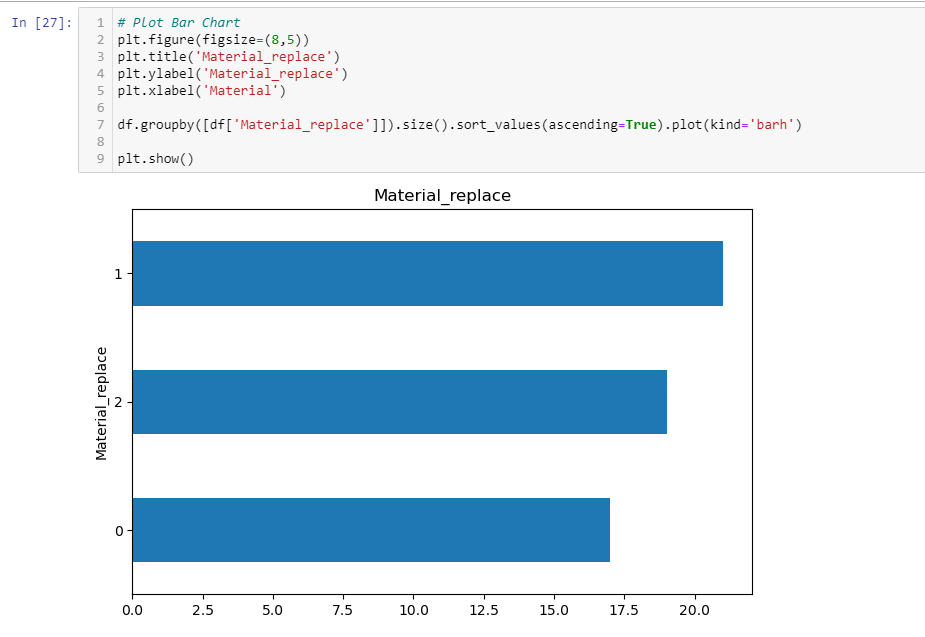
**Exploratory Data Analysis**

Exploratory data analysis is necessary step in understanding the dataset before execute the machine learning to produce insight to visualize the correlation of the features and determine the missing values in the columns.





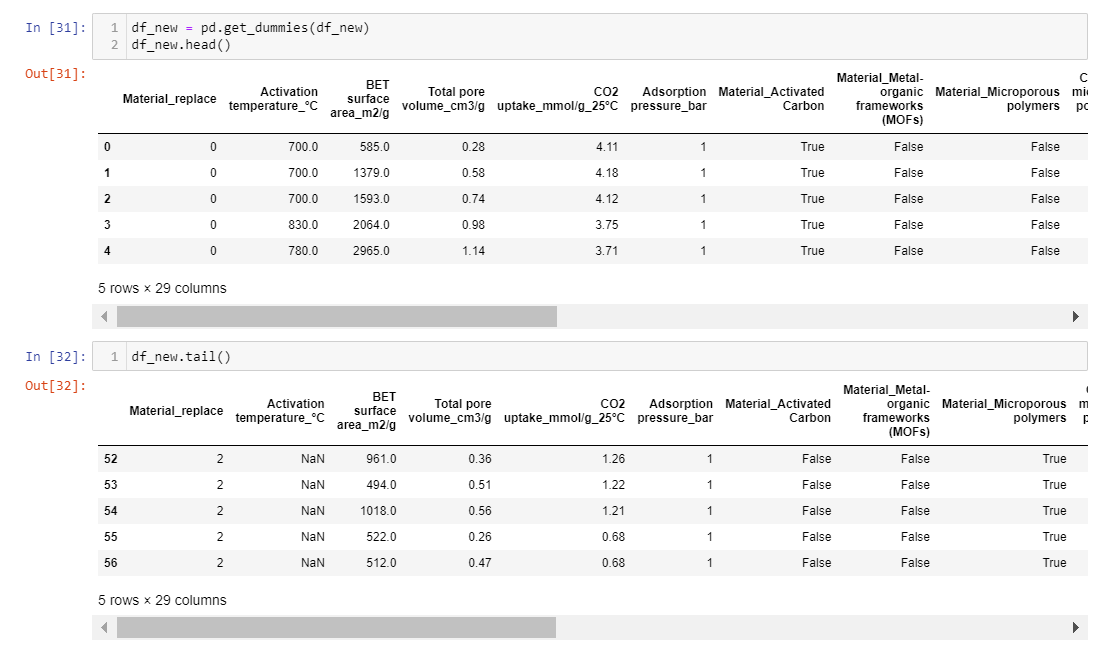
Using the libraries of import missingno as msno is the fundamental method to execute the visualization of the missing data and the correlation of the columns upon using the heatmap to determine which columns has a high result of missing values. The conjugated microporous polymer has -1 values it means that it has a lot of missing values on it. On the dendrogram it is clearly stated that there is no null values in material, material replace and CO2 uptake\_mmol/g\_25°C. The conjugated microporous polymer has more missing values than the unnamed 12 and 13 columns, activation temperature\_°C, and activation agent. The adsorption pressure\_bar and the BET surface area\_m2/g has a missing value. While the Total pore volume\_cm 3/g and Precursor has a missing value.

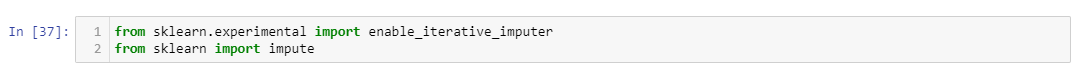


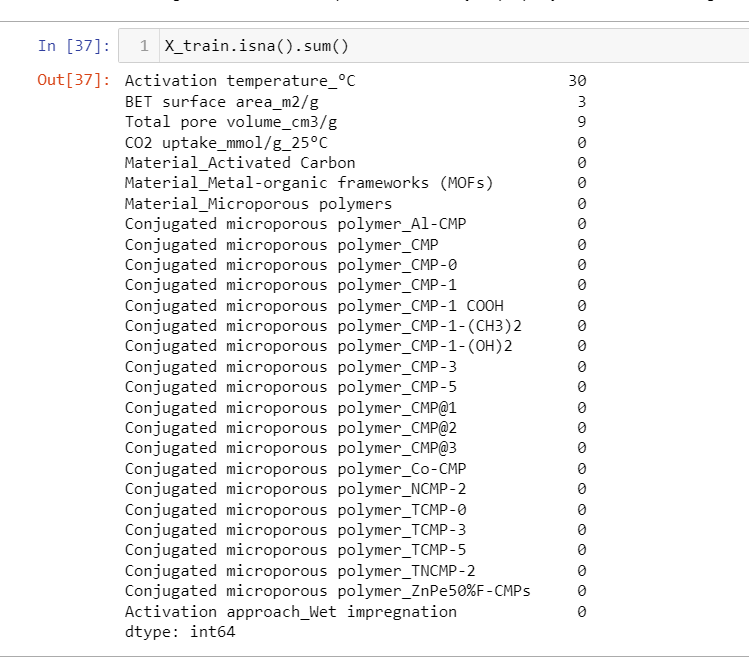
The material replace is undergo 3 processes in the y axis. In the y axis it is shows that number 1 has the high material replace with a value of 20.0 in the x axis. There is more replace process taken on the number 1 material than the other two materials.

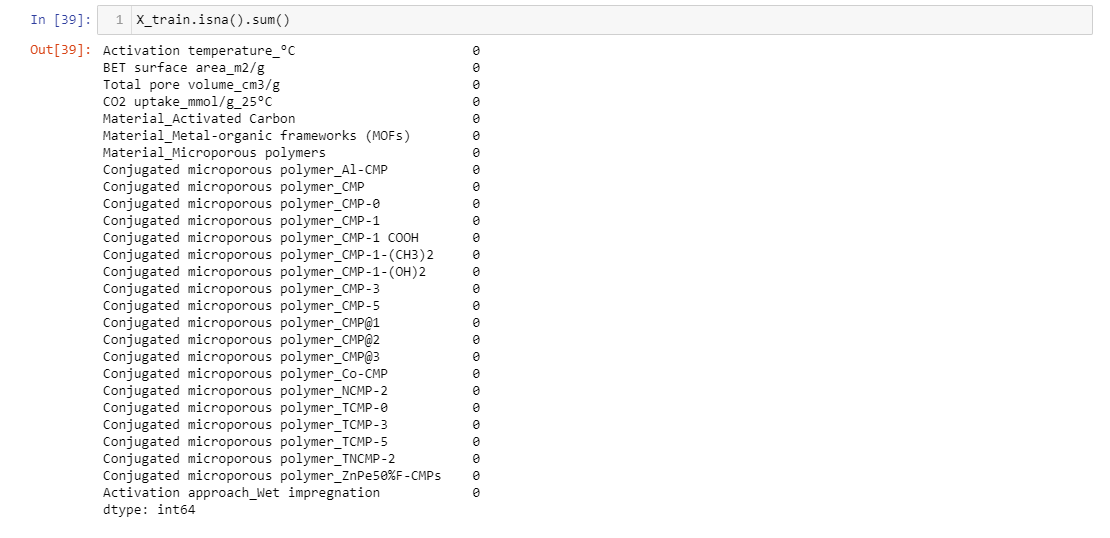
**Data Preparation**

The data preparation is very important because the dataset need to be clean first before performing the machine learning. The one hot encoding has the ability to change the categorical variables to numerical variables. But it shows that is not successfully transform to numerical variables. Because the material activated carbon and the rest of the materials are only transformed by true or false. That is why I find another library which I can use to convert the categorical columns to numerical. (Ganji, 2019)



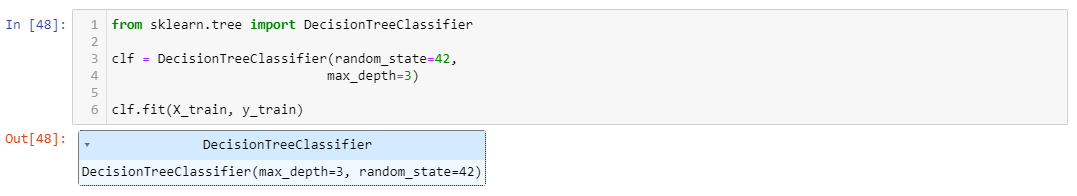


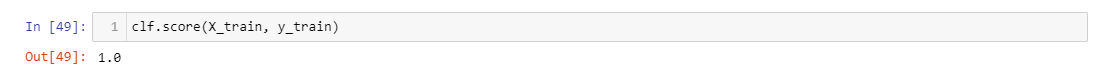




I decide to used the impute the data to transform the categorical to numerical variables before I will perform the machine learning model, because machine learning only read numbers. It is a very important to do the data cleaning first so that I will have a good test accuracy score as a result of training and testing the model. It is safe to use the impute data because it will prevent to loss data that is important columns despite the fact it is incomplete. (scikit-learn.org, n.d.)

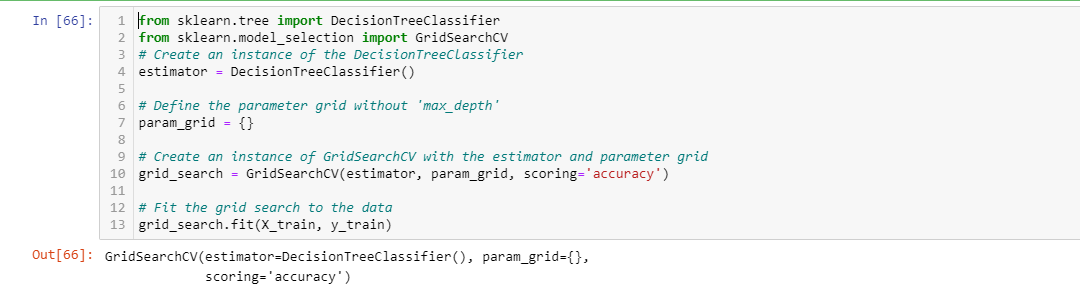
Modeling and Evaluation

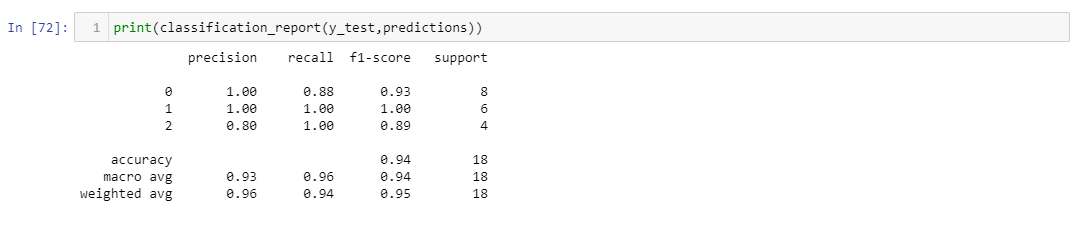


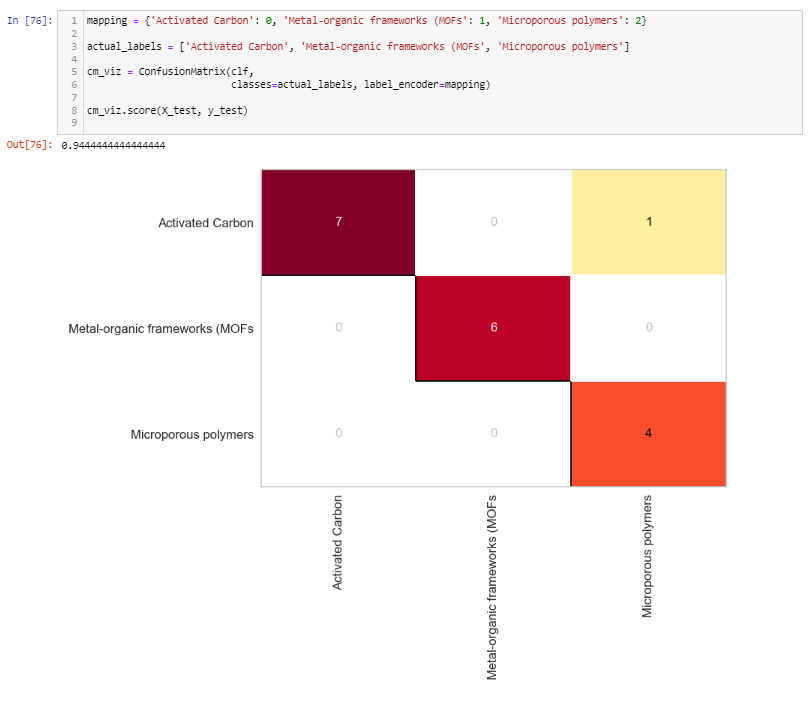




The Decision tree classifier gives me the 1.0 accuracy score. But I think this accuracy score will get my data analysis overfitting. That is why I decide to use the alternative method to help reduce the overfitting. Using the GridSearchCV and the k-folds can help reduce the overfitting of the model accuracy. As the max depth increase it will decrease the overfitting and improve the model accuracy of materials value.







This is the result of using the GridSearchCV it improves to reduce the accuracy of the materials and it is very helpful to reduce the overfitting. As stated above the evaluation accuracy reduce to 0.94 in f-1 score column from 1.0. In the confusion metrix it is stated that there is a diagonal true predicted value and one expected value.

Deployment

Since the accuracy of models evaluated that it improves and reduce the model in overfitting. I will develop a production score plan before I will put it to the production environment. The companies can make a series of this to make sure the effectiveness of the materials to reduce the CO2 adsorption before it will introduce to use to the customers.

Conclusion:

In this project, my primary goal was to develop a CO2 adsorption with an accurate accuracy using the decision tree classifier. Which is focus on the 3 materials that undergo material replace. I use the decision tree classifier and I also used the gridsearchCV enable to reduce the evaluation score of the accuracy of the 3 materials. During the first implementation of the decision tree classifier I

# References CA1

Ehsan Daneshvar *et al*. (2022). Biologically-mediated carbon capture and utilization by microalgae towards sustainable CO2 biofixation and biomass valorization – A review. *Chemical Engineering Journal*, *427*. <https://doi.org/10.1016/j.cej.2021.130884>

Furao Ren and Weijun Liu (2023). Review of CO2 Adsorption Materials and Utilization Technology. *Catalysts*, *13*(8), 1176. <https://doi.org/10.3390/catal13081176>

Muhammad Asif *et al.* (2018). Post-combustion CO2 capture with chemical absorption and hybrid system: current status and challenges. In *Greenhouse Gases: Science and Technology* (Vol. 8, Issue 6, pp. 998–1031). John Wiley and Sons Inc. https://doi.org/10.1002/ghg.1823

Thaddeus Huetteman, Michelle Bowman and Nancy Slater-Thompson (2016). Electricity. In *International Energy Outlook 2016 With Projections to 2040* (pp. 81–100). Energy Information Administration (EIA).

References CA2

Ganji, L. (2019). *ML | One Hot Encoding to treat Categorical data parameters*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/.

scikit-learn.org. (n.d.). *6.4. Imputation of missing values — scikit-learn 0.22.2 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/impute.html.

saturncloud.io. (2023). How to Count NaN Values in a Pandas DataFrame Column | Saturn Cloud Blog. [online] Available at: <https://saturncloud.io/blog/how-to-count-nan-values-in-a-pandas-dataframe-column/#:~:text=The%20isna()%20function%20returns> [Accessed 13 Dec. 2023].

[www.w3schools.com](http://www.w3schools.com/). (n.d.). Python dir() Function. [online] Available at: <https://www.w3schools.com/python/ref_func_dir.asp> [Accessed 13 Dec. 2023].