***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:** | 17th December 2023 |

**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**

**Business Understanding………………………………………………………….5**

**Business Understanding………………………………………………………….6**

**Data Understanding………………………………………………………………...7**

**Data Preparation……………………………………………………………………12**

**Modelling…………………………………………………………………………….14**

**Evaluation……………………………………………………………………………17**

**Deployment………………………………………………………………………….17**

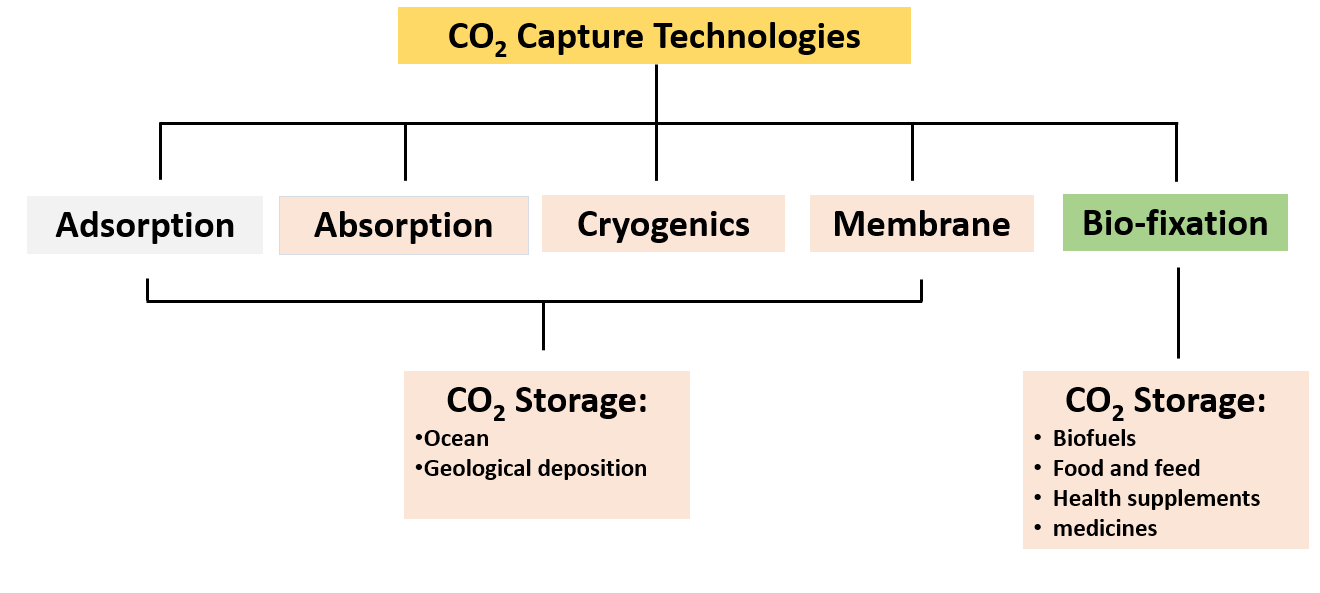
**Reference List……………………………………………………………………….18**

# 

# Introduction

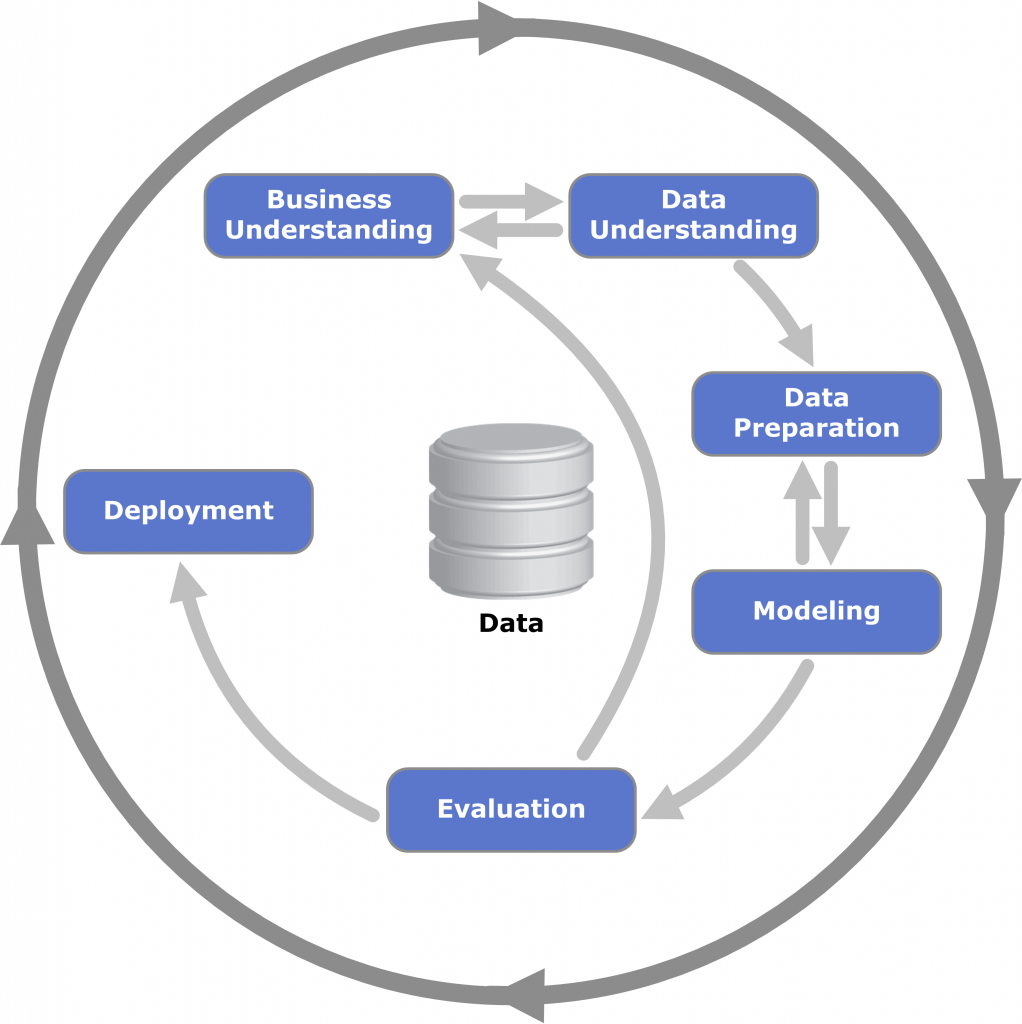
Carbon capture and storage have been of extreme importance, as the concentration of this gas has increased, and data has shown that electricity generation will increase by 69% in 2040, with coal-based generation rising by 25% until 2040, achieving high levels of carbon dioxide (CO2) emissions. (Huetteman,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



**Business Understanding**

Business understanding is the first phase of the CRISP-DM methodology. The main purpose is to well understand the business and its necessity, which is the objective and the requirements to use this project.

The principal objective of this project is to 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. The Hypothesis of this project is to understand which material is more efficient and, consequently, which one should be focused on to reduce production costs and become more economically affordable.

**The requirements of this project are:**

* 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 models.
* Implement the chosen ML model to predict CO2 adsorption in these materials.

General Goal:

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, a decision tree and k-NN.

The list of important tools and technologies that will be used in the project are:

* List of the Python Libraries:

1. Pandas
2. Numapy
3. Matplotlib
4. Seaborn
5. Missingno and etc.

* Modelling: Decision Tree Classifier- a machine learning model to test using of branching of nodes.

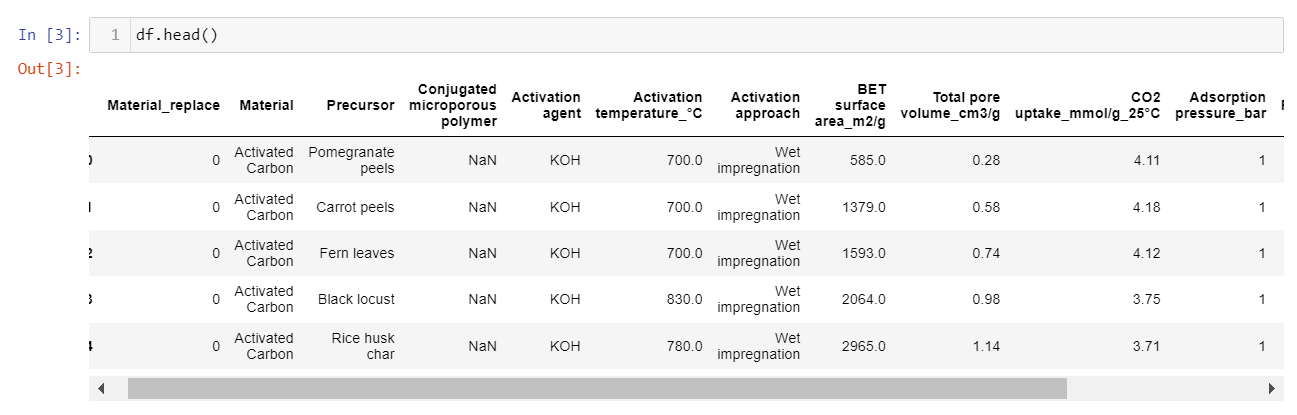
**Data Understanding**

Data understanding is the second phase of the CRISP-DM Methodology, It means to know about what I have understand to the data.

On this second stage of the CRISP-DM it is crucial to take some time to look every detail of the project data that is stated on the dataset. In order to avoid and encounter some errors when I will proceed to the data preparation. Data preparation is the vital part of the data analysis which is on the third phase of the CRISP-DM methodology. On that part I need to execute the data analysis like cleaning the data which vital before performing the modelling.

Our dataset is based on academic papers that compile at least 50 records for materials used in the adsorption CO2 capture technique and combine this data into an Excel spreadsheet. And the name of our dataset is CO2 adsorption.

As I use the df.head() function to begin to load the dataset, I see the first 5 rows of the dataframe and the 11 columns as visible in Figure 2. The conjugated microporous polymers have NAN in the first 5 rows.



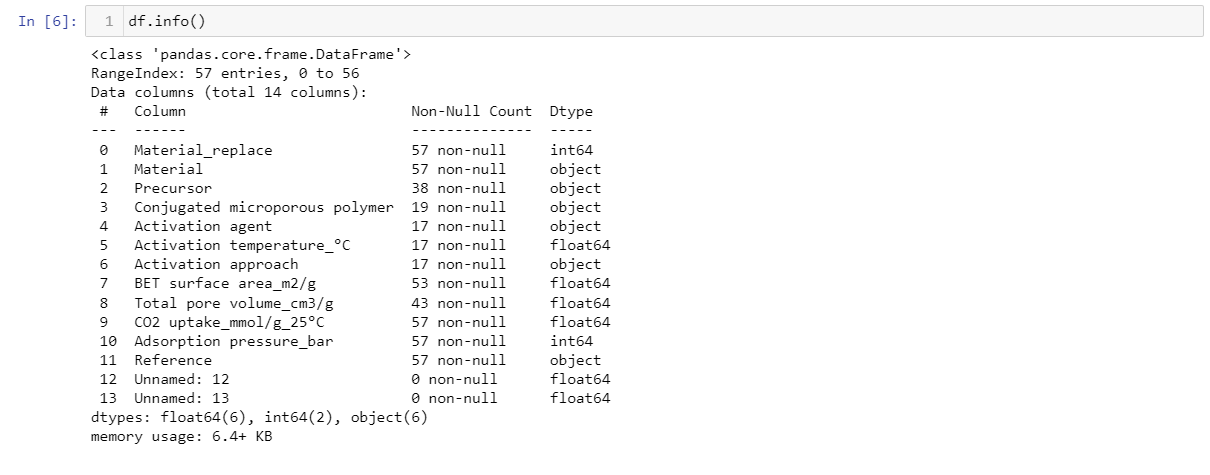
**Figure 2:** Head of the dataset.

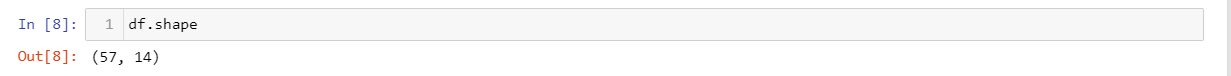
To better understand the dataset this is the data dictionary with definitions and the data type that the CO2 adsorption dataset has.

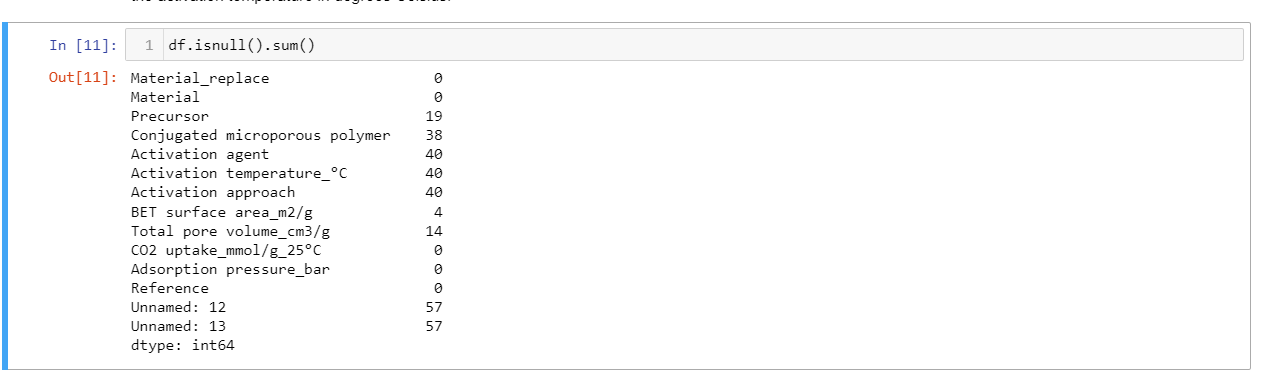
**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| 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 |

By using the df.info() function it will visualize the information about the dataframe. This will help me to understand if there are any data incompatibilities, like the unnamed columns. It also identifies the types of attributes in the dataset. There are 6 floats, 2 int64s, and 6 objects. the df.shape function will visualize the overall counts of the rows and columns of the C02 adsorption dataset, which have a total row of 57 and 14 columns.



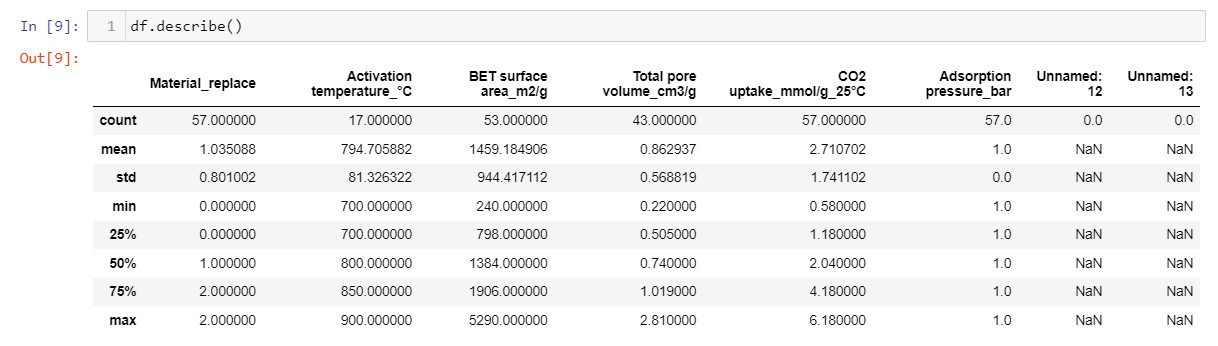
. **Figure 3:** Shape and info of the dataset.



**Figure 4:** isnull().sum() dataset.

In Figure 4 the df.isnull().sum() is used to see the total null values of each column. As stated, there are 0 null values in material replace and material. While the activation approach, activation temperature\_°C, and activation agent have 40 null values. The two unnamed columns have the complete 57 null values and this unnamed column needs to be dropped since this is not important to be used in machine learning.

Proceeding to the next step which is the summary statistics of the data frame which is the df.describe(). It is responsible for the central tendency dispersion and the std shows the amount of change in the data and determines how it expands the values that come from the mean. The min shows the values of each column from higher to lower. Also, it helped me understand the numerical columns properly. (pandas.pydata.org, n.d.).

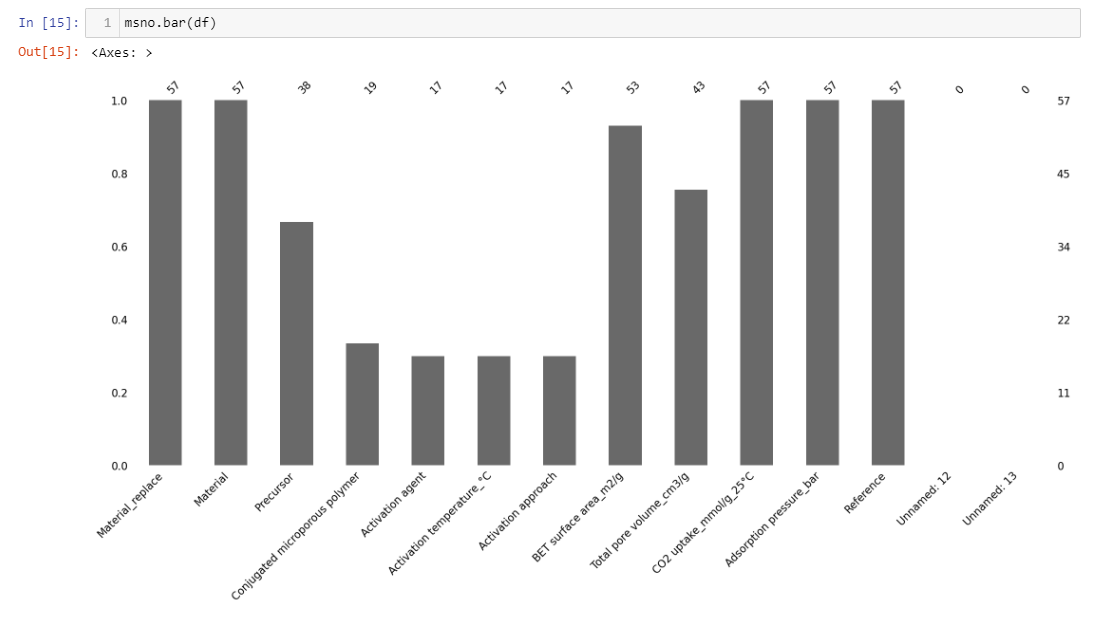


**Figure 5:** describe of the dataset.

In Figure 5, it is detailed to show that “Material\_replace” is capable of finding values with a range of 0.000000 to 2.000000. Since the min  has a value of 0.000000 value is low or small, it means that the standard deviation is closer to the mean and not scattered.

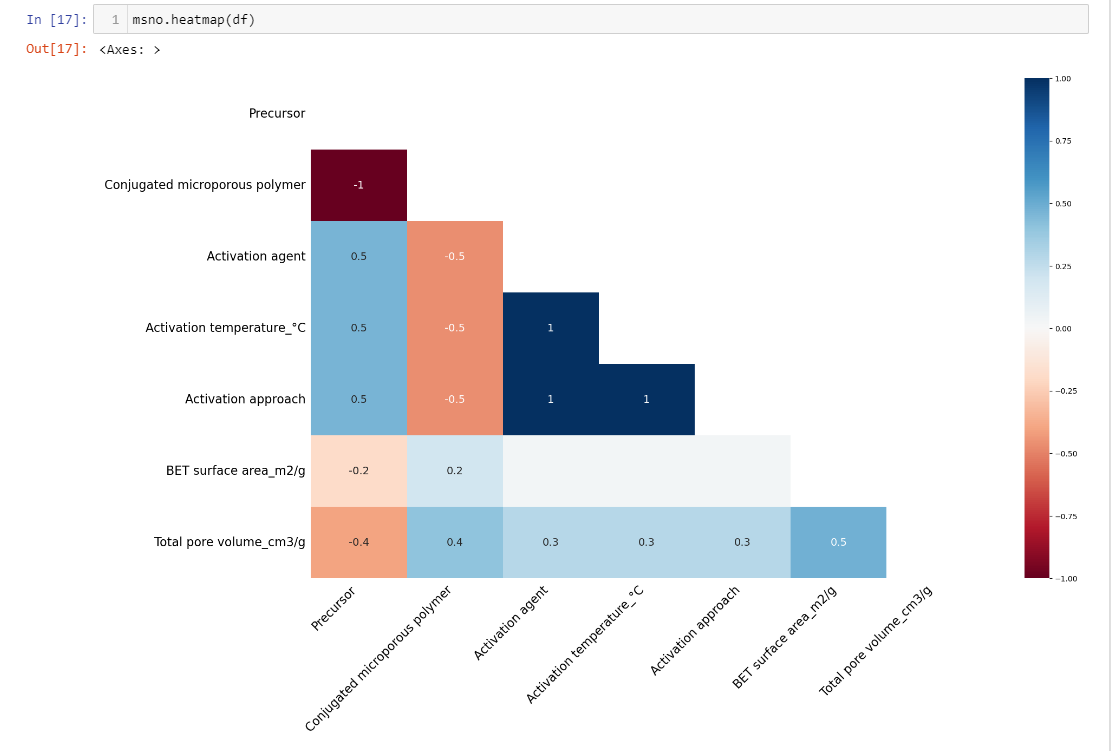
As stated in Figure 5, there are two unnamed columns that are not necessary to use for further data analysis, and these two unnamed columns need to be dropped since they have an NAN value, which is not useful for a machine learning model.

To easily understand the missing values of the CO2 adsorption dataset, here is an informative graph that will show the correlation of the missing values in the columns.



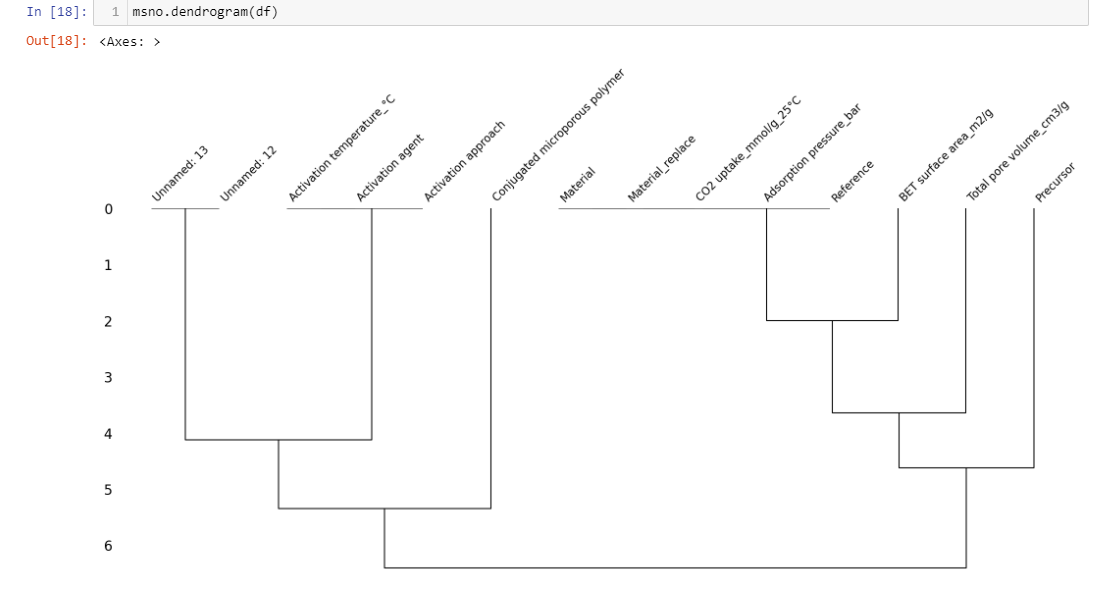
**Figure 6:** Bar graph missing value of dataset.

The dataset of CO2 adsorption has missing values. The bar graph shows that the precursor, conjugated microporous polymers, activation agent, activation temperature\_°C, activation approach, and total pore volume (cm3/g) at 25 °C have missing values. The material\_replace, material CO2 uptake\_mmol/g\_25°C, adsorption pressure\_bar, and reference have no missing values.



**Figure 7:** heatmap missing value of dataset.

The heatmap shows the correlations of the missing values of the columns, just like the columns of Bet surface area\_m2/g have missing values in the precursor columns. It is vital to illustrate and understand the correlation between the null values in each column. The conjugated microporous polymers have the highest missing values because they have a value of -1.



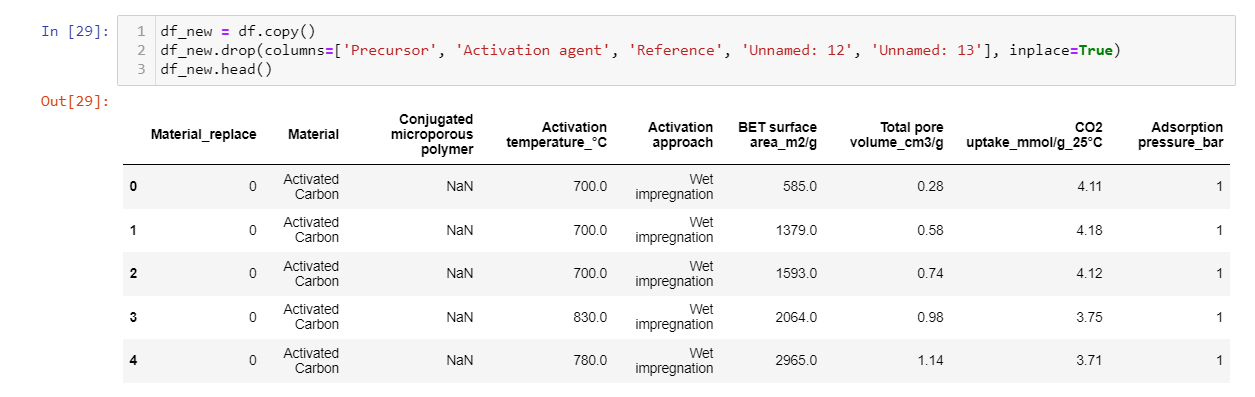
**Figure 8:** dendrogram missing values of the dataset.

On the right side of the dendrogram, the two unnamed columns have higher missing values than the activation agent columns. The conjugated microporous polymer has the most missing values in the dendrogram. There are also missing values in the adsorption pressure\_bar and the BET surface ares\_m2/g. The total pore volume cm3/g and the precursor have missing values and are correlated to conjugated microporous polymers.

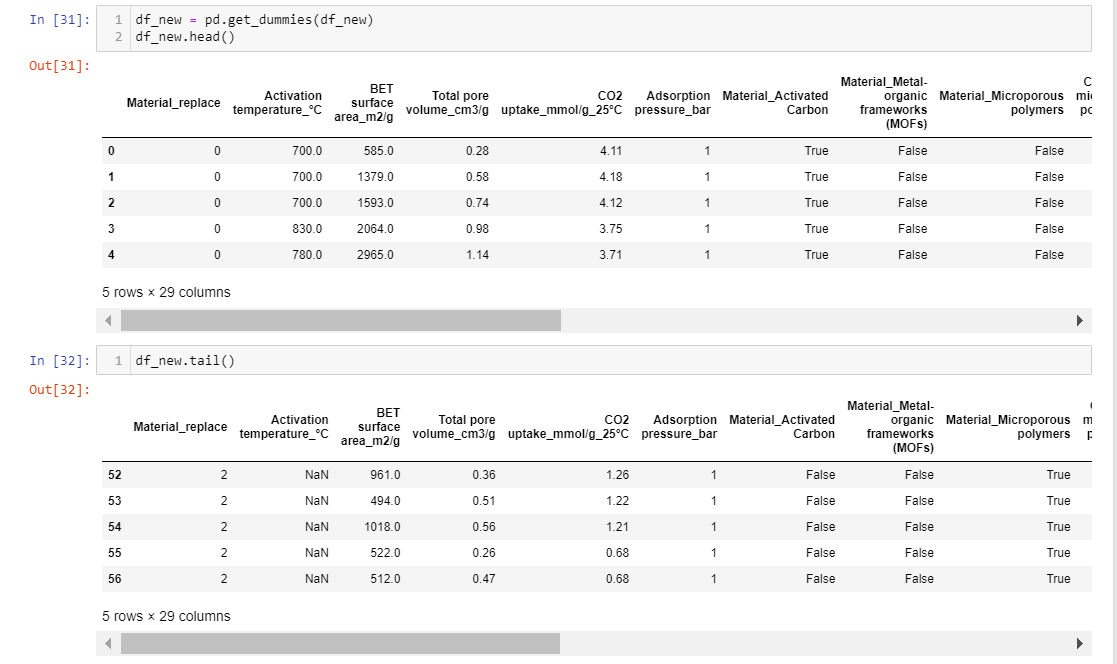
**Data Preparation**

The third phase is data preparation, which means that data cleaning should be implemented in this phase of the CRISP-DM methodology. It requires fixing the errors in the CO2 adsorption dataset before accepting the machine learning models. It has steps that need to be followed, like selecting data, cleaning the data (missing or outliers), constructing the data, integrating the data, and formatting the data. Some of the data has issues like outliers and missing values. (Saluja, 2018b).

The very important step that needs to be done in data cleaning is dropping the columns that are not important to be used for further analysis.

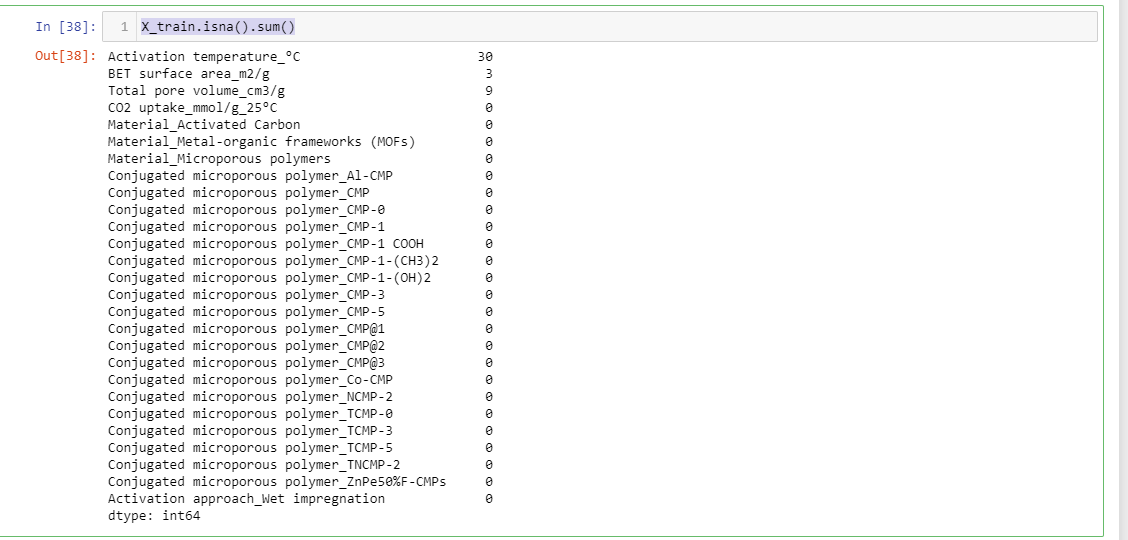


**Figure 9:** Dropping of the columns of dataset.

****

**Figure 10:** One hot coding

The CO2 adsorption dataset has categorical missing values that need to be transformed into numerical variables before performing machine learning. The one hot coding is a technique that can transform the categorical into numerical variables. Unfortunately, it is not transformed into numerical variables. But I use the iterative imputer to transform the Boolean true or false to numerical variables.



**Figure 11A:** Iterative imputer.

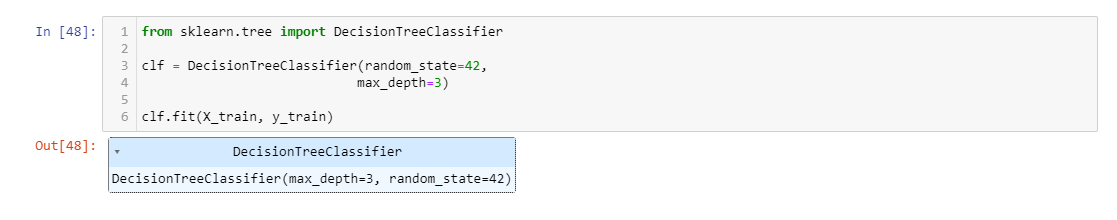


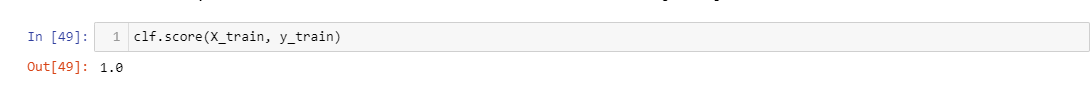
**Figure 11B:** Iterative imputer.

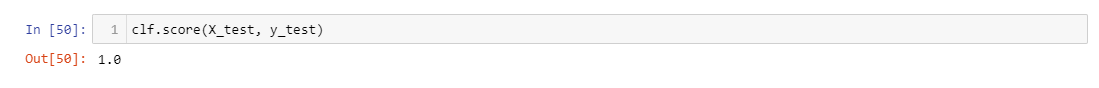
It is very important to transform the categorical variables to numerical before modelling because machine learning doesn’t read strings; machine learning only reads numbers. It must always clean the data before moving on to the next phase, which is the modelling. Also, it needs to fix all the errors in the dataset during data preparation so that when moving on to the next phase, it will provide a correct accuracy score for the modelling.

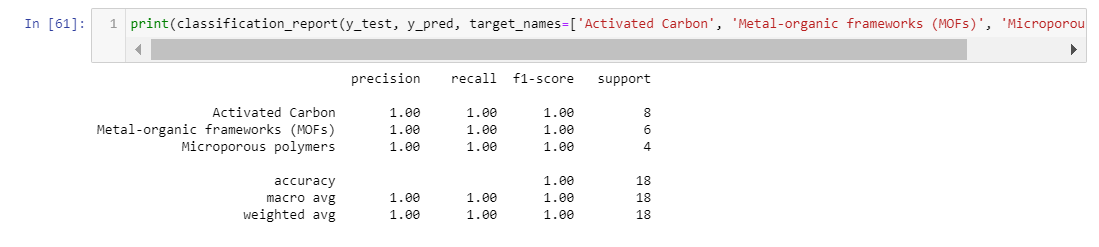
**Modelling**

The decision tree classifier is used to forecast the missing values, and it also has the ability to apprehend the non-linear pattern. It can also produce an overfit, and it is also easily affected by corrupt data. (Avinash Navlani, Fandango and Idris, 2021b). The Decision Tree Classifier is doesn’t need to do the standardize the features.



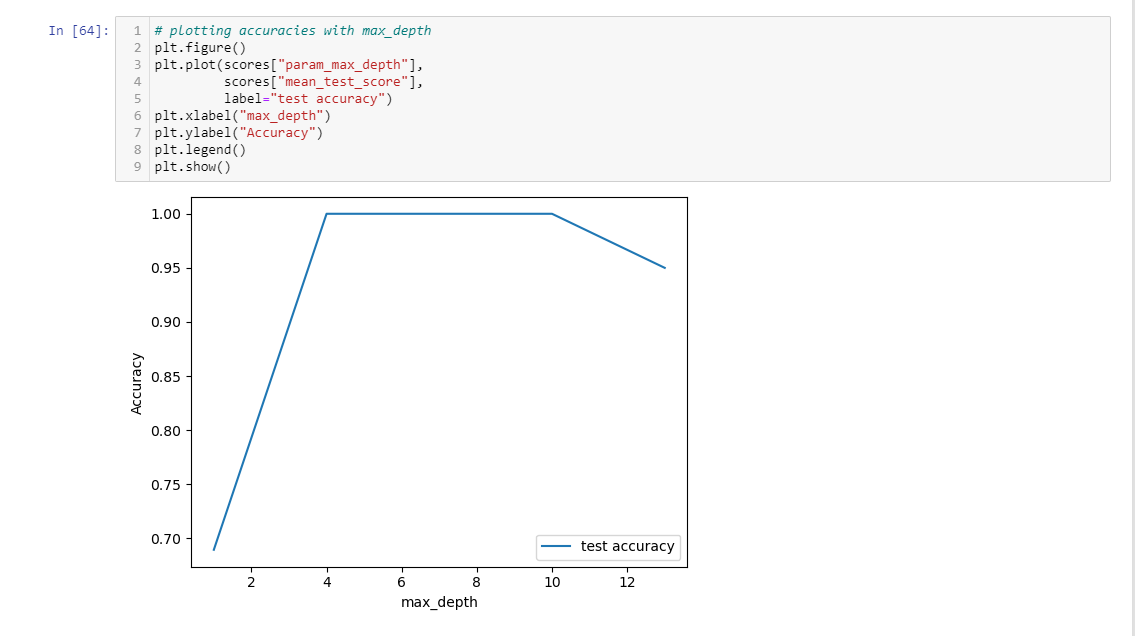
******

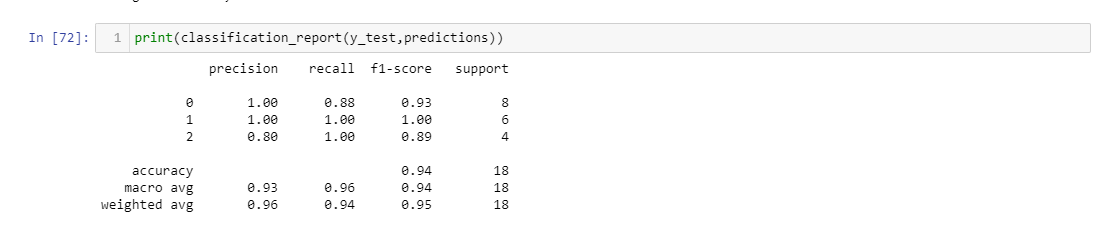


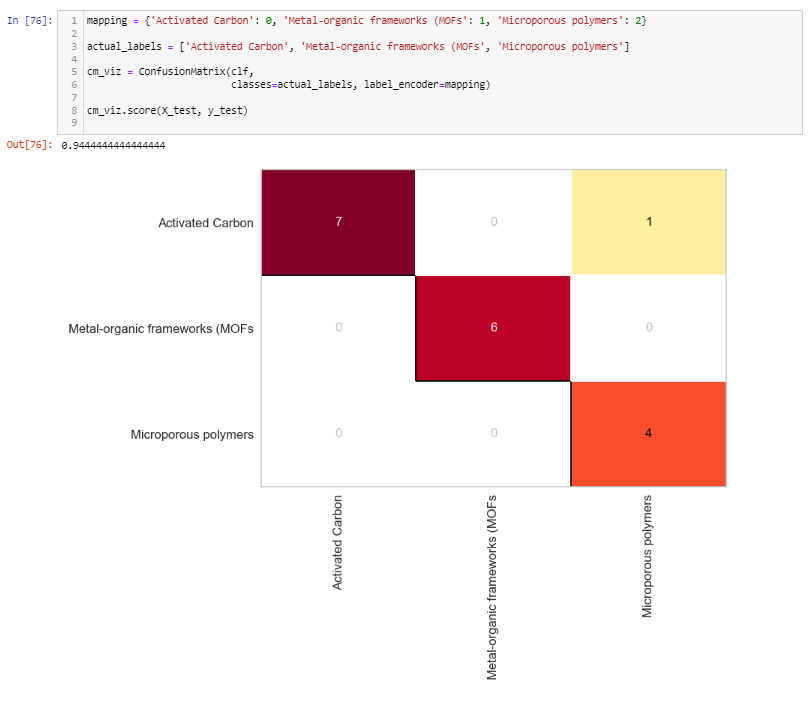


**Figure 12:** Decision Tree Classifier

The decision tree classifier gave me a result of 1.0 accuracy in precision, recall, and f1-score. I think that the 1.0 accuracy is not good because it will cause overfitting. That is why I used the hyperparameter GridSearchCV to reduce the overfitting. As the max depth increases, it will decrease the overfitting and improve the model accuracy value.

 **Figure 13:** plot of the max\_depth.





**Figure 14:** Confusion Matrix of the accuracy using GridSearchCV of the dataset.

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

**Evaluation:**

Since the accuracy score is reduced by using GridSearchCV, it already reduces the overfitting of the models. The results show that the hypothesis has the correct efficiency and improves the accuracy of the models. As stated, the f-1 score has an accuracy of 0.94, which is a good evaluation of the accuracy of the models. Therefore, I conclude that it was challenging to execute the data analysis of the CO2 adsorption dataset. But in spite of that, I really learned a lot, and I developed and gained more skills and knowledge that I can apply for the continuation of this CA2 for the next semester.

The next steps for this project will be:

* Comparing the machine learning models with my groupmate from CA1, which machine learning model has a good accuracy for the materials of the C02 adsorption dataset.
* Gather more academic papers to gain some datasets that we can use for further analysis.

**Deployment**

Rechecking this project, I can say this is more thrilled and harder than I was anticipated. It is because I was the only one figure it out how I can solve the errors that I have encounter during I do my data analysis. No matter how hard I have experienced I was still happy that I did my very best to finish the data analysis of my CA2.

**Reference List**

Avinash Navlani, Fandango, A. and Idris, I. (2021c). *Python Data Analysis*. Packt Publishing Ltd

pandas.pydata.org. (n.d.). pandas.DataFrame.describe — pandas 1.0.3 documentation. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html>.

Saluja, C. (2018b). *Data Preparation — A crucial step in Data Mining*. [online] Medium. Available at: https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281#:~:text=According%20to%20CRISP%2DDM%2C%20the..

My Github link: https://github.com/2023396RachelMaeLademora/CA2-Strategic-Thinking-.git