A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Daniela Mariano Barreto |
| *Student Number* | 2023278 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 3 Final Submission |
| *Assessment Due Date* | 10th May 2024 |
| *Date of Submission* | 10th May 2024 |

**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.

**Applied machine learning to estimate suitable material to adsorb CO2**

by

*Daniela Mariano Barreto*

*Higher Diploma in Science in Data Analytics for Business Strategic Thinking*

*James Garza*

*CCT College*

*Dublin, Ireland*

Abstract

200 – 250 WORDS

[NOTE: This section is designated for the abstract. Abstracts are not assigned page numbers and should precede the table of contents. If an abstract is unnecessary for your work, please delete this page.]

Attention: All notes must be removed from the document before submission!!

**Key-words:** k-nearest neighbours (k-NN), Gradient Boosting (GB), CRISP-DM, ….

GitHub link:

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Contents

[**Introduction** 1](#_Toc163826746)

[**Business description** 2](#_Toc163826747)

[*Hypothesis* 2](#_Toc163826748)

[*General goal* 2](#_Toc163826749)

[*Scope* 2](#_Toc163826750)

[*Success criteria/indicators* 3](#_Toc163826751)

[**Technologies** 3](#_Toc163826752)

[*Libraries* 3](#_Toc163826753)

[*Models and machine learning algorithms* 3](#_Toc163826754)

[*Hyperparameters tuning and cross validation* 3](#_Toc163826755)

[**Accomplishment** 3](#_Toc163826756)

[*Data* 3](#_Toc163826757)

[*Machine learning models* 3](#_Toc163826758)

[**Challenges** 3](#_Toc163826759)

[*Problem definition* 3](#_Toc163826760)

[**Cross Industry Standard Process - Data Mining (CRISP-DM)** 4](#_Toc163826761)

[Business understanding 5](#_Toc163826762)

[Data understanding 5](#_Toc163826763)

[Data preparation 9](#_Toc163826764)

[Machine learning implementation 9](#_Toc163826765)

[**Evaluation** 9](#_Toc163826766)

[**Conclusions** 12](#_Toc163826767)

[**Future recommendations** 12](#_Toc163826768)

[**Project's timeline overview** 13](#_Toc163826769)

[**Ethical considerations** 13](#_Toc163826770)

[**Appendices** 13](#_Toc163826771)

[**References** 16](#_Toc163826772)

# **Introduction**

Concern about the increase in carbon dioxide (CO2) in the atmosphere began decades ago. Still, this problem continues nowadays with unsatisfactory predictions for 2040, in which an increase of approximately 70% and 25% is estimated to produce electricity and coal-based, respectively (Huetteman, Bowman and Slater-Thompson, 2016, p.81; Ren and Liu, 2023, p.1).

Studies show that CO2 emissions have caused enormous global damage due to the greenhouse effect, a potential explanation for climate change. These changes probably cause a decrease in biodiversity and harm to human lives due to tsunamis and earthquakes, for example; consequently, a drastic decrease in ecosystem productivity and economic disparities (Daneshvar *et al*., 2022, p.1; Ren and Liu, 2023, p.1).

Considering all, we see why mitigating the removal of CO2 has been the focus of several institutions, and the increase in the development of technologies and products is the central reflection of these concerns and actions that have increasingly gained momentum. (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.)

A diagram of a structure

Description automatically generated

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

This capstone project proposes to focus on the adsorption technique, considering the multitude of options available within each type of technology. 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).

In light of the advancements in technologies for CO2 capture, we propose the implementation of machine learning (ML) models to determine the most suitable materials for adsorbing a desired amount of CO2 for a company. Each material possesses its own peculiarities, necessitating specific action plans that can subsequently be refined to lower production costs and enhance economic feasibility.

# **Business description**

xxxxxxxxxx

## *Hypothesis*

## *General goal*

Principal objective:

Based on the ML model results, I proposed changing the objective because each industry emits a certain concentration of CO2; therefore, this amount that needs to be adsorbed tends to differ. Thus, the model would aim to estimate which material would suit the concentration of CO2 the industry needs to adsorb.

I also suggest the title change for: Applied machine learning to estimate suitable material to adsorb CO2.

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

The first two have been executed, but the third needs to be assessed with my colleague. The last one will probably be changed to estimate suitable material given a CO2 concentration.

## *Scope*

This project aims to assess some ML models to estimate which material is suitable to adsorb a certain concentration of CO2. A compilation of data will be analyzed by descriptive statistics. Secondly, we will implement ML models, such as a decision tree, k-NN, and regression.

The first boundary would be using data from exclusively adsorption techniques. This project will be limited by data with a maximum of fifteen 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.

## *Success criteria/indicators*

# **Technologies**

## *Libraries*

## *Models and machine learning algorithms*

## *Hyperparameters tuning and cross validation*

# **Accomplishment**

## *Data*

## *Machine learning models*

# **Challenges**

Including strategies used to overcome them.

## *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 ML model.

# **Cross Industry Standard Process - Data Mining (CRISP-DM)**

CRISP-DM has been used as a project management methodology (figure 2a) with the breakdown of the phases. The project plan (figure 2b) is from gathering data to a conclusion, and all stages are interchangeable; the evaluation can generate results that can return to the preprocessing phase, for example.

The action plan will search for data with comparable experimental conditions, such as pressure and temperature. Use Python in Jupyter Notebook to clean and organize the data and create visualizations for an overview (EDA). Perform descriptive statistics and preprocessing the data if necessary.

Posteriorly, experiment with some ML models to obtain better results using cross-validation and hyperparameters adjustment. Evaluate the model’s performance and return to any previous phases if necessary. At the end, conclude the findings.



**Figure 2:** a) CRISP-DP scheme. Source: <https://healthdataminer.com/wp-content/uploads/2019/11/800px-CRISP-DM_Process_Diagram.png>. b) Project plan.

## Business understanding

The focus on sustainability has led industries to request technologies to capture CO2 from their emissions. This study aims to estimate which material should be suitable in relation to a given concentration of CO2 that the industry wishes to adsorb. This is important to provide information to create a specific plan that meets the sector's needs, avoiding unnecessary expenses.

## Data understanding

*Data source*

Data was gathered from a review article that compiled hundreds of articles, and we focused on the three most studied materials: carbon-based, metal-organic frameworks (MOFs), and polymers (Dziejarski et al., 2023, p.3,20,36,41). We increased the boundary of the study year to fifteen years otherwise, the data would be scarce.

*Data description*

The dataset is composed of 12 features, and 57 records with 3 classes of material used in the CO2 adsorption (carbon-based, metal-organic frameworks (MOFs), and polymers). Features description in table 1.

**Table 1:** Data dictionary.

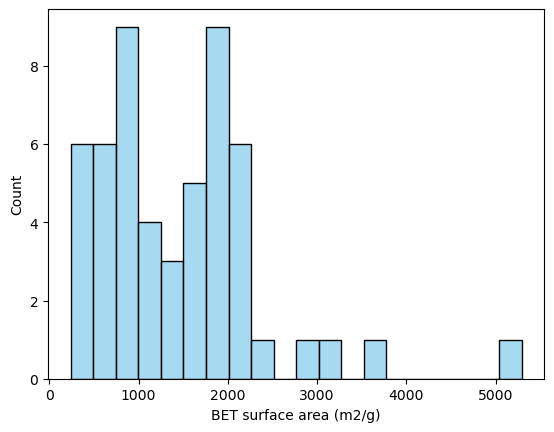
A close-up of a text

Description automatically generated

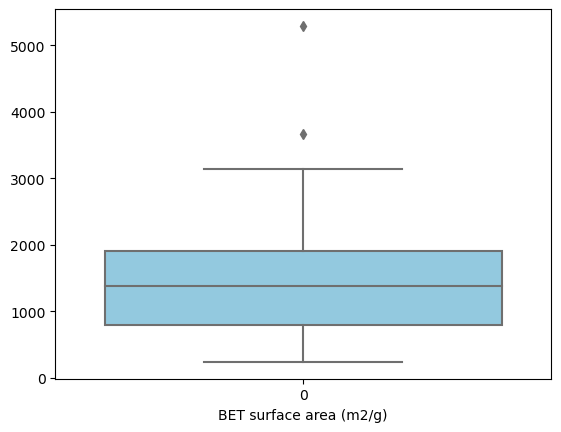
*Exploratory data analysis (EDA)*

For this analysis, I focused on these features: 'Material\_replace', 'BET surface area\_m2/g', 'Total pore volume\_cm3/g', 'CO2 uptake\_mmol/g\_25°C' and 'Adsorption pressure\_bar', because they are common to all types of materials.

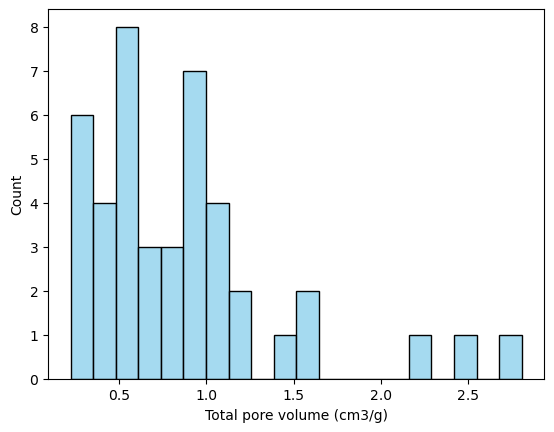
Due to missing values, I needed to impute them with interpolation values because the data is skewed and has outliers; therefore, this method is more robust than the median in this case. Figure 3 shows the data distribution and outliers.



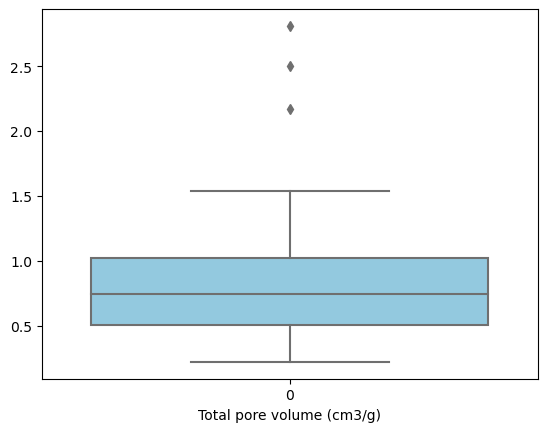
a



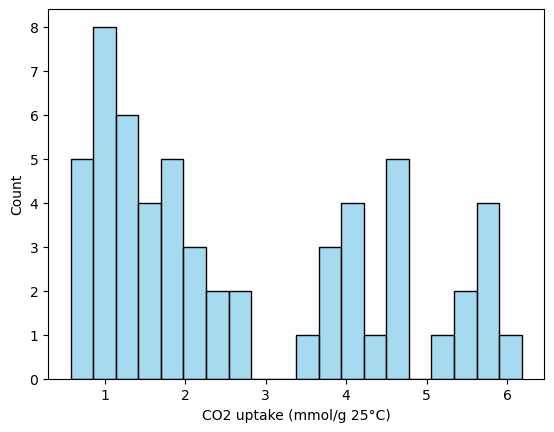
b



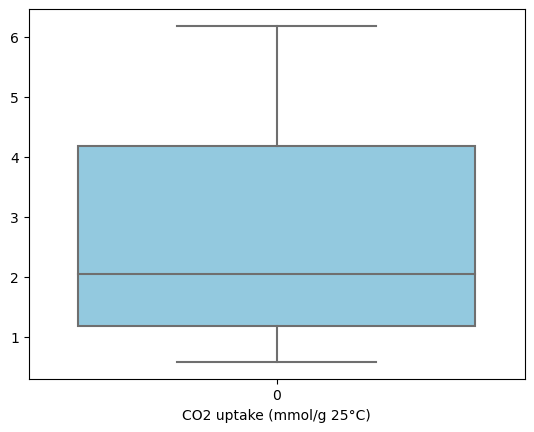
c



d



e



f

**Figure 3:** Histogram and boxplot before data imputation. 'BET surface area\_m2/g' (a and b), 'Total pore volume\_cm3/g' (c and d), ‘CO2 uptake\_mmol/g\_25°C’ (e and f).

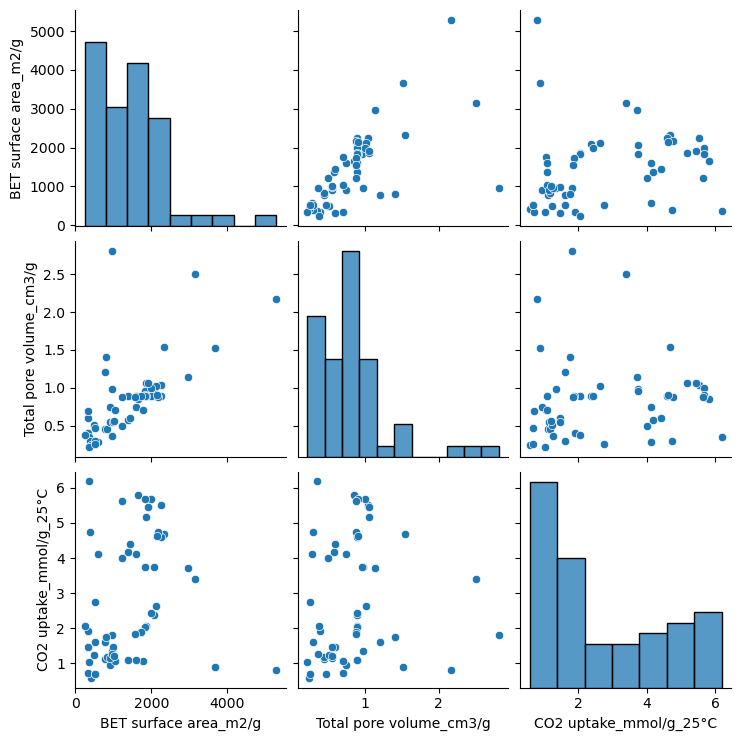
*Data Cleaning*

The dataset had no duplicates, but 7% of the values were outliers. Therefore, since it was a significant amount, I decided not to withdraw them and I will find a robust ML to handle it.

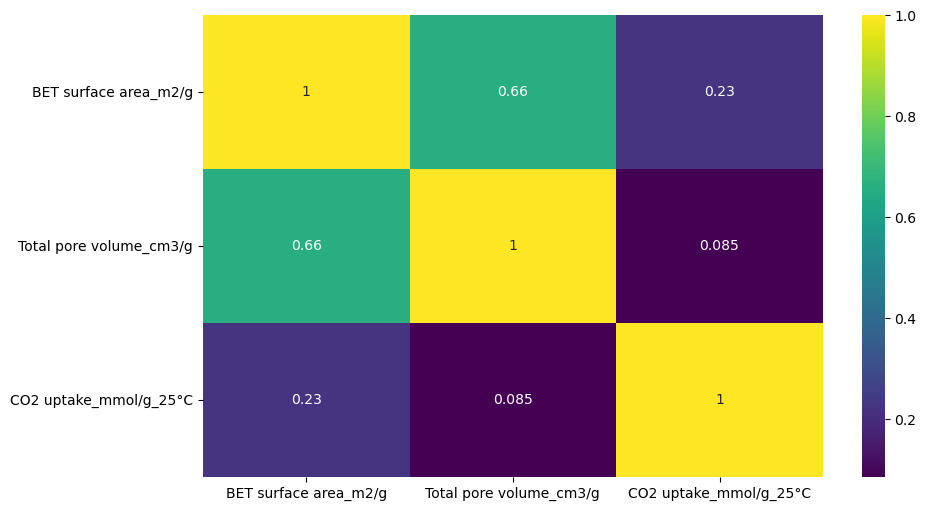
*Descriptive Statistics*

Summary statistics parameters showed that data is skewed and slightly sparse even after handling the missing values. The minimum and maximum values presented a considerable variance, and each feature's scale is very different; thus, it will be necessary to scale the data.

Figures 4 presented the pair plot and heatmap. The results showed a moderately strong positive correlation between Surface Area and Pore Volume (0.66); thus, there is a tendency to increase area at the same time as increasing volume and vice versa. However, when comparing CO2 uptake with Surface Area (0.23), the correlation is weak if compared with Area and Volume. In addition, practically, there is no correlation between CO2 and Pore Volume (0.085), meaning little or no linear correlation.



a



b

**Figure 4:** a) Pair plot, and b) heatmap of features correlation.

*Inferential Statistics*

I applied the Kruskal-Wallis test because it is appropriate when handling a non-normal distribution and small sample size (Devore, 2012, p.645). Considering a 95% confidence level, I found that there is a significant difference between the materials regarding CO2 adsorption. Therefore, the material that can adsorb more CO2 in this case is the Activated carbon, followed by MOFs (figure 5).

A graph of different colored squares

Description automatically generated

**Figure 5:** CO2 uptake in three different materials.

## Data preparation

*Preprocessing*

I applied the one-hot encoding to avoid attributing weight to the categories, ensuring that all materials have the same weight. I also standardized the data using the function RobustScaler, which is more suitable in the presence of outliers, which is essential for many ML models (scikit-learn, n.d.) and there are outliers in this dataset.

## Machine learning implementation

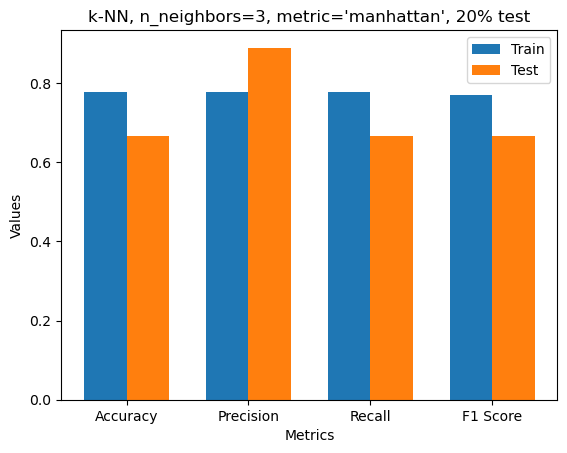
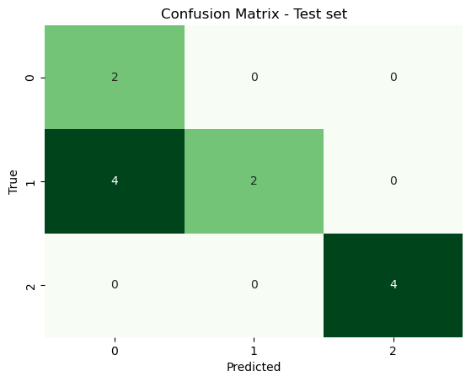
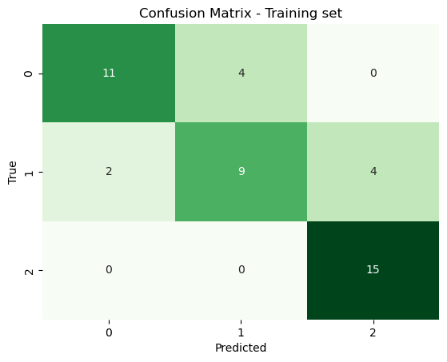
I applied the k-nearest neighbours (k-NN) because it is used with many classes and often performs reasonably without complex adjustments (Müller and Guido, 2017, p.36,44). In addition, I applied the Gradient Boosting (GB) model to improve the estimation, as it builds serial trees that try to correct the errors of the previous and provide better accuracy with the correct configuration (Müller and Guido, 2017, p .88,89). I also used cross-validation, manual hyperparameters adjustment, and Grid Search to achieve the optimal hyperparameters. SMOTE was used to deal with class imbalanced.

The target variable will be the 'Material\_replace' because I want to identify which material can be suitable given a certain concentration of CO2 uptake desired by the industry.

# **Evaluation**

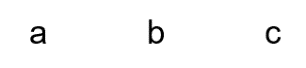
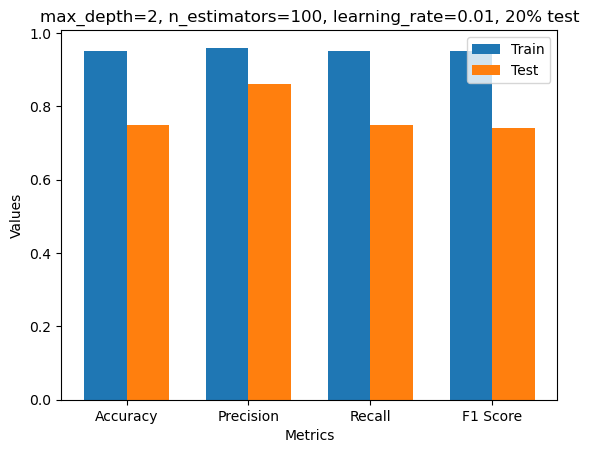
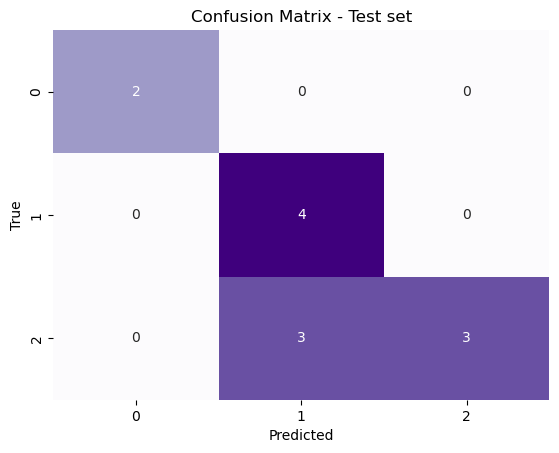
*Summary findings*

k-NN (figure 6) training and test sets presented reasonable performance, with better training results than the test, indicating that this model may be overfitting and incapable of generalizing well. SMOTE increased these differences (Jupyter Notebook file).



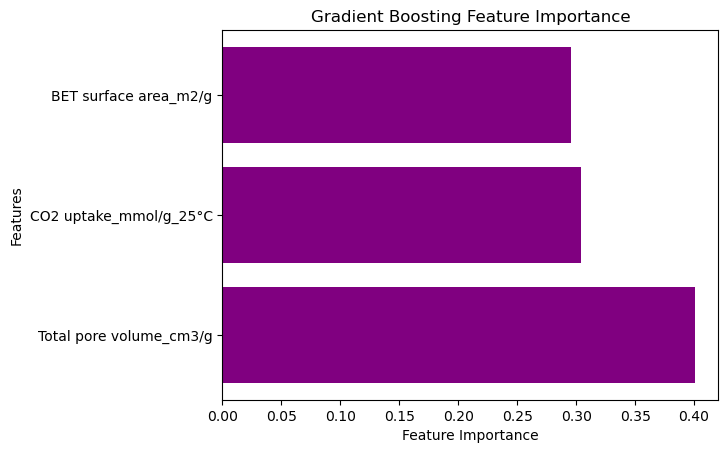
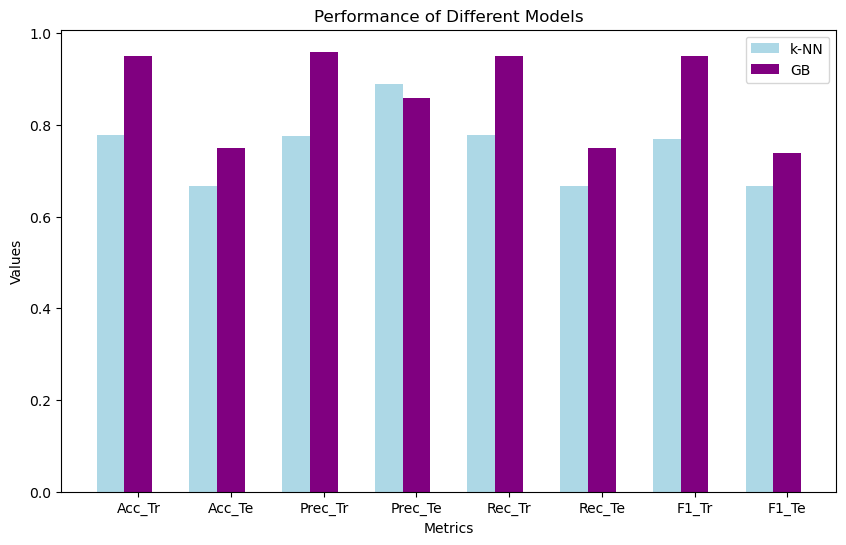
**Figure 6:** k-NN: Confusion matrix for training (a) and test (b) sets. C) Metrics for training and test sets.

GB (figure 7) improved compared to k-NN on both sets and although model training performed very well (probably overfitting), the test set may not generalize well to unseen data. Precision is high in both sets, indicating a high probability of correct prediction for positive instances in actual results, despite the test set capturing 75% of positive cases predicted (recall).



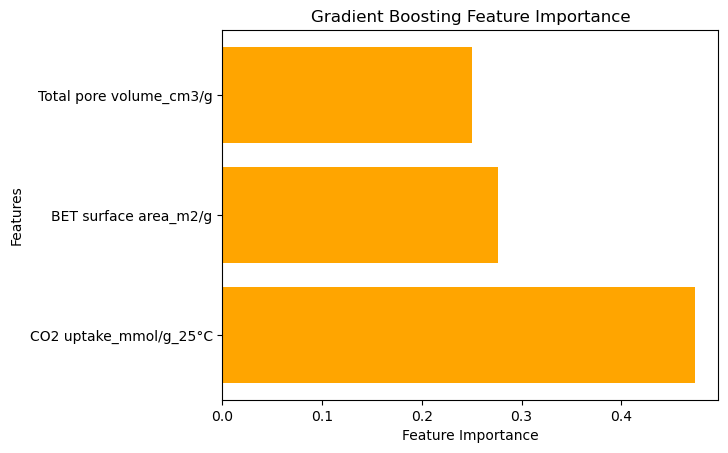
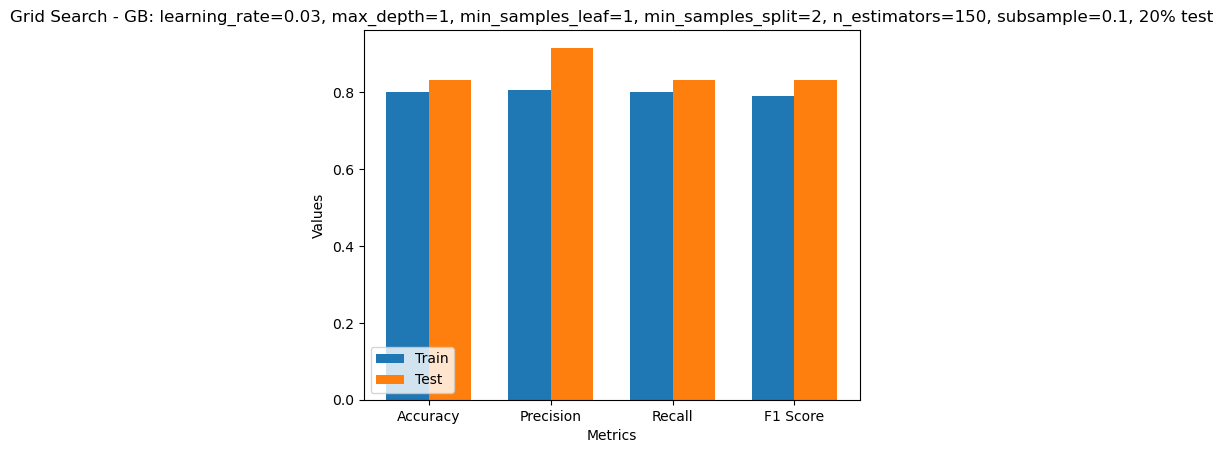
**Figure 7:** GB: Confusion matrix for training (a) and test (b) sets. C) Metrics for training and test sets.

Figure 8 summarises these two models to facilitate comparison. Neither generalised well to new data and may result in overfitting. In addition, the most important feature of the GB is 'Total pore volume', followed by the 'CO2 uptake'.



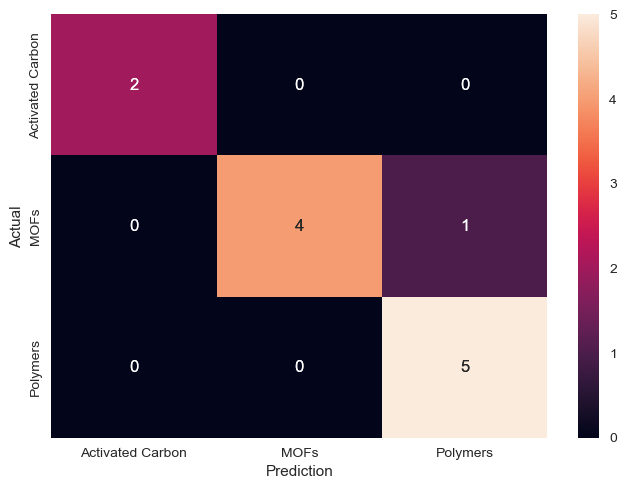
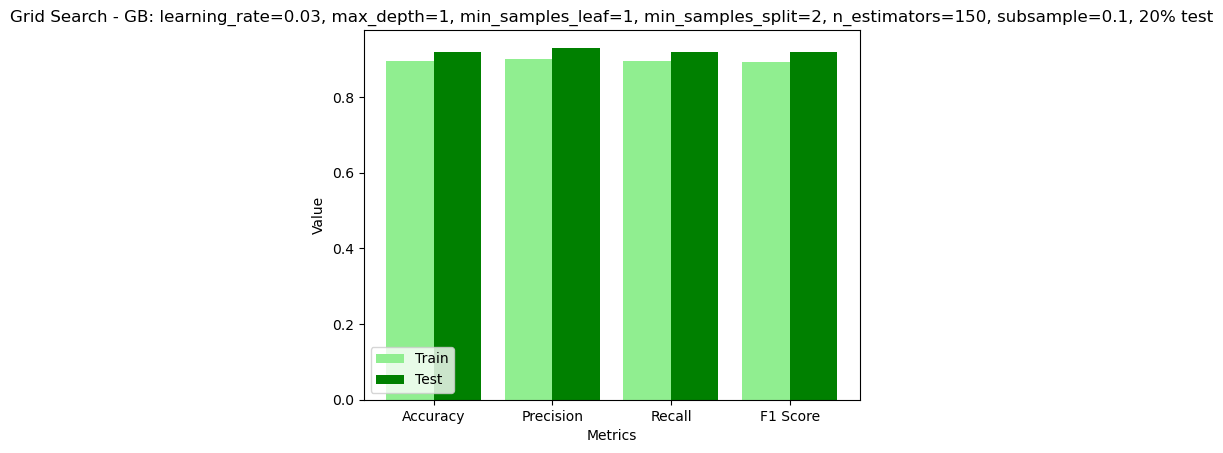
**Figure 8:** a) Metrics for k-NN (blue) and GB (purple), Tr: training set, Te: test set, Acc: accuracy, Prec: precision, Rec: recall, and F1: F1-score. b) Feature importance of GB.

The Grid Search for GB (figure 9) resulted in better accuracy in test than in the training set, which is counterintuitive for overfitting and the same happened with other metrics, thus the model may made right prediction 83% of the time. The parameter ‘subsample’ probably helped to reduce overfitting and improved the generalization because it introduces randomness to fit each tree. Inclusively, the ‘CO2 uptake’ became a feature of more importance in this model and when using the SMOTE.



**Figure 9:** a) Grid Search - GB: learning\_rate=0.03, max\_depth=1, min\_samples\_leaf=1, min\_samples\_split=2, n\_estimators=150, subsample=0.1, 20%. b) Feature importance Grid Search – GB.

Figure 10 considers the SMOTE and the class balance provided good performance in training and test sets with small difference, suggesting that is not overfitting. Consistent and slightly higher performance in all metrics on the test compared to the training suggest that the model can generalize well to new data. 93% of the samples predicted as positive were actually positive (precision), and 92% of the actual positive samples were predicted as positive (recall), indicating that the model performed well on the positive class, which is essential in this study as the objective is to identify which material can be suitable given a certain concentration of CO2 uptake desired by the industry.



**Figure 10:** a) Metrics. b) Confusion matrix. Condition: learning\_rate=0.03, max\_depth=1, min\_samples\_leaf=1, min\_samples\_split=2, n\_estimators=150, subsample=0.1, 20% test.

# **Conclusions**

The GB is robust and performed better than k-NN, and their performance was validated using cross-validation to ensure that the chosen hyperparameters generalized well to new data. Although the Grid Search with SMOTE helped to reduce the probability of overfitting and improved the generalization, there is still room for more progress. As the objective is to identify which material can be suitable given a certain concentration of CO2 uptake desired by the industry, improving the reliability of the positive class performance will be necessary, and increasing the number of records might solve it.

# **Future recommendations**

I recommend increasing the number of records per material to make the data more robust and increase the reliability of the results. And I suggest that the objective changes to answer which material would be suitable to adsorb a certain amount of CO2 instead of predicting the adsorption of CO2 in different materials, as all the materials analysed already have this capacity. In addition, it would be interesting if a company asked if I want to adsorb ‘X’ amount of CO2, what material should I use? Answering this question should help structure the company's planning.

# **Project's timeline overview**

Overall, the development of CA2 helped achieve many project milestones, particularly in gathering data, performing statistics, and implementing ML (Appendix 1). This process was satisfactory because my colleague and I will discuss our findings to reach the best option and finally deploy the model.

During the implementation phase, I faced many challenges, such as finding data, getting good metrics scores, and improving the models. But this helped me see that before proceeding with the following steps, I will have to increase the number of records to give the model more credibility.

# **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. The data will not be anonymised because they are from researchers who have done all the work and deserve to be appropriately recognised. In Appendix 2, we are presenting our Ethics Form signed.

# **Appendices**

Appendix 1:

Milestones achieved.

A diagram of a process

Description automatically generated

Phase 1

A screenshot of a computer screen

Description automatically generated

Phase 2

A screenshot of a computer screen

Description automatically generated

Phase 3

A graph with a bar chart

Description automatically generated with medium confidence

Appendix 2:

A paper with text and a checklist

Description automatically generated with medium confidence

# **References**

Chawla, N.V. (2009). Data Mining for Imbalanced Datasets: An Overview. In: *Data Mining and Knowledge Discovery Handbook*. [online] pp.875–886. doi:https://doi.org/10.1007/978-0-387-09823-4\_45. [Accessed 13 Dec. 2023].

Daneshvar, E., Wicker, R.J., Show, P.-L. and Bhatnagar, A. (2022). Biologically-mediated carbon capture and utilization by microalgae towards sustainable CO2 biofixation and biomass valorization – A review. *Chemical Engineering Journal*, 427, p.130884. doi:https://doi.org/10.1016/j.cej.2021.130884.

Devore, J.L. (2012). *Probability and statistics for engineering and the sciences*. Belmont, Calif.] Brooks/Cole, Cengage Learning.

Dziejarski, B., Serafin, J., Andersson, K. and Krzyżyńska, R. (2023). CO2 capture materials: a review of current trends and future challenges. *Materials Today Sustainability*, 24, p.100483. doi:https://doi.org/10.1016/j.mtsust.2023.100483.

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>

Müller, A. C. and Guido, S. (2017). *Introduction to machine learning with Python: a guide for data scientists*. 1st ed. United States of America. O’reilly Media.

scikit-learn. (n.d.). sklearn.preprocessing.RobustScaler. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler> [Accessed 2 Dec. 2023].

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).