APPLIED MACHINE LEARNING MODELS TO ESTIMATE MATERIAL TO ADSORB CO_2

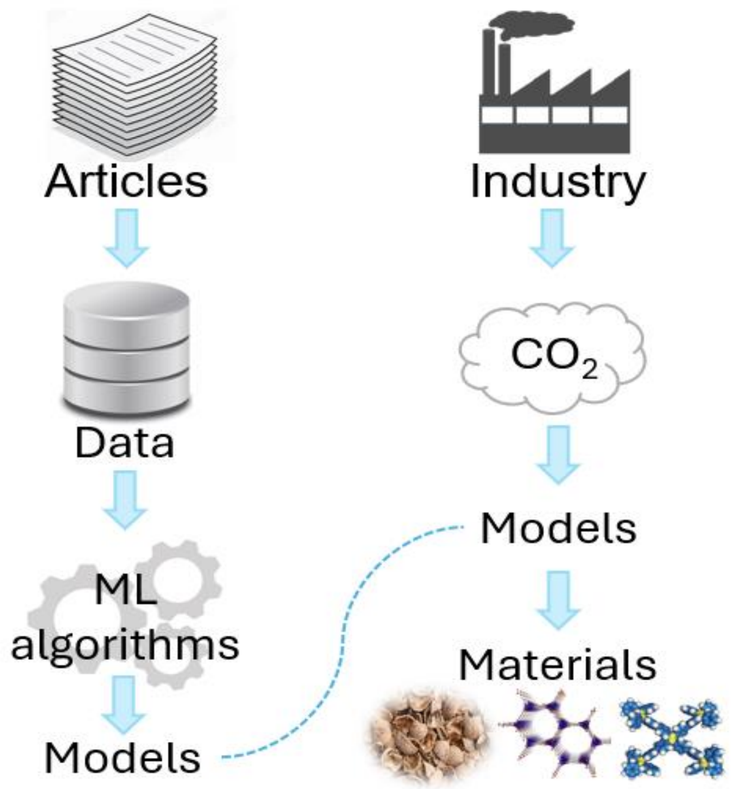
Author
Daniela M. Barreto

Supervisor
James Garza

Introduction

The increase in carbon dioxide (CO_2) in the atmosphere started decades ago, and this problem persists today, potentially explaining climate change due to the greenhouse effect. Consequently, there is a growing recognition among industries of the urgent need to adopt sustainable business practices, implementing measures to reduce and capture CO_2.

Business Understanding



(Dziejarski et al., 2023)

Research Objective

Evaluate the performance of various machine learning models to identify the most suitable materials for capturing a predetermined amount of CO_2.

Technologies

Machine learning algorithms

- Ada Boost Classifier
- Gradient Boosting Classifier (GB)
- Random Forest Classifier (RF)
- Decision Tree Classifier (DT)
- k-Nearest Neighbors Classifier
- Gaussian Naive Baye
- Support Vector Classification

Libraries

- SciPy
- Scikit-learn
- Yellobrick
- SHAP
- Shapash

Hyperparameter tuning and cross-validation

- Random Search
- Stratified K-Fold

Findings

After screening, the three tree-based models (GB, RF, and DT) showed superior performance. This improvement can be attributed to the use of the SMOTE to address class imbalance (Chawla, 2010) and the application of Stratified K-Fold cross-validation. Stratified K-Fold is particularly effective for imbalanced datasets, as it ensures that each fold maintains the original class distribution (Prusty, Patnaik, & Dash, 2022), reducing the risk of bias.

SMOTE

Stratified K-Fold

Random Search

Generalization

⇩

Success Criteria

0.87 and 0.89

Most influential feature

⇩

CO_2

Random Search was used with cross-validation for model validation and hyperparameter tuning, optimizing model performance . The results are presented in figures 1 and 2.

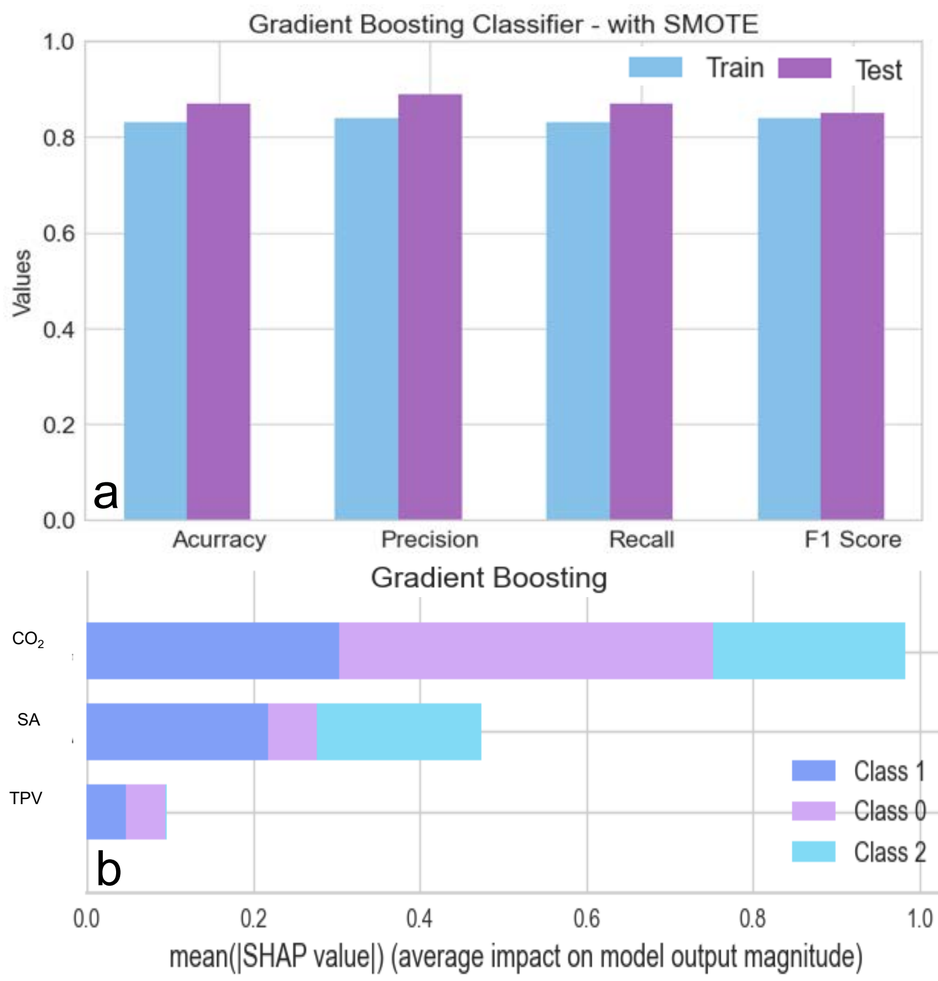


Figure 1: a) Performance metrics and **b)** SHAP summary – class influence.

The SHAP charts reveal a strong influence of CO_2 on the model's performance (Fig. 1b and 2).

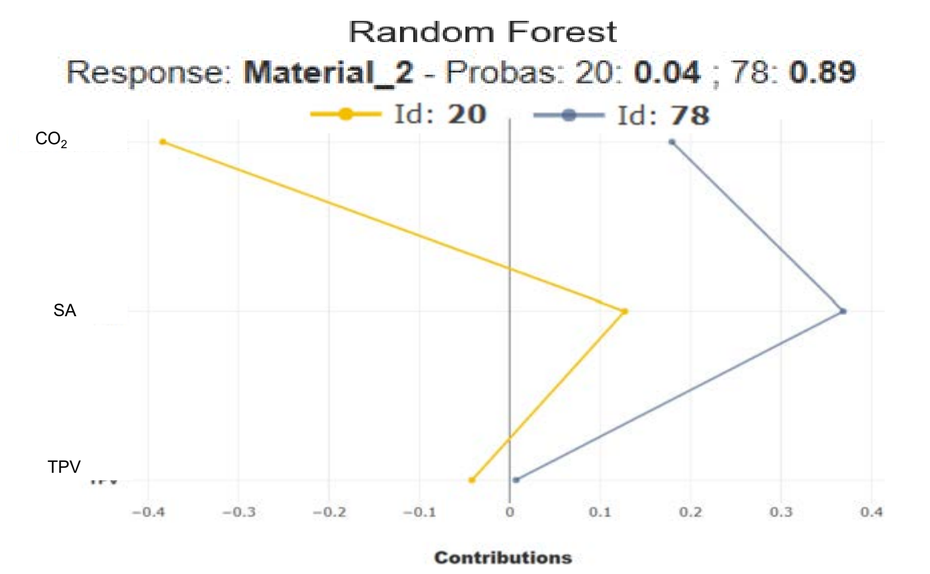


Figure 2: Model explainability for the Random Forest model deployed on two instances.

Conclusions

The three machine learning models successfully predicted the most suitable materials for capturing CO_2. This performance can be attributed to the models' robustness, enhanced by effective training using SMOTE and Stratified K-Fold cross-validation, as well as the optimization of hyperparameters.

References

Chawla, N.V. (2010). 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].

Dziejarski, B., Serafin, J., Andersson, K. and Krzyżyńska, R. (2023). CO_2 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.

Prusty, S., Patnaik, S. and Dash, S.K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4. doi:https://doi.org/10.3389/fnano.2022.972421.