ACENET
Microcredential in Advanced Computing
ISP Report

Participant name: **Sepideh Alimohammadi**

Project title: **Leveraging Machine Learning for Enhanced Underground Carbon Storage Research**

Date: July 31st, 2024

Abstract:

In this project, machine learning models were developed to predict density based on temperature, pressure, and energy using data from molecular dynamics simulations. Hyperparameter tuning with Optuna optimized model performance, demonstrating the effectiveness of ML in predicting thermodynamic properties in challenging scenarios.

1. Introduction

My objective for my Independent Study Project (ISP) is to utilize Machine Learning (ML) methodologies, using data generated from Molecular Dynamics (MD) simulations, to forecast crucial physical attributes like density in geological carbon storage. This research is significant as it can transform our approach to managing carbon storage, a key element in mitigating global warming. By leveraging ML to predict essential physical variables, we can streamline the assessment and optimization of carbon storage strategies, leading to more efficient storage systems and a better selection of parameters for carbon sequestration.

My motivation for researching this field is to improve the effectiveness of underground carbon storage by suggesting a substitute for traditional models, e.g., thermodynamics and experimental tests. Enhancing our understanding and management of carbon sequestration through ML can significantly contribute to a more sustainable future and address pressing environmental challenges.

2. Background

Geological carbon storage, also known as carbon sequestration, involves capturing carbon dioxide (CO_2) emissions from industrial sources and storing them underground to prevent their release into the atmosphere. This method is considered one of the most effective strategies for mitigating climate change by reducing greenhouse gas

concentrations. However, the success of carbon storage depends on accurately predicting and managing the physical properties of the storage sites, such as density, pressure, and temperature, which influence the stability and capacity of the storage reservoirs.

Machine Learning (ML) offers powerful tools to enhance our understanding and optimization of these properties. By analyzing large datasets generated from Molecular Dynamics (MD) simulations, ML models can predict the behaviour of CO_2 under various conditions, providing valuable insights into the feasibility and safety of storage sites. This approach not only improves the efficiency of carbon storage but also helps select optimal storage parameters, ultimately contributing to global efforts to combat climate change.

3. Analysis

3.1. Approach and Dataset Acquisition:

To achieve the objective of utilizing Machine Learning (ML) for predicting the density of geological carbon storage sites, water density, I employed Molecular Dynamics (MD) simulations to generate the necessary data. The dataset was created using LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator), a powerful open-source software for molecular dynamics simulations. The simulation parameters were carefully selected to represent various conditions of temperature, pressure, and energy levels, which are critical for understanding the behaviour of CO_2 in underground storage reservoirs.

3.2. Dataset Description:

The dataset consists of the following columns: Step: Simulation step number, Temp: Temperature at each simulation step, f_TempAve: Average temperature, Press: Pressure at each simulation step, f_PressAve: Average pressure, f_PEAve_Mol: Average potential energy per molecule, f_DensAve: Average density, which is the target variable for the ML models.

3.3. Data Preparation Process:

- Data Cleaning: The raw data obtained from the MD simulations were parsed from the LAMMPS log files and converted into CSV format for easier manipulation and analysis. This involved handling missing values and ensuring the consistency of the dataset.
- Feature Engineering: Feature selection is a crucial step in building an effective machine learning model, especially when dealing with datasets containing

multiple variables. In the context of predicting the density of geological carbon storage sites, the primary features considered were temperature (Temp), pressure (Press), and average potential energy per molecule (f_PEAve_Mol). These features were selected based on their direct influence on the physical properties of the storage medium.

To ensure the selected features were statistically significant, an Ordinary Least Squares (OLS) regression was performed. This involved fitting a linear model to the data and examining the p-values of the coefficients. Features with p-values less than 0.05 were considered statistically significant, indicating that changes in these features were likely to affect the target variable significantly. All predictors (Temp, Press, f_PEAve_Mol) are significant in predicting f_DensAve.

3.4. Choice of Analysis Method:

Given the complexity and non-linear relationships within the dataset, multiple ML models were explored to identify the best approach for predicting density:

- Random Forest Regression: Selected for its ability to handle non-linear interactions and its robustness to overfitting. It provides feature importance, helping to understand the contribution of each feature to the prediction.
- Gradient Boosting Regression: Chosen for its strong predictive performance and its capability to handle complex datasets through iterative improvement of the model.
- Support Vector Regression (SVR): Used for its effectiveness in high-dimensional spaces and its flexibility with different kernel functions.
- Additional Models: Other models such as Linear Regression, k-Nearest Neighbors (k-NN), and Ridge Regression were also considered to ensure comprehensive analysis.

3.5. Model Training and Evaluation:

Each model was trained using the training set and evaluated on the validation set. The performance metrics used were Mean Squared Error (MSE) and R-squared (R^2) to assess the accuracy and explanatory power of the models.

3.6. Hyperparameter Tuning:

All three models were tuned using Optuna, for-loops, and PSO. Optuna, a hyperparameter optimization framework based on Bayesian optimization, was employed to fine-tune the models. This process involved defining a search space for each model's hyperparameters and leveraging Optuna’s advanced optimization

algorithms to efficiently identify the best parameters that minimized validation errors. Among the methods used, Optuna demonstrated superior performance in optimizing hyperparameters and enhancing model accuracy.

3.7. Utilizing High-Performance Computing (HPC):

At present, due to the relatively small size of the dataset, high-performance computing (HPC) resources are not required. However, I plan to expand the dataset to include a larger volume of data and additional physical properties. For this purpose, I will leverage HPC resources such as Compute Canada and Siku. These remote computing facilities will provide the necessary computational power to efficiently handle and analyze the expanded dataset, allowing for more extensive and complex analyses.

4. Results

The Ordinary Least Squares (OLS) regression model for predicting density achieved an R-squared value of 0.859, indicating a strong fit between the model and the observed data, with temperature, pressure, and average molecular energy all showing significant contributions to the model. The high F-statistic and low p-values suggest that the predictors collectively have a substantial impact on density, validating the model's effectiveness in capturing the relationship between the variables.

Table 1. OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5857	0.017	34.823	0.000	0.552	0.619
Temp	0.0002	8.61e-06	17.845	0.000	0.000	0.000
Press	1.285e-05	5.4e-07	23.774	0.000	1.18e-05	1.39e-05
f_PEAve_Mol	-0.0314	0.001	-22.927	0.000	-0.034	-0.029

Omnibus:	60.285	Durbin-Watson:	0.255
Prob(Omnibus):	0.000	Jarque-Bera (JB):	442.891
Skew:	1.728	Prob(JB):	6.72e-97
Kurtosis:	12.659	Cond. No.	7.28e+04

To illustrate the analysis of outliers in your dataset, plots were generated showing the relationships between steps and variables such as temperature (T), pressure (P), energy, and density. These plots revealed that the data from the second step deviated significantly from the others, likely due to the first step corresponding to the crystal structure of water, which is not representative of the operating conditions under study. Despite this deviation, the outlier data from the second step was retained in the simulation to capture a broader range of conditions and ensure comprehensive analysis.

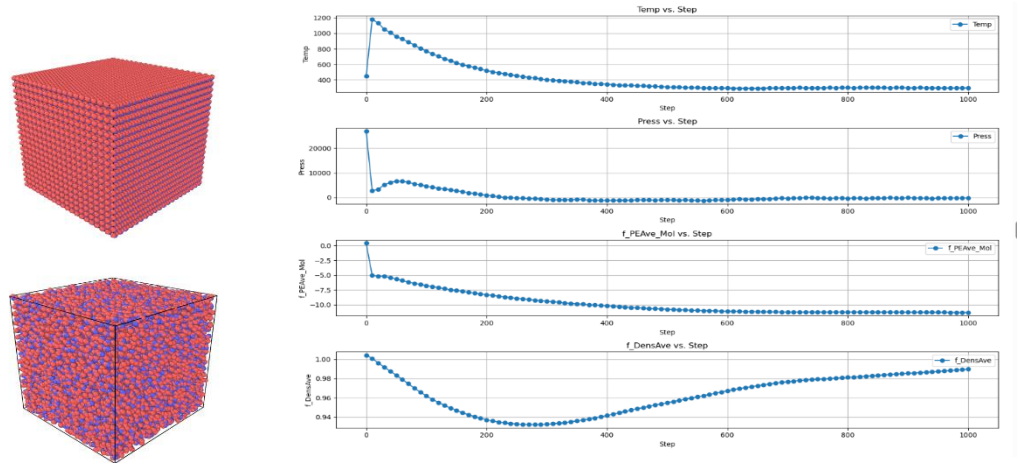

Figure 01. Plot showing the relationships between steps and variables including temperature (T), pressure (P), energy, and density.

Figure 2 illustrates the performance of the Random Forest Regressor model compared to actual data for both training and validation sets. The first subplot displays model predictions against observed values, showing the model's accuracy in capturing the density based on temperature (T), pressure (P), and energy. The second subplot plots density versus T, P, and energy, demonstrating how these features influence density predictions. The model achieved a validation RMSE of 0.00812 and an R^2 score of 0.861, indicating a strong fit and predictive capability. Note that the outlier data discussed in the previous section was retained in the dataset and included in the validation process, which led to a significant reduction in accuracy metrics.

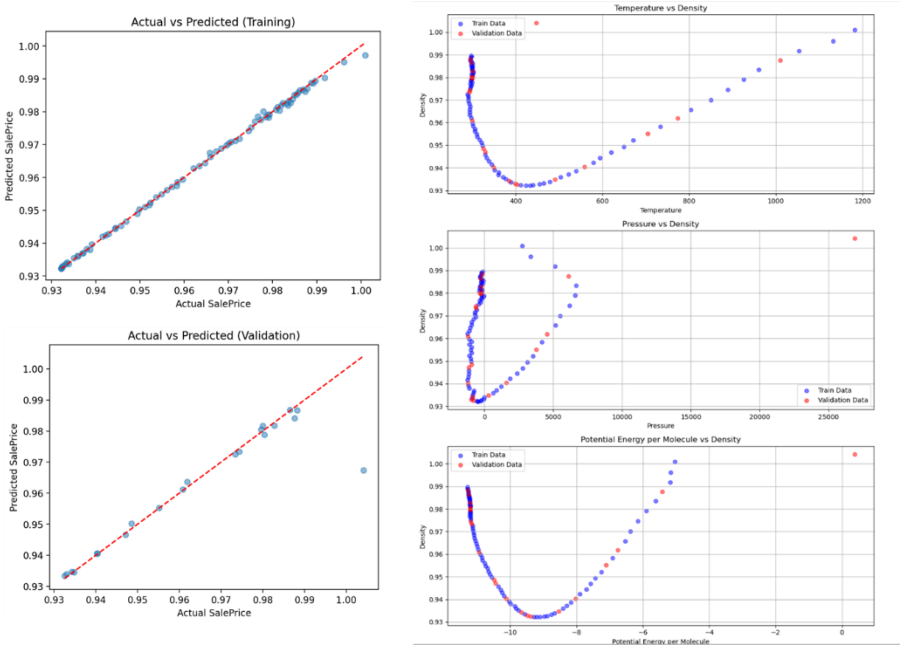

Figure 02. Model Performance and Feature Relationships

Hyperparameter tuning was performed using For-Loops, OPTUNA, and PSO to optimize the Random Forest Regressor. OPTUNA emerged as the most effective method, providing the best hyperparameters and improving model performance significantly.

Table 2. Hyperparameter Tuning

	For-Loop	OPTUNA	PSO
MSE	0.00801	0.0064	0.0080
R^2	0.8649	0.8971	0.8648

Among the models tested, Random Forest achieved the highest performance with an MSE of 0.0001 and an R^2 of 0.8706, indicating the best predictive accuracy. In contrast, SVR, Linear Regression, and Ridge Regression showed significantly lower performance with higher MSE values and negative R^2 scores.

Table 3. Comparison Between MLs

	Random Forest	Gradient Boosting	SVR	Linear Regression	Ridge Regression	K-Neighbors Regressor
MSE	0.0001	0.0001	0.0005	0.003	0.0032	0.01
R^2	0.87	0.83	-0.024	-7.0058	-5.65	0.766

5. Discussion

In this project, the Random Forest model outperformed other models with an MSE of 0.0001 and an R^2 of 0.8706, demonstrating its effectiveness in predicting density. The Gradient Boosting model also performed well with an MSE of 0.0001 and an R^2 of 0.8395, while SVR and linear regressions showed poorer results. A significant challenge was the presence of outliers in the dataset, which were retained during validation and led to reduced accuracy metrics. Optuna proved to be highly effective in tuning hyperparameters, enhancing model performance. Overall, while the Random Forest model provided the best results, addressing the outlier impact and leveraging Optuna's tuning capabilities were key aspects of the project's success.

Conclusion

Machine Learning has effectively predicted thermodynamic properties in challenging conditions, demonstrating the power of tools like Optuna for hyperparameter tuning. Random Forest was identified as the most effective model for this task. Future work will focus on expanding the dataset and utilizing remote computing resources for more comprehensive analysis.

Supplementary Materials

GitHub: https://github.com/sepideh-ali/ISP/tree/main