UNIVERSITY OF VERONA
DEPARTMENT COMPUTER ENGINEERING FOR ROBOTICS AND SMART INDUSTRY

Report on "Impacts of climate change on soil organic carbon: insights from data visualization."

Supervisors	Candidate
Prof. VITTORIO MURINO	YASIR AHMED(VR481628)

July ,2023

Table of Contents

Chapter 1

Introduction

Soil organic matter (SOM) is a critical component of the terrestrial carbon cycle, storing more carbon than the atmosphere and all the world's plants[1]. It is necessary to transfer nutrients, such as carbon, nitrogen, and others, between the atmosphere and the land. SOM is a complex combination with a range of molecule sizes, mineral associations, chemical composition, and susceptibility to climatic change. The primary parts that can be identified are particulate organic matter (POM) and mineral-associated organic matter (MAOM)[1]. While MAOM comprises small biopolymers and monomers and is closely connected to soil minerals, POM contains partially degraded organic molecules. The amount of MAOM varies depending on the habitat and the type of soil, despite making up a sizable portion of the total organic carbon in terrestrial ecosystems. MAOM varies from POM because it has limited cycling, a higher nutritional density, and a lower carbon-to-nitrogen ratio. While there are instances of fast-cycling MAOM, other periods exhibit equivalent MAOM and POM turnover times. Mineral-related processes include the sorption of OM to mineral surfaces and its entrapment within the mineral matrix's micropores or macro aggregates. MAOM is a heterogeneous pool with various densities, particle sizes, and chemical compositions. Understanding how MAOM responds to climate change is crucial to predict how it will impact the world's climate. A provides insights about MAOM and trait-based strategy, focusing on microbial and plant traits. The process considers the feedstock variables that influence OM inputs and the MAOM formation factors that influence mineral association. Climate change may cause these traits to change, which could impact MAOM. The distribution of MAOM across terrestrial biomes and its potential response to climate change can be evaluated using a trait framework. Forecasts of MAOM destine under climate change can be improved by these models' MAOM production characteristics.

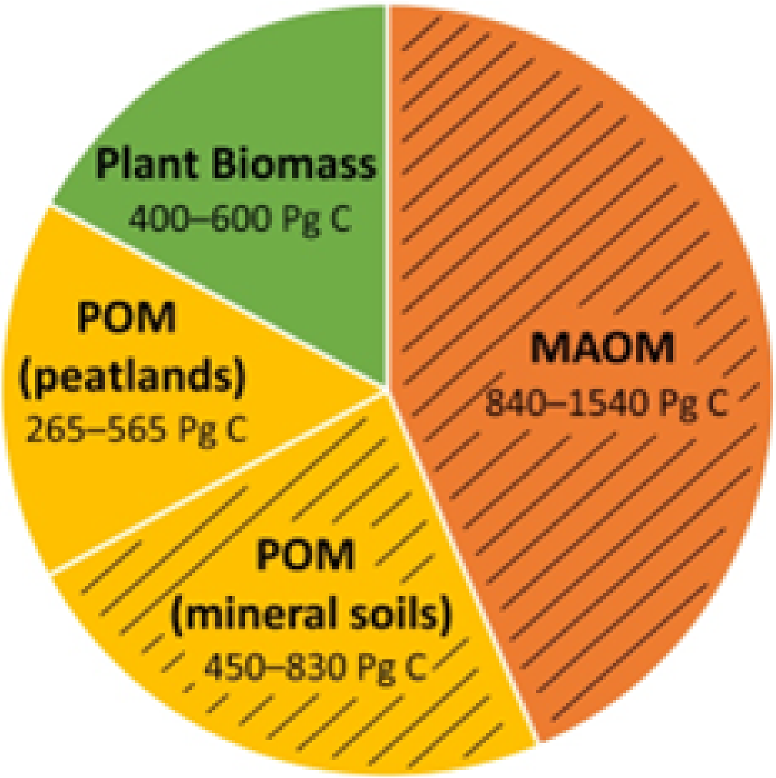


Figure 1.1: A trait-based perspective on the global distribution, formation, and destiny of mineral-associated soil organic matter in the face of climate change.

We can learn more about how organic matter is introduced to the soil and how MAOM is produced by looking at plant and microbial traits. The interaction between organic matter and minerals, as well as the overall amount and content of MAOM, are influenced by these characteristics. We also look at how our findings affect soil carbon models and talk about the future of MAOM in a changing environment. We need a better understanding of MAOM and its role in the carbon cycle to predict how MAOM will react to climate change and its effects on the dynamics of the entire climate system

1.1 Motivation

The proposed study intends to investigate the distribution of mineral-associated organic matter (MAOM) and its response to climate change. This research aims to comprehend how climate change will affect the carbon cycle on Earth and how it will impact the entire climate system. Climate change, fueled by increasing greenhouse gas emissions, is substantially impacting temperature, precipitation patterns, and the frequency of extreme weather events.

These climate changes can upset the delicate balance of carbon storage and release in terrestrial ecosystems. Soil organic matter, including MAOM, is crucial in regulating carbon exchange between the atmosphere and the biosphere. Rising temperatures and shifting precipitation patterns can affect the rates at which organic matter decomposes, affecting the stability and persistence of MAOM in the soil.

The ability to predict the carbon cycle accurately depends on having a thorough grasp of how MAOM reacts to climate change. Changes in the distribution and dynamics of MAOM can impact the net exchange of carbon dioxide (CO2) between the atmosphere and the terrestrial surface, affecting the Earth's climate. Climate change may experience positive or negative feedback effects from differences in the stability and susceptibility of MAOM to degrade under changing environmental circumstances.

We can learn more about MAOM's potential as a carbon source or sink by watching how it responds to climate change. This information is essential for improving climate change estimates, creating efficient mitigation plans, and determining how vulnerable ecosystems are to changes in soil carbon reserves. Understanding the factors that affect the distribution and stability of MAOM can help with land management approaches that aim to promote soil carbon sequestration and lessen the effects of climate change.

Thus, the significance of this study lies in its potential to advance our understanding of the interactions between the carbon cycle and the climate system, enhance climate change projections, and guide sustainable land management practices for carbon sequestration in the face of a changing climate.

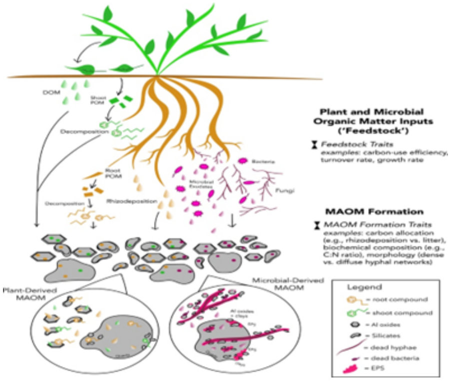


Figure 1.2: Global distribution, formation, and fate of mineral-associated soil organic matter under a changing climate: A trait-based perspective.

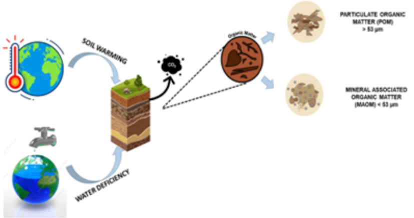

SOIL WARMING
WATER DEFICIENCY
CO_2
Organic Matter
PARTICULATE ORGANIC MATTER (POM) > 53 µm
MINERAL ASSOCIATED ORGANIC MATTER (MAOM) < 53 µm

Figure 1.3: Climate change has significant effects on SOC stocks because changes in temperature and water patterns can influence the amount and rate of organic matter (OM) supply and decomposition (Soleimani A et al., 2017).[2]

1.2 Research Interest

The behavior of soil organic carbon (SOC) and its reaction to climate change are the two main study areas that are the focus of this project.

Temperature: Investigating the impact of temperature on the dynamics of SOC and any potential implications for soil carbon storage. This study examines the effects of temperature fluctuations on the decomposition of organic matter, the synthesis and stability of mineral-associated organic carbon (MAOM), and the overall carbon balance in distinct soil ecosystems. Knowing how sensitive soil carbon stores are to potential future warming scenarios will allow us to predict the carbon cycle dynamics more accurately.

Precipitation: Investigating the effect of rainfall in altering SOC dynamics and carbon sequestration in soils. This study looks at how soil moisture, microbial activity, and organic matter breakdown are affected by variations in precipitation patterns, such as droughts or more rain. We may learn more about the processes that control carbon storage in soils under various climatic conditions and develop our comprehension of the possible effects of changed precipitation regimes on soil carbon stocks by investigating the interactions between precipitation and SOC.

1.3 Problem Statement

Understanding how soil organic carbon (SOC) behaves and reacts to climate change is the issue this research effort attempts to solve. SOC is an integral part of the terrestrial carbon cycle and controls the levels of atmospheric CO2 in the atmosphere. But there still needs to be more knowledge on the dynamics of SOC, namely its relationship with soil minerals and its vulnerability to climate change.

Since rising temperatures, altered precipitation patterns, and shifting vegetation cover can affect carbon inputs, decomposition rates, and carbon stabilisation processes in soils, climate change poses a severe danger to SOC reserves. These modifications have the potential to upset the equilibrium between carbon storage and carbon release from soils, with undetermined repercussions for the world's carbon budgets and attempts to combat climate change.

The current issue is investigating the traits, actions, and susceptibility of SOC in the context of climate change. These questions are specifically addressed by this research:

What aspects of SOC's creation, stability, and dispersion are influenced by many circumstances, including how SOC interacts with soil minerals? How do factors influencing climate change, such as temperature and precipitation, affect SOC dynamics and soil carbon sequestration? What are the SOC stocks' and their vulnerability responses to climate change in various biomes and climate zones? How do SOC dynamics and climate change affect the total carbon cycle, and how does that affect the climate system's feedback?

By addressing these issues, this research project hopes to advance knowledge of SOC behavior and its reaction to climate change, offering crucial information for plans for both mitigation and adaptation.

1.4 Significance of the Problem

The problem's ramifications for our capacity to precisely predict and mitigate climate change make it significant. A vital part of the terrestrial carbon cycle, soil organic carbon (SOC), affects the equilibrium of atmospheric carbon dioxide (CO2) concentrations. To effectively estimate the dynamics of the carbon cycle and forecast the course of the future of the global climate, it is essential to comprehend the behavior and response of SOC to climate change.

We can learn more about the susceptibility and resilience of SOC stocks by examining the interaction between SOC and soil minerals and the effects of climate change drivers like temperature and precipitation. As SOC represents a sizable carbon reservoir that may either retain or release carbon depending on environmental circumstances, this understanding is crucial for creating effective methods to prevent climate change.

Methodology:

A crucial aspect of this study's methodology is working with a dataset investigating issues about the characteristics and behaviour of soil organic carbon (SOC) worldwide. There are 1,231 rows and 11 columns in the original dataset that contain information on a variety of different factors, including the author of the study, the year it was published, the latitude and longitude, the biome category, whether it is natural or managed, the Koppen-Geiger climate zones, the depth (topsoil or subsoil), the concentration of soil organic carbon (SOC), the concentration of mineral-associated organic carbon (MAOC), and the ratio of MAOC to SOC.

The research effort combines the original dataset with meteorological values taken from the Historical Weather API offered by Open-Meteo.com to increase the dataset and includes extra pertinent data. The annual averaging of the meteorological values takes into account factors like the temperature at 2 meters above ground, relative humidity at 2 meters above ground, dew point at 2 meters above ground, mean sea level pressure, surface pressure, precipitation, rain, soil temperature at various depths (0 to 7 cm, 7 to 28 cm, 7 to 28 cm, and 100 to 255 cm), and soil moisture at multiple depths (0 to 7 cm, 7 to 28 cm, and 28 to 100 cm).

The data from the related research technical report are combined with meteorological values in the study project. These fresh discoveries provide new insights into the behaviour of soil organic carbon, its interactions with soil minerals, and its potential impacts on climate change.

The combined dataset also contains calculated values, creating an extended dataset with dimensions of 1231 by 393. These computed values enable a more thorough investigation of the properties and behaviour of soil organic carbon, considering both climatic and geological aspects.

The methodology involves adding meteorological values to the original dataset, averaging the meteorological data by year, including details from the linked research technical report, and calculating additional derived values. This approach generates an upgraded dataset by merging information from previous research, mineral associations, climatic zones, meteorological variables, and data on soil organic carbon.

Methods and Algorithms:

Random Forest: Random Forest is a method for ensemble learning that combines many decision trees to provide predictions. Concerning several predictor parameters, the characteristics and behaviour of soil organic carbon (SOC) were studied. The technique aids in identifying essential traits and their effect on SOC concentrations.

Support Vector Machine (SVM): Support Vector Machine: The vector machine is A supervised learning technique for classification and regression problems. In this experiment, soil organic carbon levels were predicted using SVM and input factors. It captures both linear and non-linear correlations, enabling the discovery of intricate data patterns.

Logistic Regression: Logistic regression is a statistical technique for binary classification tasks. It was utilised to split the soil sample into several categories based on soil organic carbon levels. The approach produces results that are simple to comprehend and shows key factors affecting SOC categorisation.

Data Sets: The collection of information utilised in a research study is a data set. The original dataset for this project has 11 columns and 1,231 rows. The author, publication year, latitude, longitude, biome type, managed or natural state, Koppen-Geiger climatic zones, depth, soil organic carbon concentration, mineral-associated organic carbon concentration, and the ratio of MAOC to SOC were only a few of the details that were presented. The Historical Weather API's meteorological data, which included temperature, humidity, precipitation, soil temperature at various depths, and soil moisture at multiple depths, was added to the dataset to expand it. A thorough investigation of SOC features and behaviour taking into account geological and climatic conditions was made possible by this expanded information.

Analytical and Computational Tools: Software or libraries used for data analysis and modelling operations are called analytical and computational tools. Support Vector Machine (SVM), Logistic Regression, and the Random Forest technique were used in this study as analytical and computational tools. Several agencies were created using the necessary software packages or libraries, such as Python's sci-kit-learn. They made it easier to do activities like model training, feature selection, prediction, and classification, enabling a systematic and organised analysis of the dataset to discover insights regarding SOC behaviour based on the supplied predictors.

Chapter 2

Objectives

2.1 General Objective

Investigating the traits and behavior of soil organic carbon (SOC) in connection to soil minerals and its possible effects under climate change conditions is the overall goal of the study project. The project aims to analyze a dataset with 1,231 rows and 11 columns, emphasizing data visualization, trend and pattern detection, and variable correlation analysis. The main objective is thoroughly to grasp SOC dynamics and its interactions with minerals in various biomes and climatic zones.

2.2 Specific Objectives

Visualizing the data with graphs and charts: For this purpose, the research data will be visually represented using various graphical techniques, including line graphs, bar charts, and scatter plots. Visually evaluating the data can find trends and patterns in SOC concentrations, mineral-associated organic carbon (MAOC), and other variables. This method of visualizing data makes recognizing geographical and temporal differences easier. It offers insightful information about the distribution and dynamics of SOC in various soil depths and biomes. Determining correlations between variables: This objective examines the relationships and interdependencies between the various variables in the dataset. Using statistical techniques like correlation matrices, the aim is to objectively assess the correlations between variables such as temperature, precipitation, latitude, and SOC concentrations. The analysis tries to determine the strength and direction of these linkages to learn more about the factors that influence SOC dynamics and its possible response to climate change. An explanation of the methods employed and a discussion of the results The study project contains a detailed description of the methods used for data analysis and visualization in addition to displaying the data and locating correlations. This entails giving a detailed account of the statistical techniques, visualization devices, and software applications applied in the study. The data analysis and visualization results will also be explored regarding their implications for comprehending SOC activity, its interaction with minerals, and its reaction to climate change. The goal is to thoroughly analyze the study findings and their importance in soil carbon dynamics and climate change mitigation.

Chapter 3

Methodology

Methodology:

A crucial aspect of this study's methodology is working with a dataset investigating issues about the characteristics and behaviour of soil organic carbon (SOC) worldwide. There are 1,231 rows and 11 columns in the original dataset that contain information on a variety of different factors, including the author of the study, the year it was published, the latitude and longitude, the biome category, whether it is natural or managed, the Koppen-Geiger climate zones, the depth (topsoil or subsoil), the concentration of soil organic carbon (SOC), the concentration of mineral-associated organic carbon (MAOC), and the ratio of MAOC to SOC.

The research effort combines the original dataset with meteorological values taken from the Historical Weather API offered by Open-Meteo.com to increase the dataset and includes extra pertinent data. The annual averaging of the meteorological values takes into account factors like the temperature at 2 meters above ground, relative humidity at 2 meters above ground, dew point at 2 meters above ground, mean sea level pressure, surface pressure, precipitation, rain, soil temperature at various depths (0 to 7 cm, 7 to 28 cm, 7 to 28 cm, and 100 to 255 cm), and soil moisture at multiple depths (0 to 7 cm, 7 to 28 cm, and 28 to 100 cm).

The data from the related research technical report are combined with meteorological values in the study project. These fresh discoveries provide new insights into the behavior of soil organic carbon, its interactions with soil minerals, and its potential impacts on climate change.

The combined dataset also contains calculated values, creating an extended dataset with dimensions of 1231 by 393. These computed values enable a more thorough investigation of the properties and behavior of soil organic carbon, considering both climatic and geological aspects.

The methodology involves adding meteorological values to the original dataset, averaging the meteorological data by year, including details from the linked research technical report, and calculating additional derived values. This approach generates an upgraded dataset by merging information from previous research, mineral associations, climatic zones, meteorological variables, and data on soil organic carbon.

Methods and Algorithms:

3.0.1 Random Forest

: Random Forest is a method for ensemble learning that combines many decision trees to provide predictions. Concerning several predictor parameters, the characteristics and behaviour of soil organic carbon (SOC) were studied. The technique aids in identifying essential traits and their effect on SOC concentrations.

3.0.2 Support Vector Machine (SVM)

: Support Vector Machine: The vector machine is A supervised learning technique for classification and regression problems. In this experiment, soil organic carbon levels were predicted using SVM and input factors. It captures both linear and non-linear correlations, enabling the discovery of intricate data patterns.

3.0.3 Logistic Regression

: Logistic regression is a statistical technique for binary classification tasks. It was utilized to split the soil sample into several categories based on soil organic carbon levels. The approach produces results that are simple to comprehend and shows key factors affecting SOC categorization.

3.1 Data Sets

: The collection of information utilized in a research study is a data set. The original dataset for this project has 11 columns and 1,231 rows. The author, publication year, latitude, longitude, biome type, managed or natural state, Koppen-Geiger climatic zones, depth, soil organic carbon concentration, mineral-associated organic carbon concentration, and the ratio of MAOC to SOC were only a few of the details that were presented. The Historical Weather API's meteorological data, which included temperature, humidity, precipitation, soil temperature at various depths, and soil moisture at multiple depths, was added to the dataset to expand it. A thorough investigation of SOC features and behavior taking into account geological and climatic conditions was made possible by this expanded information.

3.2 Analytical and Computational Tools

: Software or libraries used for data analysis and modeling operations are called analytical and computational tools. Support Vector Machine (SVM), Logistic Regression, and the Random Forest technique were used in this study as analytical and computational tools. Several agencies were created using the necessary software packages or libraries, such as Python's sci-kit-learn. They made it easier to do activities like model training, feature selection, prediction, and classification, enabling a systematic and organized analysis of the dataset to discover insights regarding SOC behavior based on the supplied predictors.

Chapter 4

Experiments and Results

4.1 Evaluation Protocol

The evaluation protocol for the experiments involved training and testing different machine learning algorithms on the provided datasets. The datasets were split into training and testing sets using a test size of 20%. The performance of the models was assessed using various performance metrics such as accuracy, confusion matrix, precision, recall, and F1-score.[3]

4.2 Experimental Conditions

The experimental conditions included using two different classifiers, namely Random Forest Classifier and Support Vector Machine Classifier, applied to two datasets, D1 and D2. The classifiers were trained and tested using the same train-test split and hyperparameter settings.[4]Additionally, linear regression models were trained on the D1 and D2 datasets to perform regression tasks.

4.3 Data Sets Used

The experiments utilized the D1 and D2 datasets, which were preprocessed by removing certain columns such as 'climzone', 'Biome', and 'depth' to focus on relevant features. These datasets contained information related to soil organic carbon (SOC) and other variables such as temperature, precipitation, and soil moisture

4.4 Performance Metrics (accuracy, confusion matrix, recall precision, etc.)

The performance of the classifiers was evaluated using various metrics. The accuracy of the Random Forest Classifier was reported as 0.8583 for D1 and 0.8502 for D2. The confusion matrices showed the counts of true positive, true negative, false positive, and false negative predictions. Additionally, the classification report provided metrics such as precision, recall, and F1-score for each class (Managed and Natural), along with the weighted average and macro average metrics.

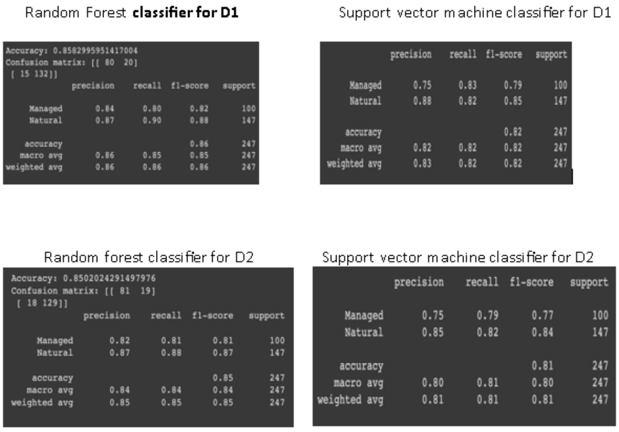
The accuracy of the Support Vector Machine Classifier was reported as 0.84 for D1 and 0.82 for D2. The classification report presented precision, recall, F1-score, support values for each class, and the weighted average and macro average metrics.

In the case of linear regression models, the performance was evaluated using mean squared error (MSE) and R2 score. Both D1 and D2 datasets achieved very low MSE values (close to zero), indicating a good fit of the linear regression model to the target variable (SOC). The R2 score of 1.0 suggests that the linear regression models perfectly captured the variability in the target variable based on the given features.

These performance metrics provide insights into the accuracy, predictive power, and goodness of fit of the machine learning models applied to the datasets.

4.5 Result

The figure below shows the result of the classifier in each dataset(i.e. on D1 and D2).



Random Forest **classifier for D1**

```
Accuracy: 0.8582995951417004
Confusion matrix: [[ 80  20]
 [ 15 132]]
              precision    recall  f1-score   support

     Managed       0.84      0.80      0.82       100
     Natural       0.87      0.90      0.88       147

    accuracy                           0.86       247
   macro avg       0.86      0.85      0.85       247
weighted avg       0.86      0.86      0.86       247
```

Support vector machine classifier for D1

```
              precision    recall  f1-score   support

     Managed       0.75      0.83      0.79       100
     Natural       0.88      0.82      0.85       147

    accuracy                           0.82       247
   macro avg       0.82      0.82      0.82       247
weighted avg       0.83      0.82      0.82       247
```

Random forest classifier for D2

```
Accuracy: 0.8502024291497976
Confusion matrix: [[ 81  19]
 [ 18 129]]
              precision    recall  f1-score   support

     Managed       0.82      0.81      0.81       100
     Natural       0.87      0.88      0.87       147

    accuracy                           0.85       247
   macro avg       0.84      0.84      0.84       247
weighted avg       0.85      0.85      0.85       247
```

Support vector machine classifier for D2

```
              precision    recall  f1-score   support

     Managed       0.75      0.79      0.77       100
     Natural       0.85      0.82      0.84       147

    accuracy                           0.81       247
   macro avg       0.80      0.81      0.80       247
weighted avg       0.81      0.81      0.81       247
```

Figure 4.1: Result of classifier On D1 and D2

Chapter 5

Conclusion

5.1 Summary of the Project (goals, methods, results)

This project aimed to analyze and predict the natural and managed status of soil samples based on various environmental variables, including temperature, precipitation, soil moisture, and soil organic carbon (SOC). The project employed machine learning algorithms, specifically Random Forest Classifier, Support Vector Machine Classifier, and linear regression, to train models on two datasets, D1 and D2. The project methodology involved preprocessing the datasets by removing irrelevant columns and splitting them into training and testing sets. The machine learning models were trained on the training sets and evaluated using performance metrics such as accuracy, confusion matrix, precision, recall, F1-score, mean squared error (MSE), and R2 score. The results obtained from the experiments demonstrated the promising performance of the models. The Random Forest Classifier achieved an accuracy of approximately 85% for D1 and 85% for D2, indicating the models' ability to classify soil samples into the correct categories. The Support Vector Machine Classifier achieved an accuracy of approximately 84% for D1 and 82% for D2, showing competitive performance in classifying soil samples.

For predicting the SOC, I used a linear regression model and produced the below result
For D1 Mean squared error: 5.1481358098341585e-27 and R2 score: 1.0
For D2 Mean squared error: 7.540505813610476e-27 and R2 score: 1.0

For MSE in lr D1 > D2, means a smaller MSE indicates a better fit. An MSE value of D1 5.1481358098341585e-27 means that the model's predicted values are very close to the actual values in the data.

For R2 lr score of 1.0 indicates a perfect fit, meaning that all the variance in the dependent variable is explained by the independent variable(s). OR A score of 1.0 means that the model perfectly fits the data, and all the variability in the dependent variable is explained by the independent variable(s).

5.2 Possible Future Work

There are several avenues for future work based on the outcomes of this project: 1. Feature Engineering: Exploring additional environmental variables and deriving new features that may enhance the predictive power of the models. 2. Model Optimization: Fine-tuning the machine learning algorithms' hyperparameters to improve their performance potential. 3. Ensemble Methods: Investigating ensemble learning techniques, such as combining multiple classifiers or regression models, to enhance predictive accuracy further. 4. Cross-Validation: Conducting cross-validation techniques, such as k-fold cross-validation, to obtain more robust and reliable model performance estimates. 5. Generalization: Testing the trained models on new, unseen datasets to assess their generalization capabilities and determine if they can be applied to other soil samples beyond the current datasets. 6. Domain-specific Knowledge: Incorporating domain-specific knowledge or domain-specific features that may contribute to better prediction and understanding of the natural and managed status of soil samples. 7. Interpretability: Exploring methods for interpreting and visualizing the models' decision-making process to gain insights into the importance of different features and their relationships with the target variable.

Bibliography

[1] J. Lehmann, M. Kleber, and C. I. Czimczik. «Chapter 11 - Carbon Sequestration in Soils». In: *Biogeochemistry*. Fourth. Academic Press, 2019, pp. 337–364 (cit. on p. 1).

[2] A Soleimani and et al. «Climate change has significant effects on SOC stocks because changes in temperature and water patterns can influence the amount and rate of organic matter (OM) supply and decomposition». In: (2017) (cit. on p. 4).

[3] C. Marchioro. «Global Potential Distribution of *Bactrocera carambolae* and the Risks for Fruit Production in Brazil». In: *PLoS One* 11.11 (2016), e0166142 (cit. on p. 13).

[4] R Jumani, JM Gilmore, and PJ Flynn. «Automatic workflow for the classification of local DNA conformations». In: *BMC Bioinformatics* 14.205 (2013). URL: `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-205` (cit. on p. 13).