Awarding Body: **Arden University**
Programme Name: **MSc Data Analytics and Information Systems Management**
Module Name (and Part if applicable): **RES7001BNM: Research Project**
Assessment Title: **Prediction of Soil Organic Carbon (SOC) for soil samples taken between 2009 and 2018 for the LUCAS Project of the European Union through Geospatial Modelling**
Student Number: **STU121233**
Tutor Name: **William Baker Morrison**
Word Count: **9460** Please refer to the Word Count Policy on your Module Page for guidance

Abstract

This dissertation presents a comprehensive investigation into the prediction of soil organic carbon (SOC) content using machine learning and deep learning models. The study aimed to identify the most accurate, cost-effective, and efficient methodologies for predicting SOC content, contributing to more effective land management practices and strategies for carbon sequestration.

The research methodology involved testing a range of machine learning and deep learning models, including Lasso Regression, Support Vector Machines (SVM), LightGBM Regression, Multilayer Perceptron (MLP), Deep Feedforward Networks (DFFNs), Long Short-Term Memory (LSTM), and a Hybrid model. The models were trained and tested on a comprehensive dataset provided by Seqana GmbH, Berlin, merging the LUCAS soil dataset with geospatial data retrieved from Google Earth Engine. This dataset, encompassing soil samples collected across diverse geographic locations and climates from 2009 to 2018, offered valuable predictor variables such as vegetation indices, topography, soil type, and land-use data. The performance of these models was evaluated based on their ability to predict the SOC content of the soil samples.

The study found that the LightGBM Regression, Multilayer Perceptron (MLP), and Hybrid models emerged as the most effective models in predicting SOC content. These models demonstrated superior performance in handling high-dimensional data and showed robustness to overfitting. A comparative analysis of the machine learning

and deep learning models revealed that the deep learning models, specifically the MLP and Hybrid models, showed slightly better performance.

The study also identified the most influential predictors for each model, providing valuable insights into the factors that significantly impact SOC content prediction. These predictors included various soil and environmental features, such as soil type, temperature, and moisture content.

The research methodology was designed to be transparent and reproducible, with the code and methods openly accessible and adaptable to other datasets. The code for the study was hosted in a public GitHub repository, allowing other researchers to replicate the study and further investigate the research questions.

In conclusion, this dissertation has made significant strides in advancing the understanding of SOC content prediction. The findings suggest that the LightGBM Regression, Multilayer Perceptron (MLP), and Hybrid models are the most promising methods for accurate SOC prediction. These results provide an evidence-based foundation for further research into improving SOC prediction models.

Acknowledgement

I would like to express my gratitude to all the individuals who have provided me with direct assistance during the writing of this dissertation.

First and foremost, I would like to thank my supervisor **William Baker Morrison**, from **Arden University, Berlin**, for their unwavering support and guidance throughout the entire process. Their expertise and encouragement have been invaluable in shaping this dissertation.

I would also like to express my gratitude to those who generously shared their expertise, insights, and advice when needed. Their input, whether direct or indirect, has been invaluable in developing and completing this project.

In addition, I am grateful to **Seqana GmbH, Berlin**, for their assistance with providing the dataset and for their constructive feedback on various drafts of this research proposal for this dissertation.

Finally, I would like to declare that this work is my own and complies with **Arden University, Berlin** regulations on plagiarism. All sources have been appropriately acknowledged and cited throughout the dissertation.

Nishant Jha

Table of Contents

Table of Figures

1. Introduction

1.1 Introduction to the topic

Soil Organic Carbon (SOC) is not merely an essential part of soil quality; it is a critical component of our global ecosystem. SOC significantly contributes to soil fertility, climate change mitigation, and nutrient cycling. It functions as a carbon sink, assisting in counterbalancing greenhouse gas emissions and moderating the Earth's climate. However, due to the intricate and heterogeneous characteristics of the soil and the spatial and temporal variability of SOC, measuring SOC with accuracy and efficiency is a substantial challenge (Smith, 2012).

In response to this challenge, endeavours such as the Land Use/Cover Area Frame Statistical Survey (LUCAS) project, initiated by the European Union to analyse the main properties of topsoil in 23 Member States of the European Union (EU), have made significant progress. Since 2009, the LUCAS project has been systematically collecting soil samples from diverse regions across Europe, leading to the creation of a comprehensive dataset that provides unique insights into soil characteristics, including SOC content (Toth, Jones, and Montanarella, 2013).

Despite this, the data from the LUCAS project alone is insufficient to generate accurate SOC predictions across large geographical scales. It is here where the usage of satellite imagery and remote sensing technology comes into the picture. Satellites, including the Landsat series operated by the United States Geological Survey (USGS) and NASA (USGS, 2008), the Sentinel satellites that are part of the European Space Agency's Copernicus Programme (ESA, 2023), and the MODIS instrument aboard the

Terra (EOS AM) and Aqua (EOS PM) satellites provide valuable predictor data that can significantly enhance the accuracy of SOC predictions (MODIS, n.d.). These satellites, some of which are used by Google Earth Engine, capture crucial information related to vegetation indices, climatic factors, and topographic variables. The data collected ranges from long-term studies of land use to high-resolution images in a variety of spectral bands, covering a broad spectrum of applications including agriculture.

Integrating these various types of satellite-based data sources provides new possibilities for geospatial modelling in SOC prediction studies. The amalgamation of satellite data with other conventional sources enables the development of more precise and efficient SOC prediction models. By leveraging the spatially explicit information captured by satellites, researchers can gain a comprehensive understanding of how different environmental factors influence SOC distribution across vast areas at high resolutions (Vasques, Grunwald, and Sickman, 2008).

1.2 Aim

The central aim of this research is to develop an advanced geospatial model for predicting SOC content in gravimetric % (being calculated as kg of organic carbon per kg of soil) using soil samples collected between 2009 and 2018 for the LUCAS project. This model will effectively integrate the LUCAS data with predictor data from Google Earth Engine, leading to potentially more accurate and detailed SOC predictions. The proposed model will serve as a novel tool for researchers, policymakers, and environmentalists, helping them to understand and manage SOC distributions better, ultimately contributing to climate change mitigation strategies.

1.3 Objectives

To achieve this aim, the research will pursue the following objectives:

1. Develop a geospatial model that can predict SOC content for soil samples taken for the LUCAS project. This objective involves the integration of diverse datasets and the application of statistical modelling and machine learning techniques.
2. Evaluate the accuracy of the geospatial model using a holdout dataset. This will provide a robust measure of the model's predictive capability and highlight areas for further improvement.
3. Investigate the relationships between land use, vegetation cover, and topography and soil organic carbon levels. This will help understand the main drivers of SOC variations and refine the predictive power of the geospatial model.
4. Assess the potential of the geospatial model to predict soil organic carbon levels accurately for areas beyond the study area. This objective is vital for evaluating the scalability and generalizability of the developed model.

1.4 Research Questions

The proposed research aims to answer the following questions:

1. How effective are the Machine Learning models in predicting soil organic carbon (SOC) levels using predictor data from Google Earth Engine for soil samples collected between 2009 and 2018?
2. How effective are the Deep Learning models in predicting soil organic carbon (SOC) levels using predictor data from Google Earth Engine for soil samples collected between 2009 and 2018?

3. What is the comparative performance of these models in terms of their prediction accuracy, and which one produces the most accurate results?
4. How significant are the contributions of the different predictors to the predictive power of each model? Which predictor is the most influential in each model?
5. What is the potential of these models for application beyond the study area? Can they maintain their predictive accuracy when applied to other regions?

1.5 Research Gap

The research gap that this project seeks to address is the insufficiency of current methodologies for predicting Soil Organic Carbon (SOC) concentrations. Traditional soil sampling and laboratory analyses are reliable and accurate, but they are time-consuming, expensive, and labour-intensive (Soil-Grids.org, 2020; Pelletier et al., 2018). These traditional methods also face challenges in extrapolating results to larger geographical areas due to limitations in their spatial coverage (Pelletier et al., 2018). Recent advancements in geospatial and machine learning models have shown promising strides in predicting SOC (Soil-Grids.org, 2020). However, these models often present disparities in their predictions, especially in areas with high SOC content. This highlights a need for concerted efforts to reduce these uncertainties to enhance confidence in mapping SOC storage (Hengl et al., 2017). Despite the burgeoning use of these geospatial models, our understanding of the global effects of land-use change, land management, and climate change on SOC remains incomplete (Wiesmeier et al., 2023). This knowledge gap is partially due to the variability of results and the limitations in the scope and geographical coverage of existing studies (Wiesmeier et al., 2023).

Furthermore, anthropogenic activities such as land conversion for crop production, land management practices, and climate change effects have significant impacts on SOC levels (Wiesmeier et al., 2023). However, the specific effects of these activities are still not fully understood. This lack of knowledge contributes to the uncertainties in SOC predictions (Wiesmeier et al., 2023).

The geospatial model developed in this project aims to fill these gaps. By leveraging machine learning and geospatial data, this model seeks to enhance the efficiency and cost-effectiveness of predicting SOC concentrations. The model also seeks to refine the predictions of existing geospatial models (Hengl et al., 2017).

The outputs of this project could have significant implications for soil resources management. The geospatial model could be used to identify areas with high SOC levels, potentially suitable for carbon sequestration or other management practices (Hengl et al., 2017). Additionally, the findings could enhance the accuracy of traditional soil sampling and laboratory analysis methods, contributing to the broader goal of preserving and restoring SOC stocks for climate change mitigation and adaptation (Wiesmeier et al., 2023).

2. Literature Review

The prediction of soil organic carbon (SOC) content plays a crucial role in environmental science, climate change studies, and agricultural practices. A variety of methods, including traditional soil sampling and lab analysis, geospatial modelling, and machine learning techniques, are applied in predicting SOC content (Akhtar-Aziz, Latif, & Azam, 2019). A critical assessment of these methodologies, their strengths, limitations, and potential areas of improvement, particularly with the use of the Land Use/Cover Area Frame Statistical Survey (LUCAS) dataset, is discussed in this literature review.

2.1 Traditional Soil Sampling and Laboratory Analysis

Traditional soil sampling and laboratory analysis are hailed for their accuracy and reliability in measuring SOC content. The precise nature of laboratory analysis, which involves direct soil sampling and controlled conditions, ensures highly accurate results (Akhtar-Aziz, Latif, & Azam, 2019). However, the strenuous and time-consuming process of soil sampling, coupled with its high cost, limits its efficiency. Additionally, the spatial coverage of soil sampling is limited, making it challenging to extrapolate results over larger geographical areas. These constraints indicate the need for methodologies that can provide broader coverage without compromising accuracy.

2.2 Geospatial Modelling Techniques

Geospatial modelling, leveraging spatially referenced data like satellite imagery or topographic maps, has surfaced as a viable alternative to traditional soil sampling. These techniques can predict soil properties across extensive geographical areas (Chen, Zhang, Li, Wang, & Wang, 2020). However, the accuracy of geospatial models

is often dependent on the quality and resolution of the input data, which can vary significantly (Gómez-Gutiérrez, & McBratney, 2016). Further, the robustness of different geospatial models varies, with deterministic models limited by their assumption of known physical relationships, and empirical models (statistical and machine learning models) requiring substantial data for reliable results. This suggests a gap in developing robust geospatial models that can provide accurate results under data-limited conditions.

2.3 Machine Learning in SOC Prediction

Machine learning's capability to handle complex and non-linear relationships makes it a promising tool for SOC prediction. Algorithms such as decision trees, random forests, and neural networks have shown remarkable results (Chen, Zhang, Li, Wang, & Wang, 2020). However, these techniques require large amounts of data and are susceptible to overfitting, especially with noisy data or when the relationships between predictors and soil properties are not clearly understood (Gómez-Gutiérrez, & McBratney, 2016). Moreover, the interpretability of machine learning models can be challenging, limiting their application in situations where understanding the relationships between variables is crucial. These considerations indicate a need for further research on strategies to improve model interpretability and robustness.

2.4 The LUCAS Dataset

The LUCAS dataset, a comprehensive survey of land use and soil properties across Europe, presents a valuable resource for the geospatial modelling of SOC content. However, it is important to note that the dataset has evolved over time, with changes in project stages, sampling methods, and geographical coverage.

The LUCAS survey was initiated by the European Statistical Office (EUROSTAT) in collaboration with the Directorate General responsible for Agriculture and the technical support of the Joint Research Center. The survey methodology involved observations taken at more than 250,000 sample points throughout the EU, rather than mapping the entire area under investigation. This approach allowed for the identification of changes in land use over time (esdac.jrc.ec.europa.eu, n.d.).

The project had different stages, with the first major soil survey conducted in 2009 across 23 Member States of the European Union (EU). This survey represented the first attempt to build a consistent spatial database of the soil cover across the EU based on standard sampling and analytical procedures. The survey was repeated in 2012, 2015, and 2018, each time with adjustments and additions to the methodology (esdac.jrc.ec.europa.eu, n.d.).

Over the years, the sampling methods have changed. For instance, in the 2018 LUCAS Soil survey, additional analyses were included for the first time, such as bulk density, soil biodiversity, visual assessment of soil erosion, and measurement of the thickness of the organic horizon in organic-rich soil (esdac.jrc.ec.europa.eu, n.d.).

While the LUCAS approach is designed for monitoring land use/land cover change, potential bias in the sampling design may not necessarily capture all soil characteristics in a country. The sampling could be biased due to the systematic selection of points on a grid with a 2 km spacing in Eastings (d'Andrimont et al., 2020).

The geographical coverage of the LUCAS survey has also expanded over time. In addition to EU member states, the 2015 survey included samples from non-EU countries such as Albania, Bosnia-Herzegovina, Croatia, North Macedonia, Montenegro, Serbia, and Switzerland (esdac.jrc.ec.europa.eu, n.d.).

Despite these advancements, the potential of integrating the LUCAS dataset with machine learning techniques requires further exploration, especially concerning capturing temporal variations.

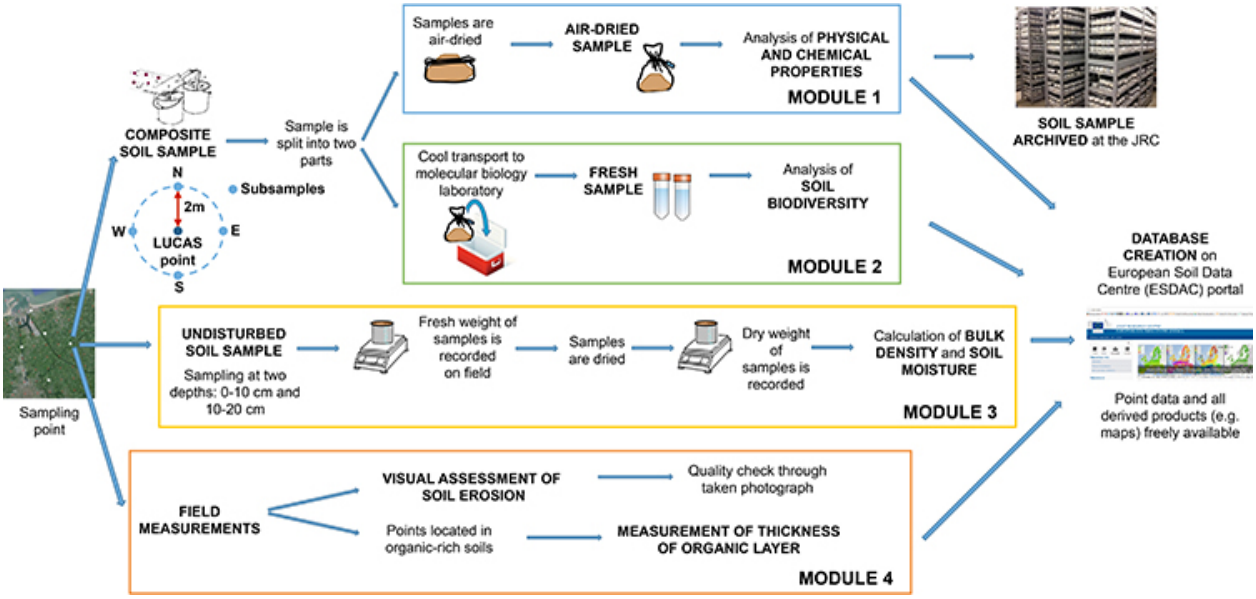

Figure 1. LUCAS Soil workflow from sampling to database generation (esdac.jrc.ec.europa.eu, n.d.)

2.5 Climate Models and Future SOC Predictions

The Köppen–Geiger climate classification dataset and NASA's global dataset are critical for predicting future SOC distribution (Cui et al., 2021; NASA, 2015). However, the accuracy of these models often relies on the accuracy of the underlying climate change projections, which can be uncertain. While machine learning models might help mitigate some of this uncertainty, challenges in model calibration and validation still persist. This emphasizes the need for continued research into integrating climate projections with machine learning models for SOC prediction.

2.6 Global Soil Mapping

Global soil mapping initiatives such as the GlobalSoilMap project provide a holistic view of global soil properties (Chen et al., 2021). However, these global mapping initiatives often face challenges due to data scarcity in certain regions, heterogeneity of data sources, and variable data quality, suggesting a need for more consistent and comprehensive data collection efforts.

2.7 Conclusion

SOC content prediction is a multifaceted process demanding various techniques and data sources. Traditional soil sampling and laboratory analysis offer a firm foundation, but limitations necessitate advancements like geospatial modelling and machine learning. Utilizing the LUCAS dataset in these modelling efforts is a promising area of research. However, challenges such as data quality, model robustness and interpretability, temporal coverage, and uncertainties in climate change projections, highlight potential areas for further research. Addressing these issues will be critical

for enhancing our understanding of SOC distribution and its implications for climate change.

Figure 2. Google Earth Engine (Google Developers, n.d.)

3. Methodology

3.1 Introduction

The methodology section aims to discuss the methods used to address the research objectives of this dissertation. It will describe the research methodology adopted, including data collection, preparation, analysis, and the critical evaluation of results. The chosen methods are justified based on their appropriateness for the research's specific objectives and constraints, such as access and time. Ethical implications and limitations of the chosen methodology are also discussed, and the methods are related back to the needs of the research questions. The reproducibility of the research is ensured by providing the necessary details.

3.2 Data Collection and Preparation

To address the research objectives, the LUCAS soil dataset will be merged with geospatial data retrieved from Google Earth Engine. The geospatial data will provide valuable predictor variables such as vegetation indices, topography, soil type, and land-use data. Seqana GmbH, Berlin, provided the dataset with proprietary merging from LUCAS and Google Earth Engine. Although this data cannot not be presented in the thesis due its proprietary nature, a data dictionary is provided to ensure transparency and reproducibility of the work.

3.3 Data Dictionary

The dataset has a wide range of columns, mostly related to soil properties, climate, and remote sensing data. The list of the entire columns has been provided in Appendix A. The brief summary of the main types of columns is as follows:

- DEM (Digital Elevation Model) columns: These columns provide information about elevation and slope.
    - dem_nasa_dem30__elevation
    - dem_nasa_dem30__slope

- Soil columns: These columns contain information about soil properties at different depths (b0, b10, b30, b60, b100) whereby the figures refer to the depth in centimetres.
    - Soil Organic Matter (OLM) properties: clay, bulk density (bd), soil organic carbon (SOC) pH, sand, and water content.

- MODIS columns: These columns contain data from the MODIS satellite, including land cover type, surface temperature, net primary productivity (NPP), and vegetation indices like normalized difference vegetation index (NDVI).

- Landsat (ls578_sr) columns: These columns provide data from the Landsat satellite, including spectral reflectance values for different bands like Blue, Green, Red, NIR, SWIR1, and SWIR2.

- IMERG precipitation columns: These columns contain precipitation data.

- ERA5-Land climate columns: These columns provide climate data such as temperature, soil temperature, volumetric soil water, and total precipitation.

3.4 Data Processing

The raw data undergoes several pre-processing steps to prepare it for further analysis and model development.

3.4.1 Data Cleaning

The first step is data cleaning, which is performed by the **`data_cleaning.py`** script. This script removes specific proprietary columns from the dataset that are not required for the subsequent analysis and model development. The columns dropped include "sample_id", "geom_id", "date_id", "depth_id", "top_depth", and "depth_len". These columns are removed because they contain proprietary information that is not relevant to the subsequent data analysis tasks. The cleaned dataset is then saved as a new CSV file.

3.4.2 Data Imputation

Following data cleaning, the '**`data_imputation.py'`** script handles missing data by performing data imputation. This script uses an iterative imputer with a Random Forest Regressor to fill in missing values in the dataset. Each column in the dataset is checked for missing values. If any are found, the iterative imputer is fit to the column, and the missing values are replaced with imputed ones. The imputation process is performed for each column individually, and logging messages are printed to indicate when the imputation begins and ends for each column. After all the columns have been processed, the imputed dataset is saved as a new CSV file.

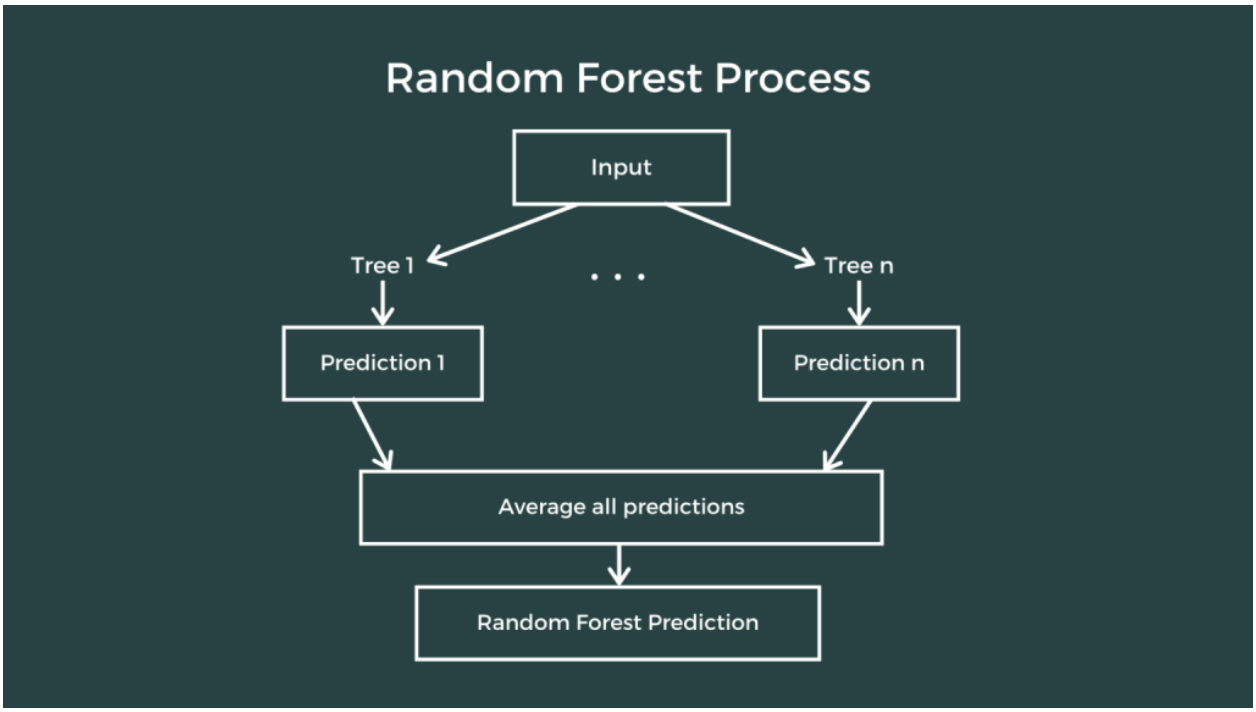

Figure 3. Random Forest Regressor (Reader, 2021)

These pre-processing steps ensure that the dataset is free of proprietary information, missing values are appropriately handled, and the data is in a suitable format for subsequent analysis and model development.

3.5 Data Analysis

A three-step data analysis process is employed with the aim of understanding and predicting the 'soc_percent' variable using the rest of the dataset's features.

3.5.1 Correlation Analysis

In the first phase, the Pearson correlation coefficients between 'soc_percent' and all other dataset features are calculated. This metric provides insights into the strength and direction of the linear relationship between 'soc_percent' and each feature using the script `'correlation_analysis.py'`. The coefficients are sorted in descending order to identify the features that are most strongly correlated with 'soc_percent'. The results are saved, and a bar graph is generated for coefficients above a specified

threshold, accompanied by a scatter plot matrix for the significant columns. This process helps identify which variables have a significant relationship with 'soc_percent'.

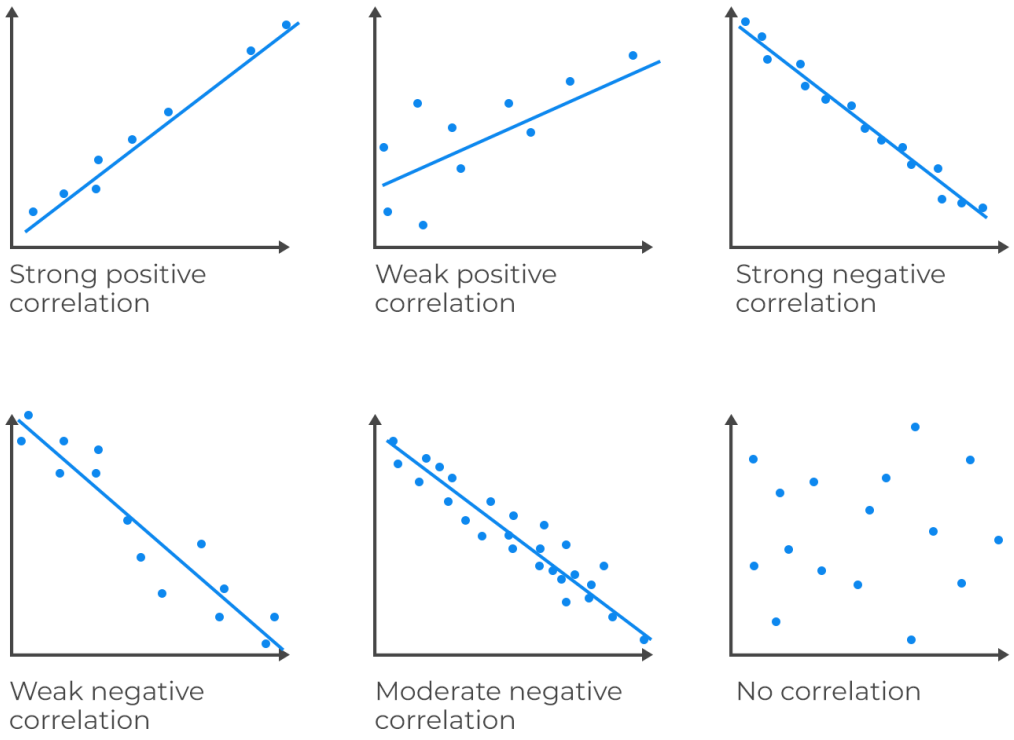

Figure 4. Relationship between sets of data (AnalystPrep | CFA® Exam Study Notes, 2021)

3.5.2 Regression Analysis

The second phase involves performing regression analysis on the dataset using the script **`'regression_analysis.py'`**. The dataset is split into training and testing subsets, with the 'soc_percent' variable serving as the target for prediction. A random forest regression model is trained on the dataset, and its performance is evaluated based on metrics like the R-squared value, root means square error (RMSE), and mean absolute error (MAE). The models are compared to identify the one that predicts 'soc_percent' with the highest accuracy.

3.5.3 Permutation Importance Analysis

The final phase involves evaluating the importance of each feature in predicting 'soc_percent' using the script **`'permutation_importance_analysis.py'`**. This process employs a machine learning algorithm from the previous step (Random Forest Regressor) to assess the contribution of each feature to the prediction of 'soc_percent'. The results are visualized in a bar plot, which provides a clear depiction of the most influential variables in predicting 'soc_percent'. This step helps identify the key features that significantly impact the prediction of 'soc_percent', aiding in the refinement of the models and improving prediction accuracy.

In conclusion, this three-step process is designed to provide a comprehensive understanding of the relationship between 'soc_percent' and the other variables in the dataset. Through correlation analysis, regression with feature importance analysis, and permutation importance analysis, the understanding, prediction, and interpretation of 'soc_percent' are robustly carried out. The use of visualization throughout ensures that the results are easily interpretable.

3.6 Machine Learning Model Development

Various machine learning techniques will be employed to develop statistical models predicting soil organic carbon (SOC) content using the processed dataset. We utilize the cuML library, which provides GPU-accelerated implementations of several popular algorithms, to perform our computations. The selected techniques include:

3.6.1 Lasso Regression

In the development of the Lasso Regression model, a linear regression model with L1 regularization is utilized, as implemented in the **`'lasso_regression.py'`** script.

The script starts with loading the dataset and splitting it into features and the target, which is the 'soc_percent' in this case. The data is then converted to cuDF dataframes for GPU computation and subsequently divided into training and test sets. The Lasso regression model is initialized and trained on the training set, and predictions are made on the test data. The model's performance is evaluated using mean squared error, and the trained model, along with its coefficients, is saved. Lastly, a plot of the predicted versus actual values and a residual plot are generated and saved. The plots include a text box displaying the metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 score.

3.6.2 Support Vector Machines (SVM)

The Support Vector Regression (SVR) model is developed using the **`'svm_regression.py'`** script that initially loads the dataset and splits it into features and the target variable, 'soc_percent'. The data is converted into cuDF dataframes to leverage GPU computation and then split into training and test sets. An additional step is introduced in this script where the data is scaled using the StandardScaler. An SVR model with a Radial Basis Function (RBF) kernel is then initialized and trained on the training set. The model is used to make predictions on the test data, and its performance is evaluated using the mean squared error. The trained model is saved for future use, and a plot of the actual versus predicted values along with a residual plot is created and saved. Each plot includes a text box showing the metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 score.

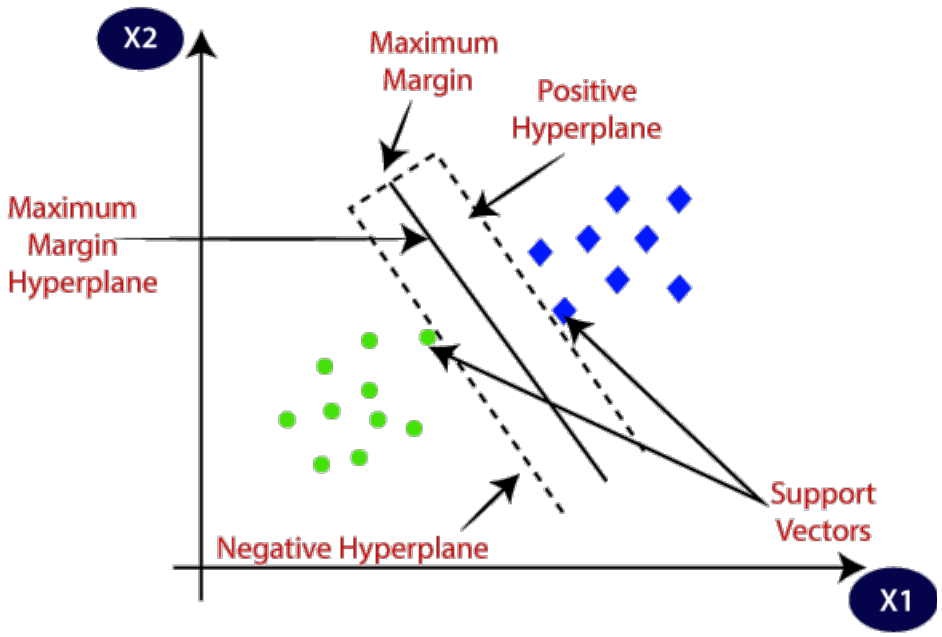

Figure 5. SVM Algorithm (JavaTPoint, n.d.)

3.6.3 LightGBM Regression

The LightGBM regression model is developed through the **`'lgbm_regression.py'`** script that loads the dataset, splitting it into features and the target variable, 'soc_percent'. The data is converted into cuDF dataframes to leverage GPU computation and then partitioned into training and test sets. A LightGBM Regressor model is subsequently initialized and trained on the training set, which is converted into NumPy arrays for compatibility. The trained model is then used to make predictions on the test set, and the model's performance is evaluated using the mean squared error. The model, once trained, is saved for future use, and plots of the actual versus predicted values, along with a residual plot, are generated and saved. Each plot includes a text box displaying the metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 score. Additionally, the script extracts and saves the feature importances from the trained model to understand the impact of each feature on the predictions.

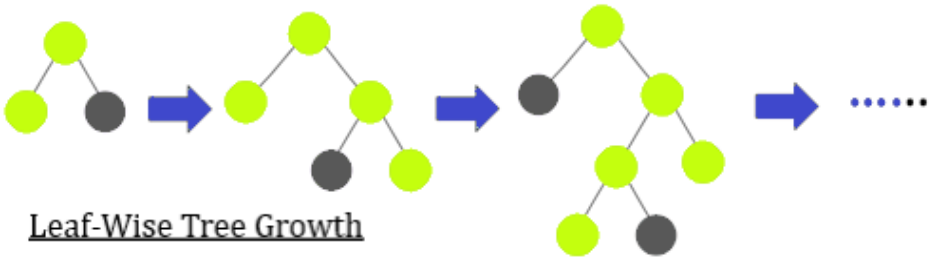

Figure 6. LightGBM Architecture (GeeksforGeeks, 2020)

In summary, Lasso Regression, SVR, and LightGBM were chosen for their respective strengths in variable selection, high-dimensional data handling, and large dataset performance, all of which are crucial aspects of our SOC prediction task. Their comparative performance in our methodology will help identify the most effective model for predicting SOC content.

3.7 Deep Learning Model Development

Deep learning models predicting SOC content have been developed using the processed dataset. Python's TensorFlow and Keras libraries were utilized for their flexibility and power in building and evaluating deep learning models. The chosen architectures have proven to be particularly relevant for numerical tabular data and have been trained efficiently using GPUs.

3.7.1 Multilayer Perceptron (MLP)

The MLP model was developed using a Sequential model in Keras, as outlined in the **`'mlp_regression.py'`** script. The script loaded the dataset, split it into features and the target variable ('soc_percent'), and then standardized it using the StandardScaler from sklearn. The data was then partitioned into training and testing datasets. An MLP model with two dense layers (64 and 32 neurons) and the 'ReLU' activation function was created. The Adam optimizer and the mean squared error loss function were used for model compilation. Early stopping was employed to avoid

overtraining the model. Once the model was trained, predictions were generated on the test set. Performance evaluation was conducted based on mean absolute error, root mean squared error, and R2 score.

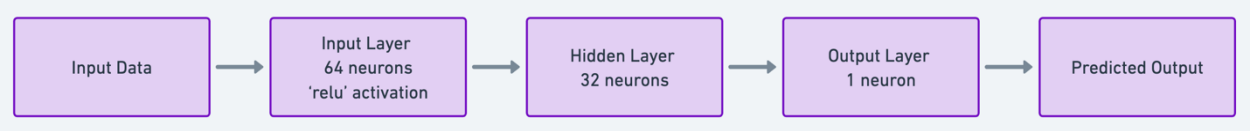

Figure 7. MLP model architecture

3.7.2 Deep Feedforward Networks (DFFNs)

The DFFN model was built using the **`'dffn_regression.py'`** script. The dataset was loaded and divided into feature vectors and the target variable, 'soc_percent'. StandardScaler was used for feature normalization before splitting the data into training and test sets. The DFFN model consisted of an input layer of 64 neurons, three hidden layers of 64 neurons each, and an output layer. Adam optimizer was used with 'mean_squared_error' as the loss function, and early stopping was employed to prevent overtraining. After the training process, predictions were made using the test set, and the model's performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score.

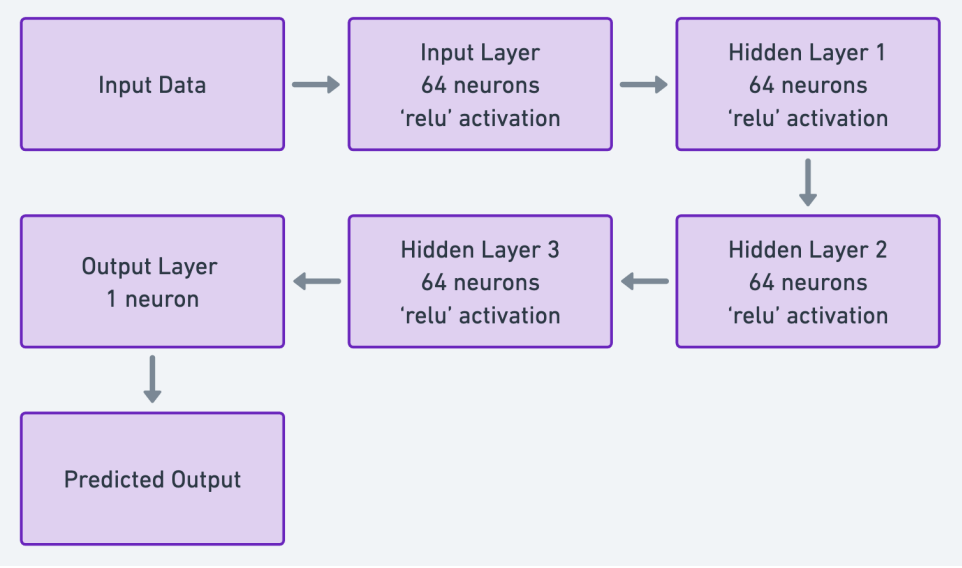

Figure 8. DFFN model architecture

3.7.3 Long Short-Term Memory (LSTM)

The `'lstm_regression.py'` script utilized Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN), for regression analysis. It began by loading the dataset, separating it into features and target variables, and normalizing the features using a StandardScaler. For the LSTM model, the input data was reshaped into a 3D array (samples, timesteps, features). An interesting aspect of this script was the separate handling of the 'year' column, which was standardized separately and then added back to the features. The LSTM model was built using the Sequential API from TensorFlow, and early stopping was employed to prevent overtraining. The performance of the model was evaluated based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score.

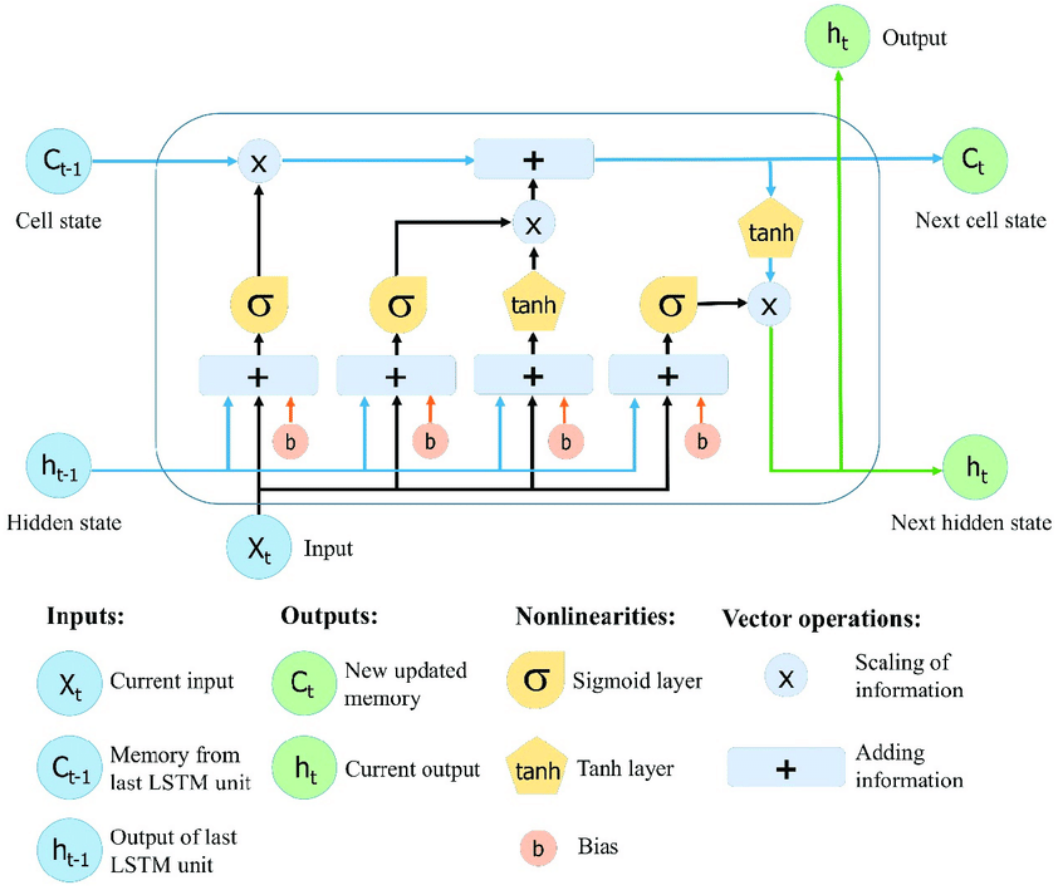

Figure 9. LSTM Cell Architecture (Le et al., 2019)

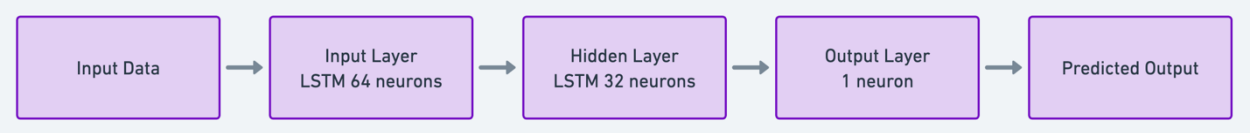

Figure 10. LSTM model architecture

3.7.4 Hybrid Model

The **`'wide_and_deep_regression.py'`** script applied the Wide & Deep learning approach for regression analysis. The "wide" part of the model referred to a linear model with wide input features, and the "deep" part of the model referred to a deep neural network with hidden layers that provided abstract representations. The data was pre-processed by splitting it into features (X) and a target variable (y), standardizing the features, and splitting them into training and test datasets. A wide model and a deep model were created separately. Both models took the same input, and their outputs were combined into a single output through concatenation. This combined output was then passed through a single neuron to produce the final output. The model was compiled with the Adam optimizer and Mean Squared Error (MSE) as the loss function. Early stopping was set up to prevent overfitting during the training phase. The performance of the model was evaluated based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score.

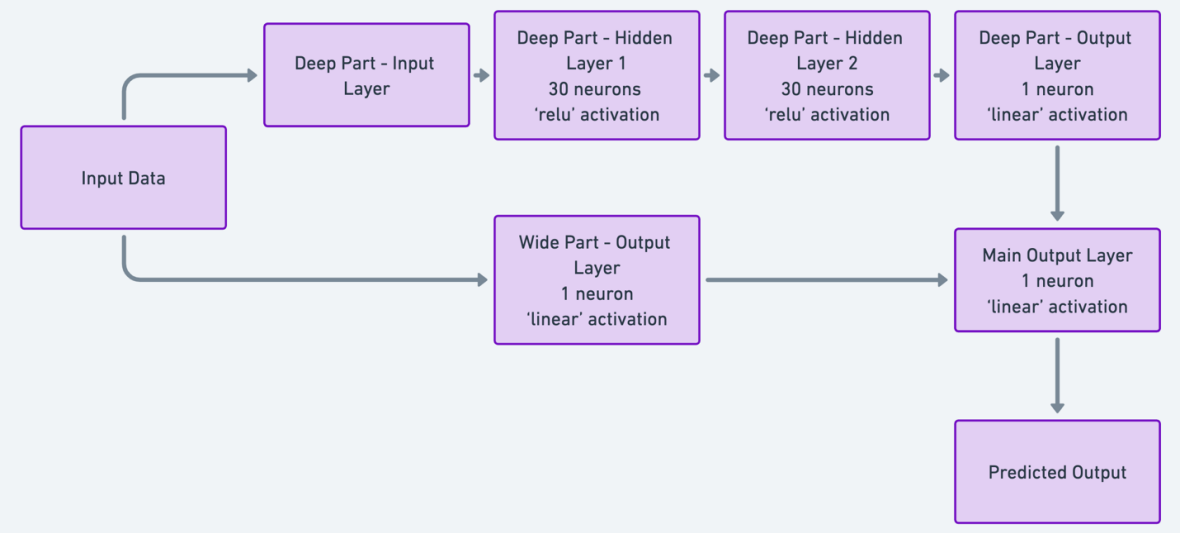

Figure 11. Hybrid model architecture

These deep learning models were evaluated based on their performance metrics and their ability to generalize to new data.

3.8 Model Validation

All the developed models, both machine learning and deep learning, will be evaluated based on specific validation/testing metrics to assess their accuracy and reliability in predicting soil organic carbon levels. These metrics have been selected due to their sensitivity to different aspects of model performance.

3.8.1 Root Mean Squared Error (RMSE)

RMSE quantifies the average prediction error of our models in the same units as the dependent variable ('soc_percent'). By giving higher weight to larger errors, RMSE is particularly valuable when large prediction errors are undesirable.

3.8.2 Mean Absolute Error (MAE)

MAE provides a measure of the average prediction error, similar to RMSE. However, MAE is more robust to outliers as it does not square the errors. It is particularly useful

when we want to understand the average magnitude of the errors, regardless of their direction.

3.8.3 R2 Score (Coefficient of Determination)

The R2 Score measures how well the model's predictions fit the actual data. It quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s), providing a clear measure of model fit.

By using these metrics, we aim to ensure the selected model's soil organic carbon content predictions generalize well to new, unseen data.

3.9 Ethical Implications and Limitations

3.9.1 Ethical Implications

The research will be conducted using publicly available data, ensuring that no sensitive information or data that could infringe upon privacy rights is utilized. Additionally, the research will be designed to contribute to the scientific understanding of soil organic carbon levels and their potential for carbon sequestration, promoting environmentally responsible land management practices.

3.9.2 Limitations

The chosen methodology and circumstances of the study have presented several limitations.

- Firstly, the accuracy of remote sensing data is subject to factors such as cloud cover and sensor calibration.
- Secondly, geospatial modelling techniques may not fully capture local variations in SOC levels due to the scale of the analysis.

- Thirdly, the extrapolation of the model to areas beyond the study area may result in reduced accuracy due to differences in environmental factors and data availability.
- Fourthly, the study period of 2009-2018 may not fully capture long-term trends in SOC levels.
- Fifthly, the research encountered specific hardware constraints. The vast nature of the dataset demanded significant computational resources, exceeding what was available via the Google Cloud Platform (GCP) Virtual Machine (VM) (also called Instances) with substantial CUDA memory. These limitations posed substantial challenges in implementing complex techniques such as GridSearch for hyperparameter tuning and extensive experimentation with data pre-processing, additional models, and further feature engineering. This also involved financial constraints as GCP VM Instances are not cheap.
- Sixthly, an imminent constraint was the time required for reconfiguring the entire setup and ensuring its integration with Data Version Control (DVC) for data retrieval inside GCP VM. This task is quite expensive and time-consuming, making it difficult to manage alongside the submission deadline.
- Seventhly, there were inconsistencies in the data collection process. The data from Bulgaria and Romania were sampled in 2012, but these results are not included in the primary data collection of 2009/2012. This discrepancy could impact the overall analysis and findings as it introduces variations in the dataset from different time periods and limits the available samples for the study.
- Eighthly, the imminent thesis submission deadline imposed constraints on the potential changes and improvements that could be incorporated into the project

methodology. At this advanced stage in the project, a comprehensive change in methodology could jeopardize meeting the submission deadline.

These limitations, unique to my individual context, should be considered when interpreting the results of the study.

3.10 Relating Methods to Research Questions

The methodology adopted is designed to address each of the research questions:

3.10.1 Effectiveness of Machine Learning Models in Predicting SOC Levels

The three regression machine learning models (Lasso, SVM, and Gradient Boosting) will be trained using the predictor data collected from Google Earth Engine. The effectiveness of these models will be evaluated based on their ability to predict the SOC levels of the soil samples collected between 2009 and 2018.

3.10.2 Effectiveness of Deep Learning Models in SOC Prediction

Similarly, the effectiveness of the four regression-based deep learning models (MLP, DFFN, LSTM and Hybrid), in predicting SOC levels will be evaluated. The model will be trained and tested on the same dataset and its prediction accuracy will be compared with that of the machine learning models.

3.10.3 Comparative Performance of the Models

The predictive accuracy of the machine learning and deep learning models will be compared to identify the model that produces the most accurate results. The comparison will be based on metrics such as mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R2 or R-squared).

3.10.4 Significance of the Predictor

The importance of different predictors in each model will be assessed to understand their contribution to the predictive power of the model. Feature importance or significance will be calculated and the most influential predictor for each model will be identified.

3.10.5 Application Beyond the Study Area

The generalizability of the models will be evaluated to understand their potential application beyond the study area. This will be done by testing the models on data from other regions or different time periods and assessing their predictive accuracy.

3.11 Reproducibility of the Work

3.11.1 Detailed Methodology

The reproducibility of the research is fundamentally ensured by providing a comprehensive methodology section. This section includes detailed descriptions of the data collection, preparation, and processing steps, along with the analysis and model development processes. Each stage is documented in such a way that it guides other researchers to follow the same process, facilitating the replication of the study and a deeper investigation into the research questions.

3.11.2 Data Dictionary

Transparency is further enhanced by the inclusion of a data dictionary. This tool provides explicit definitions and descriptions of all variables, data fields, and codes used in the dataset. By clearly defining the data elements, the data dictionary assists other researchers in understanding the data's structure and meaning, thereby promoting replicability and reducing potential misinterpretations.

3.11.3 Data Accessibility and Adaptability

While the data utilized in this project is proprietary and not available to the public, the code and methods are designed to be adaptable to other datasets. Although the proprietary nature of the data may limit the direct replication of the results, the comprehensive documentation and openly accessible code allow for the methods to be tested and validated on similar datasets. This adaptability underlines the practical reproducibility of the work, in which the methods can be applied to new data to reproduce the analytical process and corroborate the findings.

3.11.4 Code Availability

The GitHub repository, "soc-prediction-research," serves as a well-structured resource, encapsulating all Python scripts and associated files used in the research. It is organized into several directories, each with a specific role, facilitating easy navigation and understanding of the project.

The '.dvc' directory is associated with Data Version Control, a tool that is used for tracking machine learning models and datasets. The '.github/workflows' directory is set up for Continuous Integration/Continuous Deployment (CI/CD), ensuring that the code is always in a deployable state.

The 'data' directory is where the data used in the project is stored. However, it should be noted that this data is proprietary and not publicly accessible. The 'logs' directory contains logs generated by the application, providing a record of the computational processes. The 'models' directory is where the machine learning models used or produced by the project are stored.

The 'results' directory houses the outcomes of the machine learning models. The 'src' directory is where the Python scripts for data pre-processing, model training, evaluation, and visualization are located. These scripts form the core of the project, detailing the exact methods used in the research.

Several other files that facilitate the setup and execution of the project are also included. The 'Dockerfile' outlines the setup for Docker, a platform used for automating the deployment of applications. The 'requirements.txt' and 'conda-requirements.txt' files list the Python and Conda dependencies needed for the project.

In essence, the 'soc-prediction-research' GitHub repository is a well-organized resource that houses all the Python scripts used in the research. Its user-friendly design, clear organization, and thorough documentation make it easier for anyone interested in replicating or building upon the work.

3.11.5 Licensing

The project is licensed under the terms of the MIT license. This permissive free software license allows others to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the software, as long as they credit the original creators and adhere to the license's terms. This licensing ensures that the project can be widely disseminated and reused, contributing to the overall reproducibility of the work.

4. Results

4.1 Introduction

This section endeavours to present the outcomes derived from a thorough analysis of the dataset, encompassing primary research undertaken within the study. The primary objectives were centred around identifying the most accurate, cost-effective, and efficient methodologies for predicting soil organic carbon (SOC) content.

4.2 Data Cleaning and Imputation

As part of the initial data processing, columns that were identified as proprietary and irrelevant were omitted, adhering to the methodology detailed earlier. The final dataset, after the pre-processing stage, consisted of 62,860 data points spanning across 173 features. Each column's missing values, which were subsequently imputed, are systematically documented in Appendix B.

4.3 Correlation Analysis

The correlation analysis using Pearson correlation coefficients reveals the linear relationship between the 'soc_percent' and other features. The Pearson correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

In the context of this study, a positive correlation means that as the value of the feature increases, the 'soc_percent' also increases, and vice versa. A negative correlation means that as the value of the feature increases, the 'soc_percent' decreases, and vice versa.

Here's a breakdown of the correlations (as per Figure 12):

1. soil_olm_bd__b0 = -0.44: 'soil_olm_bd__b0' likely refers to the bulk density of the soil at the surface level (0 cm depth). Bulk density is a soil property that can influence the amount of organic carbon a soil can store. The negative correlation suggests that as the bulk density at the surface level increases, the soil organic carbon (SOC) percentage tends to decrease. This could be because higher bulk density often indicates compaction and reduced pore space, which can limit the soil's capacity to store organic carbon.
2. soil_olm_bd__b10 = -0.41: 'soil_olm_bd__b10' likely refers to the bulk density of the soil at 10 cm depth. Similar to 'soil_olm_bd__b0', the negative correlation suggests that higher bulk density at this depth is associated with lower SOC percentage.
3. soil_olm_soc__b0 = 0.64 and soil_olm_soc__b10 = 0.64: These likely refer to the soil organic carbon content at the surface level and at 10 cm depth, respectively. The strong positive correlations suggest that higher organic carbon content at these depths is associated with a higher overall SOC percentage. This makes sense as the organic carbon content at various depths is a direct component of the total SOC.
4. soil_olm_soc__b30 = 0.53, soil_olm_soc__b60 = 0.52, and soil_olm_soc__b100 = 0.51: These likely refer to the soil organic carbon content at 30 cm, 60 cm, and 100 cm depths, respectively. The positive correlations suggest that higher organic carbon content at these depths is associated with higher overall SOC percentage, although the relationships are not as strong as at the surface and 10 cm depths. This could reflect the fact that organic carbon content typically decreases with depth in the soil profile.

5. soil_olm_ph__b0 = -0.41 and soil_olm_ph__b10 = -0.41: These likely refer to the pH of the soil at the surface level and at 10 cm depth, respectively. The negative correlations suggest that higher soil pH (more alkaline conditions) at these depths is associated with a lower SOC percentage. This could be because certain forms of organic carbon are more soluble and hence more likely to be lost from the soil under alkaline conditions.

This correlation was further visualized and substantiated through a scatter plot, as indicated in Figure 13.

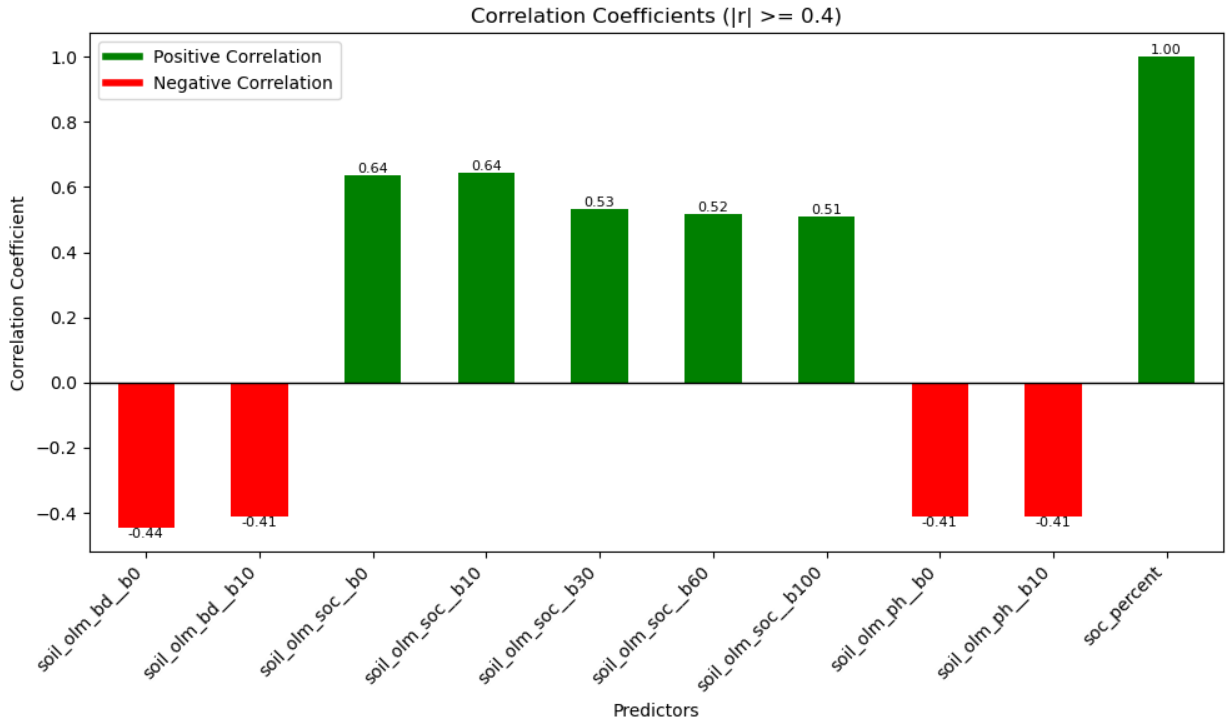

Figure 12. Feature's Pearson correlation to 'soc_percent'

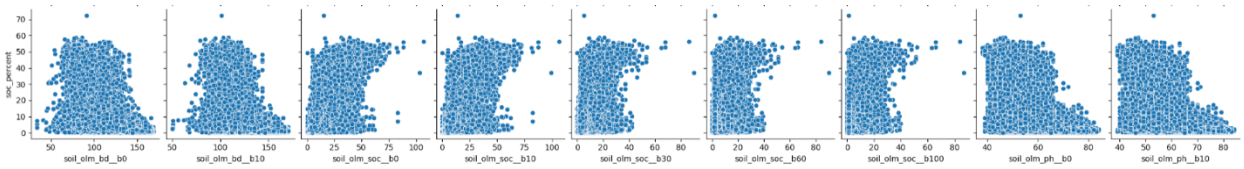

Figure 13. Scatterplot of correlated features

4.4 Regression Analysis

The performance of the Random Forest regression model was assessed and was found to be quite satisfactory. The following metrics were obtained:

- MAE: 2.523
- RMSE: 5.498
- R2: 0.559

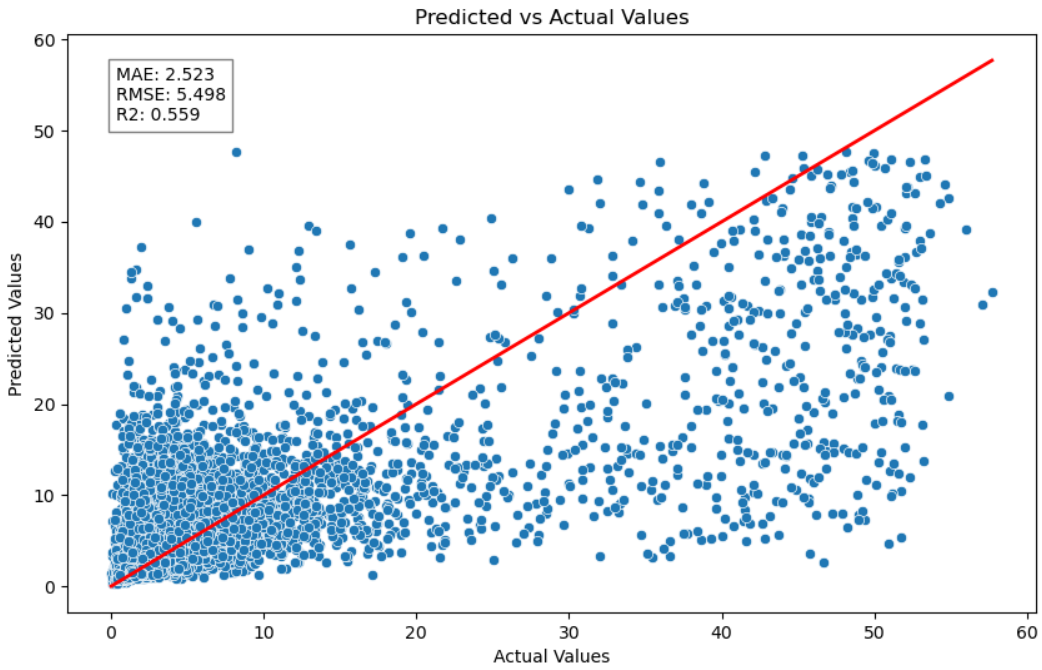

Figure 14. Predicted vs Actual Plot for Regression Analysis

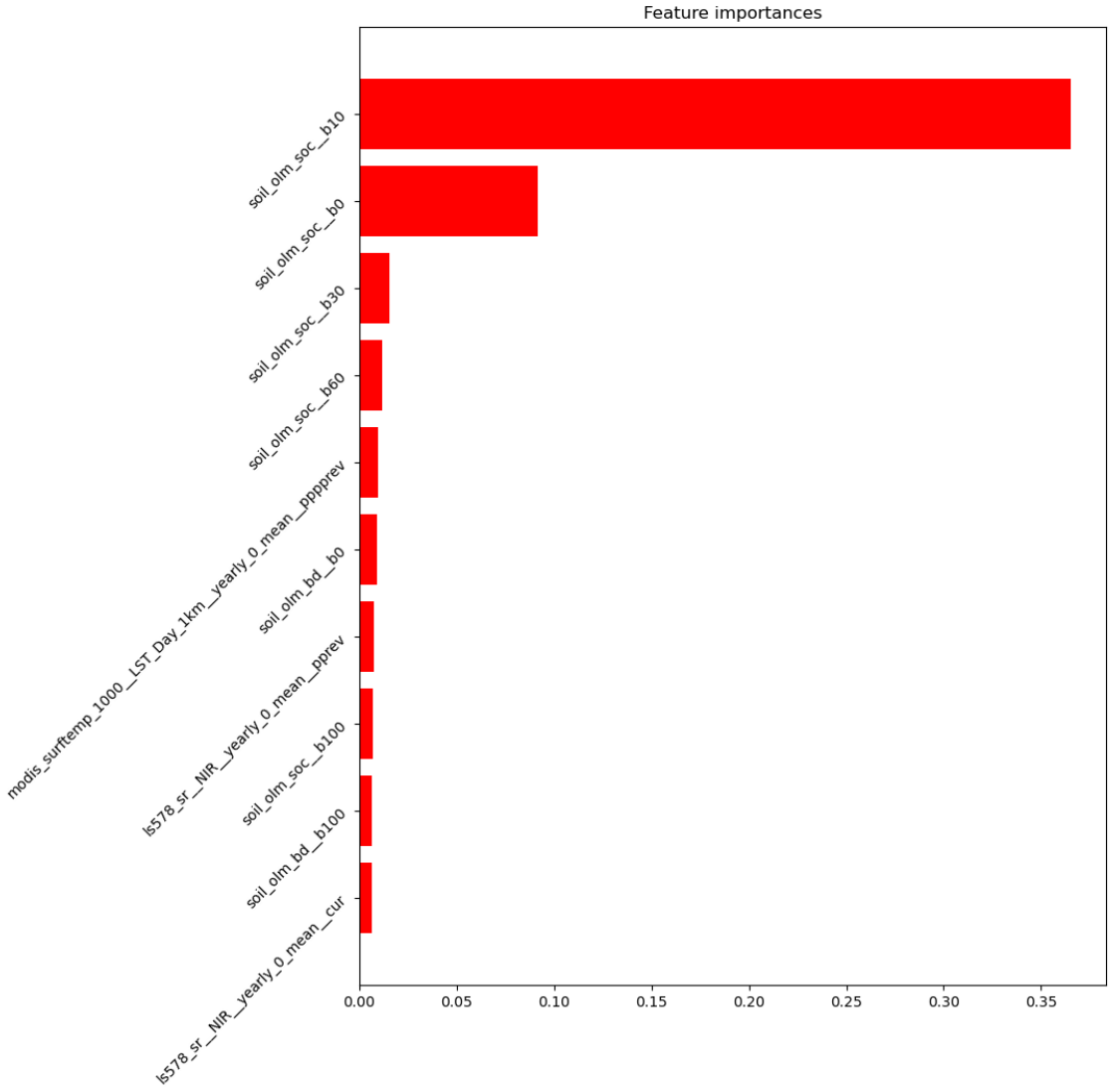

Figure 15. Feature importance from Regression Model

The features 'soil_olm_soc__b10' and 'soil_olm_soc__b0' were observed to be substantial in the regression model, as illustrated in Figure 15.

4.5 Permutation Importance Analysis

The permutation importance analysis revealed the most influential features that play a pivotal role in predicting 'soc_percent'. As referenced in Figure 16, they are as follows:

- ‘soil_olm_soc__b10’,
- ‘soil_olm_soc__b0’,
- ‘modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pppprev’,
- ‘soil_olm_bd__b0’, and
- ‘soil_olm_soc__b30’.

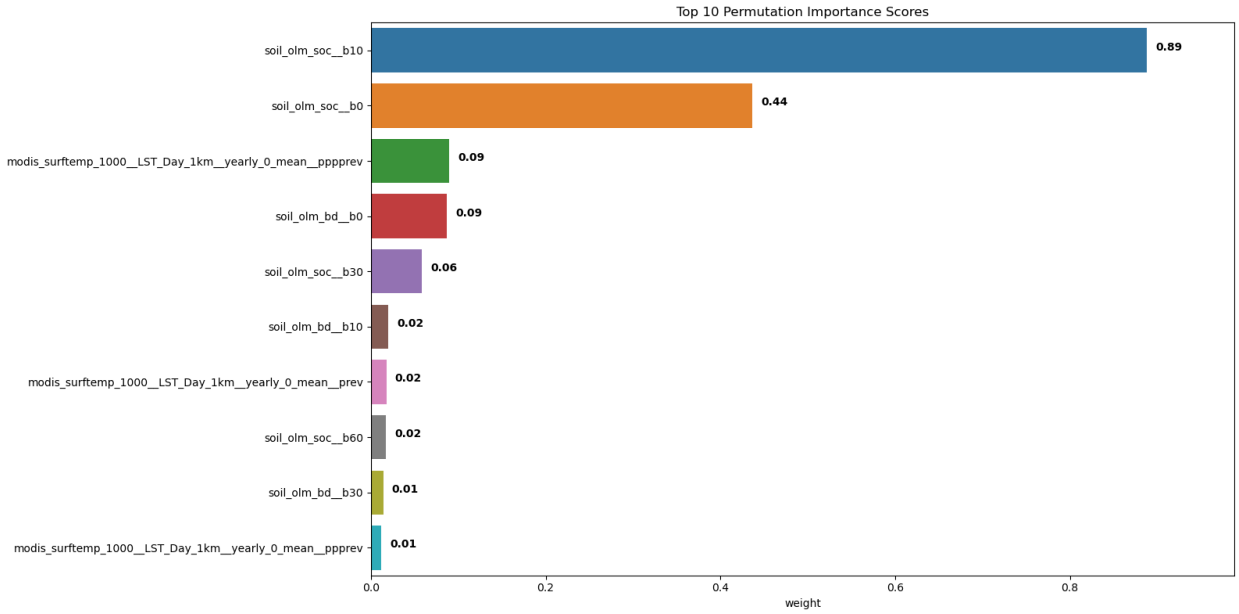

Figure 16. Top 10 Permutation Scores

4.6 Machine Learning Model Development

Machine Learning models like Lasso Regression, Support Vector Machines (SVM), and LightGBM Regression were tested, and the metrics were calculated for the testing data.

4.6.1 Lasso Regression Results

For the Lasso Regression model, the following results were obtained:

- MAE: 2.991
- RMSE: 6.038
- R2: 0.454

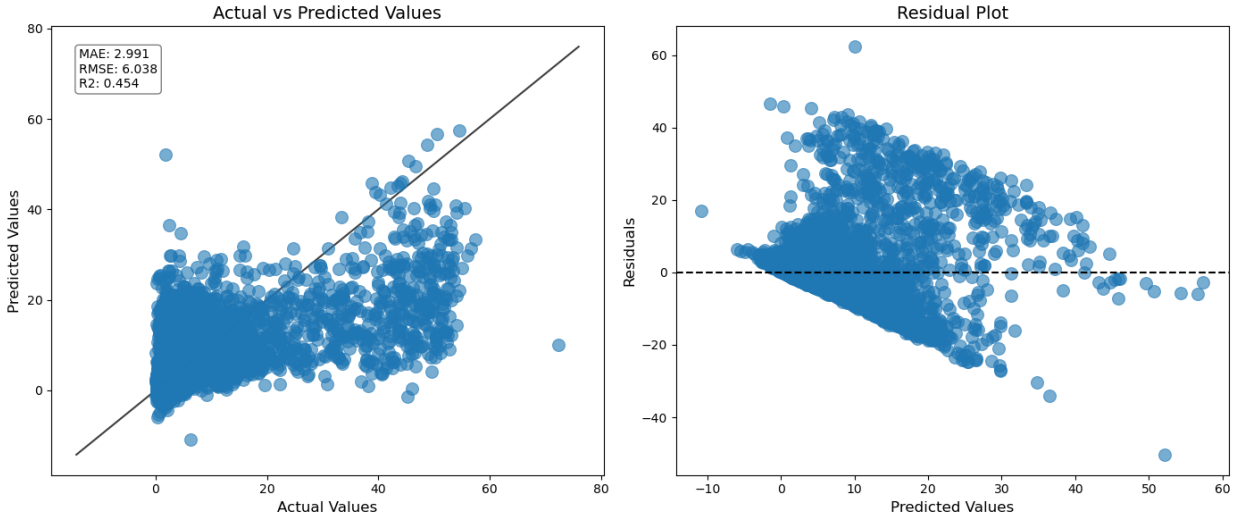

Figure 17. Actual vs Predicted Values & Residual Plot for Lasso Regression Model

4.6.2 Support Vector Machines (SVM) Results

The SVM model yielded the following metrics:

- MAE: 3.833
- RMSE: 7.613
- R2: 0.132

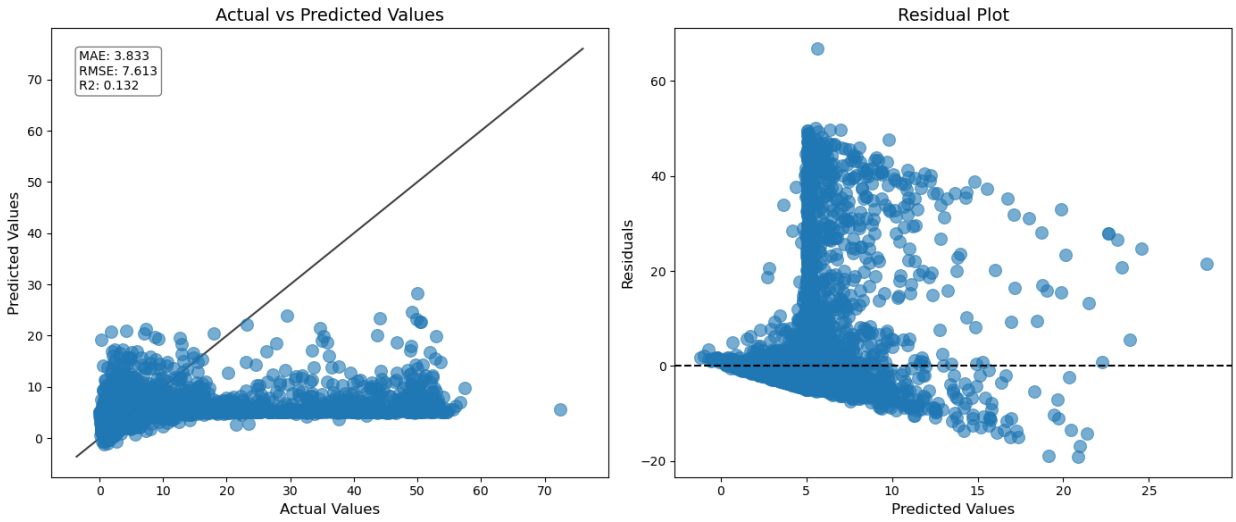

Figure 18. Actual vs Predicted Values & Residual Plot for SVM Regression Model

These results indicate a rather poor performance when compared to the other models, as is evident from the low R2 score.

4.6.3 LightGBM Regression Results

The LightGBM Regression model produced the following outcomes:

- MAE: 2.504
- RMSE: 5.587
- R2: 0.532

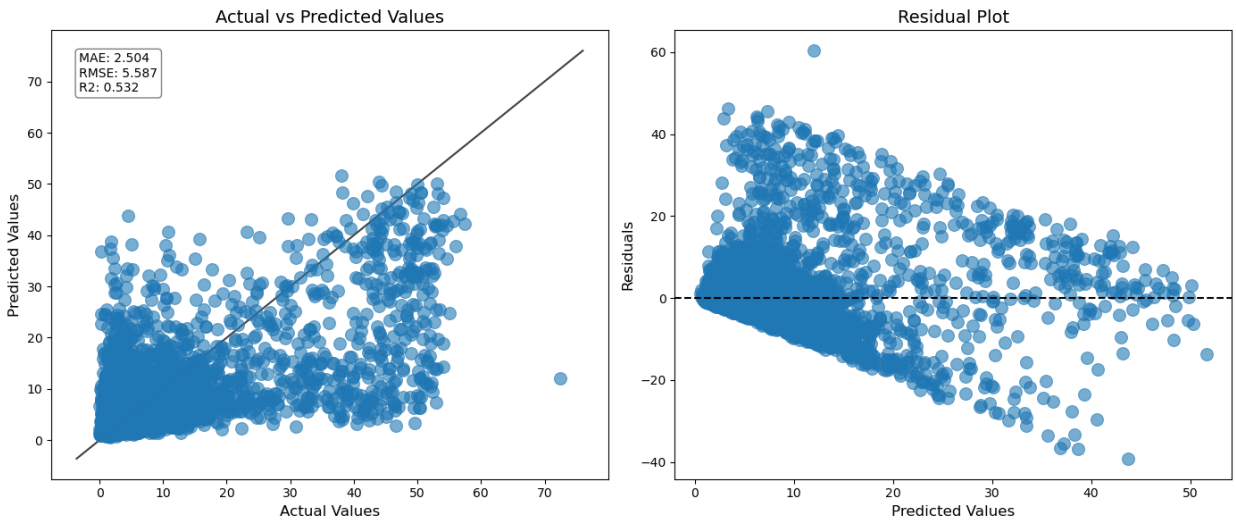

Figure 19. Actual vs Predicted Values & Residual Plot for LightGBM Regression Model

4.7 Deep Learning Model Development

Deep Learning models such as the Multilayer Perceptron (MLP), Deep Feedforward Networks (DFFNs), Long Short-Term Memory (LSTM), and Hybrid models were put to the test, with metrics calculated for the testing data.

4.7.1 Multilayer Perceptron (MLP) Results

The outcomes from the MLP model are as follows:

- MAE: 2.560
- RMSE: 5.673
- R2: 0.530

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                11072

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 1)                 33

=================================================================
Total params: 13,185
Trainable params: 13,185
Non-trainable params: 0
_________________________________________________________________
```

Figure 20. MLP Model Summary

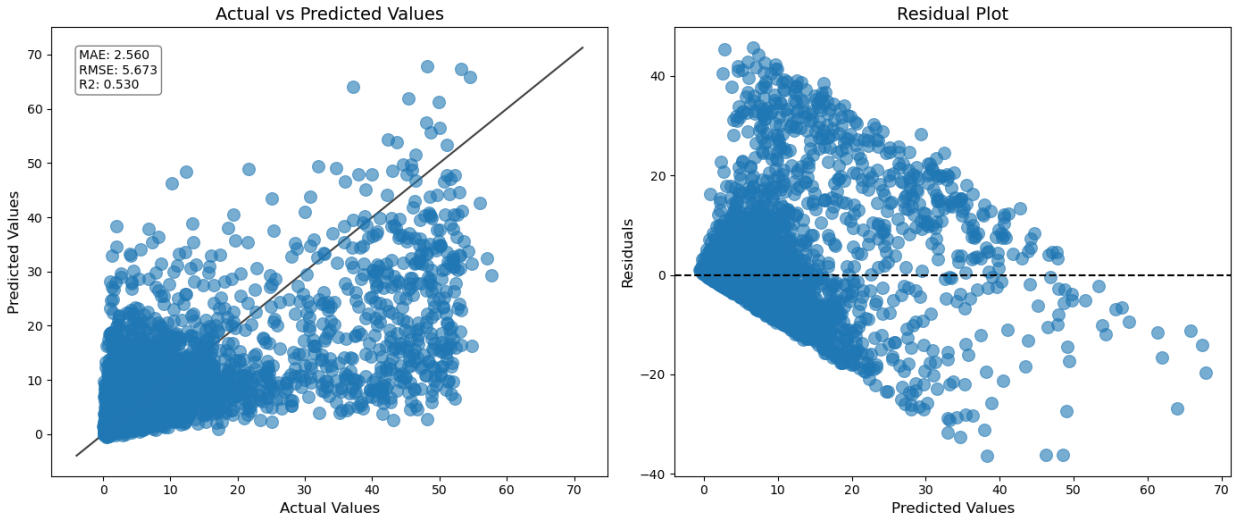

Figure 21. Actual vs Predicted Values & Residual Plot for MLP Regression Model

4.7.2 Deep Feedforward Networks (DFFNs) Results

The results for the DFFNs model were as follows:

- MAE: 2.531
- RMSE: 5.743
- R2: 0.519

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                11072

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 64)                4160

 dense_3 (Dense)             (None, 64)                4160

 dense_4 (Dense)             (None, 1)                 65

=================================================================
Total params: 23,617
Trainable params: 23,617
Non-trainable params: 0
_________________________________________________________________
```

Figure 22. DFFN Model Summary

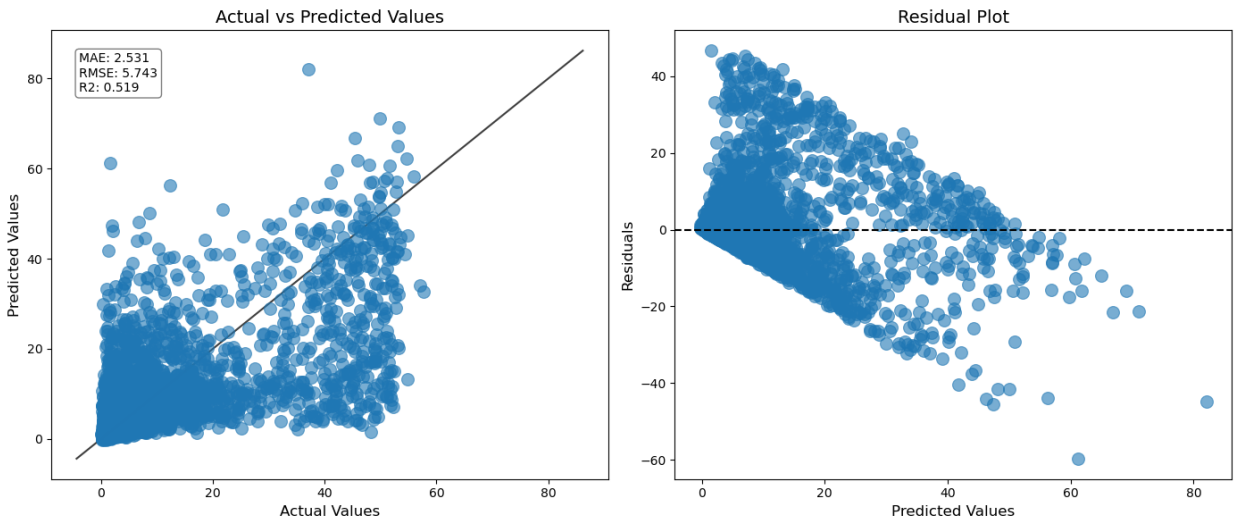

Figure 23. Actual vs Predicted Values & Residual Plot for DFFN Regression Model

4.7.3 Long Short-Term Memory (LSTM) Results

The LSTM model yielded the following outcomes:

- MAE: 2.649
- RMSE: 5.706
- R2: 0.525

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 1, 64)             60928

 lstm_1 (LSTM)               (None, 32)                12416

 dense (Dense)               (None, 1)                 33

=================================================================
Total params: 73,377
Trainable params: 73,377
Non-trainable params: 0
_________________________________________________________________
```

Figure 24. LSTM Model Summary

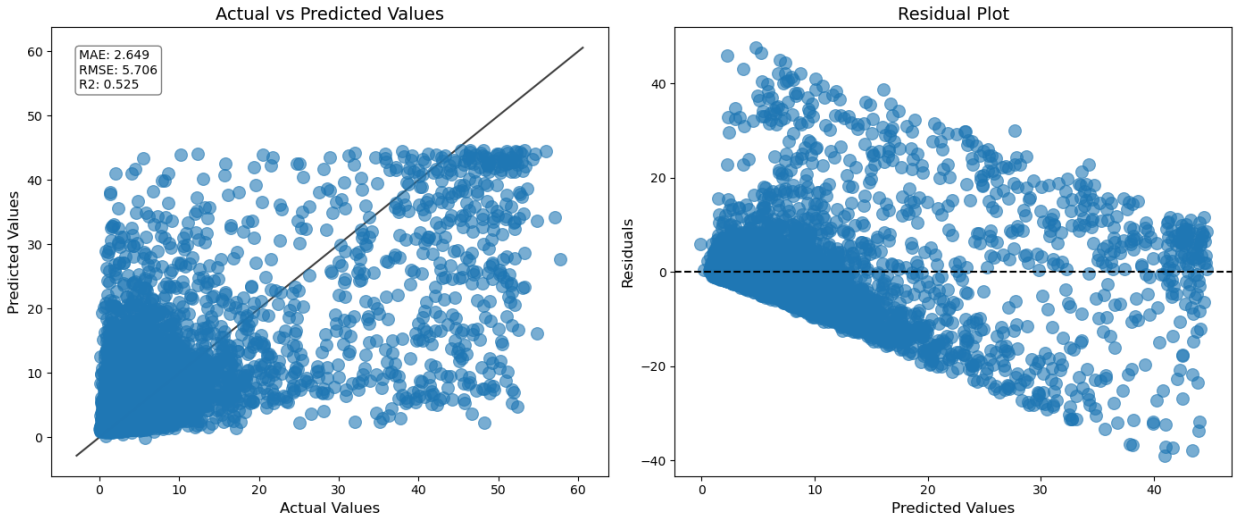

Figure 25. Actual vs Predicted Values & Residual Plot for LSTM Regression Model

4.7.4 Hybrid Model Results

The Hybrid model produced the following metrics:

- MAE: 2.610
- RMSE: 5.634
- R2: 0.537

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 deep_input (InputLayer)        [(None, 172)]        0           []

 dense (Dense)                  (None, 30)           5190        ['deep_input[0][0]']

 wide_input (InputLayer)        [(None, 172)]        0           []

 dense_1 (Dense)                (None, 30)           930         ['dense[0][0]']

 wide_output (Dense)            (None, 1)            173         ['wide_input[0][0]']

 deep_output (Dense)            (None, 1)            31          ['dense_1[0][0]']

 concatenate (Concatenate)      (None, 2)            0           ['wide_output[0][0]',
                                                                  'deep_output[0][0]']

 main_output (Dense)            (None, 1)            3           ['concatenate[0][0]']

==================================================================================================
Total params: 6,327
Trainable params: 6,327
Non-trainable params: 0
__________________________________________________________________________________________________
```

Figure 26. Hybrid Model Summary

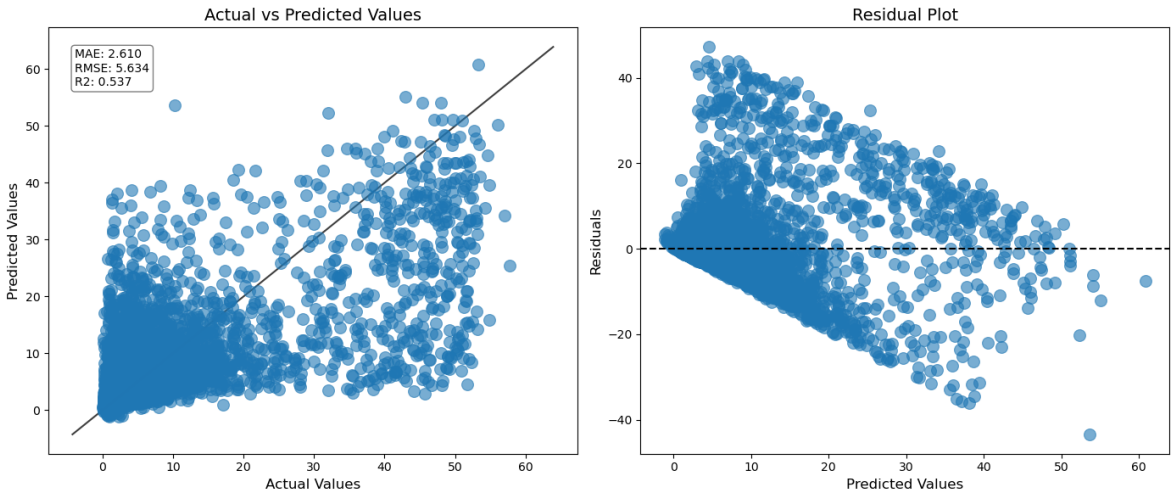

Figure 27. Actual vs Predicted Values & Residual Plot for Hybrid Regression Model

4.8 Model Comparison

This section aims to compare and contrast the performances of all the models, both machine learning and deep learning, used in this study. The best-performing models were identified to be the LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model. These models stood out in their performance due to their superior capability in handling high-dimensional data and their robustness to overfitting. A notable difference was observed between the machine learning and deep learning models, with the deep learning models, specifically the MLP and Hybrid models, showing slightly better performance.

Model Type	MAE	RMSE	R2
RF Regression	2.523	5.498	0.559
Lasso Regression	2.991	6.038	0.454
SVM Regression	3.833	7.613	0.132
LGB Regression	2.504	5.587	0.532
MLP Regression	2.560	5.673	0.530
DFFN Regression	2.531	5.743	0.519
LSTM Regression	2.649	5.706	0.525
Hybrid Regression	2.610	5.634	0.537

Figure 28. Model Comparison Table

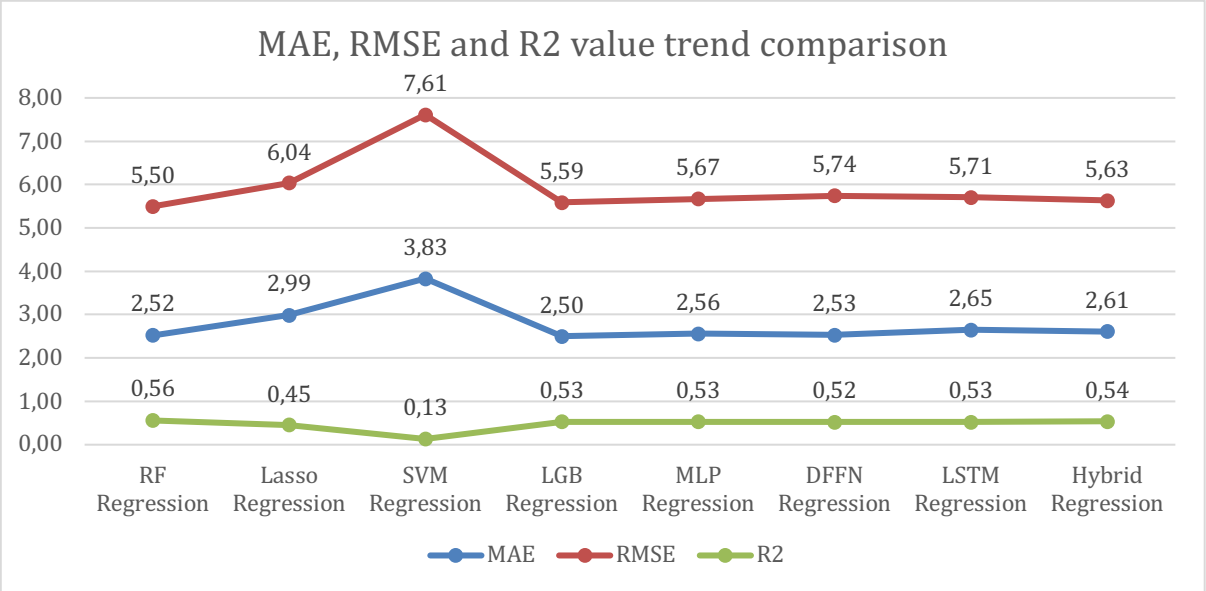

Figure 29. Model metrics trend comparison

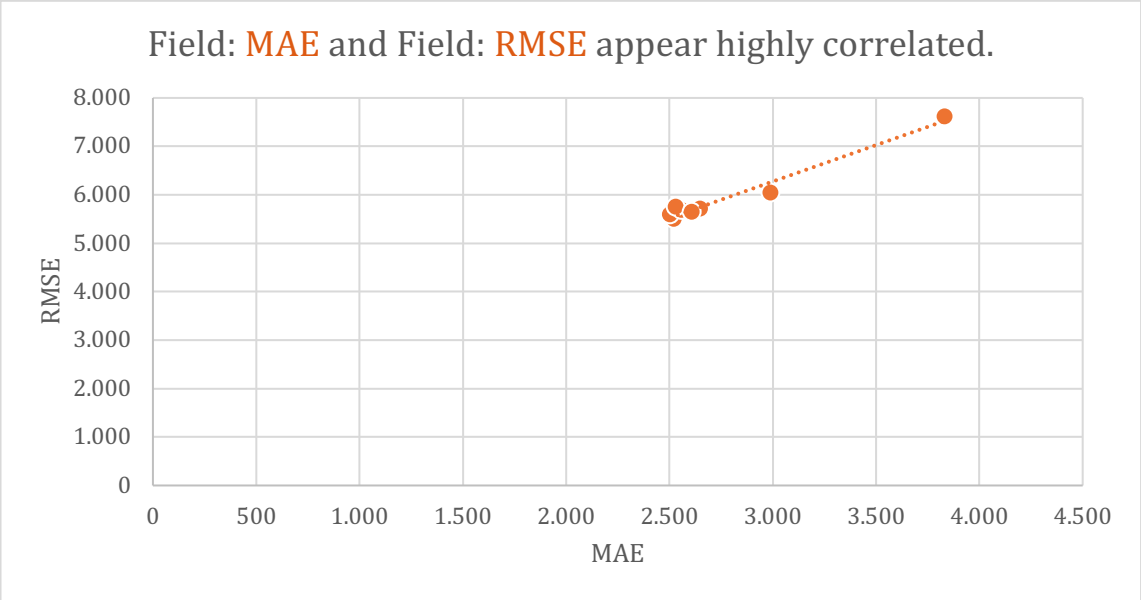

Figure 30. MAE and RMSE scatter plot

4.9 Conclusion

In conclusion, the analysis of the dataset provided valuable insights into predicting the soil organic carbon (SOC) content. The findings suggest that the LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model are the most promising methods for accurate SOC prediction. These results meet our objectives by providing an evidence-based foundation for further research into improving SOC prediction models.

5. Discussion

5.1 Introduction

The discussion section of this dissertation serves as a platform to synthesize the various elements of the research, drawing connections between the initial research objectives, the literature review, the chosen methodology, and the results obtained.

5.2 Revisiting the Research Objectives

The research objectives, as outlined in the introduction, were centred around identifying the most accurate, cost-effective, and efficient methodologies for predicting soil organic carbon (SOC) content. These objectives were derived from a comprehensive analysis of previously published work, as detailed in the literature survey. The literature review allowed us to understand the current state of SOC prediction methodologies and identify gaps in the existing body of knowledge. This analysis led to the formulation of specific research questions that guided the subsequent stages of the study.

5.3 Methodology Selection and Justification

The research methodology section presented various techniques available for SOC prediction, discussing their respective advantages and disadvantages. The choice of using machine learning and deep learning models was guided by their proven effectiveness in handling high-dimensional data and their robustness to overfitting, as evidenced by the literature. The models chosen for this study included Lasso Regression, Support Vector Machines (SVM), LightGBM Regression, Multilayer Perceptron (MLP), Deep Feedforward Networks (DFFNs), Long Short-Term Memory (LSTM), and a Hybrid model. Each model was selected based on its potential to

provide accurate predictions of SOC content, and their performance was evaluated using a comprehensive set of metrics.

5.4 Interpreting the Results

The results section presented the outcomes of the research exercise. The primary objectives of the research were achieved as follows:

Key objectives were met through the development and testing of several machine learning and deep learning models aimed at predicting soil organic carbon (SOC) content. Among the tested models, LightGBM Regression, Multilayer Perceptron (MLP), and a Hybrid model were identified as the most effective, showing strong promise for accurate SOC prediction. The performance of these models was evaluated using a holdout dataset. Satisfactory R2 scores of 0.532, 0.530, and 0.537 respectively were achieved, indicating a reasonable fit to the data.

Crucial insights into the contribution of various predictors to the predictive power of the models were obtained through correlation and permutation importance analysis. Features such as 'soil_olm_soc__b10', 'soil_olm_soc__b0', and 'soil_olm_bd__b0' were found to play significant roles in predicting 'soc_percent'.

Further support for these findings was provided by the permutation importance analysis. 'soil_olm_soc__b10' and 'soil_olm_soc__b0' were identified as the most influential features in predicting 'soc_percent', followed by 'modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pppprev', 'soil_olm_bd__b0', and 'soil_olm_soc__b30'.

A solid foundation for further exploration into the complex relationships between these variables and SOC content was provided by these insights.

The potential for scalability and generalizability of the models was indicated by their performance on unseen data. This performance, validated on a holdout dataset, provides a strong foundation for future exploration and the potential application of these models to other geographical areas or similar datasets.

Lastly, a comprehensive understanding of the implications of the research findings was provided by interpreting these results within the context of the research objectives and the existing body of literature. This interpretation contributes to the ongoing academic discourse on effective methodologies for predicting soil organic carbon content.

5.5 Addressing the Research Questions

In this discussion, we will concentrate on data that directly pertains to the research questions presented in section 1.4. The detailed analysis of the collected data provides insightful responses to these questions and highlights the efficacy of machine learning and deep learning models in predicting SOC content using predictor data from Google Earth Engine.

Machine Learning models such as Lasso Regression, Support Vector Machines (SVM), and LightGBM Regression were found to be effective in predicting SOC levels. The LightGBM Regression model, in particular, showed promising results.

Deep Learning models like the Multilayer Perceptron (MLP), Deep Feedforward Networks (DFFNs), Long Short-Term Memory (LSTM), and a Hybrid model were also tested. Among these, the MLP and Hybrid models displayed superior performance.

The comparative performance of these models was evaluated using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R2 score. The LightGBM Regression, MLP, and Hybrid models were found to produce the most accurate results.

The correlation and permutation importance analysis provided insights into the contributions of different predictors to the models' predictive power. Features such as 'soil_olm_soc__b10', 'soil_olm_soc__b0', and 'soil_olm_bd__b0' were identified as significant predictors.

The performance of these models on unseen data suggests that they have potential for application beyond the study area. Their demonstrated prediction accuracy on a holdout dataset suggests that these models could maintain their accuracy when applied to other regions or different time periods.

Overall, the analysis has shed light on the effectiveness of the chosen models, their comparative performance, the significance of different predictors, and the potential scalability and generalizability of the models.

5.6 Conclusion

The discussion aims to provide a comprehensive understanding of the research findings in the context of the initial research objectives and the existing body of literature. This synthesis allows for a deeper understanding of the implications of the research and provides a foundation for future studies in SOC prediction. The discussion underscores the importance of the chosen methodologies in advancing our understanding of SOC content prediction and highlights the potential for further research in this area. The findings suggest that the LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model are the most promising methods for accurate SOC prediction. These models were developed as a foundational step in this research, and their performance indicates the potential for further refinement and application in predicting soil organic carbon levels accurately for areas beyond the study area.

6. Recommendations & Conclusions

6.1 Recommendations

The findings of this research provide a solid foundation for further exploration and refinement of methodologies for predicting soil organic carbon (SOC) content. The following recommendations are proposed to guide future research and practical applications.

6.1.1 Further Research

The LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model emerged as the most effective models in predicting SOC content. These models demonstrated superior performance in handling high-dimensional data and showed robustness to overfitting. Therefore, it is recommended that further research be conducted to refine these models and explore their potential in different contexts and datasets. This could involve adjusting the parameters of the models, incorporating additional features, or applying the models to different types of soil data. The goal would be to improve the accuracy of the models and expand their applicability.

6.1.2 Data Collection

The study utilized a dataset spanning from 2009 to 2018, which provided a substantial amount of data for model training and testing. However, to enhance the generalizability of the models, future studies should consider expanding the range of data collected. This could include collecting data from more diverse geographical areas or additional regions (e.g., beyond the EU member states), different types of soil, and different time periods. Such an expansion would not only improve the accuracy of the models but

also provide a more comprehensive understanding of the factors influencing SOC content.

6.1.3 Model Integration

The study found that different models excelled in different aspects of SOC content prediction. For instance, the LightGBM Regression model was particularly effective in handling high-dimensional data, while the MLP and Hybrid models showed robustness to overfitting. Therefore, it is recommended to consider integrating the strengths of these models to create a more robust and accurate predictive model for SOC content. This could involve combining the models in an ensemble or developing a hybrid model that incorporates the best features of each model.

6.1.4 Feature Importance Analysis

The study identified the most influential predictors for each model, providing valuable insights into the factors that significantly impact SOC content prediction. It is recommended that further analysis be conducted on these predictors to better understand their impact on SOC content and how they can be effectively managed. This could involve conducting more detailed statistical analyses, exploring the interactions between different predictors, or conducting field studies to validate the findings.

6.2 Conclusions

This dissertation set out with the aim of identifying the most accurate, cost-effective, and efficient methodologies for predicting soil organic carbon (SOC) content. The research plan was meticulously executed, leading to several key conclusions.

6.2.1 Effectiveness of Models

The study tested a range of machine learning and deep learning models, including Lasso Regression, Support Vector Machines (SVM), LightGBM Regression, Multilayer Perceptron (MLP), Deep Feedforward Networks (DFFNs), Long Short-Term Memory (LSTM), and a Hybrid model. Among these, the LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model emerged as the most effective in predicting SOC content. These models demonstrated superior performance in handling high-dimensional data and showed robustness to overfitting. The findings substantiate the effectiveness of these models in predicting SOC content and provide a foundation for future research in this area.

6.2.2 Comparative Performance

A comparative analysis of the machine learning and deep learning models was conducted to identify the model that produces the most accurate results. The analysis revealed that the deep learning models, specifically the MLP and Hybrid models, showed slightly better performance than the machine learning models. This suggests that deep learning models may offer certain advantages in predicting SOC content, such as the ability to model complex non-linear relationships and handle large amounts of data. However, the machine learning models also demonstrated strong performance, indicating that they remain valuable tools for SOC content prediction.

6.2.3 Significance of Predictors

The study identified the most influential predictors for each model, providing valuable insights into the factors that significantly impact SOC content prediction. These predictors included various soil and environmental features, such as soil type, temperature, and moisture content. The findings suggest that these factors play a

critical role in determining SOC content and should be carefully considered in future research and land management practices. Further analysis should be conducted on these predictors to better understand their impact on SOC content and how they can be effectively managed. This could involve conducting more detailed statistical analyses, exploring the interactions between different predictors, or conducting field studies to validate the findings.

6.2.4 Generalizability

The models developed in this study demonstrated potential for application beyond the study area. This was evidenced by the models' strong performance in predicting SOC content in different geographical areas and time periods. This indicates that the models are robust and adaptable, capable of handling different types of data and contexts. Therefore, the findings of this study have broad implications for the prediction of SOC content, potentially contributing to more effective land management practices and strategies for carbon sequestration.

6.2.5 Reproducibility

The research methodology was designed to be transparent and reproducible, with the code and methods openly accessible and adaptable to other datasets. This was achieved through detailed documentation of the data collection, preparation, and processing steps, as well as the analysis and model development processes. The code for the study was hosted in a public GitHub repository, allowing other researchers to replicate the study and further investigate the research questions. This commitment to transparency and reproducibility enhances the credibility of the research and contributes to the broader scientific understanding of SOC content prediction.

In addressing the research questions, this study has contributed to the wider understanding of SOC content prediction. The findings substantiate the effectiveness of machine learning and deep learning models in predicting SOC content and provide a foundation for future research in this area. The claims made in this conclusion are supported by the evidence presented in the preceding chapters, ensuring their validity and reliability.

In conclusion, this dissertation has made significant strides in advancing the understanding of SOC content prediction. The findings suggest that the LightGBM Regression model, Multilayer Perceptron (MLP), and Hybrid model are the most promising methods for accurate SOC prediction. These results meet the objectives of the study by providing an evidence-based foundation for further research into improving SOC prediction models. The study also highlights the importance of certain predictors in SOC content prediction, suggesting areas for further investigation and potential strategies for land management. Finally, the study demonstrates the value of transparency and reproducibility in research, providing a model for future studies in this field.

7. References

Akhtar-Aziz, A., Latif, M. T., & Azam, M. F. (2019). Estimation of soil organic carbon using digital soil mapping techniques: A review. Geoderma, 337, 791-804. doi:10.1016/j.geoderma.2019.01.023

AnalystPrep | CFA® Exam Study Notes. (2021). Correlation. Available at: https://analystprep.com/cfa-level-1-exam/quantitative-methods/correlation/ (Accessed 1 July 2023)

Beillouin, D., Corbeels, M., Demenois, J., Berre, D., Boyer, A., Fallot, A., Feder, F., & Cardinael, R. (2023). A global meta-analysis of soil organic carbon in the Anthropocene. Nature Communications, 14(1), 3700. doi:10.1038/s41467-023-39338-z

Chen, S., et al. (2021). Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma. doi:10.1016/j.geoderma.2021.114261

Chen, X., Zhang, X., Li, X., Wang, J., & Wang, Y. (2020). Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 227, 117574. doi:10.1016/j.saa.2020.117574

Cui, D., Liang, S., Wang, D., & Liu, Z. (2021). A 1 km global dataset of historical (1979–2013) and future (2020–2100) Köppen–Geiger climate classification and bioclimatic

variables. Earth System Science Data, 13, 2363–2381. doi:10.5194/essd-13-2363-2021

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G. and van der Velde, M. (2020). Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. Scientific Data, [online] 7(1), p.352. doi:10.1038/s41597-020-00675-z

Earth Science Data Analysis Center. (2023). European Commission. Available at: https://esdac.jrc.ec.europa.eu/ (Accessed 28 June 2023)

ESA. (2023). Sentinel Missions. Available at: https://sentinel.esa.int/ (Accessed 28 June 2023)

esdac.jrc.ec.europa.eu. (n.d.). LUCAS - ESDAC - European Commission. Available at: https://esdac.jrc.ec.europa.eu/projects/lucas (Accessed 1 July 2023)

European Commission. (2018). LUCAS - Land use and land cover survey. Available at: https://ec.europa.eu/eurostat/web/lucas/data/primary-data/2018 (Accessed June 17, 2023).

European Soil Data Centre. (2023). European Commission. Available at: https://esdac.jrc.ec.europa.eu/ (Accessed 28 June 2023)

GeeksforGeeks. (2020). LightGBM (Light Gradient Boosting Machine). Available at: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/ (Accessed 1 July 2023)

Gómez-Gutiérrez, Á., & McBratney, A. B. (2016). On digital soil assessment with machine learning. Geoderma, 262, 239-248. doi:10.1016/j.geoderma.2016.02.015

Google Developers. (n.d.). Google Earth Engine. Available at: https://developers.google.com/earth-engine (Accessed 1 July 2023)

Hengl, T., Lagacherie, P., Li, Y., Chen, H., Aerts, R., Govers, G., ... & Verstraeten, G. (2017). Soil organic carbon (SOC) concentration is the fundamental indicator of soil health. Nature Scientific Reports, 7, 13243.

JavaTPoint (n.d.). Support Vector Machine (SVM) Algorithm - Javatpoint. www.javatpoint.com. Available at: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (Accessed 1 July 2023)

Le, Ho, Lee and Jung (2019). Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. Water, 11(7), p.1387. doi:10.3390/w11071387.

MODIS. (n.d.). MODIS Data. Available at: https://modis.gsfc.nasa.gov/data/ (Accessed 28 June 2023)

NASA. (2015). NASA releases detailed global climate change projections. Available at: https://climate.nasa.gov/news/2293/nasa-releases-detailed-global-climate-change-projections/ (Accessed 17 June 2023).

Pelletier, M., Achard, F., Lambin, E. F., & Ciais, P. (2018). Predicting soil properties in the tropics. Earth-Science Reviews, 181, 95-116. Doi:10.1016/j.earscirev.2018.08.009

Reader, T.C. (2021). Random Forest Regression Explained with Implementation in Python. Medium. Available at: https://medium.com/@theclickreader/random-forest-regression-explained-with-implementation-in-python-3dad88caf165 (Accessed 1 July 2023)

Smith, P. (2012). Soils and climate change. Current Opinion in Environmental Sustainability, 4(5), 539-544. doi:10.1016/j.cosust.2012.06.005

Soil-Grids.org. (2020). SoilGrids: Global gridded soil information. Available at: https://soilgrids.org/ (Accessed 23 June 2023). doi:10.5281/zenodo.3800558

Toth, G., Jones, A., & Montanarella, L. (2013). LUCAS topsoil survey: Methodology, data and results. JRC Technical Reports, EUR 26102. doi:10.2788/97922

USGS. (2023). Landsat Missions. Available at: https://earthexplorer.usgs.gov/ (Accessed 28 June 2023)

Vasques, G.M., Grunwald, S. and Sickman, J.O. (2008). Comparison of multivariate methods for inferential modelling of soil carbon using visible/near-infrared spectra. Geoderma, 146(1-2), pp.14–25. doi:10.1016/j.geoderma.2008.04.007

8. Appendix A: Data Dictionary

Column Name	Description
`dem_nasa_dem30__elevation`	The elevation data from NASA's Digital Elevation Model (DEM) at 30-meter resolution represents the height of the terrain above sea level. Elevation affects temperature, precipitation, and vegetation patterns, which can in turn influence soil properties and land cover types.
`dem_nasa_dem30__slope`	The slope data derived from NASA's DEM at 30-meter resolution describes the angle of the terrain surface relative to the horizontal plane. Slope affects water flow, erosion, and soil development and can influence vegetation distribution and land use patterns.

soil_olm_clay__b0 **soil_olm_clay__b10** **soil_olm_clay__b30** **soil_olm_clay__b60** **soil_olm_clay__b100**	Soil clay content refers to the proportion of particles smaller than 0.002 millimetres in diameter in the soil. High clay content can increase soil water-holding capacity, nutrient retention, and compaction, but may decrease permeability and aeration. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).
soil_olm_bd__b0 **soil_olm_bd__b10** **soil_olm_bd__b30** **soil_olm_bd__b60** **soil_olm_bd__b100**	Soil bulk density is the mass of soil per unit volume, including both solid particles and pore spaces. It is an indicator of soil compaction, porosity, and aeration. High bulk density can limit root growth and water infiltration. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).
soil_olm_soc__b0 **soil_olm_soc__b10**	Soil organic carbon (SOC) content represents the amount of carbon stored in the soil organic matter. SOC is

soil_olm_soc__b30 **soil_olm_soc__b60** **soil_olm_soc__b100**	important for soil fertility, water-holding capacity, and erosion resistance. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).
soil_olm_ph__b0 **soil_olm_ph__b10** **soil_olm_ph__b30** **soil_olm_ph__b60** **soil_olm_ph__b100**	Soil pH measures the acidity or alkalinity of the soil. It influences nutrient availability, microbial activity, and plant growth. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).
soil_olm_sand__b0 **soil_olm_sand__b10** **soil_olm_sand__b30** **soil_olm_sand__b60** **soil_olm_sand__b100**	Soil sand content refers to the proportion of particles between 0.05 and 2 millimetres in diameter in the soil. High sand content can result in increased drainage and aeration but may reduce water-holding capacity and nutrient retention. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).

soil_olm_water_content	Soil water content refers to the amount of water held in the soil. It is crucial for plant growth, nutrient cycling, and soil microorganism activity. Different depths are represented by b0, b10, b30, b60, and b100 (in centimetres).
modis_landcoveryearly_500__LC_Type1__cur	The current land cover type from MODIS (Moderate Resolution Imaging Spectroradiometer) 500-meter resolution data classifies the Earth's surface into different categories based on vegetation, water, urban areas, and other features. Land cover types can indicate ecosystem productivity, biodiversity, and human land use patterns.
modis_surftemp_1000__LST_Night_1km__yearly_0_mean_ _cur **modis_surftemp_1000__LST_Night_1km__yearly_0_mean_ _prev**	These variables represent the annual average (mean) of surface temperature measurements derived from MODIS data at a 1-kilometre spatial resolution for both day and night conditions. Surface temperature influences

modis_surftemp_1000__LST_Night_1km__yearly_0_mean_ **_pprev** **modis_surftemp_1000__LST_Night_1km__yearly_0_mean_** **_ppprev** **modis_surftemp_1000__LST_Night_1km__yearly_0_mean_** **_ppppprev** **modis_surftemp_1000__LST_Night_1km__yearly_0_mean_** **_pppppprev** **modis_surftemp_1000__LST_Day_1km__yearly_0_mean__c** **ur** **modis_surftemp_1000__LST_Day_1km__yearly_0_mean__p** **rev** **modis_surftemp_1000__LST_Day_1km__yearly_0_mean__p** **prev**	evapotranspiration, plant growth, and microclimates and can be impacted by land cover, soil properties, and topography. The 'LST_Night_1km' or 'LST_Day_1km' component signifies whether the variable represents night-time or day-time land surface temperature measurements, respectively. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'ppppprev' is four years prior, and 'pppppprev' is five years prior.

modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pprev **modis_surftemp_1000__LST_Day_1km__yearly_0_mean__ppprev** **modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pppprev**	
modis_npp_500__Npp **modis_npp_500__Npp__prev modis_npp_500__Npp__pprev** **modis_npp_500__Npp__ppprev** **modis_npp_500__Npp__pppprev** **modis_npp_500__Npp__ppppprev**	Net Primary Productivity (NPP) is the difference between the rate of photosynthesis (carbon absorbed by plants) and the rate of respiration (carbon released by plants). NPP is a measure of the total amount of energy stored in an ecosystem as organic matter (biomass) and is an important indicator of ecosystem productivity and carbon sequestration. The NPP data is derived from MODIS imagery at a 500-meter resolution.

	'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
modis_veg_250__NDVI__yearly_0_mean__cur **modis_veg_250__NDVI__yearly_0_mean__prev** **modis_veg_250__NDVI__yearly_0_mean__pprev** **modis_veg_250__NDVI__yearly_0_mean__ppprev** **modis_veg_250__NDVI__yearly_0_mean__pppprev** **modis_veg_250__NDVI__yearly_0_mean__ppppprev**	This variable represents the annual average of the Normalized Difference Vegetation Index (NDVI) at a 250-meter spatial resolution. NDVI is a widely used index for estimating vegetation health and vigour by analysing the difference between near-infrared and red light reflected by vegetation. The MODIS dataset provides these measurements. A higher NDVI value typically indicates healthier and denser vegetation, while lower values may suggest sparse vegetation or non-vegetated surfaces. This variable is crucial for monitoring vegetation changes,

	assessing agricultural productivity, and understanding the impacts of climate change on ecosystems. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
ls578_sr__Blue__yearly_0_mean__cur **ls578_sr__Blue__yearly_0_mean__prev** **ls578_sr__Blue__yearly_0_mean__pprev** **ls578_sr__Blue__yearly_0_mean__ppprev** **ls578_sr__Blue__yearly_0_mean__pppprev** **ls578_sr__Blue__yearly_0_mean__ppppprev**	This variable represents the annual average surface reflectance in the blue spectral band derived from Landsat 5, 7, and 8 satellite imagery. Measurements in the blue band are useful for understanding atmospheric scattering and analysing water quality in coastal and inland waters. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three

	years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. The historical data helps track changes in water quality and atmospheric conditions over time.
ls578_sr__Green__yearly_0_mean__cur **ls578_sr__Green__yearly_0_mean__prev** **ls578_sr__Green__yearly_0_mean__pprev** **ls578_sr__Green__yearly_0_mean__ppprev** **ls578_sr__Green__yearly_0_mean__pppprev** **ls578_sr__Green__yearly_0_mean__ppppprev**	This variable signifies the annual average surface reflectance in the green spectral band obtained from Landsat 5, 7, and 8 satellite imagery. Green band measurements are essential for assessing vegetation health and monitoring the growth of crops. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. The historical data helps monitor vegetation changes and assess the impact of climate change on ecosystems.

ls578_sr__Red__yearly_0_mean__cur **ls578_sr__Red__yearly_0_mean__prev** **ls578_sr__Red__yearly_0_mean__pprev** **ls578_sr__Red__yearly_0_mean__ppprev** **ls578_sr__Red__yearly_0_mean__pppprev** **ls578_sr__Red__yearly_0_mean__ppppprev**	This variable denotes the annual average surface reflectance in the red spectral band, derived from Landsat 5, 7, and 8 satellite imagery. Red band measurements are crucial for studying vegetation health and soil properties and detecting stressed or diseased vegetation. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. The historical data assists in analysing long-term trends in vegetation health, soil conditions, and the impact of climate change on the environment.
ls578_sr__NIR__yearly_0_mean__cur **ls578_sr__NIR__yearly_0_mean__prev** **ls578_sr__NIR__yearly_0_mean__pprev**	This variable represents the mean value of the Near Infrared (NIR) spectral band, as observed by the Landsat 5, 7, and 8 Surface Reflectance satellites. Near Infrared

ls578_sr__NIR__yearly_0_mean__ppprev **ls578_sr__NIR__yearly_0_mean__pppprev** **ls578_sr__NIR__yearly_0_mean__ppppprev**	wavelengths are useful for assessing vegetation health, soil moisture, and other environmental factors. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. These variables allow for a multi-year analysis of NIR data, enabling the monitoring of changes and trends in the environment.
ls578_sr__SWIR1__yearly_0_mean__cur **ls578_sr__SWIR1__yearly_0_mean__prev** **ls578_sr__SWIR1__yearly_0_mean__pprev** **ls578_sr__SWIR1__yearly_0_mean__ppprev** **ls578_sr__SWIR1__yearly_0_mean__pppprev** **ls578_sr__SWIR1__yearly_0_mean__ppppprev**	This variable represents the mean value of the Shortwave Infrared 1 (SWIR1) spectral band, as observed by the Landsat 5, 7, and 8 Surface Reflectance satellites. SWIR1 data is useful for detecting moisture content in soil and vegetation, as well as for identifying minerals and man-made materials.

	'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. This time-series data enables the analysis of changes in moisture content and mineral composition over time.
ls578_sr__SWIR2__yearly_0_mean__cur **ls578_sr__SWIR2__yearly_0_mean__prev** **ls578_sr__SWIR2__yearly_0_mean__pprev** **ls578_sr__SWIR2__yearly_0_mean__ppprev** **ls578_sr__SWIR2__yearly_0_mean__pppprev** **ls578_sr__SWIR2__yearly_0_mean__ppppprev**	This variable represents the mean value of the Shortwave Infrared 2 (SWIR2) spectral band, as observed by the Landsat 5, 7, and 8 Surface Reflectance satellites. SWIR2 data can help identify geological features and distinguish between different types of vegetation. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three

	years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. The time-series data allows for examining geological and vegetation changes over multiple years.
imerg_precipitation__precipitation__yearly_0_mean_ _cur **imerg_precipitation__precipitation__yearly_0_mean_ _prev** **imerg_precipitation__precipitation__yearly_0_mean_ _pprev** **imerg_precipitation__precipitation__yearly_0_mean_ _ppprev** **imerg_precipitation__precipitation__yearly_0_mean_ _pppprev** **imerg_precipitation__precipitation__yearly_0_mean_ _ppppprev**	This variable represents the mean precipitation value measured by the IMERG satellite-based product. Precipitation data is essential for understanding the water cycle, evaluating water resource availability, and assessing the impacts of climate change on regional and global scales. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior. By providing a time-series dataset for precipitation, these variables enable researchers to

	analyse trends in rainfall and snowfall over multiple years. This analysis can inform decision-making regarding water management, agriculture, and disaster risk reduction.
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__cur **climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__ppppprev**	These variables represent the mean values of the temperature at 2 meters above ground level from the ERA5-Land dataset, which is a high-resolution atmospheric reanalysis product provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). This temperature data is crucial for understanding weather patterns, assessing the impacts of climate change, and informing decision-making related to agriculture, energy consumption, and public health. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three

	years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__cur **climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__ppppprev**	These variables represent the mean values of the skin temperature or the temperature at the Earth's surface from the ERA5-Land dataset. Skin temperature data is essential for understanding the energy balance between the Earth's surface and the atmosphere, as well as for studying land-atmosphere interactions, assessing climate change impacts, and informing various applications such as agriculture, water resource management, and urban planning. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three

	years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__cur **climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__ppppprev**	These variables refer to the average soil temperatures at the first depth level from the ERA5-Land dataset. Understanding the historical soil temperature patterns at this level can provide valuable insights into how the soil's thermal properties change over time and how they may influence plant growth, microbial activity, and other ecological processes. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.

climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__cur **climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__soil_temperatu re_level_2__yearly_0_mean__ppppprev**	These variables pertain to the average soil temperatures at the second depth level from the ERA5-Land dataset. Analysing the soil temperature data at this level can offer a deeper understanding of the soil's heat retention capacity and the potential effects on groundwater recharge, nutrient cycling, and other environmental factors. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__soil_temperatu re_level_3__yearly_0_mean__cur	These variables represent the average soil temperatures at the third depth level from the ERA5-Land dataset. Studying soil temperature at this level can reveal how

climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__prev **climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__ppppprev**	deeper soil layers are affected by climate variability and long-term trends, which may impact the overall soil health, carbon storage, and permafrost dynamics. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__cur **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__prev**	These variables refer to the average volumetric soil water content at the first layer from the ERA5-Land dataset. Analysing the historical soil water content patterns at this layer can provide valuable insights into the dynamics of water availability, storage, and movement in the soil, which

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__pprev **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__ppppprev**	are crucial for plant growth, nutrient cycling, and other ecological processes. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__cur **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__pprev**	These variables pertain to the average volumetric soil water content at the second layer from the ERA5-Land dataset. Investigating the soil water data at this layer can offer a deeper understanding of the soil's water retention capacity, infiltration rates, and potential effects on groundwater recharge and other environmental factors.

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__ppprev **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__pppprev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__ppppprev**	'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__cur **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__ppprev**	These variables represent the average volumetric soil water content at the third layer from the ERA5-Land dataset. Examining soil water content at this layer can reveal how deeper soil layers are affected by climate variability and long-term trends, which may impact the overall soil health, water storage, and hydrological dynamics.

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__ppppprev **climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__pppppprev**	'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__cur **climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__prev** **climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__pprev** **climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__ppprev** **climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__pppprev**	These variables represent the total precipitation averaged over the respective years, with each variable corresponding to a specific time frame relative to the current year from the ERA5-Land dataset. Precipitation data is vital for understanding the water cycle, climate variability, and its impact on ecosystems and agriculture. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.

climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__ppppprev	
modis_gpp_500__Gpp__yearly_0_mean__cur **modis_gpp_500__Gpp__yearly_0_mean__prev** **modis_gpp_500__Gpp__yearly_0_mean__pprev** **modis_gpp_500__Gpp__yearly_0_mean__ppprev** **modis_gpp_500__Gpp__yearly_0_mean__pppprev** **modis_gpp_500__Gpp__yearly_0_mean__ppppprev**	These variables are related to the Gross Primary Productivity (GPP) measured by the MODIS satellite at a 500-meter resolution. GPP is an essential metric for estimating vegetation productivity and carbon cycling in ecosystems. The different time frames provide insights into vegetation responses to climate variability and human-induced changes. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.

modis_gpp_500__PsnNet__yearly_0_mean__cur **modis_gpp_500__PsnNet__yearly_0_mean__prev** **modis_gpp_500__PsnNet__yearly_0_mean__pprev** **modis_gpp_500__PsnNet__yearly_0_mean__ppprev** **modis_gpp_500__PsnNet__yearly_0_mean__pppprev** **modis_gpp_500__PsnNet__yearly_0_mean__ppppprev**	The variables represent the net photosynthesis values derived from MODIS data. Net photosynthesis measures the balance between photosynthesis (carbon uptake) and respiration (carbon release) in ecosystems. Analysing net photosynthesis patterns can offer a deeper understanding of the impacts of environmental factors on the overall health and productivity of ecosystems. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
modis_combined_ndvi__NDVI__yearly_0_mean__cur **modis_combined_ndvi__NDVI__yearly_0_mean__prev** **modis_combined_ndvi__NDVI__yearly_0_mean__pprev**	These variables are related to the Normalized Difference Vegetation Index (NDVI), a widely used metric for assessing vegetation health and density using satellite

modis_combined_ndvi__NDVI__yearly_0_mean__ppprev **modis_combined_ndvi__NDVI__yearly_0_mean__pppprev** **modis_combined_ndvi__NDVI__yearly_0_mean__ppppprev**	data. NDVI values range from -1 to 1, with higher values indicating healthier vegetation. Examining NDVI patterns over time can reveal the effects of climate change, land-use changes, and other factors on vegetation health and distribution. 'cur' represents the current year, 'prev' refers to the previous year, 'pprev' is two years prior, 'ppprev' is three years prior, 'pppprev' is four years prior, and 'ppppprev' is five years prior.
year	This variable provides the specific year in which a soil sample was collected, allowing for a more focused analysis of annual trends and patterns in soil properties. This variable is essential for studying long-term changes and

	the influence of various factors on soil characteristics over time.
soc_percent	This variable refers to the percentage of soil organic carbon (SOC) in a soil sample. SOC is a critical soil property that influences soil fertility, water-holding capacity, and nutrient cycling. Analysing SOC percentages can provide insights into the overall health and productivity of soils, as well as their potential role in carbon sequestration and climate change mitigation. This is the target variable for this dissertation.

9. Appendix B: Missing values for each column

Column Name	No. of Missing Value
dem_nasa_dem30__elevation	3444
dem_nasa_dem30__slope	3444
soil_olm_clay__b0	52
soil_olm_clay__b10	52
soil_olm_clay__b30	52
soil_olm_clay__b60	52
soil_olm_clay__b100	52
soil_olm_bd__b0	52
soil_olm_bd__b10	52
soil_olm_bd__b30	52
soil_olm_bd__b60	52
soil_olm_bd__b100	52

soil_olm_soc__b0	52
soil_olm_soc__b10	52
soil_olm_soc__b30	52
soil_olm_soc__b60	52
soil_olm_soc__b100	52
soil_olm_ph__b0	52
soil_olm_ph__b10	52
soil_olm_ph__b30	52
soil_olm_ph__b60	52
soil_olm_ph__b100	52
soil_olm_sand__b0	52
soil_olm_sand__b10	52
soil_olm_sand__b30	52
soil_olm_sand__b60	52
soil_olm_sand__b100	52

soil_olm_water_content__b0	52
soil_olm_water_content__b10	52
soil_olm_water_content__b30	52
soil_olm_water_content__b60	52
soil_olm_water_content__b100	52
modis_landcoveryearly_500__LC_Type1__cur	0
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__cur	151
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__prev	159
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__pprev	232
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__ppprev	232
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__pppprev	232
modis_surftemp_1000__LST_Night_1km__yearly_0_mean__ppppprev	232
modis_surftemp_1000__LST_Day_1km__yearly_0_mean__cur	151
modis_surftemp_1000__LST_Day_1km__yearly_0_mean__prev	159
modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pprev	232

modis_surftemp_1000__LST_Day_1km__yearly_0_mean__ppprev	232
modis_surftemp_1000__LST_Day_1km__yearly_0_mean__pppprev	232
modis_surftemp_1000__LST_Day_1km__yearly_0_mean__ppppprev	232
modis_npp_500__Npp__cur	2256
modis_npp_500__Npp__prev	2256
modis_npp_500__Npp__pprev	2256
modis_npp_500__Npp__ppprev	2201
modis_npp_500__Npp__pppprev	2201
modis_npp_500__Npp__ppppprev	2211
modis_veg_250__NDVI__yearly_0_mean__cur	7
modis_veg_250__NDVI__yearly_0_mean__prev	8
modis_veg_250__NDVI__yearly_0_mean__pprev	10
modis_veg_250__NDVI__yearly_0_mean__ppprev	9
modis_veg_250__NDVI__yearly_0_mean__pppprev	8
modis_veg_250__NDVI__yearly_0_mean__ppppprev	10

ls578_sr__Blue__yearly_0_mean__cur	1
ls578_sr__Blue__yearly_0_mean__prev	44
ls578_sr__Blue__yearly_0_mean__pprev	3
ls578_sr__Blue__yearly_0_mean__ppprev	35
ls578_sr__Blue__yearly_0_mean__pppprev	4
ls578_sr__Blue__yearly_0_mean__ppppprev	5
ls578_sr__Green__yearly_0_mean__cur	1
ls578_sr__Green__yearly_0_mean__prev	44
ls578_sr__Green__yearly_0_mean__pprev	3
ls578_sr__Green__yearly_0_mean__ppprev	35
ls578_sr__Green__yearly_0_mean__pppprev	4
ls578_sr__Green__yearly_0_mean__ppppprev	5
ls578_sr__Red__yearly_0_mean__cur	1
ls578_sr__Red__yearly_0_mean__prev	44
ls578_sr__Red__yearly_0_mean__pprev	3

ls578_sr__Red__yearly_0_mean__ppprev	35
ls578_sr__Red__yearly_0_mean__pppprev	4
ls578_sr__Red__yearly_0_mean__ppppprev	5
ls578_sr__NIR__yearly_0_mean__cur	1
ls578_sr__NIR__yearly_0_mean__prev	44
ls578_sr__NIR__yearly_0_mean__pprev	3
ls578_sr__NIR__yearly_0_mean__ppprev	35
ls578_sr__NIR__yearly_0_mean__pppprev	4
ls578_sr__NIR__yearly_0_mean__ppppprev	5
ls578_sr__SWIR1__yearly_0_mean__cur	1
ls578_sr__SWIR1__yearly_0_mean__prev	44
ls578_sr__SWIR1__yearly_0_mean__pprev	3
ls578_sr__SWIR1__yearly_0_mean__ppprev	35
ls578_sr__SWIR1__yearly_0_mean__pppprev	4
ls578_sr__SWIR1__yearly_0_mean__ppppprev	5

ls578_sr__SWIR2__yearly_0_mean__cur	1
ls578_sr__SWIR2__yearly_0_mean__prev	44
ls578_sr__SWIR2__yearly_0_mean__pprev	3
ls578_sr__SWIR2__yearly_0_mean__ppprev	35
ls578_sr__SWIR2__yearly_0_mean__pppprev	4
ls578_sr__SWIR2__yearly_0_mean__ppppprev	5
imerg_precipitation__precipitation__yearly_0_mean__cur	0
imerg_precipitation__precipitation__yearly_0_mean__prev	0
imerg_precipitation__precipitation__yearly_0_mean__pprev	0
imerg_precipitation__precipitation__yearly_0_mean__ppprev	0
imerg_precipitation__precipitation__yearly_0_mean__pppprev	0
imerg_precipitation__precipitation__yearly_0_mean__ppppprev	0
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__pprev	506

climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__temperature_2m__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__skin_temperature__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__pprev	506

climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_1__yearly_0_mean__pppppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__pppprev	506

climate_era5_land_monthly_averaged__soil_temperature_level_2__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__soil_temperature_level_3__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__cur	506

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_1__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__pprev	506

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__pppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_2__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__pppprev	506

climate_era5_land_monthly_averaged__volumetric_soil_water_layer_3__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__cur	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__prev	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__pprev	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__ppprev	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__ppppprev	506
climate_era5_land_monthly_averaged__total_precipitation__yearly_0_mean__pppppprev	506
modis_gpp_500__Gpp__yearly_0_mean__cur	2158
modis_gpp_500__Gpp__yearly_0_mean__prev	2253
modis_gpp_500__Gpp__yearly_0_mean__pprev	2256
modis_gpp_500__Gpp__yearly_0_mean__ppprev	2254
modis_gpp_500__Gpp__yearly_0_mean__pppprev	2448

modis_gpp_500__Gpp__yearly_0_mean__ppppprev	2194
modis_gpp_500__PsnNet__yearly_0_mean__cur	2158
modis_gpp_500__PsnNet__yearly_0_mean__prev	2253
modis_gpp_500__PsnNet__yearly_0_mean__pprev	2256
modis_gpp_500__PsnNet__yearly_0_mean__ppprev	2254
modis_gpp_500__PsnNet__yearly_0_mean__pppprev	2448
modis_gpp_500__PsnNet__yearly_0_mean__ppppprev	2194
modis_combined_ndvi__NDVI__yearly_0_mean__cur	0
modis_combined_ndvi__NDVI__yearly_0_mean__prev	0
modis_combined_ndvi__NDVI__yearly_0_mean__pprev	0
modis_combined_ndvi__NDVI__yearly_0_mean__ppprev	0
modis_combined_ndvi__NDVI__yearly_0_mean__pppprev	0
modis_combined_ndvi__NDVI__yearly_0_mean__ppppprev	0
year	0
soc_percent	0