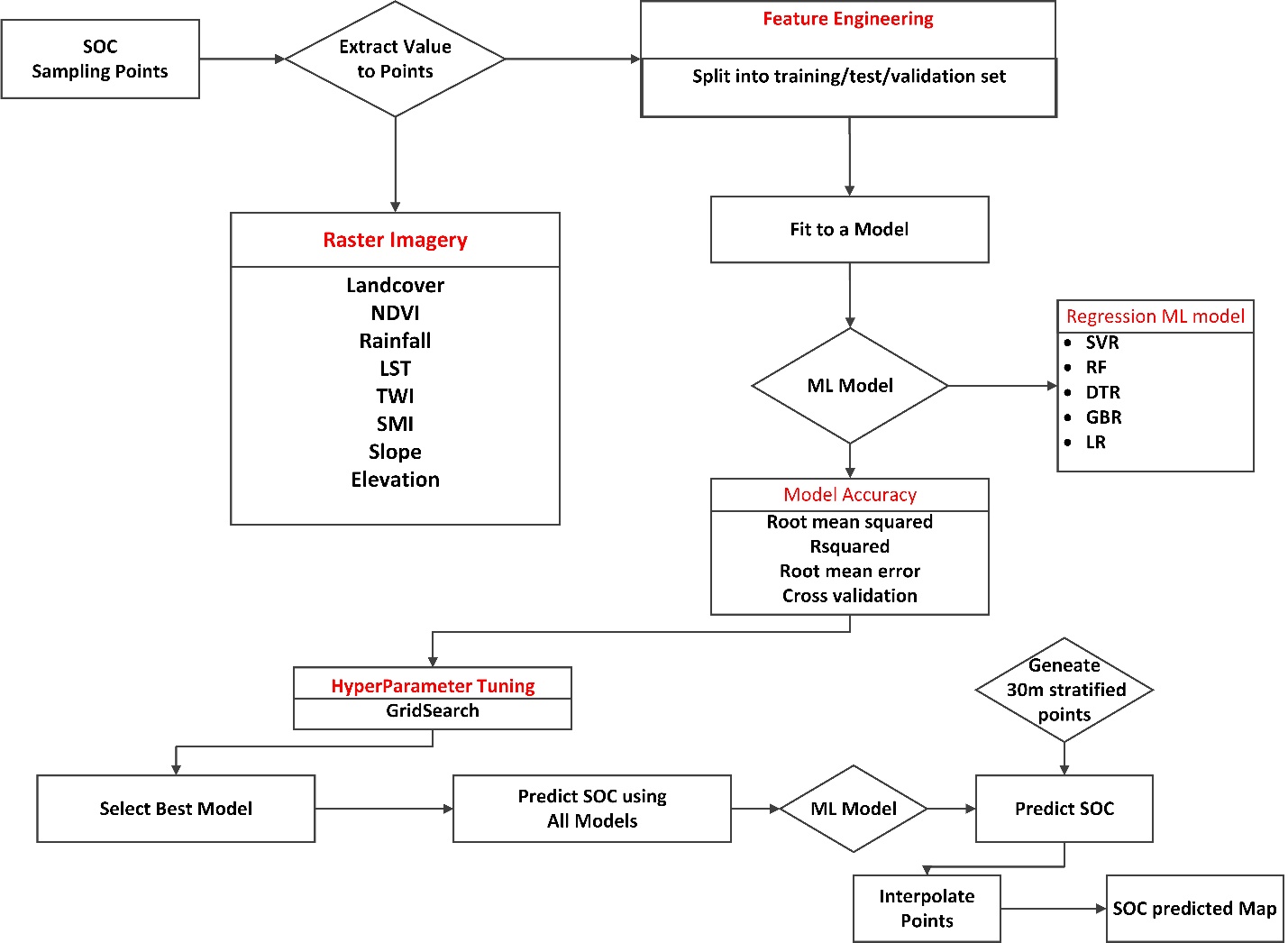
**Comparative Analysis of Machine Learning Algorithms for Predicting and Mapping Soil Organic Carbon Utilizing Remote Sensing and Field Sampling Data in Amazon Rainforest (A Case Study in Tonantins)**

**Methodology**



**Study area**

This study was conducted in the Amazon Rainforest, specifically within the Tonantins region, chosen for its ecological diversity and variability in soil properties, which are essential for analyzing soil organic carbon (SOC).

**Dataset**

The primary focus was on predicting and mapping SOC using machine learning models, combining field-collected SOC data with remotely sensed environmental variables. The soil organic carbon data was collected through field sampling, utilizing a stratified random sampling method to capture the spatial variability of SOC across different land use types and topographical features at 30cm depth. The SOC content in the collected soil samples was measured using the dry combustion method, which is known for its accuracy in determining organic carbon content in soils.

In addition to the SOC data, a variety of remote sensing data were gathered to serve as predictor variables for the machine learning models. These variables included the Normalized Difference Vegetation Index (NDVI), which was derived from Landsat imagery to estimate vegetation cover, and the Land Surface Temperature (LST) obtained from MODIS data. The Soil Moisture Index (SMI) was also calculated using a combination of NDVI and LST, providing insights into soil moisture conditions across the study area. Rainfall data, another critical environmental variable influencing SOC, was acquired from the CHIRPS.

**Data Sources**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Resolution | Source | Derivatives | references |
| DEM | 30m | SRTM | Elevation, Slope, TWI | Jarvis et al, 2008 |
| Landcover | 10m | ESA LULC |  | Zanaga, et. al, 2020 |
| Sampling Point |  | Field sampling | Bulk Density, SOC |  |
| Soil | 30cm depth | Openland | Soil PH, Soil texture, | Tomislav, 2018 |
| Landsat 8 OLI | 30m | USGS | NDVI, LST, SMI |  |
| Precipitation | 5000m | CHIRPS Daily | Average Precipitation | Funk et al., 2015 |

Data Processing

**Normalized Difference Vegetation Index (NDVI)**

The NDVI is the most widely used vegetation index for monitoring global forestry. NDVI is calculated using the high reflectance of vegetation in the Near-Infrared (NIR) spectrum and the high absorption in the Red spectrum. The following formula is used to calculate NDVI:

NDVI =

**Soil Texture and Soil Ph**

Soil texture and soil pH data were extracted from the OpenLand soil dataset using the Google Earth Engine data catalogue platform. This dataset provides valuable insights into the distribution of different soil types, while soil pH data reveals information about soil acidity. Together, these factors are essential for understanding soil properties and their influence on land management and agricultural practices.

**Land Surface Temperature (LST)**

In this study, Top of Atmosphere (TOA) Tier 1 Level 2 Landsat imagery, which is already processed and calibrated, was used. Band 10 of this processed Landsat data was utilized to compute the temperature of each location in Kelvin. The formula below was then applied to convert these values from Kelvin to Celsius.

LST(c) ​= BAND 10 ​− 273.15

**Soil Moisture Index (SMI)**

To generate the Soil Moisture Index (SMI) to assess soil moisture conditions across the study area. The SMI was derived using the normalized difference between LST and vegetation index (NDVI), as these parameters are known to correlate with soil moisture levels. The formula below was used to compute SMI using the following formula:

where LST*min*​ and LST*max* ​ are the minimum and maximum LST values for the area of interest.

**Topography**

Topographical data, including elevation, slope, and the Topographic Wetness Index (TWI), were derived from a Digital Elevation Model (DEM) of the study area. Elevation and slope were directly calculated from the DEM, while TWI was computed to account for both slope and upstream contributing areas, which are important for understanding water movement and accumulation in the landscape.

**Climatic Data**

Rainfall data from the CHIRPS Daily dataset was extracted for the required date range. To obtain a representative overview of rainfall distribution across the area over time, precipitation data was averaged over the period from January 1, 2012, to December 31, 2022. The CHIRPS dataset combines 0.05° resolution satellite imagery with in-situ station data, creating gridded rainfall time series. (Funk et al., 2015)

**Machine Learning Models**

Five machine learning algorithms were employed in this study to model soil organic carbon (SOC) in the Tocantins region: Random Forest Regressor (RFR), Support Vector Regressor (SVR), Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), and Linear Regression (LR). Each algorithm was selected for its unique approach to regression, offering a comprehensive comparison of methods for SOC prediction.

The model with the best performance, as determined by the highest R² and accuracy score and the lowest MSE and MAE, was selected for final SOC prediction and mapping. This model was applied across the entire Tocantins region, producing a spatial distribution map of SOC. This map offers valuable insights into the spatial variability of SOC, contributing to more informed environmental management and conservation strategies in the Amazon Rainforest. Finally, the accuracy of the SOC predictions was validated using an independent set of field data. A sensitivity analysis was also conducted to assess the impact of each predictor variable on the SOC predictions, helping to identify the key environmental factors influencing SOC distribution in the Amazon Rainforest. The selected model was applied to the entire study area to generate a spatial distribution map of SOC. This map, created using GIS software, revealed the spatial variability of SOC in the region, providing valuable insights for environmental monitoring and land management. To reduce multicollinearity among predictor variables, Pearson correlation analysis was performed, and highly correlated variables were removed.

**Random Forest Regressor (RFR)**

The RFR is an ensemble learning technique that constructs multiple decision trees during training and outputs the mean prediction of these trees. By creating a "forest" of decision trees, each trained on a random subset of the data and features, RFR reduces overfitting and enhances model generalization. This approach is particularly effective for handling large datasets with numerous predictor variables, which is advantageous given the complexity of environmental data in this study. Additionally, RFR provides an estimate of feature importance, enabling an understanding of which variables most influence SOC predictions.

**Support Vector Regressor (SVR)**

SVR a regression variant of the Support Vector Machine (SVM), operates by finding a hyperplane that best fits the data within a specified margin of error. SVR is particularly suited to scenarios where the relationship between features and the target variable is non-linear. It achieves this through kernel functions, such as the radial basis function, which map input features into a higher-dimensional space where a linear regression can be applied. The choice of hyperparameters, including the penalty parameter and kernel type, is critical for SVR's performance, necessitating thorough hyperparameter tuning.

**Decision Tree Regressor (DTR)**

The DTR builds a model in the form of a tree structure, where each node represents a decision based on a feature value, and the leaves represent predicted SOC values. The model recursively splits the dataset, aiming to create the most homogeneous subgroups possible at each node. The primary advantage of DTR is its interpretability, as the decision-making process is straightforward to understand and visualize. However, DTR is prone to overfitting, especially in noisy datasets, making it less robust when used in isolation compared to ensemble methods like RFR and GBR.

**Gradient Boosting Regressor (GBR)**

GBR is another ensemble learning method, but unlike RFR, it builds models sequentially. Each new model attempts to correct the errors made by the previous models, improving overall prediction accuracy. GBR is highly flexible, allowing the optimization of various loss functions, and is particularly effective in handling complex datasets. However, GBR can be computationally intensive and prone to overfitting if the hyperparameters, such as the number of trees and learning rate, are not carefully tuned.

**Model Parameter Tuning**

To optimize the performance of each model, hyperparameter tuning was performed using grid search with cross-validation. Grid search involves systematically exploring a predefined set of hyperparameter values to identify the combination that yields the best model performance. Cross-validation, typically k-fold, was used to ensure that the model's performance was consistent across different subsets of the data, thus providing a reliable estimate of its generalization ability.

The models were evaluated using several key performance metrics. The R-squared (R²) value was used to assess the proportion of variance in SOC explained by the model, with higher R² values indicating better model performance. The accuracy score, while more commonly associated with classification tasks, was adapted to measure the proportion of correct predictions within a defined error margin. Mean Squared Error (MSE) was employed to calculate the average squared difference between observed and predicted SOC values, with lower MSE values indicating more accurate predictions. Mean Absolute Error (MAE) was also used to measure the average magnitude of prediction errors, providing a straightforward interpretation of model accuracy. All the experiments were conducted on a local Jupyter lab notebook, equipped with a 20cores and 16 GB RAM. The hyperparameter tuning ranges of the models are shown in Table 2

Table 2. Hyperparameters of the models in the experiment.

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameter | Range |
| RF | n\_estimators  max\_depth  min\_samples\_split | [50, 100, 200]  [None, 10, 20, 30]  [2, 5, 10] |
| SVR | C  gamma  kernel | [0.1,1,10]  [1e-3, 1e-4, 'scale']  [‘linear’,’rbf’] |
| GBR | n\_estimators  learning\_rate  max\_depth' | [50, 100, 200]  [0.01, 0.1, 0.2]  [3, 5, 7] |
| DTR | max\_depth  min\_samples\_split | [None, 10, 20, 30]  [2, 5, 10] |

**Model Evaluation**

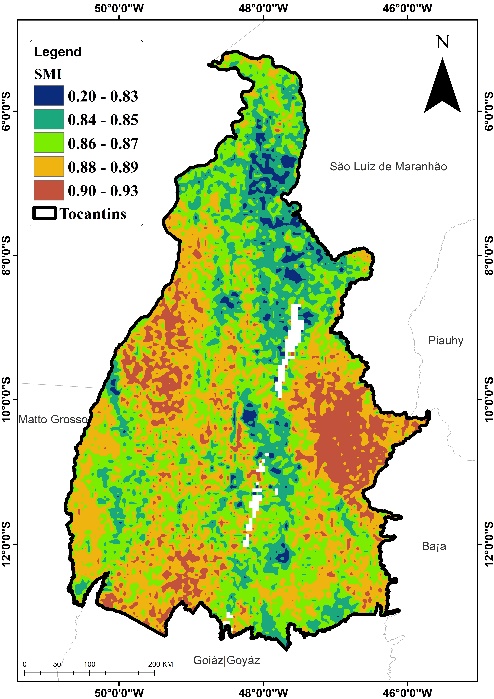
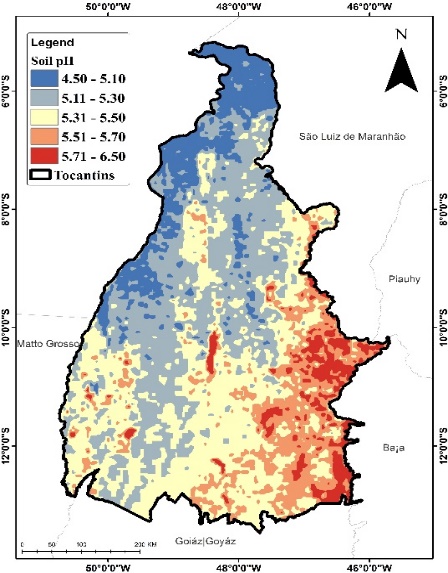
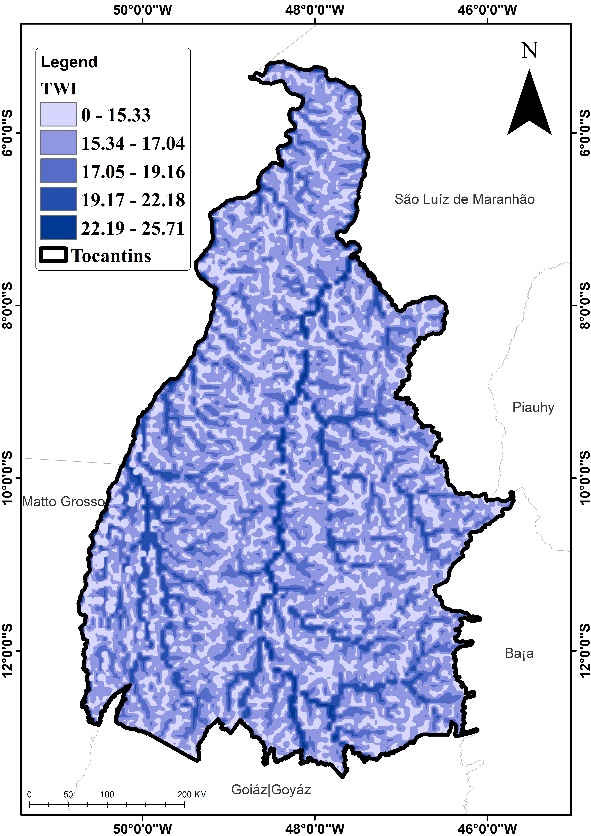
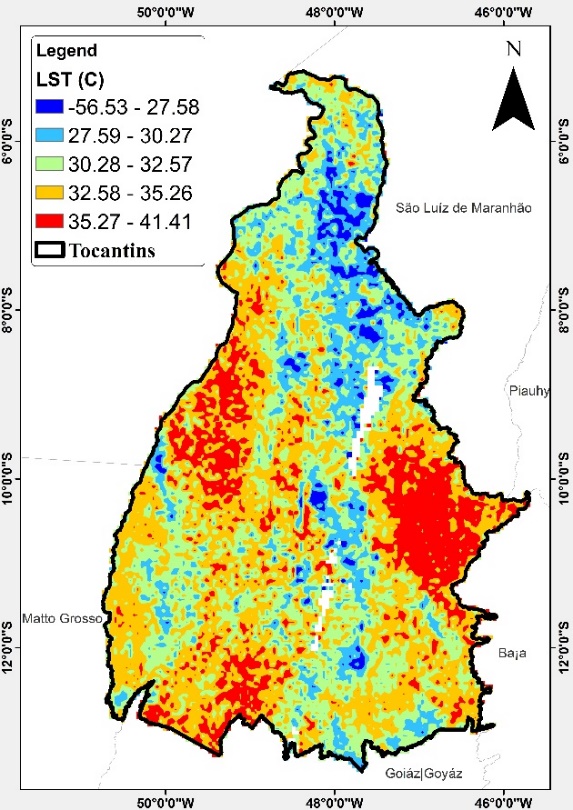
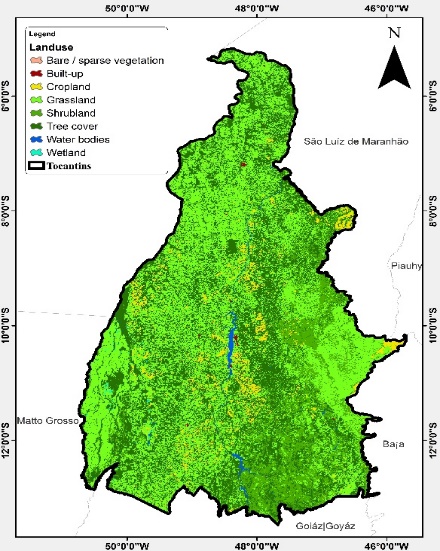
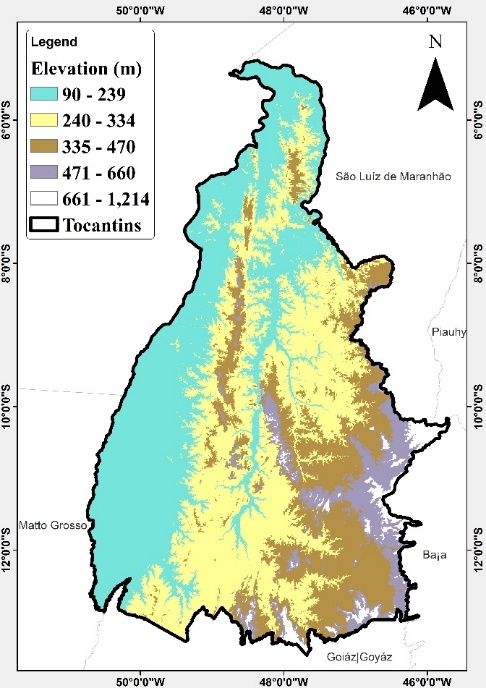
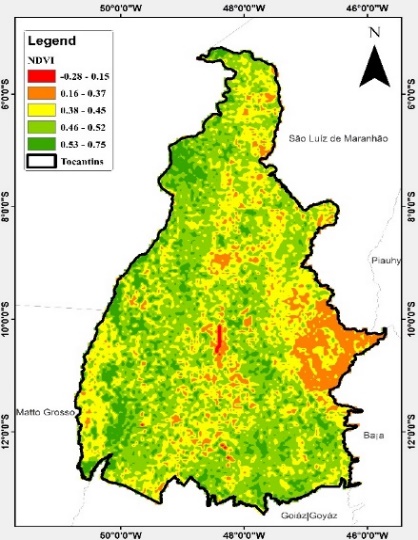
This study selected three evaluation metrics to quantify the accuracy of the different models in predicting soil moisture. The evaluation metrics used were the coefficient of determination (R2 ), root mean square error (RMSE), and mean absolute error (MAE). The calculation formulas are as follows:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Explanation** | **Formula** |
| MAE (Mean Absolute Error) | Measures the average magnitude of errors between predicted and actual values. Lower MAE values indicate better model accuracy. |  |
| MSE (Mean Squared Error) | Measures the average of the squared differences between predicted and actual values. Lower MSE values indicate better model accuracy. |  |
| R² (Coefficient of Determination) | Indicates how well the model explains the variability of the target variable. R² values range from 0 to 1, with higher values indicating better model performance. |  |

**Results and Discussion**

**Predictive variables**

The predictive variables used to model soil organic carbon (SOC) in the study area are depicted in **Figure**, each highlighting key environmental features. Figure (a) presents the slope, which ranges from 0.04 to 19.51 degrees, with most of the area having a slope less than 5 degrees, indicating relatively low topography. In figure (b), the elevation varies significantly from 90 to 1,214 meters, suggesting that the area has a diverse altitude range. Figure (c) shows the Normalized Difference Vegetation Index (NDVI), reflecting the health and density of vegetation, with values between -0.28 and 0.75, indicating a strong presence of vegetation, particularly in areas of dense tree cover and shrubs. The land surface temperature (LST) in figure (d) varies dramatically, from -56.53 to 41.41 degrees Celsius, further emphasizing the area's environmental variability. The land use map reveals six distinct classes: barelands, built-up areas, croplands, grasslands, shrub/tree cover, water bodies, and wetlands, the distribution is shown in **Table** with shrubs and tree cover being the dominant classes. The topographic wetness index (TWI) map suggests the presence of water flow patterns, with water moving downward through the study area. The soil pH map shows a range from 4.50 to 6.50, with lower pH values concentrated in the southern and southeastern parts of the area. Lastly, the soil moisture index ranges from 0.2 to 0.93, with higher values found in the southern region, indicating greater soil moisture in that part of the study area.



**Figure showing (a) Slope (b) Elevation (c) NDVI (d) LST (e) landuse (f) TWI (g) Soil pH (h) SMI**

**Table: Land use Classes Distribution**

|  |  |  |  |
| --- | --- | --- | --- |
| s/n | land use | Area (SQKM) | Percentage |
| 1 | Grassland | 128,754 | 45.34% |
| 2 | Tree cover | 102,348 | 36.04% |
| 3 | Shrubland | 38,243 | 13.47% |
| 4 | Cropland | 10,819 | 3.81% |
| 5 | water bodies | 3,033 | 1.07% |
| 6 | wetland | 396 | 0.14% |
| 7 | Built-up | 345 | 0.12% |
| 8 | Bare / sparse vegetation | 49 | 0.02% |
|  | **TOTAL** | **283,987** | **100** |

**Exploratory Data Analysis (EDA)**

The heatmap visualizes the correlation between soil organic carbon (SOC) and various environmental variables. In the heatmap, positive correlations (closer to 1) indicate that as the environmental factor increases, SOC also increases, while negative correlations (closer to -1) indicate an inverse relationship. Strong positive correlations are seen with soil pH (0.37), soil moisture index (0.27), elevation (0.26), and land surface temperature (0.27), meaning these factors tend to increase SOC levels. On the other hand, NDVI (-0.40), rainfall (-0.39), and latitude (-0.29) have strong negative correlations, suggesting these factors decrease SOC. Weaker correlations, such as TWI (0.06) and soil texture (-0.07), show minimal influence on SOC. This heatmap helps to easily identify which factors have the strongest impact on SOC.

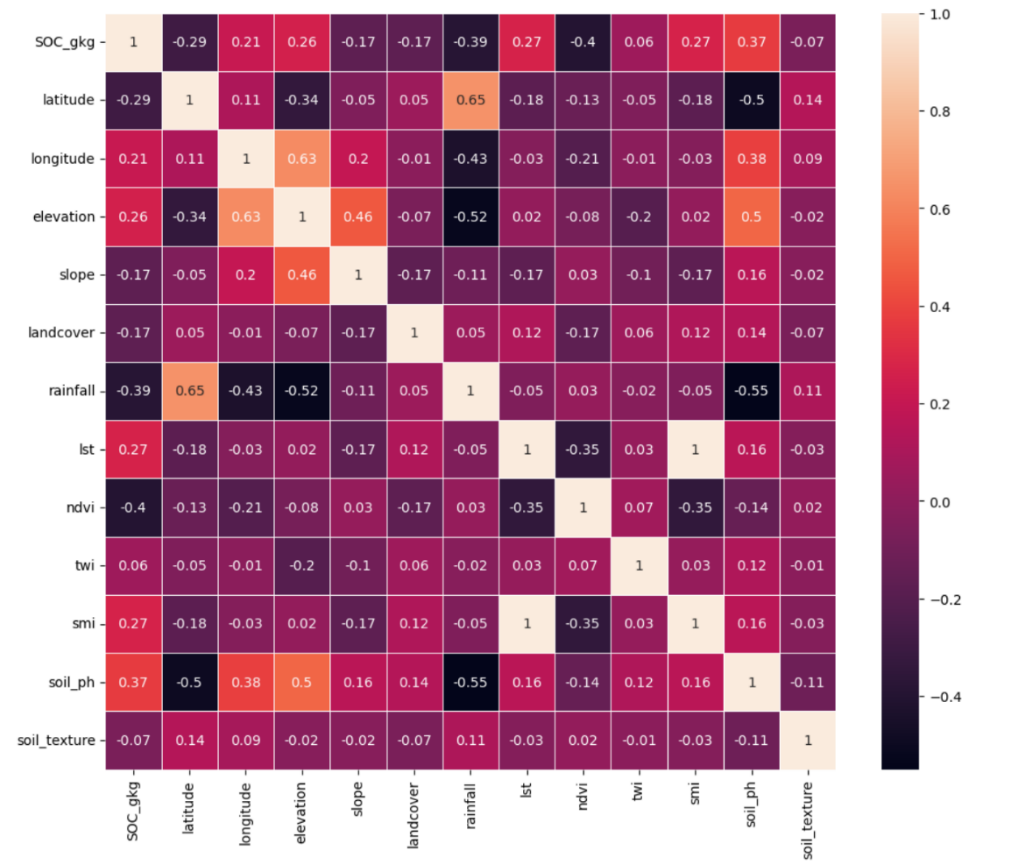


Figure: Correlation analysis of various indicators

The Ordinary Least Squares (OLS) regression results provide insights into the relationship between various environmental variables and soil organic carbon (SOC) in kilograms per gram (SOC\_gkg). The model's R-squared value of 0.473 indicates that approximately 47.3% of the variability in SOC can be explained by the predictors in the model, suggesting a moderate fit. The Adjusted R-squared value of 0.462 accounts for the number of predictors, indicating that the model is reasonably effective despite the inclusion of multiple variables. The F-statistic of 42.50 and the associated p-value (Prob (F-statistic)) of 2.01×10−65 show that the overall model is statistically significant, meaning at least one of the predictors contributes significantly to explaining SOC.

Examining the coefficients of individual predictors reveals several significant relationships. Elevation has a positive coefficient (0.0031) and a p-value of 0.000, indicating that higher elevations are associated with increased SOC levels. Conversely, slope (-0.3457) and landcover (-0.0372) have negative coefficients, suggesting that steeper slopes and certain landcover types may decrease SOC. The rainfall variable also shows a negative correlation with SOC (coefficient: -0.0002), which is statistically significant (p = 0.001). Notably, NDVI has a strong negative coefficient of -7.2484 (p < 0.0001), indicating that areas with higher vegetation density may be correlated with lower SOC levels, which may require further investigation to understand the underlying ecological dynamics. Additionally, soil pH has a positive effect on SOC (coefficient: 1.3839, p < 0.0001), emphasizing the importance of soil chemistry in organic carbon storage. The model does indicate some potential multicollinearity issues, as noted by the small eigenvalue, which could affect the reliability of the coefficient estimates.

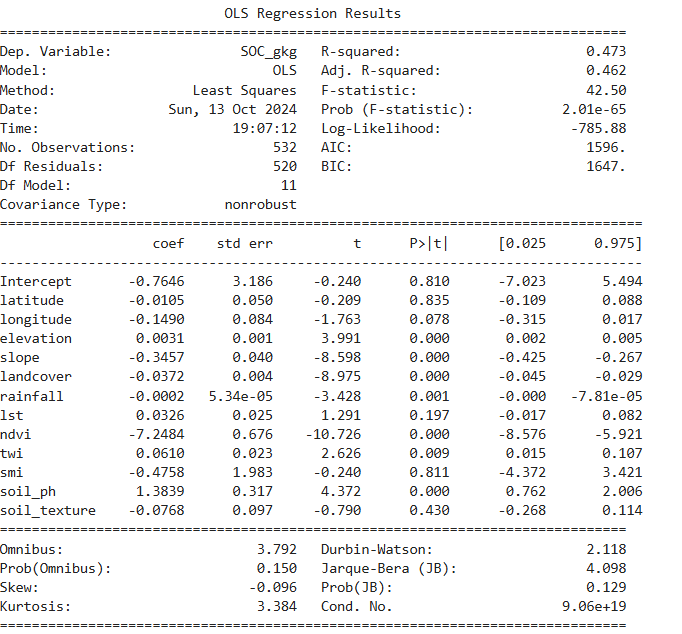


Figure:

**4.2. Model Performance**

During the model training, redundancy was observed in the characteristics of soil texture, soil pH, and land use distribution led to the removal of these properties from the model's predictions through a forward selection approach. This decision was made to enhance the model's efficiency and ensure that the included features provided unique and valuable information, ultimately improving the overall predictive performance. The table summarizes the performance of four machine learning models—Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting Regressor (GBR), and Decision Tree Regressor (DTR)—using three evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R². Among these, the Random Forest (RF) model performs the best, with the lowest MAE (0.52) and MSE (0.47), and the highest R² (0.72). This indicates that RF provides the most accurate predictions and explains 72% of the variance in the data, making it the most effective model for this task. Both SVR and GBR show moderate performance, with MAEs of 0.52 and 0.57, and R² values of 0.65. While these models are fairly accurate, they underperform compared to RF in both error reduction and explained variance. The Decision Tree Regressor (DTR), however, shows the weakest performance, with significantly higher MAE (0.85) and MSE (1.26), and a low R² of 0.25. This indicates that DTR struggles to capture the patterns in the data and explains only 25% of the variance, making it the least effective model for this analysis. Overall, RF emerges as the most reliable model, while DTR is the least suitable for this predictive task.

Table: Performance metrics (MAE, MSE, R²) comparison of four machine learning models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | MSE | R2 |
| RF | 0.52 | 0.47 | 0.72 |
| SVR | 0.52 | 0.58 | 0.65 |
| GBR | 0.57 | 0.59 | 0.65 |
| DTR | 0.85 | 1.26 | 0.25 |

A graph of different sizes of bars

Description automatically generated with medium confidence

Figure:

A group of graphs with different colored dots

Description automatically generatedFigure: Scatter plot of the actual and predicted soil organic carbon values on the testing set for different machine-learning methods. The predicted value is represented by the color of the point.

**Feature Importance**

Feature importance plays a vital role in understanding how various input features influence the predictive capabilities of machine learning models. The bar plot illustrating feature importance indicates that rainfall emerges as the most influential variable across the Random Forest, Gradient Boosting, and Decision Tree models, with importance scores of approximately 0.325, 0.322, and 0.338, respectively. This underscores its significant impact on prediction accuracy. Additionally, the Normalized Difference Vegetation Index (NDVI) demonstrates considerable importance, particularly within the Random Forest and Decision Tree frameworks, with scores of around 0.193 and 0.195. Other features, including latitude and elevation, show moderate levels of importance, whereas Topographic Wetness Index (TWI) and slope are associated with lower importance scores, indicating their lesser contribution to the models' predictive performance.

On the other hand, Support Vector Regression (SVR) does not provide feature importance scores directly. This limitation arises from the nature of the model, which seeks to identify a hyperplane that best fits the dataset rather than making decisions based on specific feature splits as tree-based models do. Moreover, the complexity of non-linear kernels complicates the assessment of feature importance since the relationships between inputs and outputs can become less straightforward.

A graph of different colored bars

Description automatically generated

Figure:

**Soil Organic Carbon Prediction**

A map of a large island

Description automatically generatedA map of a country with green and red areas

Description automatically generated

**I need the map for svr and dtr**

**Conclusion and recommendation**

**Limitation 1:** To measure feature importance in SVR, alternative techniques such as permutation importance or SHAP (Shapley Additive Explanations) can be utilized. These methods offer valuable insights into how each feature contributes to the model’s predictions. Thus, while tree-based models present clear measures of feature importance, SVR necessitates different methodologies to extract similar insights.

References

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell & Joel Michaelsen. "The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes". Scientific Data 2, 150066. doi:10.1038/sdata.2015.66 2015.

Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara. 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database: [https://srtm.csi.cgiar.org](https://srtm.csi.cgiar.org/).

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, Linlin, Tsendbazar, N.E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100. ([doi:10.5281/zenodo.5571936](https://doi.org/10.5281/zenodo.5571936))

Tomislav Hengl. (2018). Soil pH in H2O at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Version v02) [Data set]. Zenodo. [10.5281/zenodo.1475459](https://doi.org/10.5281/zenodo.1475459)