**Predicting the Soil Moisture Content in the Soil**

**Problem Statement:**

Train a regression model to predict the soil moisture content in the soil. Use the given training data( split into train and test samples) and try both ML/DL models to predict the response variable.

**Training data:**

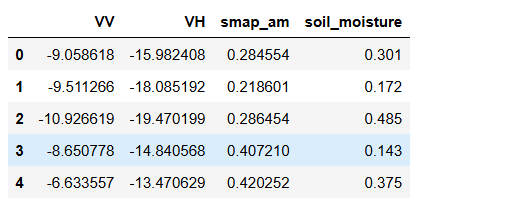
1.Dependent variable: soil\_moisture

2. Independent variables: VV, VH, smap\_am

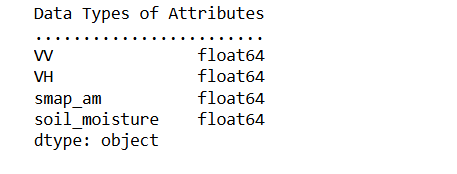
**Data Set:**

<https://drive.google.com/file/d/1vzbuiyOL5ddeCSBY2ZTospiKt8qYnT1v/view>

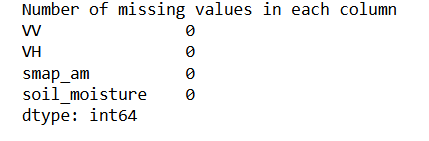
**First Five Rows of Data Set:**

****

**Data Types of Features:**

****

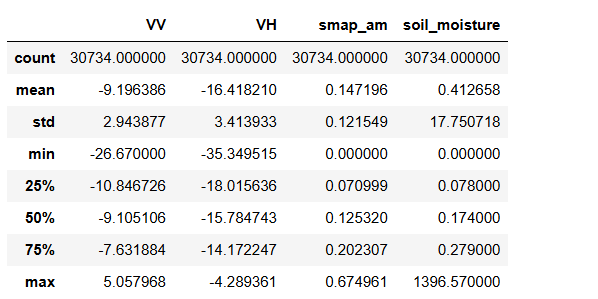
**Data Cleaning**Number of duplicated rows:13 and these are dropped using drop\_duplicates()

Null Values in each column:  


Can be filled using fillna() otherwise dropped using **dropna()**

**Exploratory Data Analysis**

**Statistical description of numerical columns:**

****

VV and VH:

VV (mean ≈ -9.2) being generally higher than VH (mean ≈ -16.4). They exhibit a moderate spread with standard deviations around 2.9–3.4.

smap\_am:

Mean = 0.147 (Range: 0.0–0.675)

75% of values ≤ 0.20, suggesting right-skewed distribution.

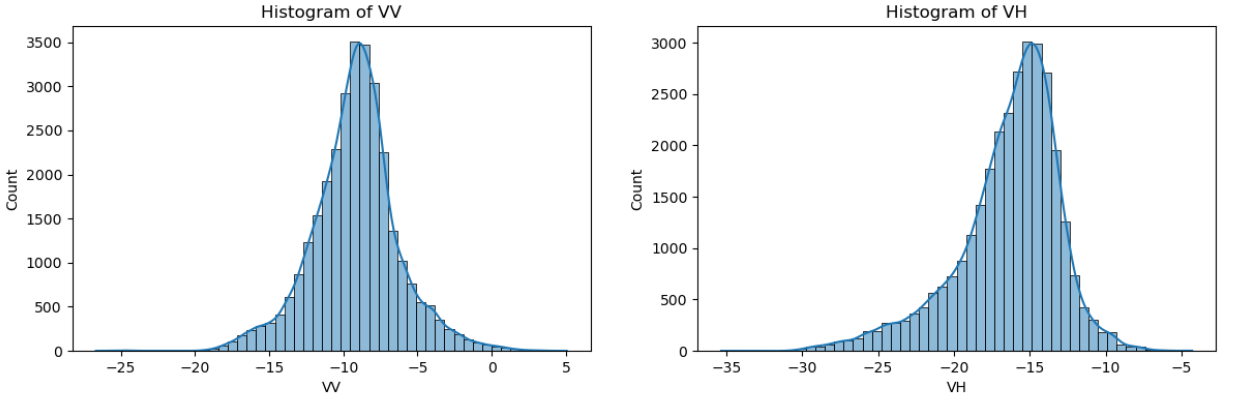
soil\_moisture:

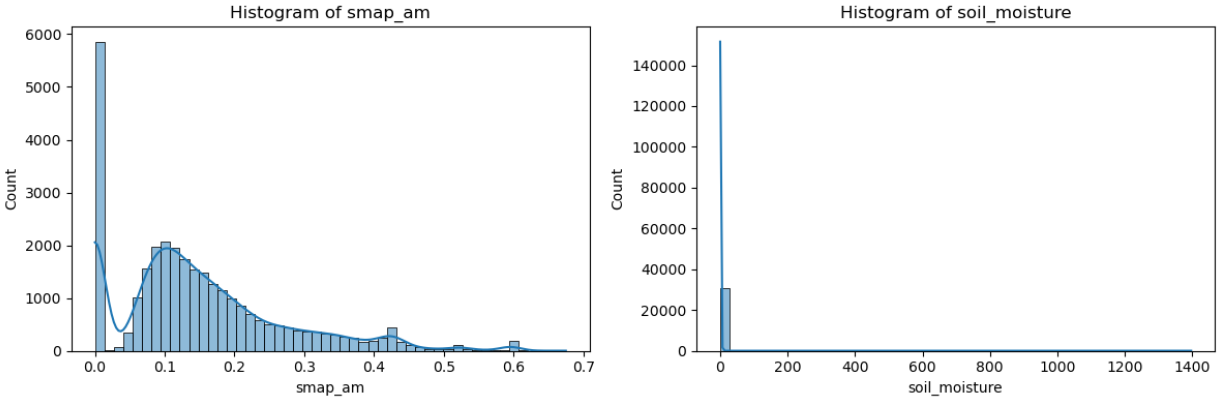
Highly skewed distribution with a mean of 0.41 but a maximum value of 1396.57, indicating strong outliers

The 25th, 50th, and 75th percentiles for soil\_moisture are 0.078, 0.174, and 0.279, suggesting that the bulk of valid values lie below 0.3, and extreme values are likely anomalies.

High std for soil\_moisture (17.75) confirms outlier influence

**Histogram Plots before outlier removal:**

****

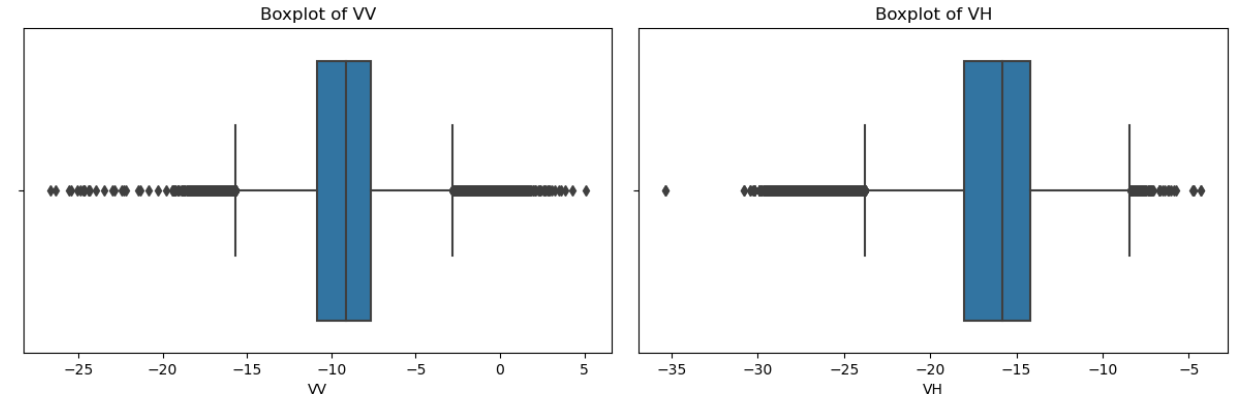
****

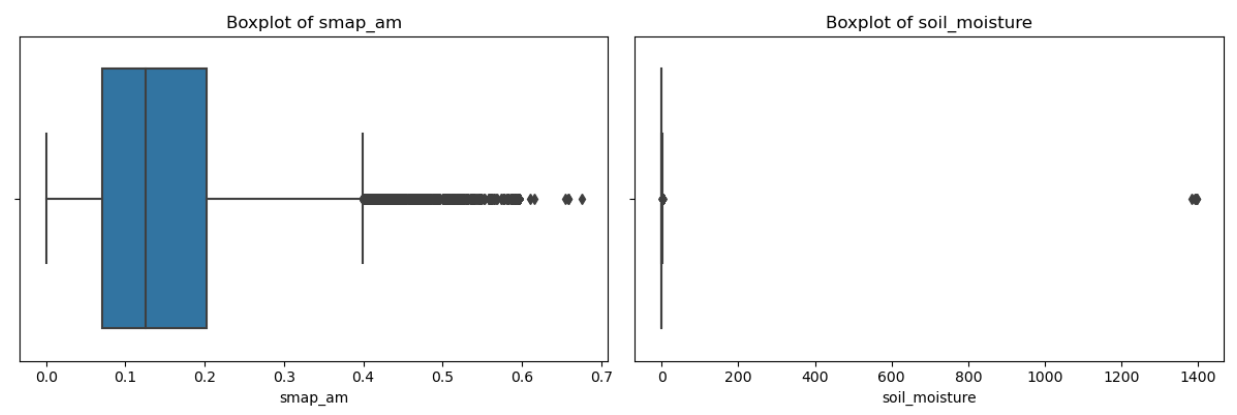
VV: normally distributed VH:slightly left skewed distribution

smap\_am: The distribution is moderately right-skewed, indicating a concentration of low SMAP values.

soil\_moisture: The distribution is highly right-skewed with visible extreme values (outliers).

**Box Plots before outlier removal:**

****

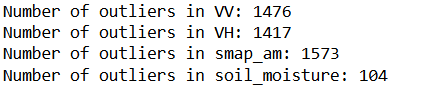
****

VV & VH: Both features show some mild outliers

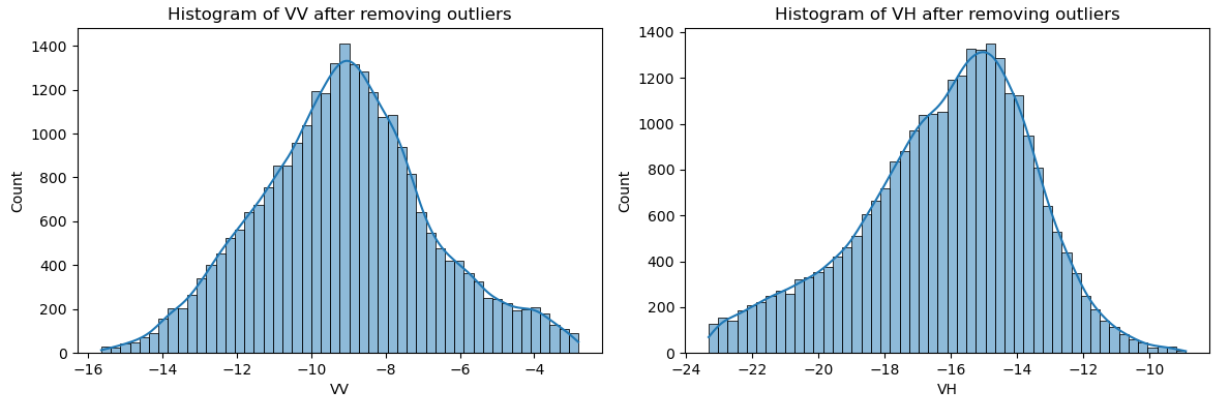
smap\_am: This feature has a few moderate outliers on the higher end

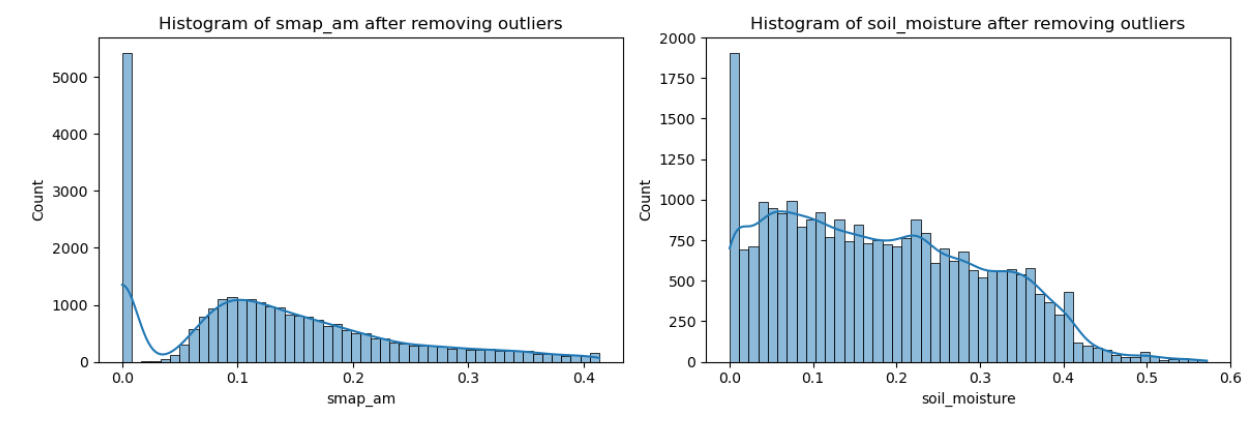
soil\_moisture: Shows significant presence of extreme outliers with values far beyond the upper whisker.

**Number of Outliers in Each Feature:**

****

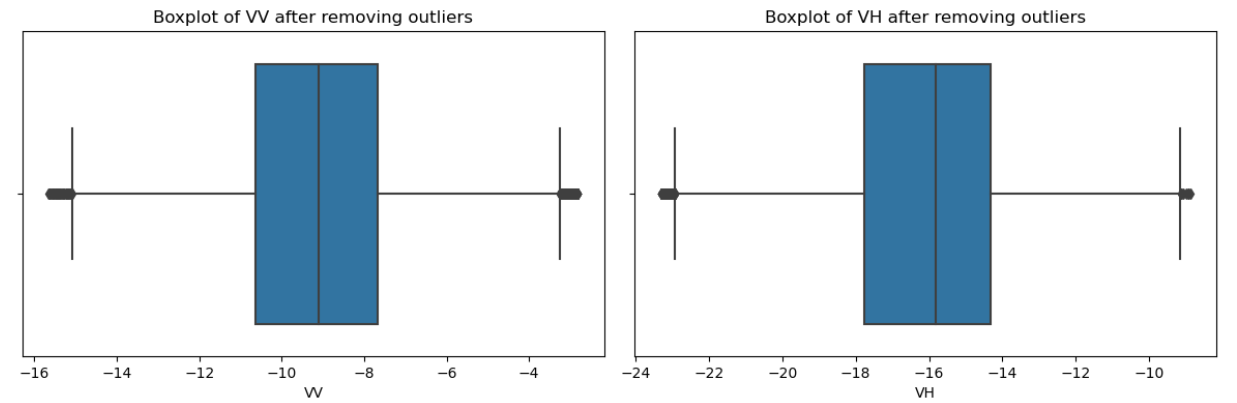
**Histogram Plots after outlier removal:**

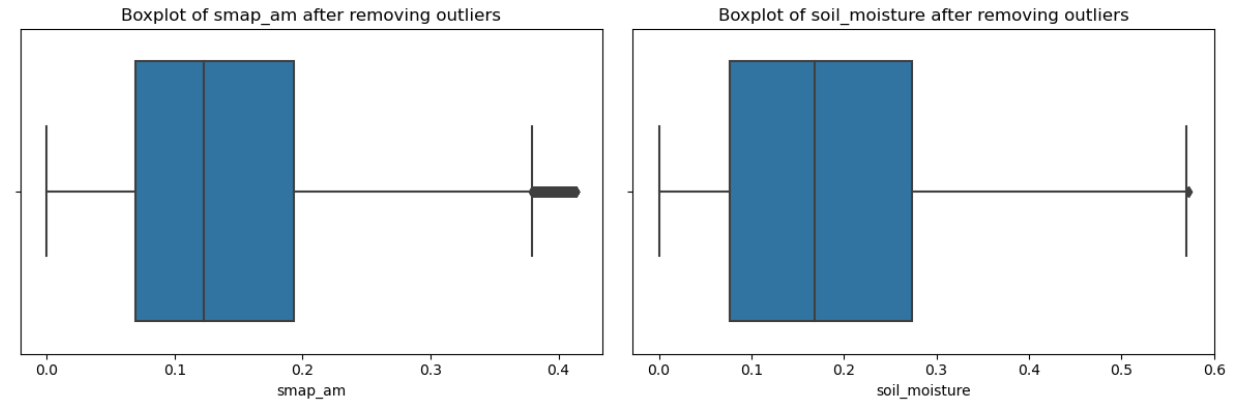
****

****

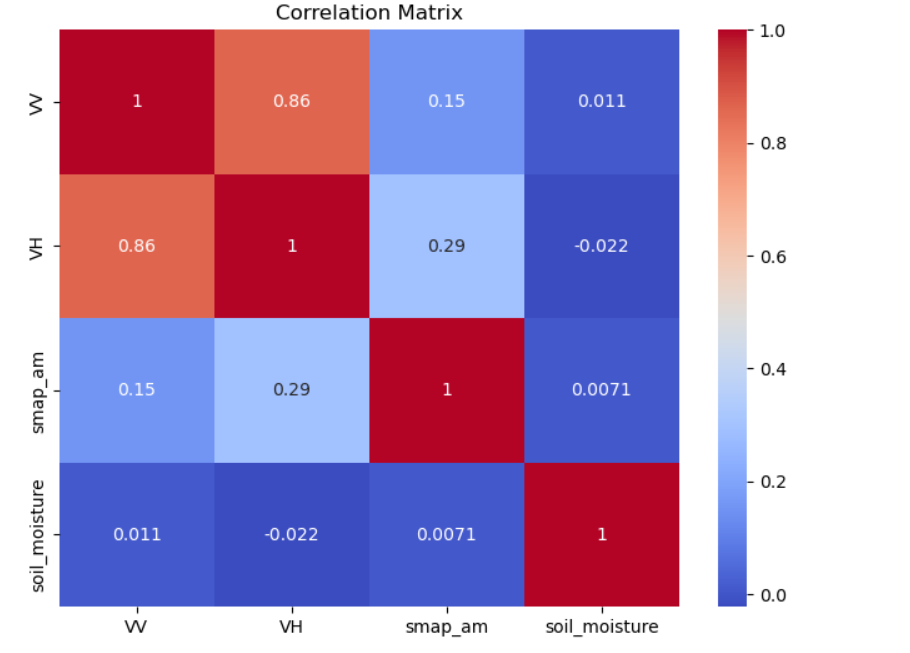
The distributions of smap\_am, soil\_moisture are right skewed and vv,vh have most likely normal distributions.

**Box Plots after outlier removal:**

****



**Correlation Matrix:**

****

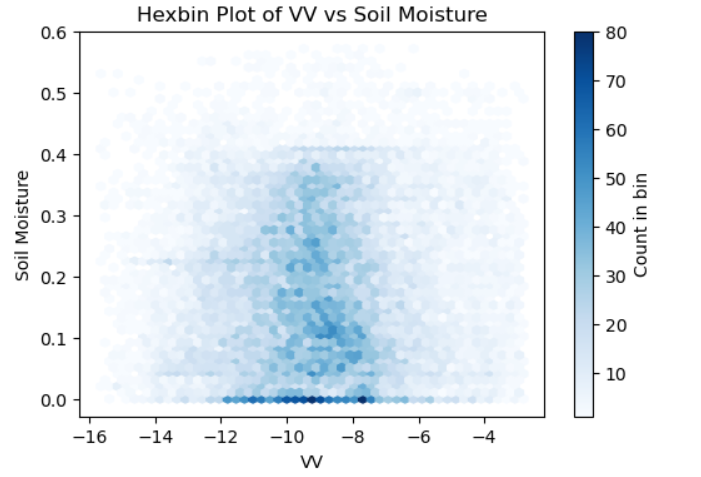
VV and VH are highly correlated (0.86), suggesting possible multicollinearity.

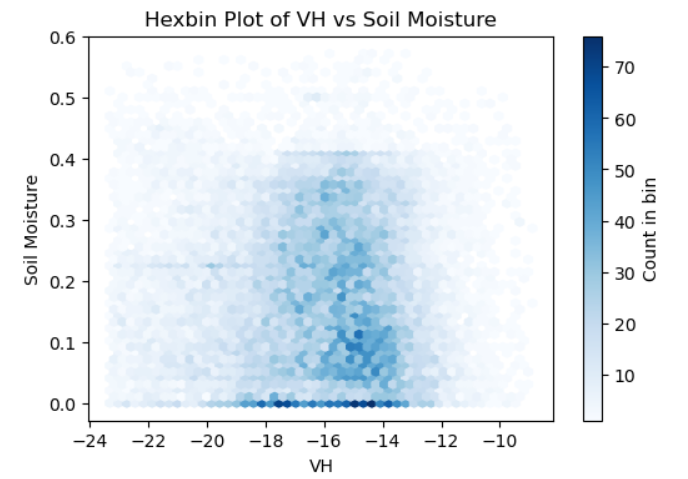
smap\_am shows a low positive correlation with VV (0.15) and VH (0.29), indicating a weak but existent relationship.

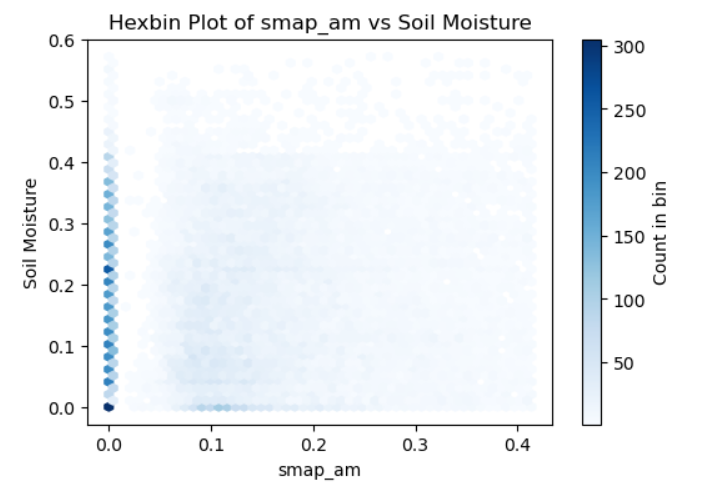
soil\_moisture has very weak or near-zero correlation with all input features VV( 0.011), VH(-0.022), smap\_am(0.0071)

Note:Low correlation with soil\_moisture doesn’t mean the features are useless. The relationship might be non-linear, which models like decision trees or neural networks can still learn effectively.

**Hexbin Plots Analysis :**

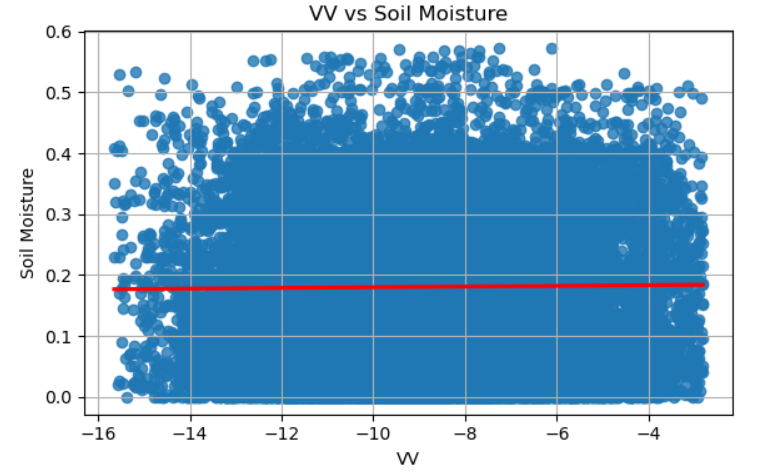
****

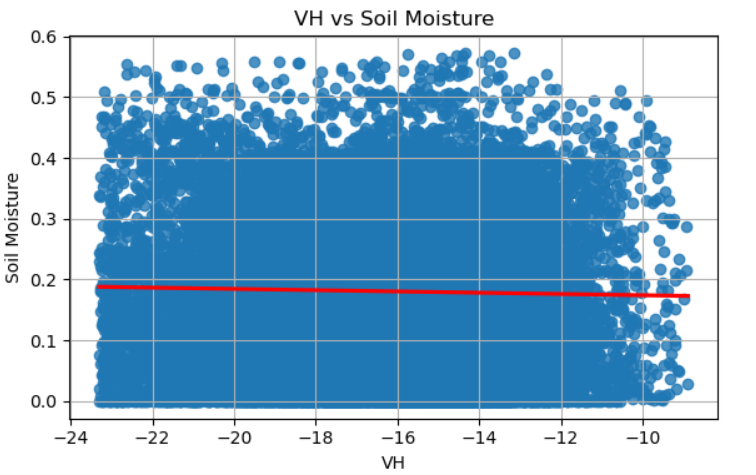
****

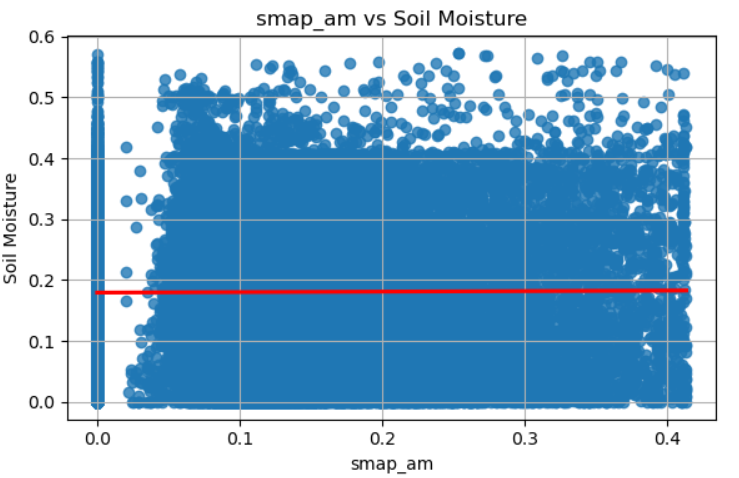
****

The data is widely scattered with no strong linear or clear non-linear patterns. The distribution of points appears random in most cases, indicating weak correlation between individual features and soil moisture values.Simple linear models may not be sufficient to capture the underlying relationships.Even non-linear models may struggle due to the lack of a strong, consistent trend in the input data.

**Linear Plot Analysis:**

****





The scatter plots with linear regression lines show that the features VV, VH, and smap\_am have no strong linear relationship with the target variable soil\_moisture.

The red regression lines are nearly horizontal, indicating very low slope.

The data points are randomly scattered around the line with no clear trend.

This suggests that Simple linear models are unlikely to capture the underlying patterns.[¶](http://localhost:8888/notebooks/Downloads/DataScience_casestudy/soil_moisture_prediction.ipynb#This-suggests-that-Simple-linear-models-are-unlikely-to-capture-the-underlying-patterns.)

**Models Building and Evaluation**

**Feature Engineering:**

Initial visualizations (hexbin and regression plots) and correlation analysis showed weak and scattered relationships between raw VV, VH, smap\_am and soil moisture, indicating poor linear correlation.

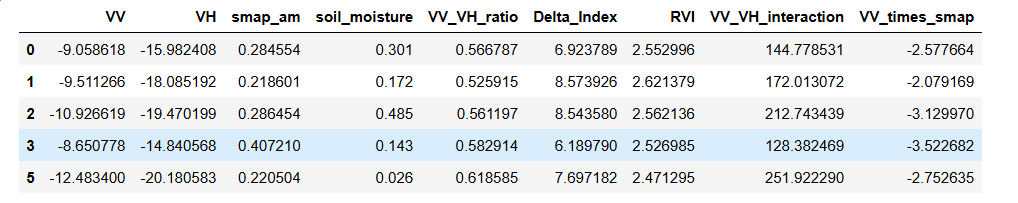
To address this, I engineered features—such as ratios, differences, vegetation indices, and interaction terms—to introduce non-linearity and enhance the model’s ability to capture complex patterns in the data

VV\_VH\_ratio=VV/VH  
Delta\_Index=VV−VH

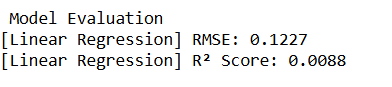
RVI=4×VH/VV+VH

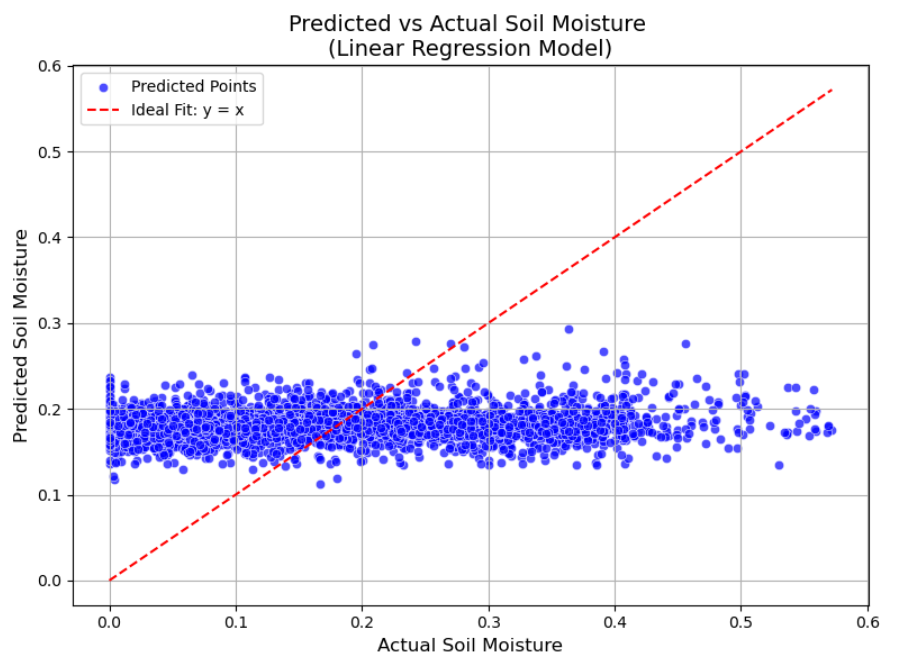
VV\_VH\_interaction=VV×VH

VV\_times\_smap=VV×smap\_am​​

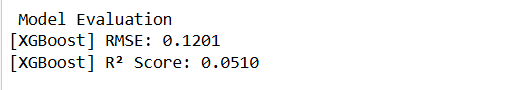


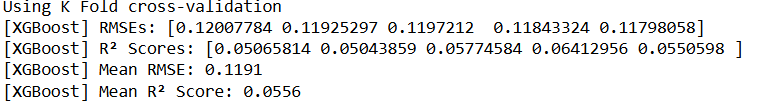
**LinearRegressor:**

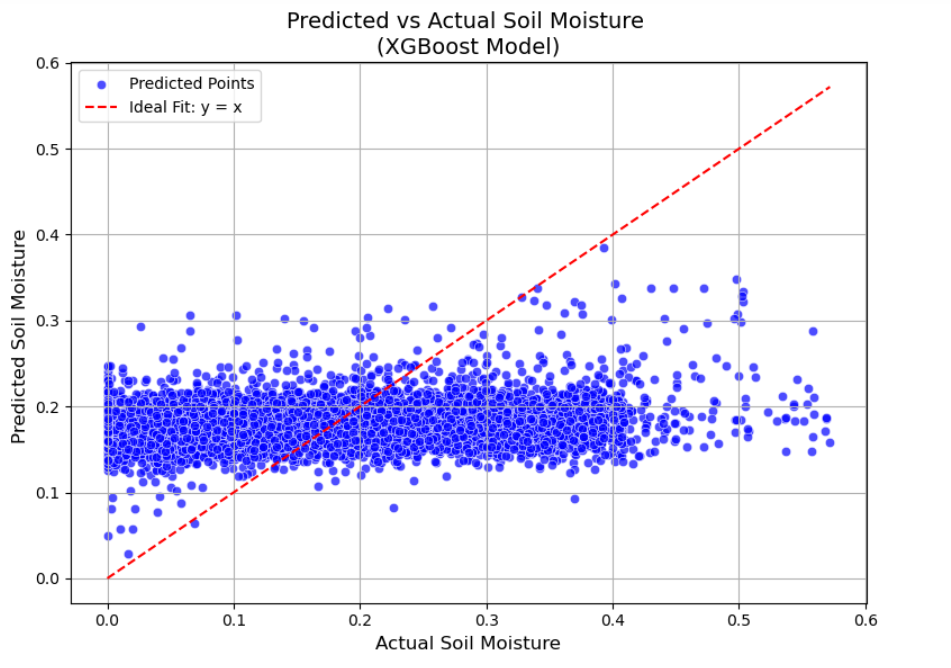
****

****

**XGBRegressor:**

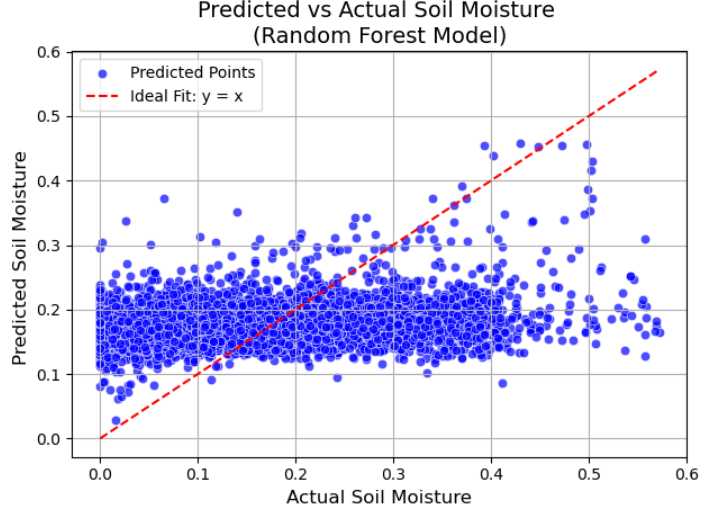
****





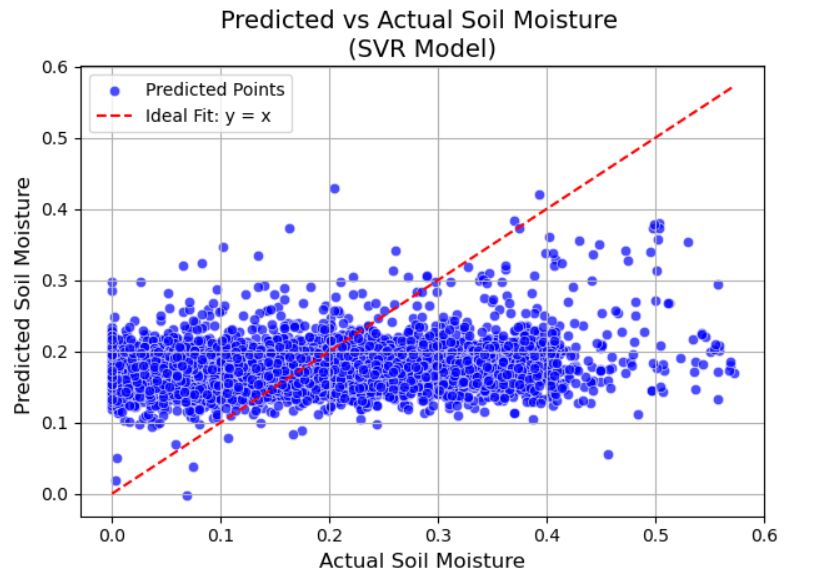
**Random Forest Regressor:**

****

****

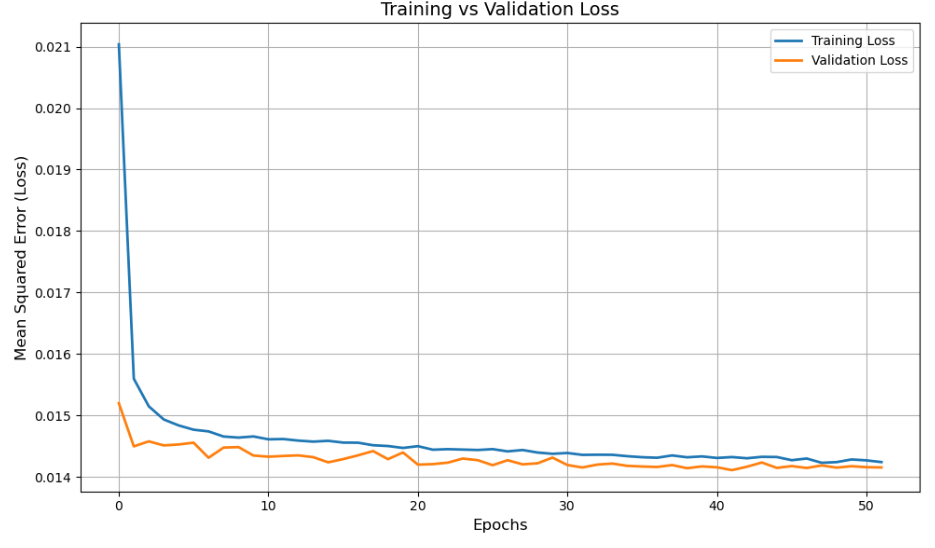
**Support Vector Regressor:**

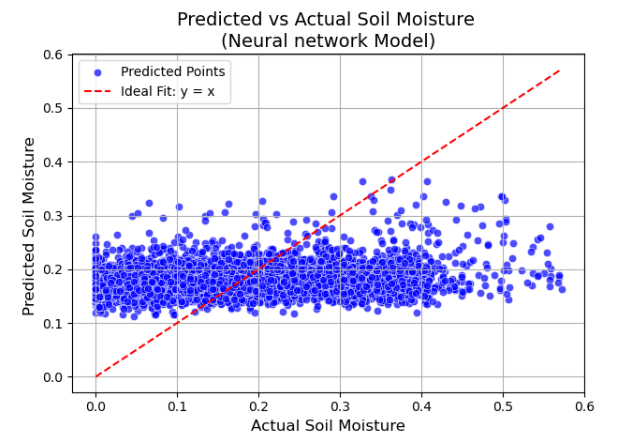
****

****

**Neural Network Model:**

****

****

****

**Conclusion**

In this study, I explored the application of multiple machine learning models to predict soil moisture using remote sensing data and derived features. The models evaluated include Linear Regression, XGBoost, Random Forest, Support Vector Regression (SVR), and a Neural Network.

* **Linear Regression**, although fast and interpretable, showed the **lowest performance** with an R² score of just **0.0088**, indicating it fails to capture the non-linear patterns in the data.
* **XGBoost** and **Random Forest** performed better, leveraging ensemble learning and non-linear capabilities, with R² scores around **0.05–0.055**.
* **SVR** and the **Neural Network** showed similar performance to the tree-based models, though with slightly higher RMSE and slightly lower R².

All models showed relatively low R² scores, indicating that they were only able to explain a small portion of the variation in soil moisture. This is likely because soil moisture is influenced by many complex and hidden environmental factors that are not fully captured by the available features. Additionally, the data appears to be scattered, making it difficult for the models to identify clear linear or non-linear relationships between the input variables and the target. Despite the low R² scores, the models demonstrated consistent performance across training and validation, suggesting they were learning meaningful (but limited) patterns.