KSAT Quest: Regression Runoff_Soilunitionist

Members Name: Shreya Gupta, Yash Joshi

A. Introduction

The goal of this project is to develop a robust machine learning model for predicting soil saturated hydraulic conductivity (Ksat) using the UKSAT dataset, a publicly available soil dataset hosted on HydroShare. Accurately estimating Ksat is vital in hydrology, agriculture, and environmental engineering, as it directly influences water movement through soil and supports effective land and water resource management.

This project follows a complete data science pipeline encompassing:

- Data Cleaning to ensure quality and consistency,
- Feature Selection to identify the most informative variables,
- Model Selection and Training using appropriate machine learning techniques,
- Hyperparameter Tuning to optimize performance, and
- Subset Experiments to evaluate model robustness under varying data sizes.

Performance is assessed using Root Mean Squared Logarithmic Error (RMSLE) and R² Score (Coefficient of Determination), providing insight into both prediction accuracy and explanatory power.

Additionally, a series of randomized subset experiments—each repeated 50 times—will simulate real-world scenarios where only limited data is available, testing the model's generalization capacity under data constraints.

This report outlines all critical stages of the modeling workflow, including preprocessing decisions, model development, evaluation results, and visualizations, with the aim of creating a replicable and explainable machine learning solution for Ksat prediction.

B. Abstract

This project aims to build a machine learning model to predict soil saturated hydraulic conductivity (Ksat) using the UKSAT dataset. The workflow involves data cleaning, feature selection, model training, and hyperparameter tuning. To assess robustness, we conduct 50 randomized subset experiments by reducing the dataset size in steps of 2,000 samples. Models are evaluated using Root Mean Squared Logarithmic Error (RMSLE) and R² Score,

with performance visualized as a function of training size. The results offer insights into model stability and accuracy under varying data conditions.

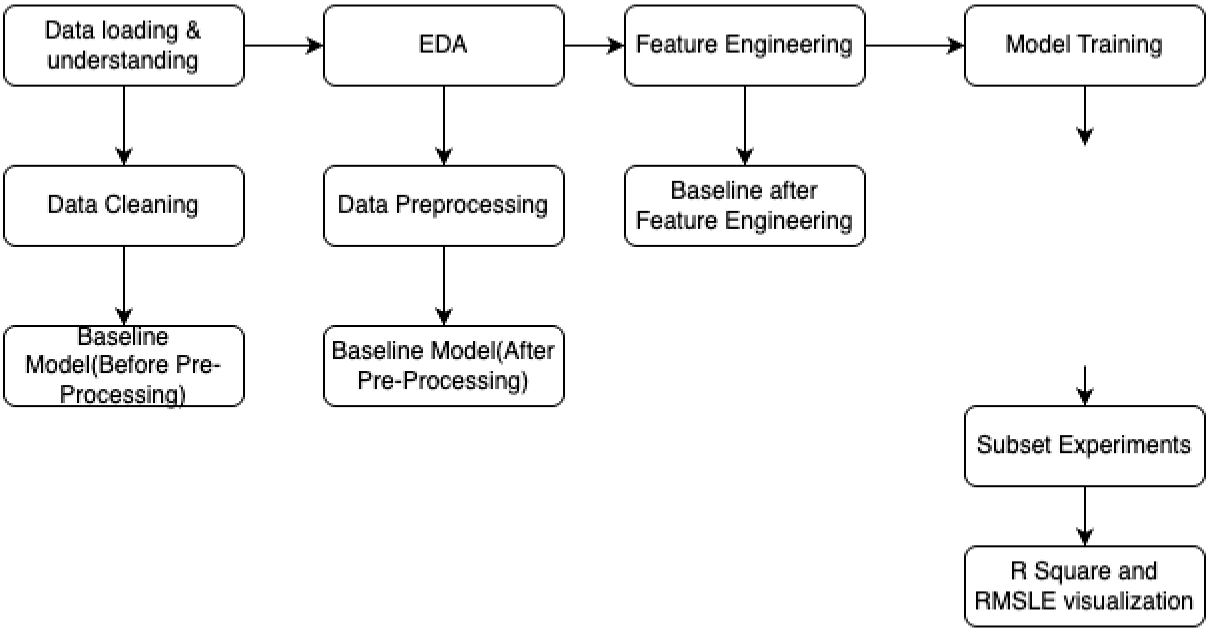
C. Problem Statement

Predicting soil saturated hydraulic conductivity (Ksat) accurately is essential for applications in hydrology, agriculture, and environmental modeling. Traditional measurement methods are time-consuming, resource-intensive, and often impractical for large-scale studies.

The objective of this project is to develop a machine learning model that can reliably estimate Ksat values using the UKSAT dataset, which contains various soil physical and chemical properties. The challenge lies in identifying relevant features, handling potentially high-dimensional data, and ensuring model performance remains stable across varying dataset sizes.

This project also investigates how training data size affects model performance through repeated subset experiments, providing insight into the minimum data requirements for effective Ksat prediction.

D. Methodology


Data loading & understanding
EDA
Feature Engineering
Model Training
Data Cleaning
Data Preprocessing
Baseline after Feature Engineering
Baseline Model(Before Pre-Processing)
Baseline Model(After Pre-Processing)
Subset Experiments
R Square and RMSLE visualization

Explanation of All Steps with Plots and Outputs

A. Data Loading & Understanding

The dataset was loaded from an Excel file using pandas. Initial exploration revealed several issues:

- Presence of unnamed and empty columns.
- Several features had incorrect data types (e.g., numeric fields stored as strings).
- The target variable was identified as Ksat (Saturated Hydraulic Conductivity).

Summary statistics and .info() output highlighted inconsistencies in the structure and suggested the need for thorough cleaning and preprocessing.

B. Data Cleaning

Key cleaning steps included:

- Removing unnamed/empty columns.
- Standardizing column names: extra spaces, special characters, and formatting inconsistencies were fixed for better usability.
- Fixing data types: numeric values stored as strings were converted properly using typecasting.
- Handling missing values: A summary of missing value percentages was created. Columns with excessive missing values were dropped, and others were handled with imputation or row removal.

After cleaning, a baseline model was trained using the cleaned data (before further preprocessing or feature engineering) to serve as a reference.

C. EDA & Data Preprocessing

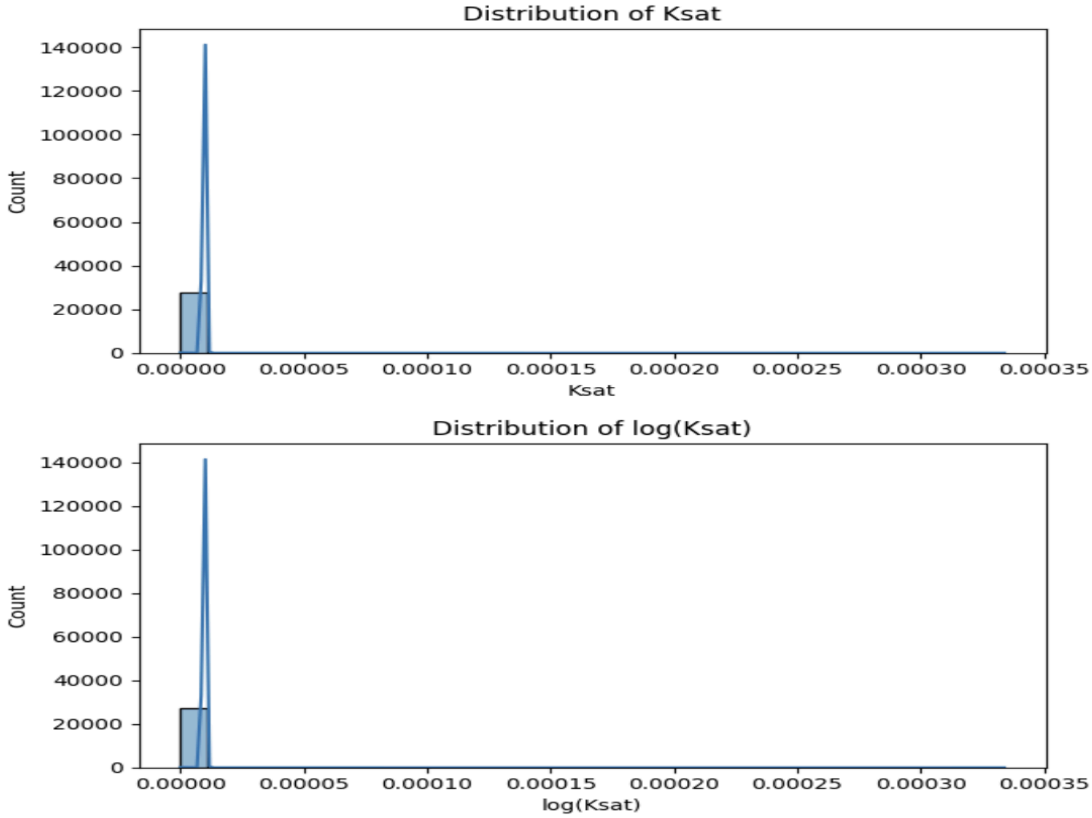
Exploratory Data Analysis (EDA) involved:

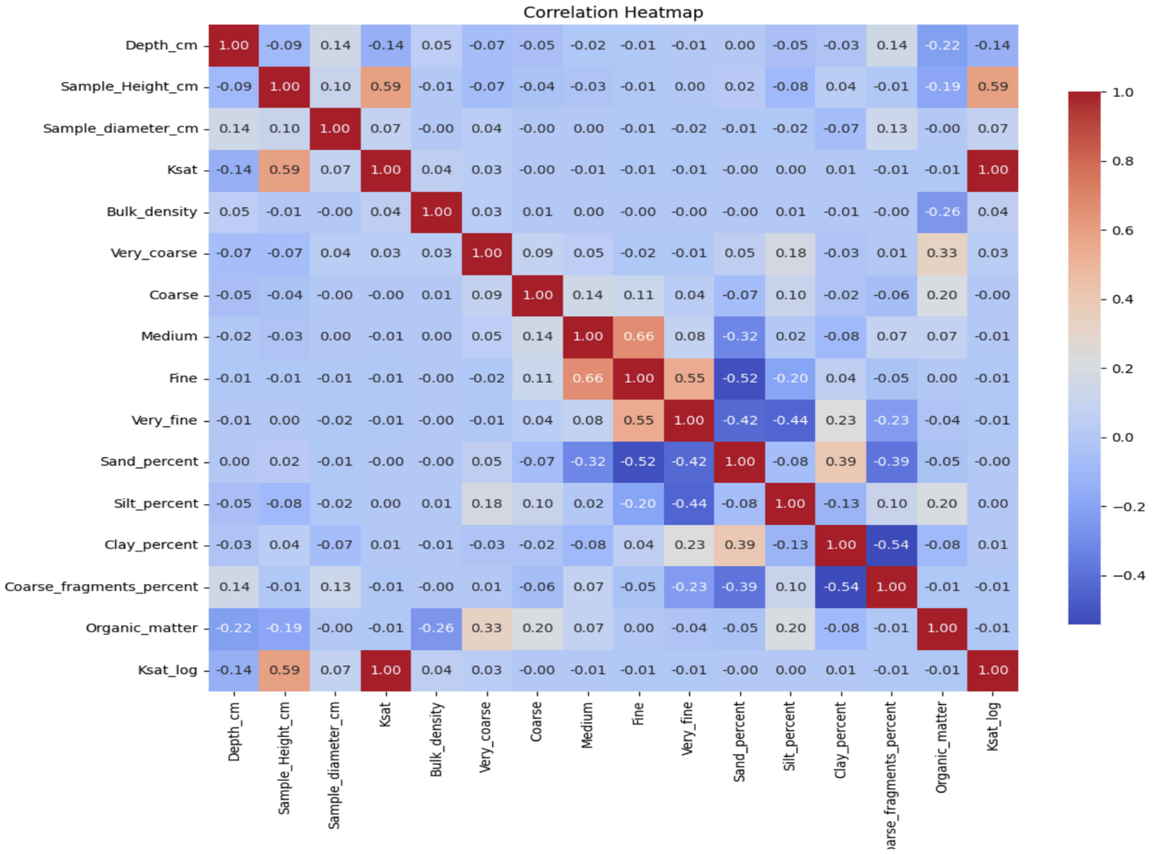
- Visualizing feature distributions and correlations with Ksat.

- Detecting outliers and skewed features.
- Understanding the relationships between soil characteristics and Ksat using scatterplots and heatmaps.

Preprocessing included:

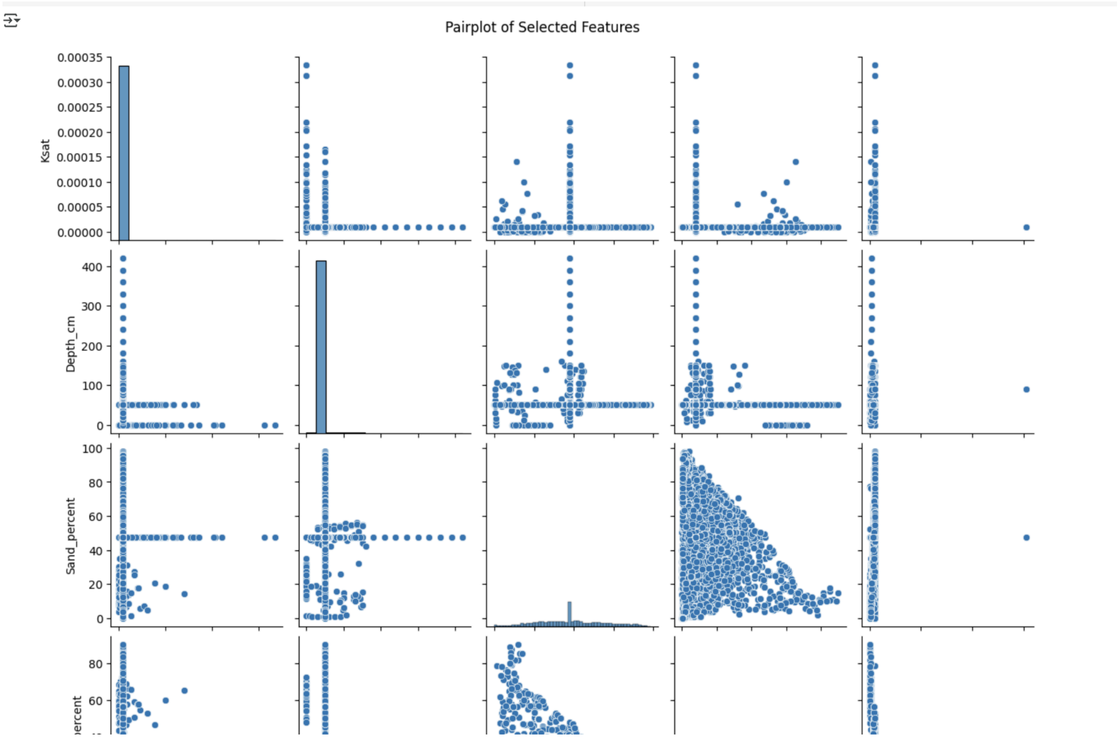
- Feature standardization or transformation, where necessary.
- Removing multicollinearity by examining correlation matrices.
- Building a second baseline model after preprocessing to measure the gain in performance due to cleaning and transformation.





	Depth_cm	Sample_Height_cm	Sample_diameter_cm	Ksat	Bulk_density	Very_coarse	Coarse	Medium	Fine	Very_fine	Sand_percent	Silt_percent	Clay_percent	arse_fragments_percent	Organic_matter	Ksat_log
Depth_cm	1.00	-0.09	0.14	-0.14	0.05	-0.07	-0.05	-0.02	-0.01	-0.01	0.00	-0.05	-0.03	0.14	-0.22	-0.14
Sample_Height_cm	-0.09	1.00	0.10	0.59	-0.01	-0.07	-0.04	-0.03	-0.01	0.00	0.02	-0.08	0.04	-0.01	-0.19	0.59
Sample_diameter_cm	0.14	0.10	1.00	0.07	-0.00	0.04	-0.00	0.00	-0.01	-0.02	-0.01	-0.02	-0.07	0.13	-0.00	0.07
Ksat	-0.14	0.59	0.07	1.00	0.04	0.03	-0.00	-0.01	-0.01	-0.01	-0.00	0.00	0.01	-0.01	-0.01	1.00
Bulk_density	0.05	-0.01	-0.00	0.04	1.00	0.03	0.01	0.00	-0.00	-0.00	-0.00	0.01	-0.01	-0.00	-0.26	0.04
Very_coarse	-0.07	-0.07	0.04	0.03	0.03	1.00	0.09	0.05	-0.02	-0.01	0.05	0.18	-0.03	0.01	0.33	0.03
Coarse	-0.05	-0.04	-0.00	-0.00	0.01	0.09	1.00	0.14	0.11	0.04	-0.07	0.10	-0.02	-0.06	0.20	-0.00
Medium	-0.02	-0.03	0.00	-0.01	0.00	0.05	0.14	1.00	0.66	0.08	-0.32	0.02	-0.08	0.07	0.07	-0.01
Fine	-0.01	-0.01	-0.01	-0.01	-0.00	-0.02	0.11	0.66	1.00	0.55	-0.52	-0.20	0.04	-0.05	0.00	-0.01
Very_fine	-0.01	0.00	-0.02	-0.01	-0.00	-0.01	0.04	0.08	0.55	1.00	-0.42	-0.44	0.23	-0.23	-0.04	-0.01
Sand_percent	0.00	0.02	-0.01	-0.00	-0.00	0.05	-0.07	-0.32	-0.52	-0.42	1.00	-0.08	0.39	-0.39	-0.05	-0.00
Silt_percent	-0.05	-0.08	-0.02	0.00	0.01	0.18	0.10	0.02	-0.20	-0.44	-0.08	1.00	-0.13	0.10	0.20	0.00
Clay_percent	-0.03	0.04	-0.07	0.01	-0.01	-0.03	-0.02	-0.08	0.04	0.23	0.39	-0.13	1.00	-0.54	-0.08	0.01
Coarse_fragments_percent	0.14	-0.01	0.13	-0.01	-0.00	0.01	-0.06	0.07	-0.05	-0.23	-0.39	0.10	-0.54	1.00	-0.01	-0.01
Organic_matter	-0.22	-0.19	-0.00	-0.01	-0.26	0.33	0.20	0.07	0.00	-0.04	-0.05	0.20	-0.08	-0.01	1.00	-0.01
Ksat_log	-0.14	0.59	0.07	1.00	0.04	0.03	-0.00	-0.01	-0.01	-0.01	-0.00	0.00	0.01	-0.01	-0.01	1.00

Insight:



D. Feature Engineering

Feature engineering involved:

- Creating interaction terms between relevant features (if any).
- Selecting top features using feature importance from tree-based models.
- Reducing dimensionality by dropping features with low relevance.

A third baseline model was trained after feature engineering to assess the impact of selecting the most informative predictors.

E. Model Training, Hyperparameter Tuning & Subset Experiments

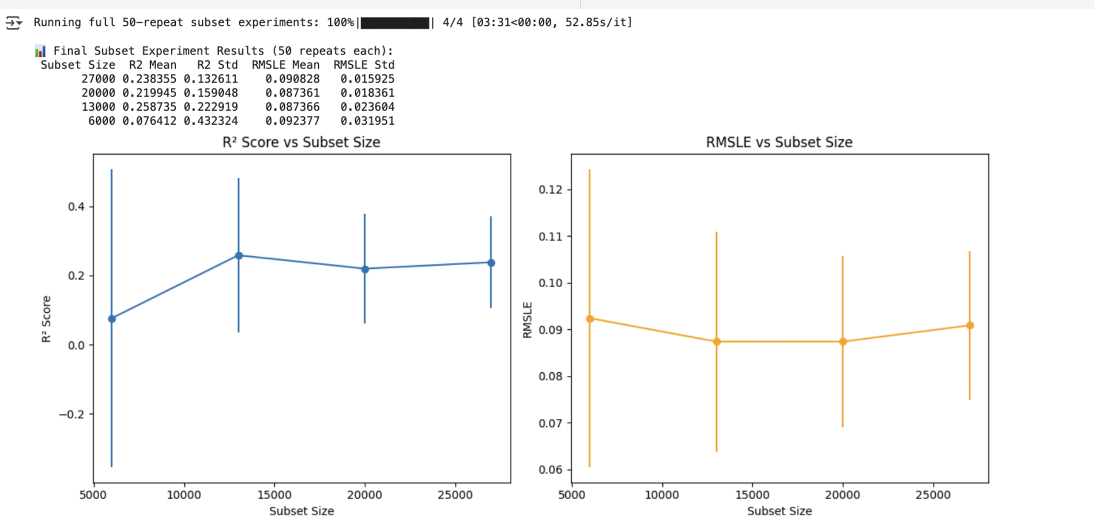
A Random Forest Regressor was selected as the final model (XGBoost was excluded as per project rules). The process included:

- Splitting the dataset into 80% training and 20% testing.

- Applying cross-validation to ensure robustness.
- Performing hyperparameter tuning using grid search for optimal values of n_estimators and max_depth.

Subset Experiments:

- The full modeling process was repeated on randomly sampled subsets of decreasing sizes (from ~27,000 to 6,000).
- For each subset size, the experiment was repeated 50 times to obtain reliable average metrics.
- This approach tested the model's generalization performance under constrained data scenarios.



```
Running full 50-repeat subset experiments: 100%|██████████| 4/4 [03:31<00:00, 52.85s/it]

 📊 Final Subset Experiment Results (50 repeats each):
 Subset Size  R2 Mean   R2 Std  RMSLE Mean  RMSLE Std
       27000 0.238355 0.132611    0.090828   0.015925
       20000 0.219945 0.159048    0.087361   0.018361
       13000 0.258735 0.222919    0.087366   0.023604
        6000 0.076412 0.432324    0.092377   0.031951
```

Subset Size vs R² Score

The highest average R² is at 13,000 samples.

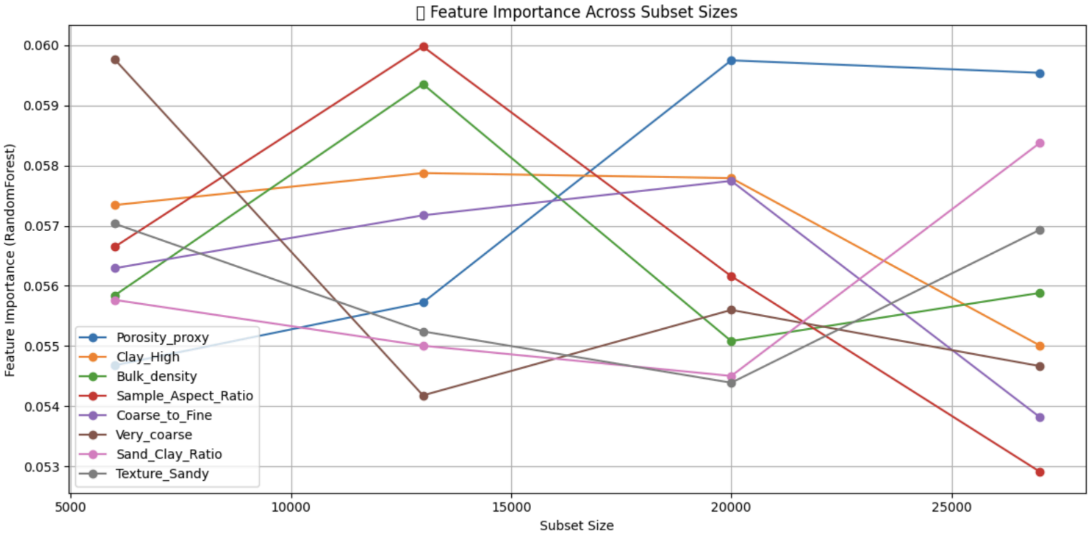
However, smaller subsets (6000) show high variance, making them less reliable.

The R² stabilizes as size increases beyond 13,000 but doesn't improve dramatically.

Subset Size vs RMSLE

RMSLE improves slightly at medium sizes (13k–20k), suggesting these subsets generalize better.

Small subset (6000) has high error variance, despite a decent mean RMSLE.



Feature Importance Across Subset Sizes.

📊 Plot Summary: Feature Importance Across Subsets

Porosity_proxy is consistently the top feature and increases in importance with larger subset sizes — showing stability and robustness.

Clay_High maintains steady importance until it slightly dips at the largest subset size.

Sample_Aspect_Ratio is volatile — peaking at 13000 and then significantly dropping at 27000.

Bulk_density and Sand_Clay_Ratio show moderate fluctuations, indicating moderate dependence on data volume.

Very_coarse and Texture_Sandy exhibit more noise and variability — they're likely less stable predictors.

✅ Insights:

Stable features across sizes: Porosity_proxy, Clay_High, Bulk_density.

Subset-size sensitive features: Sample_Aspect_Ratio, Very_coarse, Texture_Sandy.

We can prioritize stable features when building models for smaller datasets or constrained environments.

E. Results (Model results, cv results, best features, plots)

Our best model is SelectKBest with polynomial features.

📊 Performance after Polynomial + SelectKBest (Top 50 Features):

R² Score : 0.5734 RMSLE : 0.0739

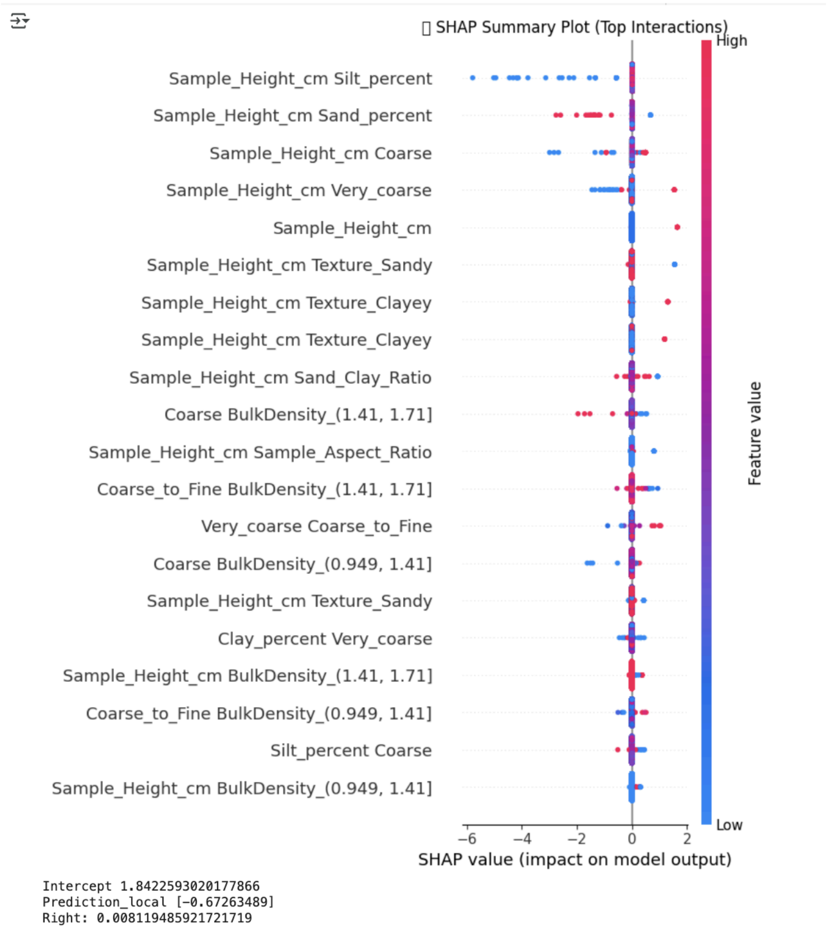
Metric Value 🚦 Interpretation

R² Score 0.5734

-Our model explains 57.3% of Ksat variance → strong performance for real-world regression

RMSLE 0.0739

-Log-scale error is very low, indicating great predictive stability



```
Intercept 1.8422593020177866
Prediction_local [-0.67263489]
Right: 0.008119485921721719
```

Shap Value for feature Importance

🧠 SHAP Summary Plot: Global Feature Importance

🔍 What it shows: The overall contribution of each interaction feature across all test samples.

Color = feature value (red = high, blue = low).

Position = impact on prediction (right = increase, left = decrease).

✅ Key Takeaways:

Sample_Height_cm is the most influential feature, appearing in many top interactions.

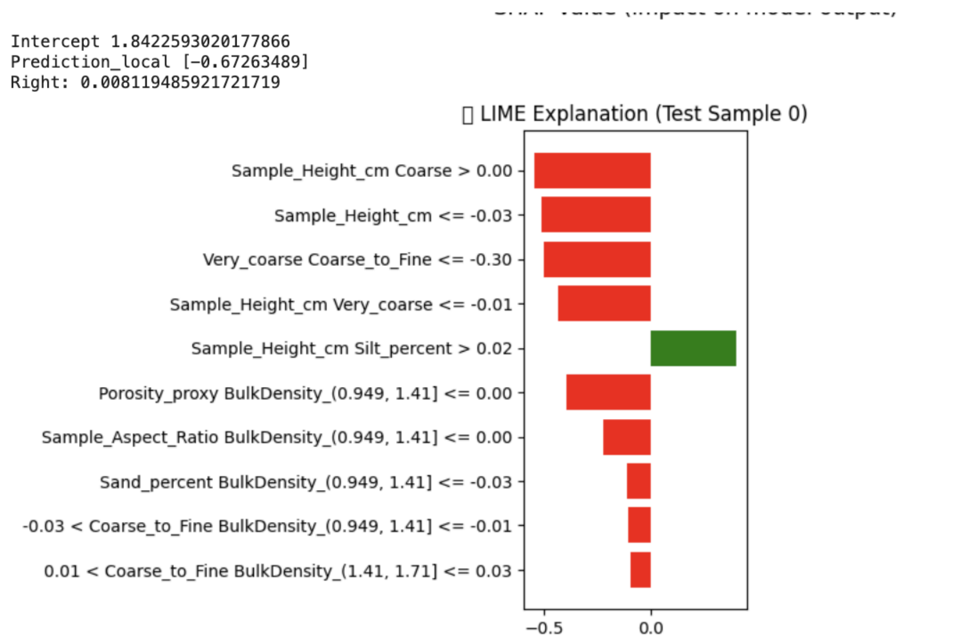
🔸 Sample_Height_cm × Silt_percent, Sample_Height_cm × Sand_percent, and Sample_Height_cm × Coarse were the top 3 interactions driving the predictions.

High values of Sand_percent or Coarse with tall sample height generally led to increased Ksat.

High Silt_percent with height had a mixed effect — likely due to its inverse relationship with permeability.

Texture classes (Texture_Sandy, Texture_Clayey) combined with height also showed clear nonlinear effects.

Engineered features like Sand_Clay_Ratio also contributed strongly, validating our feature engineering.



```
Intercept 1.8422593020177866
Prediction_local [-0.67263489]
Right: 0.008119485921721719
```

▯ LIME Explanation (Test Sample 0)
Sample_Height_cm Coarse > 0.00
Sample_Height_cm <= -0.03
Very_coarse Coarse_to_Fine <= -0.30
Sample_Height_cm Very_coarse <= -0.01
Sample_Height_cm Silt_percent > 0.02
Porosity_proxy BulkDensity_(0.949, 1.41] <= 0.00
Sample_Aspect_Ratio BulkDensity_(0.949, 1.41] <= 0.00
Sand_percent BulkDensity_(0.949, 1.41] <= -0.03
-0.03 < Coarse_to_Fine BulkDensity_(0.949, 1.41] <= -0.01
0.01 < Coarse_to_Fine BulkDensity_(1.41, 1.71] <= 0.03
−0.5
0.0

Lime Explanation

🧪 **LIME Plot: Local Explanation for Test Sample 0**

🔍 What it shows:

Top 10 features contributing to one individual prediction (Test Sample 0).

Red = decreases prediction, Green = increases prediction.

The bars show how much each feature pushed the prediction up or down.

✅ Key Takeaways: The model lowered the prediction for this sample due to:

Low Sample_Height_cm

High Coarse × Very_coarse and Very_coarse × Coarse_to_Fine combinations.

The only positive driver was moderate Silt_percent.

Porosity and texture-based bins (from our feature engineering) had minor but visible contributions.

Confirms that physical dimensions + particle interactions drive the decision locally.
