Ksat Prediction Modeling Report

1. Dataset Overview

The dataset used for Ksat prediction contains 27,483 entries and 35 columns, representing soil physical properties from lab samples collected at varying depths across different sites and soil horizons. This data is essential in environmental studies and soil science, especially for understanding how water moves through different soil types.

Target Variable: Ksat (Saturated Hydraulic Conductivity)

- Measures the rate at which water moves through soil pores.
- Expressed in cm/s or cm/hr.
- Critical for hydrological modeling, agriculture planning, and ecological sustainability.

2. Data Preprocessing and Cleaning

A thorough cleaning process was applied to standardize and prepare the dataset for modeling:

- Only the unified "Combined Data" sheet was used. All individual reference sheets were discarded.
- Manually filled missing Horizon values using Reference 1 and verified for completeness.
- Converted all measurements (depth, diameter, height) to centimeters.
- Standardized Ksat values to a consistent unit (cm/hr) and cleaned formats including scientific notation and invalid strings.
- Removed malformed or invalid entries such as "95.89.4 cm/day".

Dropped Columns
Twelve columns were removed due to excessive missing values or irrelevance, including:

arduino

CopyEdit

"Units.1", "Stdev_Bulk density", "Stdev_Sand (%)", "Stdev_Silt (%)", "Stdev_Clay (%)", "Stdev_Coarse_fragments (%)", "Textural class.1",

"Stdev_Organic carbon", "Other potentally relevant info",

"Unnamed: 33", "Unnamed: 34", "Textural class"

Additionally, Soil name and Organic matter were dropped due to high null counts.

Row Filtering

- All rows with missing values in the Ksat column (target variable) were removed.

Type Conversion and Imputation

- Converted key columns to numeric format with invalid values coerced to NaN.
- Used global medians to fill missing values in Sample Height (cm) and Sample diameter (cm).
- For other features, filled missing values using the median within each Method group, with a fallback to the global median.

Feature Encoding

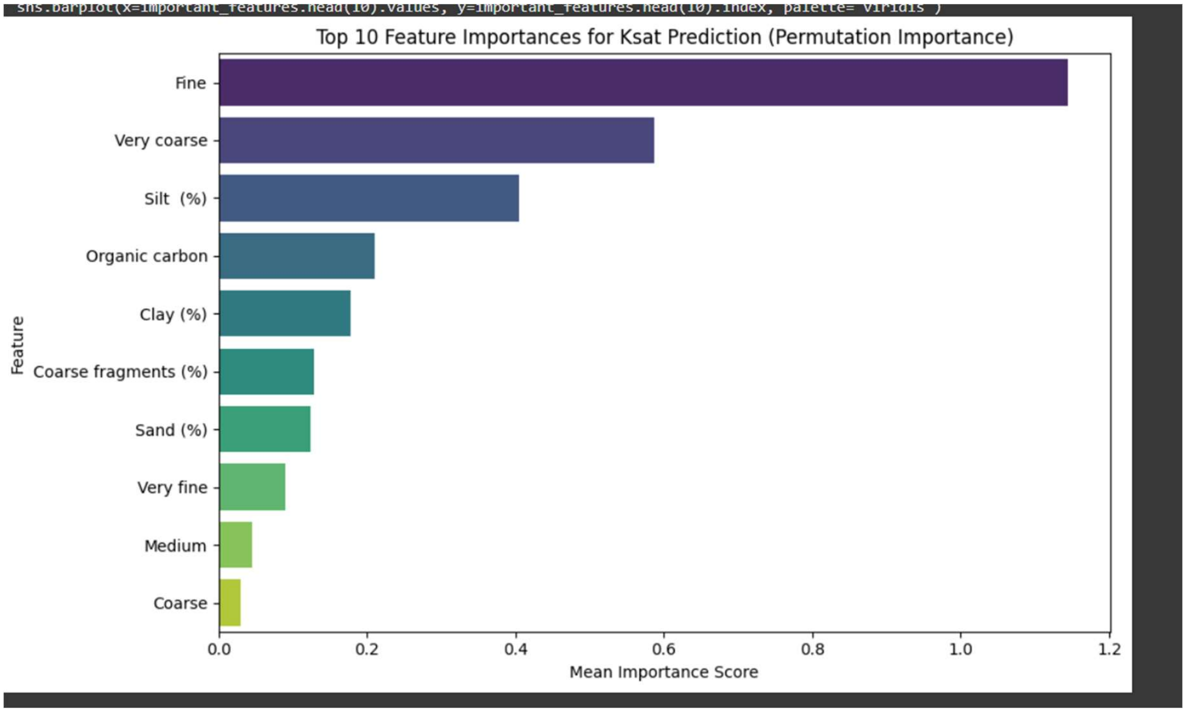
- The Method column was encoded as categorical integers.
- Dropped Source reference as it was non-informative for modeling.

The result was a clean, numeric dataset suitable for model training.

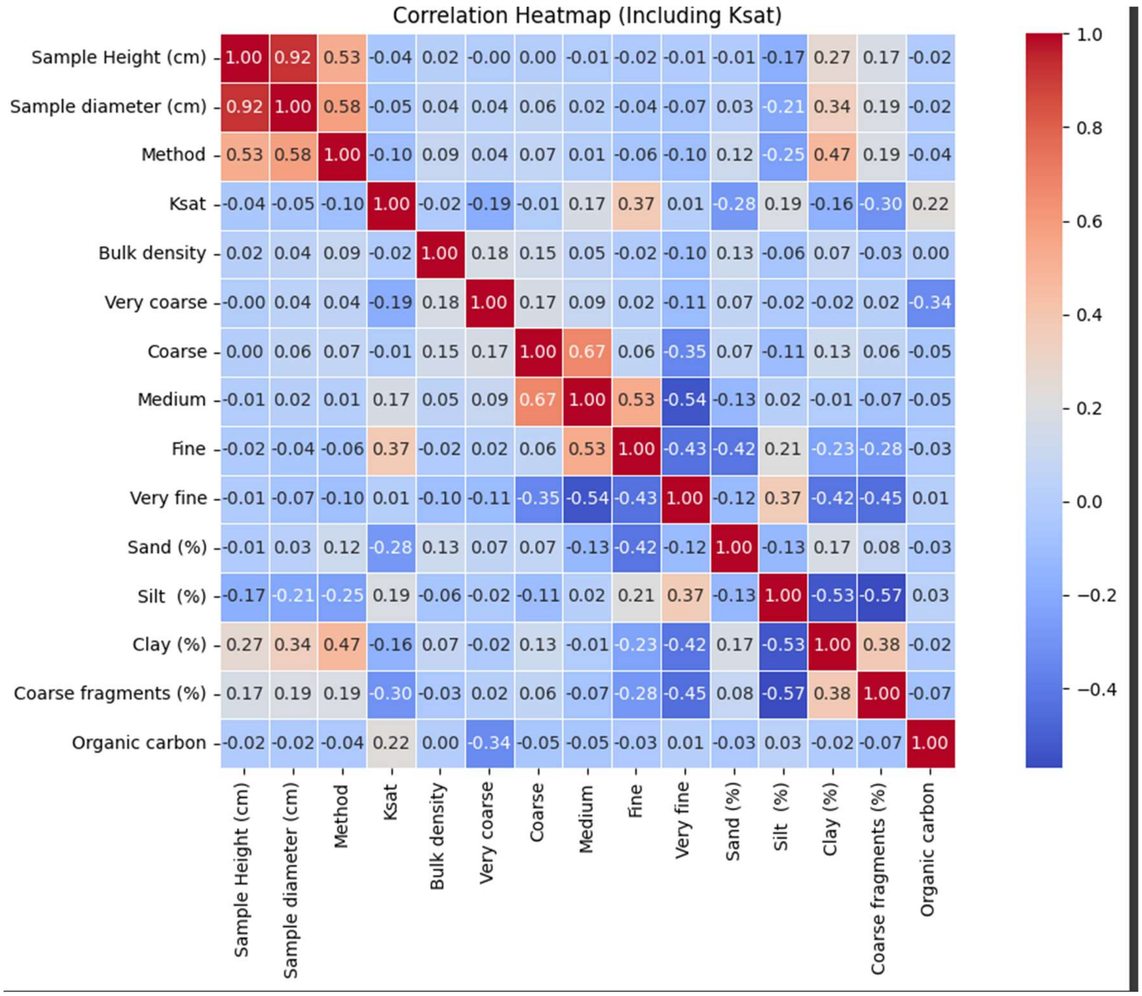
---

3. Feature Selection Methodology

- Removed non-predictive columns like Depth (cm) and Source reference.
- Employed Permutation Importance via a Random Forest model to identify the most relevant features.
- Selected the following top 10 features:
    - Sand (%), Coarse fragments (%), Method, Clay (%), Organic carbon, Bulk density, Very fine, Silt (%), Medium, Fine



```
sns.barplot(x=important_features.head(10).values, y=important_features.head(10).index, palette='viridis')
```



Correlation Heatmap (Including Ksat)

	Sample Height (cm)	Sample diameter (cm)	Method	Ksat	Bulk density	Very coarse	Coarse	Medium	Fine	Very fine	Sand (%)	Silt (%)	Clay (%)	Coarse fragments (%)	Organic carbon
Sample Height (cm)	1.00	0.92	0.53	-0.04	0.02	-0.00	0.00	-0.01	-0.02	-0.01	-0.01	-0.17	0.27	0.17	-0.02
Sample diameter (cm)	0.92	1.00	0.58	-0.05	0.04	0.04	0.06	0.02	-0.04	-0.07	0.03	-0.21	0.34	0.19	-0.02
Method	0.53	0.58	1.00	-0.10	0.09	0.04	0.07	0.01	-0.06	-0.10	0.12	-0.25	0.47	0.19	-0.04
Ksat	-0.04	-0.05	-0.10	1.00	-0.02	-0.19	-0.01	0.17	0.37	0.01	-0.28	0.19	-0.16	-0.30	0.22
Bulk density	0.02	0.04	0.09	-0.02	1.00	0.18	0.15	0.05	-0.02	-0.10	0.13	-0.06	0.07	-0.03	0.00
Very coarse	-0.00	0.04	0.04	-0.19	0.18	1.00	0.17	0.09	0.02	-0.11	0.07	-0.02	-0.02	0.02	-0.34
Coarse	0.00	0.06	0.07	-0.01	0.15	0.17	1.00	0.67	0.06	-0.35	0.07	-0.11	0.13	0.06	-0.05
Medium	-0.01	0.02	0.01	0.17	0.05	0.09	0.67	1.00	0.53	-0.54	-0.13	0.02	-0.01	-0.07	-0.05
Fine	-0.02	-0.04	-0.06	0.37	-0.02	0.02	0.06	0.53	1.00	-0.43	-0.42	0.21	-0.23	-0.28	-0.03
Very fine	-0.01	-0.07	-0.10	0.01	-0.10	-0.11	-0.35	-0.54	-0.43	1.00	-0.12	0.37	-0.42	-0.45	0.01
Sand (%)	-0.01	0.03	0.12	-0.28	0.13	0.07	0.07	-0.13	-0.42	-0.12	1.00	-0.13	0.17	0.08	-0.03
Silt (%)	-0.17	-0.21	-0.25	0.19	-0.06	-0.02	-0.11	0.02	0.21	0.37	-0.13	1.00	-0.53	-0.57	0.03
Clay (%)	0.27	0.34	0.47	-0.16	0.07	-0.02	0.13	-0.01	-0.23	-0.42	0.17	-0.53	1.00	0.38	-0.02
Coarse fragments (%)	0.17	0.19	0.19	-0.30	-0.03	0.02	0.06	-0.07	-0.28	-0.45	0.08	-0.57	0.38	1.00	-0.07
Organic carbon	-0.02	-0.02	-0.04	0.22	0.00	-0.34	-0.05	-0.05	-0.03	0.01	-0.03	0.03	-0.02	-0.07	1.00

4. Model Selection and Training

The following machine learning models were trained and evaluated:

1. Random Forest Regressor (baseline and tuned)
2. LightGBM Regressor
3. CatBoost Regressor
4. Histogram-Based Gradient Boosting (HGB)

All models used an 80/20 train-test split and were trained on the top 10 selected features.

5. Hyperparameter Tuning Strategy

For Random Forest, RandomizedSearchCV was used with the following parameter space:

python

CopyEdit

```
'n_estimators': [100, 200, 300]

'max_depth': [None, 10, 20]

'min_samples_split': [2, 5, 10]
```

5-fold cross-validation was applied using RMSLE as the scoring metric.

6. Subset Experiment Details

To assess robustness, the entire modeling process was repeated across subsets of decreasing size (from 20,000+ to 2,000 records, reduced in steps of 2,000). For each subset:

- 10 random samples were drawn.
- Each sample was split into train/test sets.
- Models were retrained and evaluated on each run.

Metrics used:

- RMSLE (Root Mean Squared Logarithmic Error)

- R^2 (Coefficient of Determination)

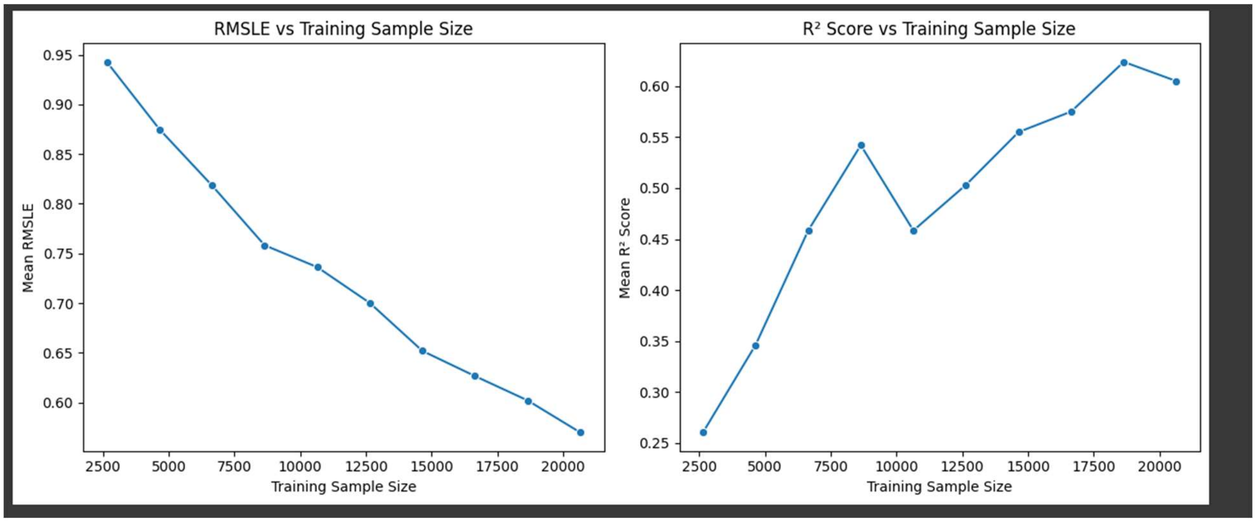
Models evaluated:

- Random Forest
- LightGBM
- CatBoost
- HGB

---

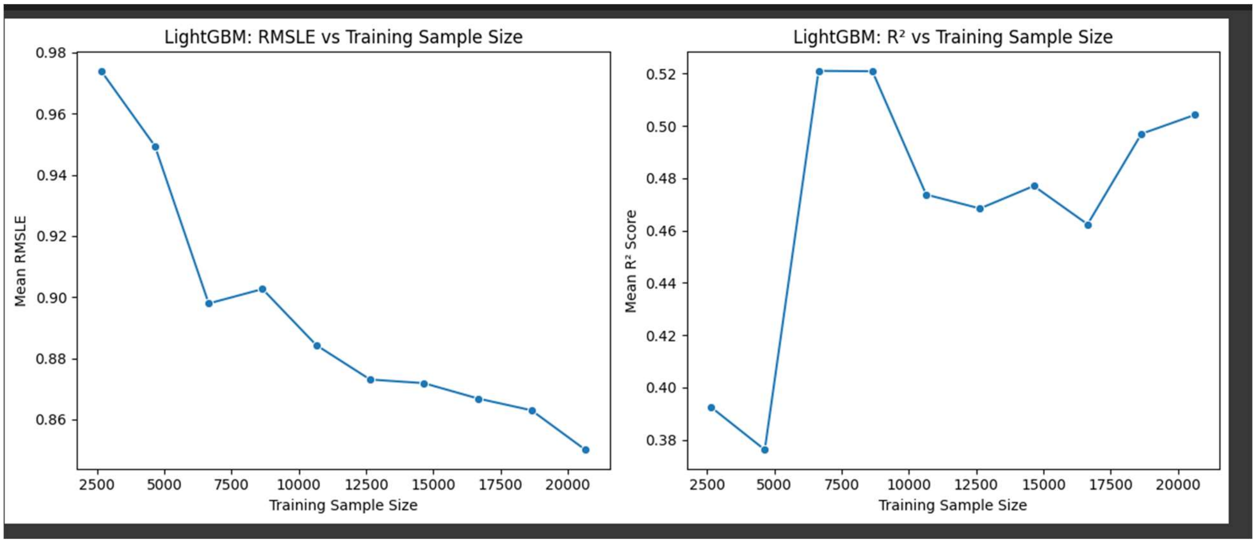
7. Analysis and Interpretation of Results

Random Forest Cross-validation:

- RMSLE: 0.6451 ± 0.0252
- R^2: 0.5789 ± 0.0894
- 

LightGBM Cross-validation:

- RMSLE: 0.8773 ± 0.0073
- R^2: 0.4550 ± 0.0583

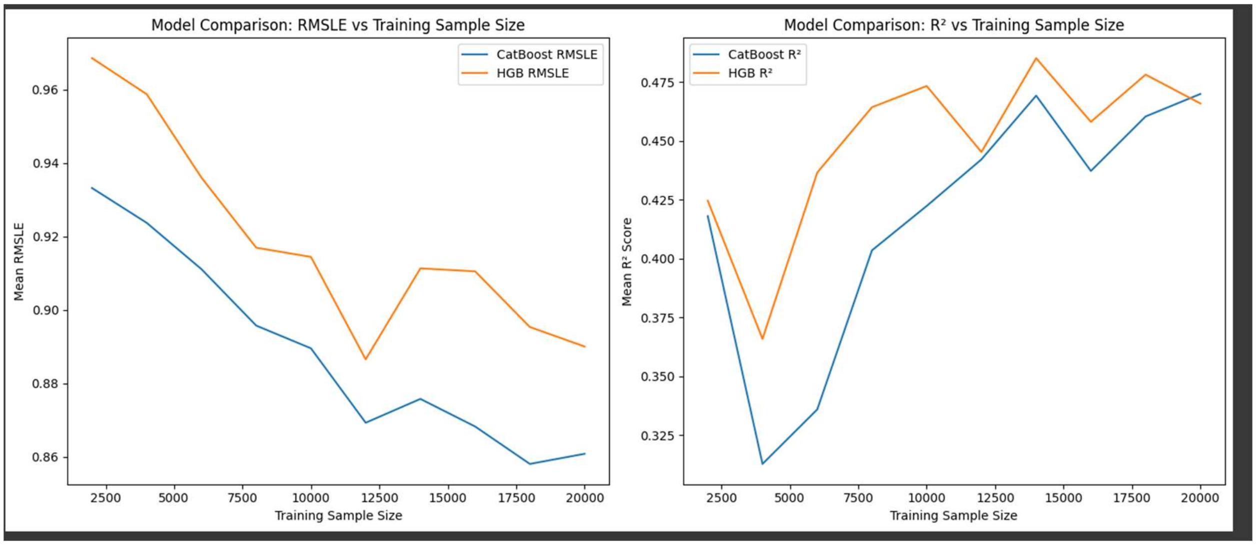


CatBoost Evaluation:

- RMSLE: 0.6468
- R^2: 0.5796

HGB Evaluation:

- RMSLE: 0.6661
- R^2: 0.5642



Interpretation:

- CatBoost delivered the best generalization with the lowest RMSLE and highest R^2 across all tests.

- All models showed consistent improvement in performance with increasing sample size.

---

8. Visualizations Summary

1. **RMSLE vs Training Sample Size**:
    - Model errors decreased as sample size increased.
2. **R^2 Score vs Training Sample Size**:
    - Prediction accuracy improved with more data.
3. **CatBoost vs HGB Comparison**:
    - CatBoost outperformed HGB consistently across all subset sizes.

Conclusion

This report documents a comprehensive pipeline for predicting Ksat using soil physical properties. After intensive data cleaning, robust feature selection, and model experimentation, CatBoost emerged as the most accurate and reliable model, especially in limited data settings. The results support the use of ensemble learning techniques for hydrological prediction tasks in soil science.