Covertype Classification

Presented By Group C

Our Team

Khaing Hsu Wai

Khaing Nyo

Khin Yadanar Aung

Our Mentor - Ko Thant Zin Bo

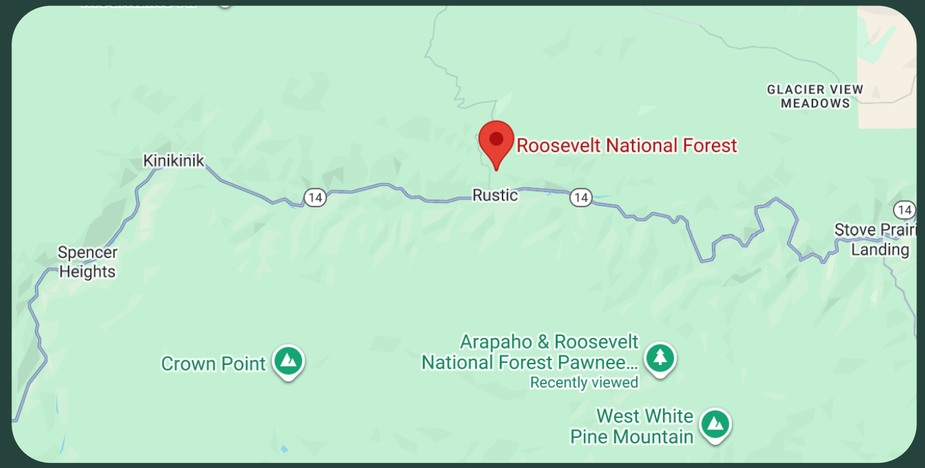
Project Introduction

The actual forest cover type was determined from **US Forest Service (USFS) Region 2 Resource Information System (RIS) data.**

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

By studying forest data from places like the USA can help us understand and learned pattern from other countries to predict Myanmar's forest types and can support further studies.

Forest cover_type Classification



Project Introduction

In this project, we will be predicting 3 forest covertype,

3 - Ponderosa_Pine

6 - Douglas_fir

7 - Krummholz

Ponderosa_Pine | Douglas_fir | Krummholz

Dataset Description

Dataset Information

Source - UCI Machine Learning Repository

Link - Covertype

Donated on 31 July 1998

Dataset Statistic

Dataset - **581,012**
Feature - **54**
Target Variable - **7Covertype**

Dataset - **73,631**
Feature - **54**
Target Variable - **3Covertype (3,6,7)**

Feature Description

Numerical : Elevation, Aspect, Slope, Horizontal/Vertical Distance to Hydrology, Hillshade indices, Horizontal Distance to Roadways/Fire Points

One Hot encoded(0,1) : Wilderness Areas(4 area: Northern, Northwest , Central, Estern), Soil Type (40 Types)

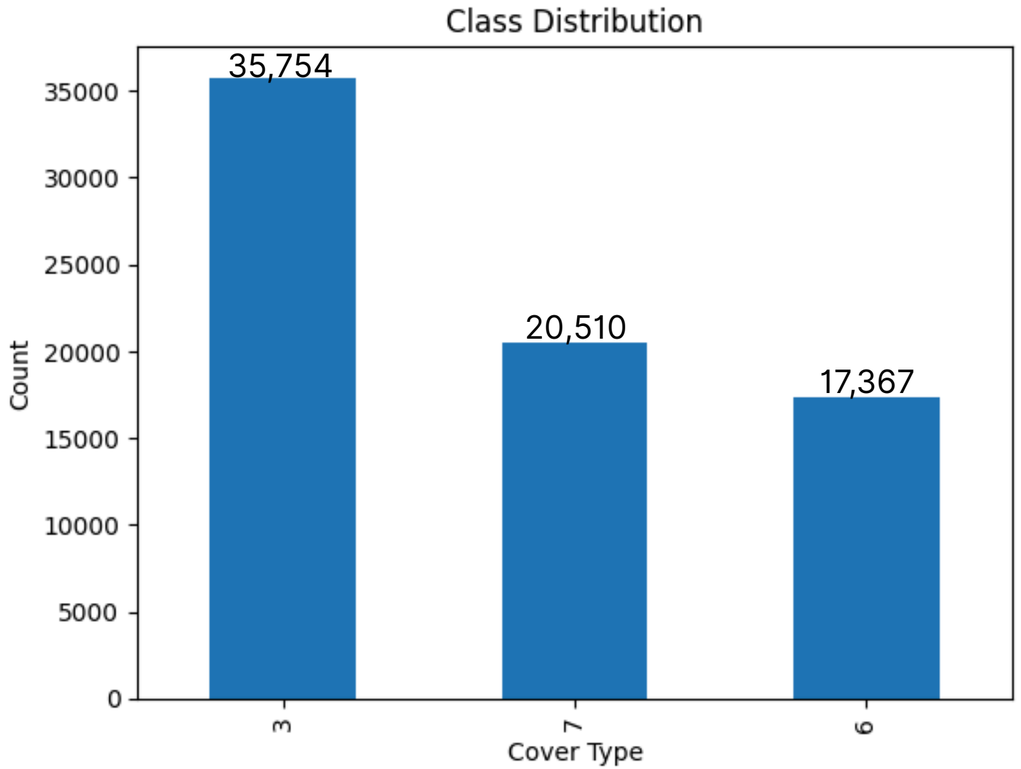
Target: Forest Cover Type (1-7)

Among 7 cover type, the three covertype that have the least data is selected in the data in order to predict which covertype forest is in a certain given feature.

Next

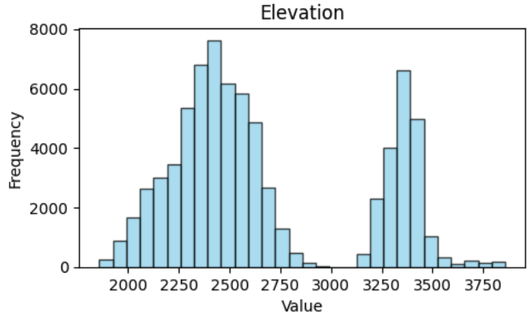
Data Preprocessing

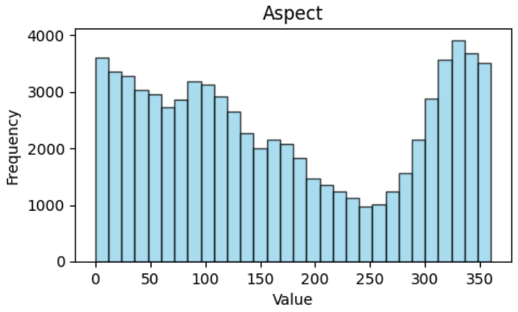
Class	Covertype
2	283,301
1	211,840
3	**35,754**
7	**20,510**
6	**17,367**
5	9,493
4	2,747
Total	581,012

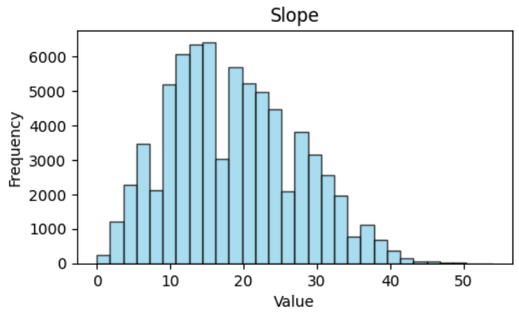


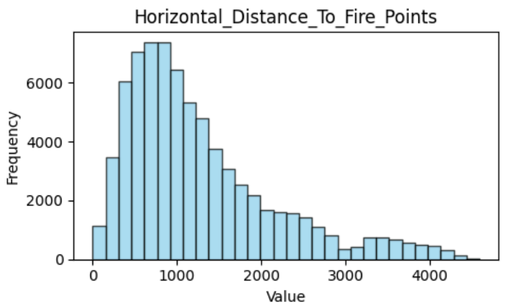
The dataset contains 581,012 samples, but the distribution of forest cover types is highly imbalanced. Classes 1 and 2 (Spruce/Fir and Lodgepole Pine) dominate the dataset, while Classes 4 and 5 (Cottonwood/Willow and Aspen) are severely underrepresented. To ensure model fairness and reduce bias, predicting will focus on the Classes 3, 6, and 7 (Ponderosa Pine, Douglas-fir, and Krummholz).

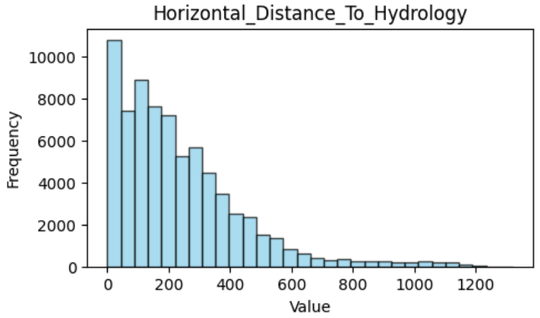
Dataset Analysis

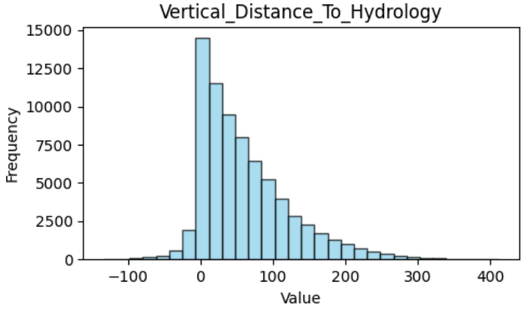


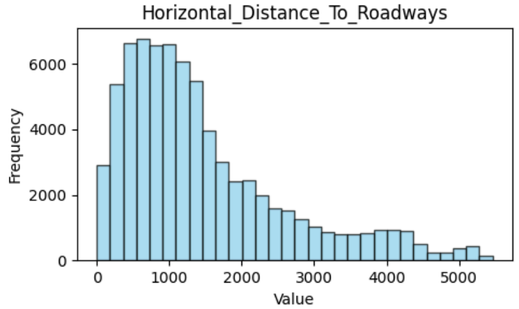


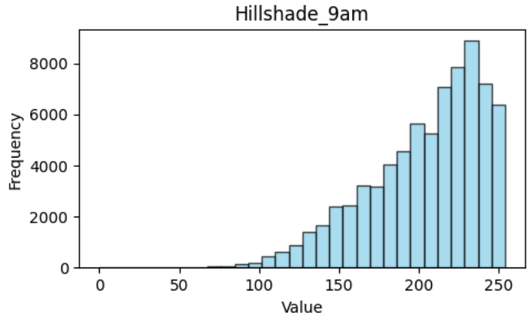


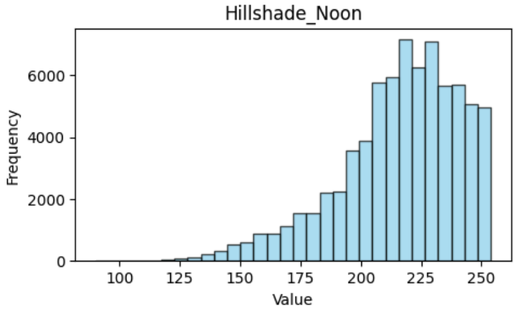


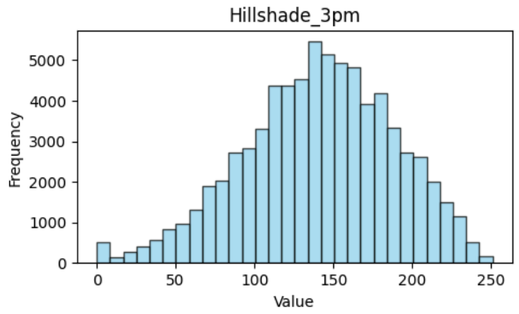




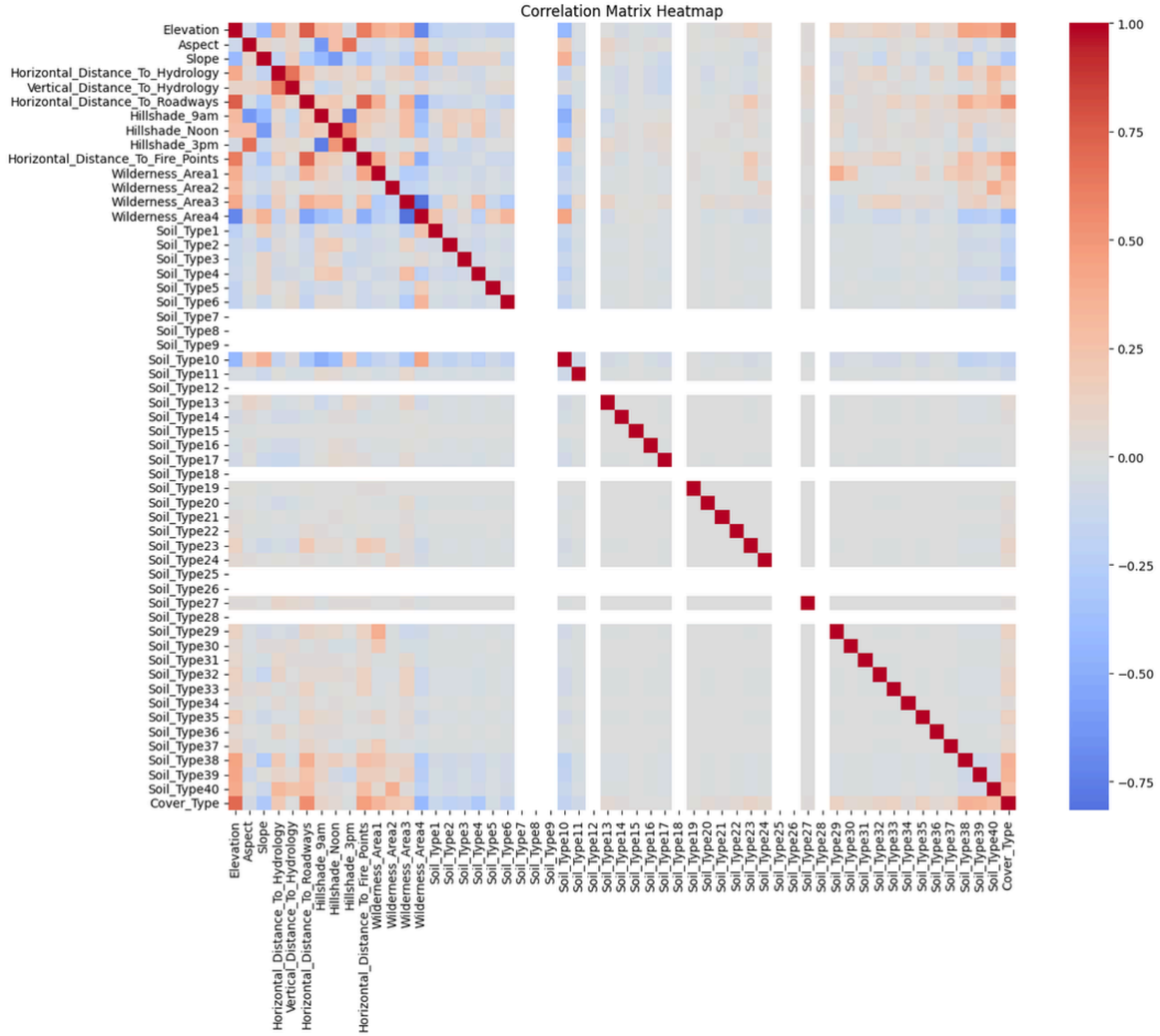


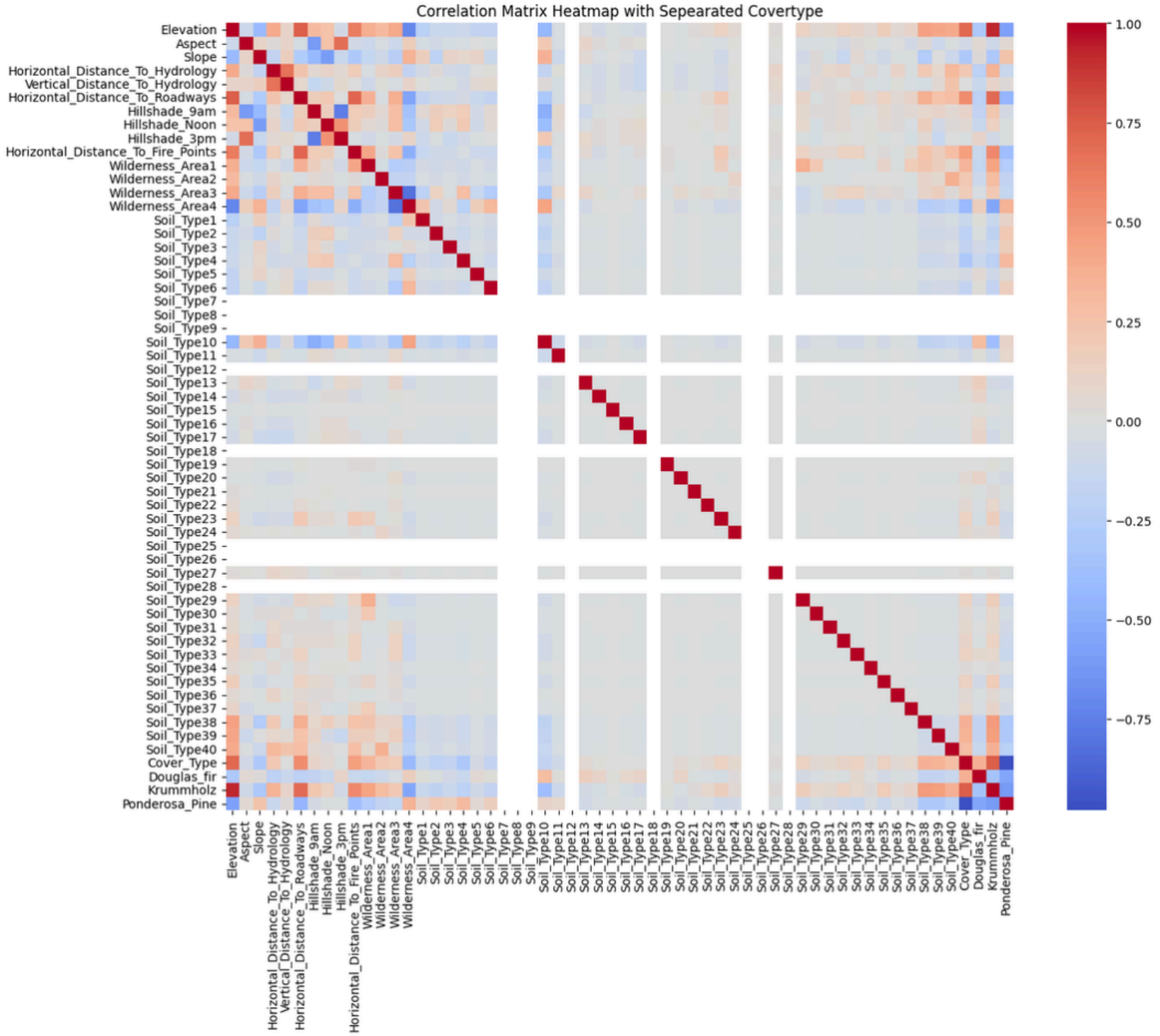






Data Preprocessing





Model Selection

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Support Vector Machine

Library

Pandas, Numpy, Seaborn, Matplotlib, Scikit-lear, Scipy

Data Preparation

Pandas and Numpy

Exploratory Analysis

seaborn + matplotlib

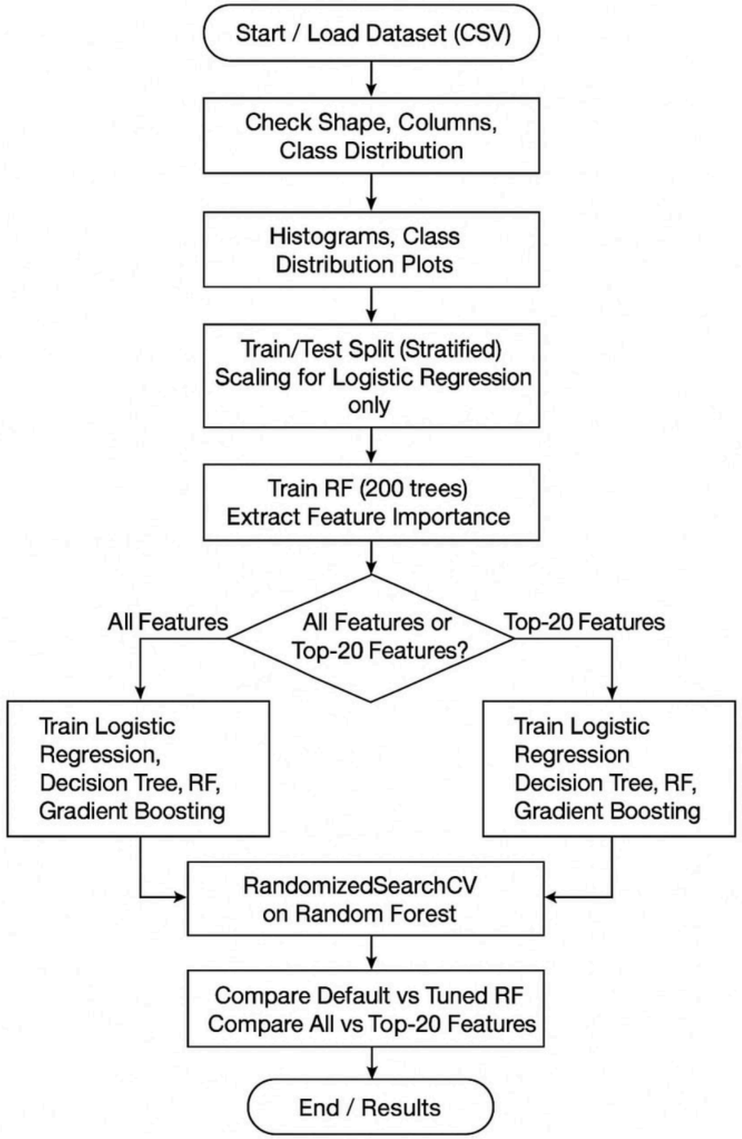
Model Building, Statistical Validation

scikit-learn models, scipy.stats

Hyperparameter Tuning

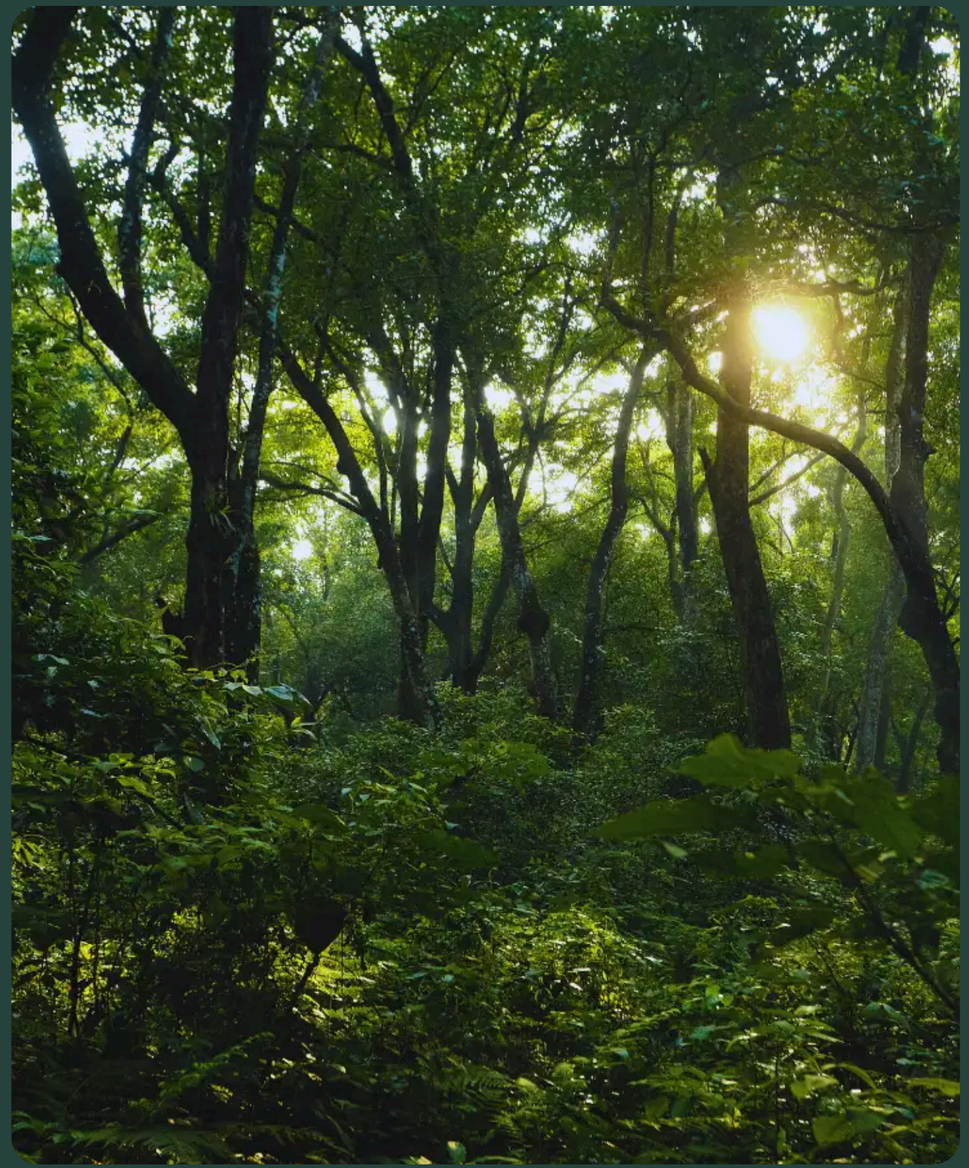
GridSearchCV

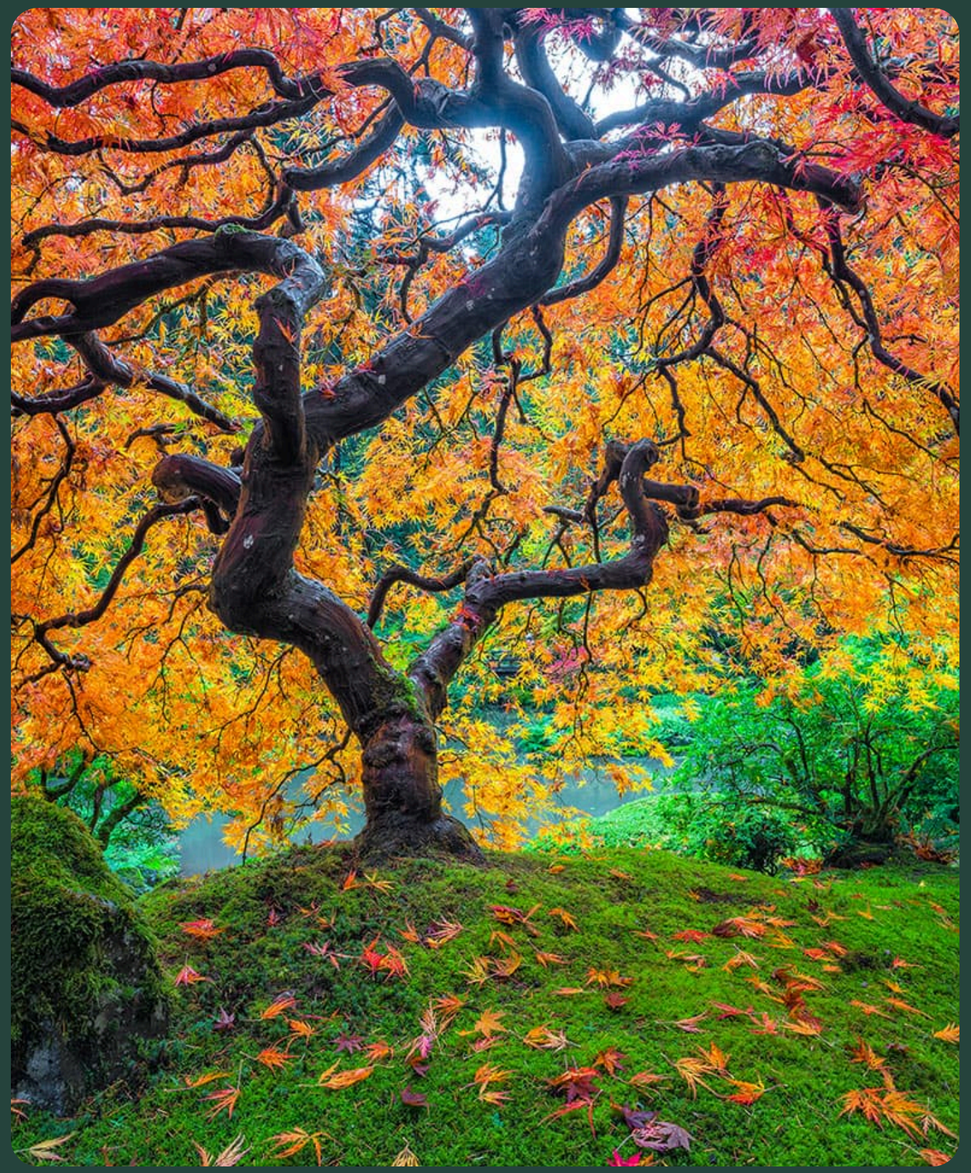
Implementation Flow Chart



Implementation Process

- Split Data (80% Train, 20% Test)
- Feature Selection (Top-20 features)
- Preprocessing (applied StandardScaler)
- Model Training
- Evaluation





Evaluation Metric

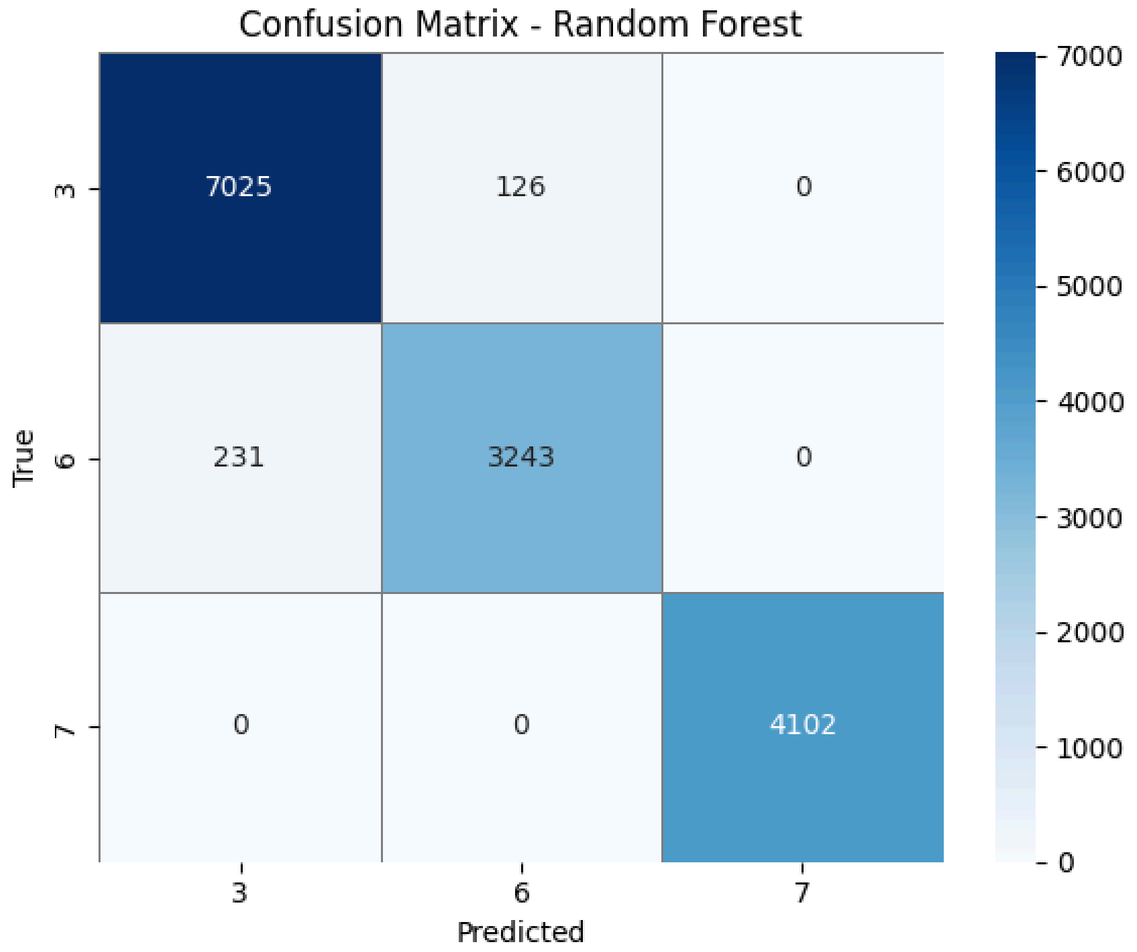
F1 score, Accuracy, Precision, Recall

As all of the three class is balanced important for this covertype prediction, model is selected mainly based on the accuracy and F1 score while the metric F1 score, accuracy, precision, recall are focus on selected model tuning and evaluation.

Evaluation

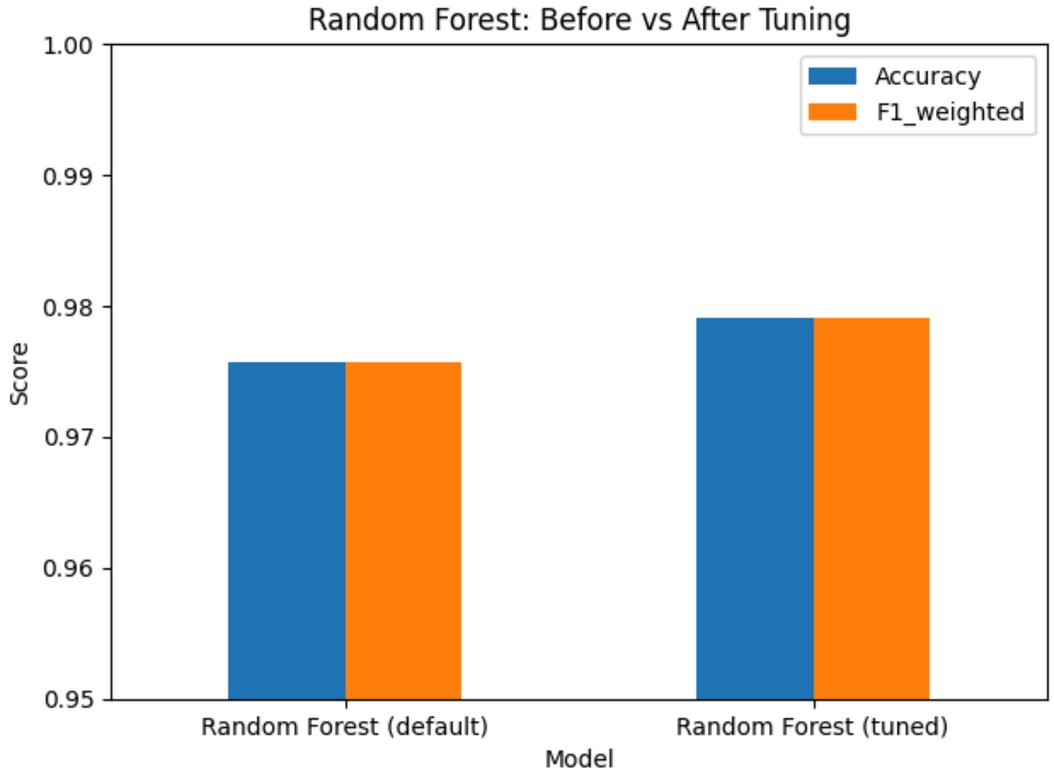
Model	Accuracy	Weighted F1
Random Forest	**0.976**	**0.976**
Decision Tree	0.957	0.957
Gradient Boosting	0.873	0.869
Logistic Regression	0.815	0.805
Linear SVM	0.813	0.798

```
=== Classification Report (Random Forest) ===
              precision    recall  f1-score   support

           3     0.9682    0.9824    0.9752      7151
           6     0.9626    0.9335    0.9478      3474
           7     1.0000    1.0000    1.0000      4102

    accuracy                         0.9758     14727
   macro avg     0.9769    0.9720    0.9744     14727
weighted avg     0.9757    0.9758    0.9757     14727
```



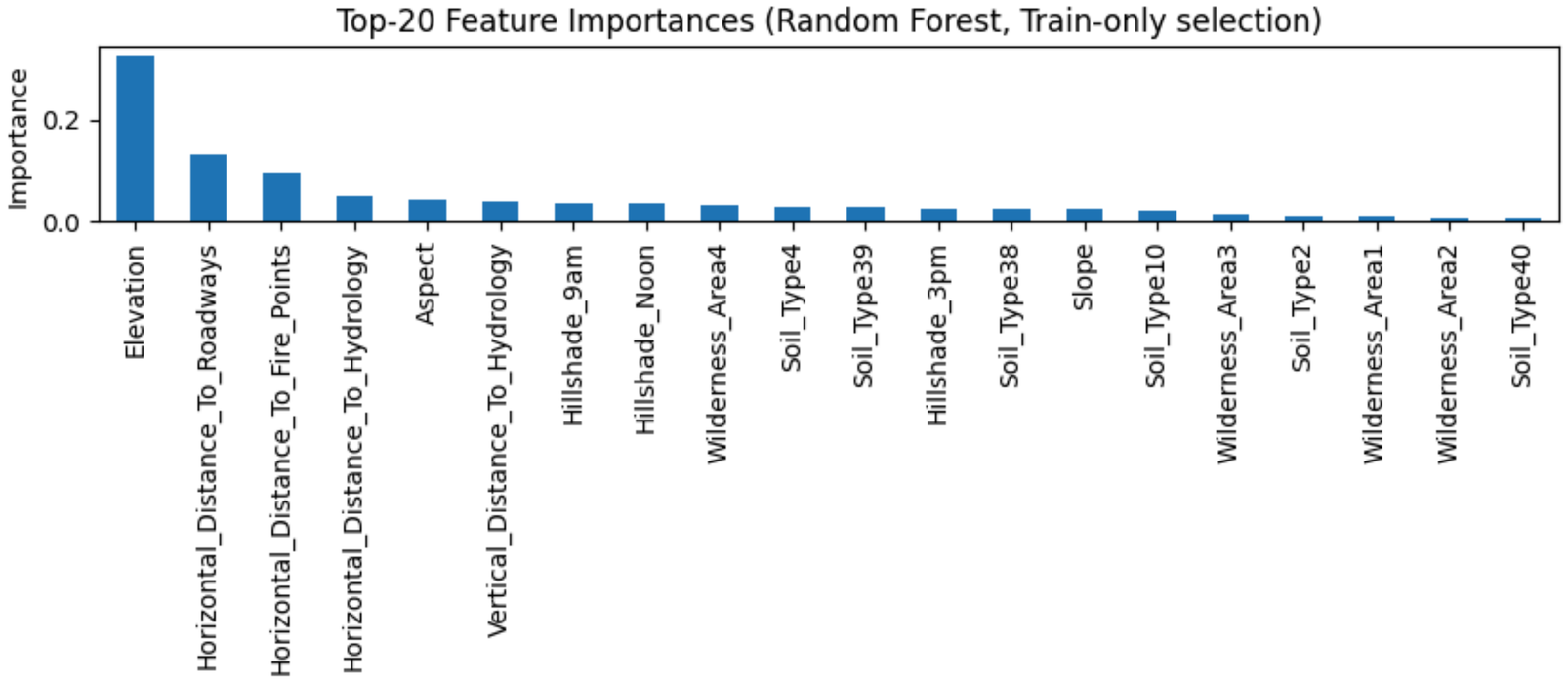
Dataset Preprocessing (Parameter Tuning)

```
=== Classification Report (Random Forest - Tuned) ===
              precision    recall  f1-score   support

           3     0.9745    0.9828    0.9786      7151
           6     0.9640    0.9470    0.9554      3474
           7     1.0000    1.0000    1.0000      4102

    accuracy                         0.9792     14727
   macro avg     0.9795    0.9766    0.9780     14727
weighted avg     0.9791    0.9792    0.9791     14727
```

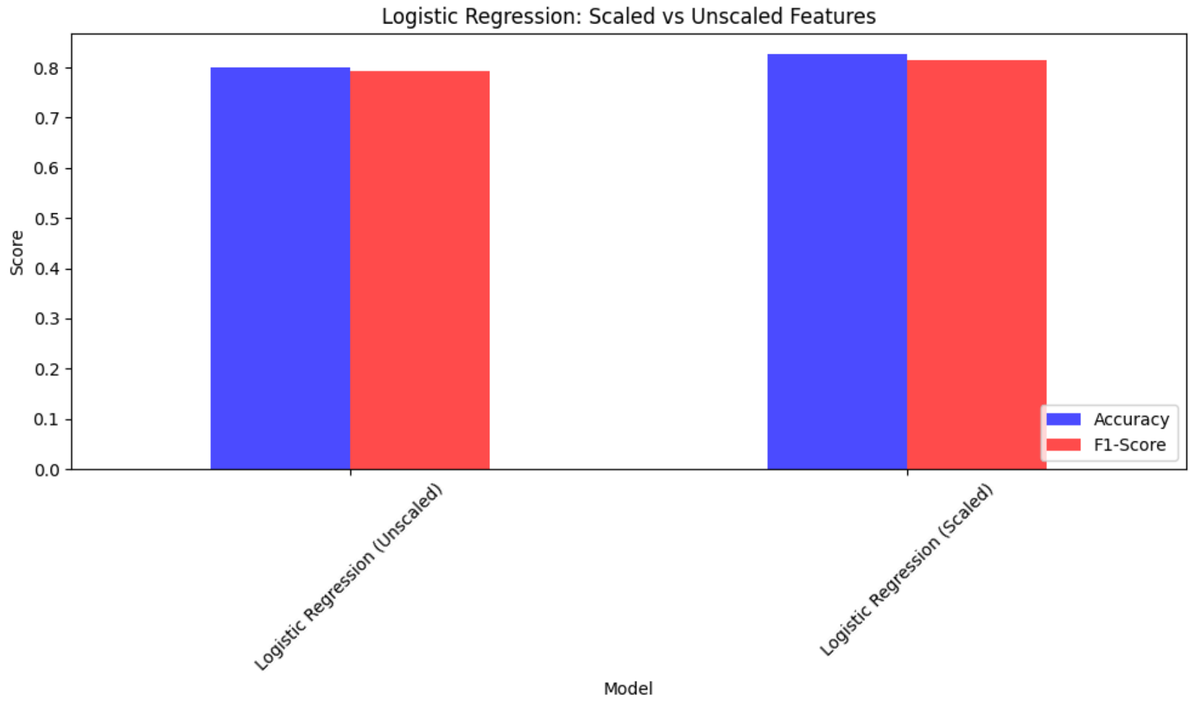



```
Fitting 3 folds for each of 30 candidates, totalling 90 fits
Best parameters: {'max_depth': 24, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 249}
Accuracy (tuned RF): 0.9791539349494126
F1-weighted (tuned RF): 0.9791058822292433
```

Results and Findings (Top 20 Features)

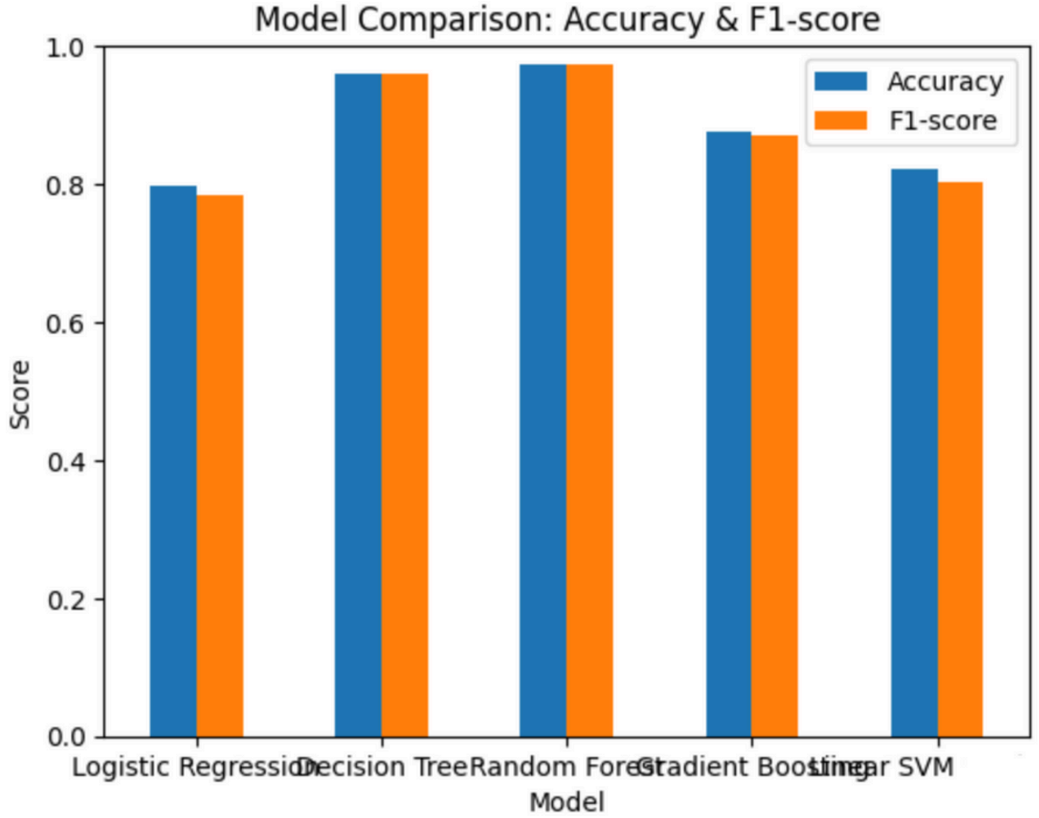


Results and Findings (Model)

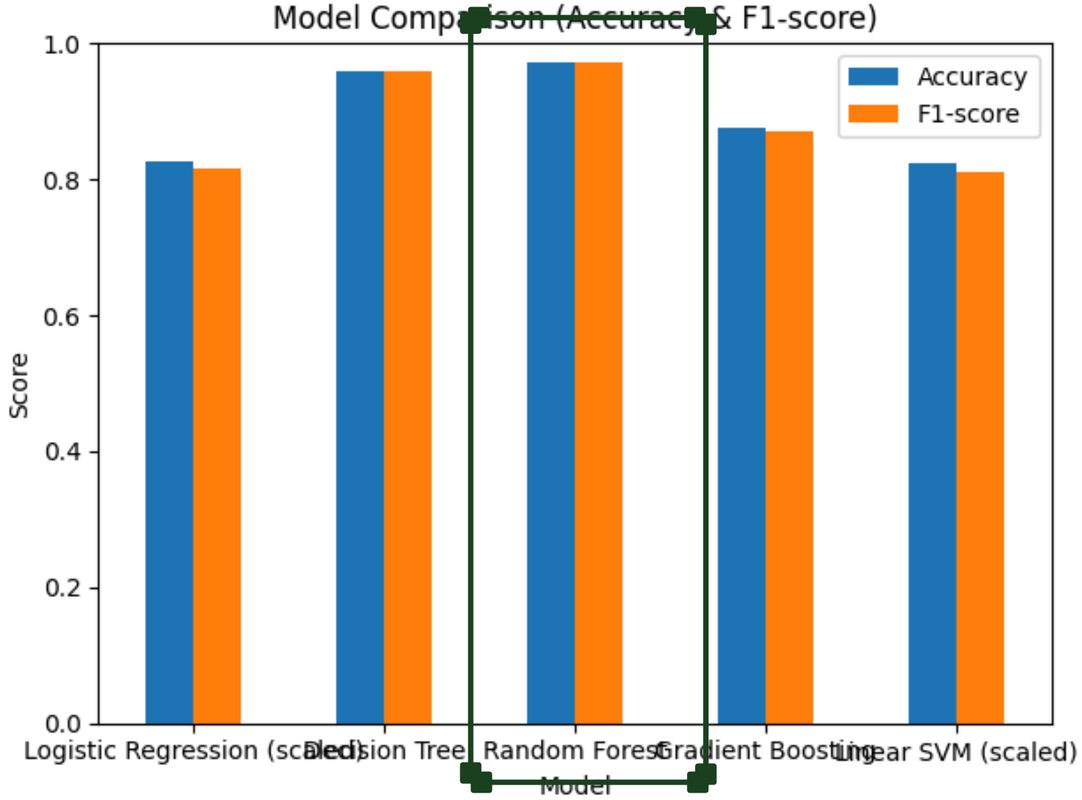


Model	Accuracy	F1-Score
Logistic Regression	0.799688	0.791836
Logistic Regression (Scaled)	0.825762	0.815497

Results and Findings (Model)

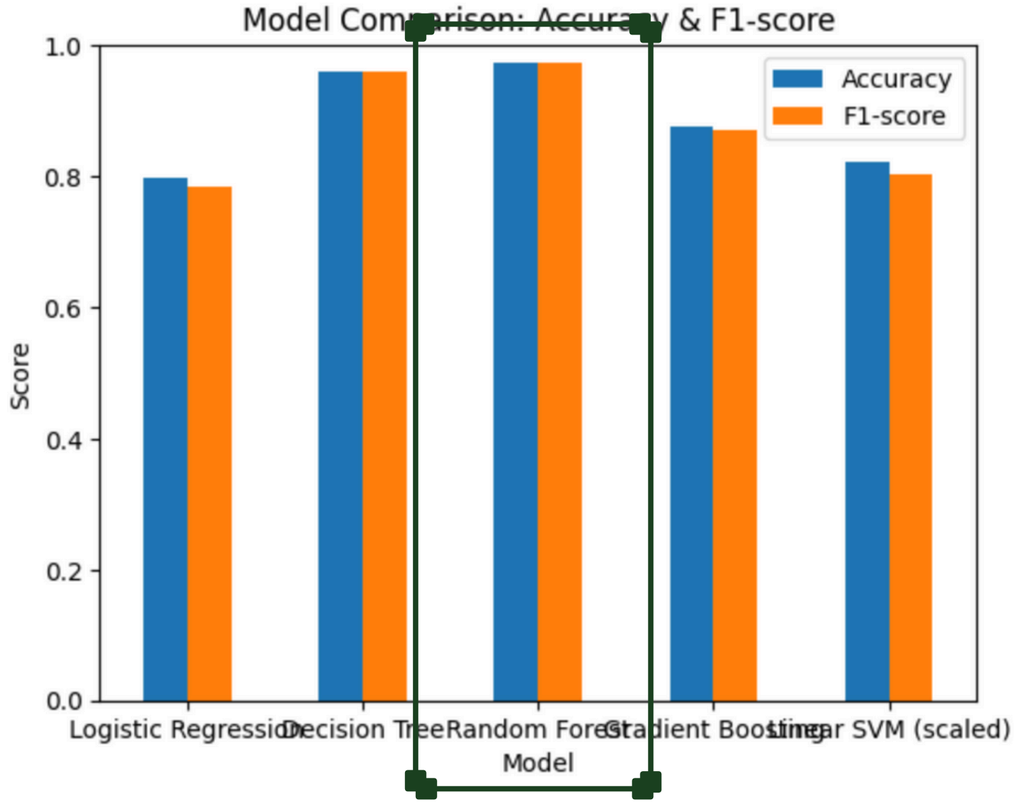


Model	Accuracy	F1-Score
Logistic Regression	0.7997	0.7918
Decision Tree	0.9599	0.9598
Random Forest	0.9726	0.9725
Gradient Boosting	0.8748	0.8702
Linear SVM (scaled)	0.8217	0.8030

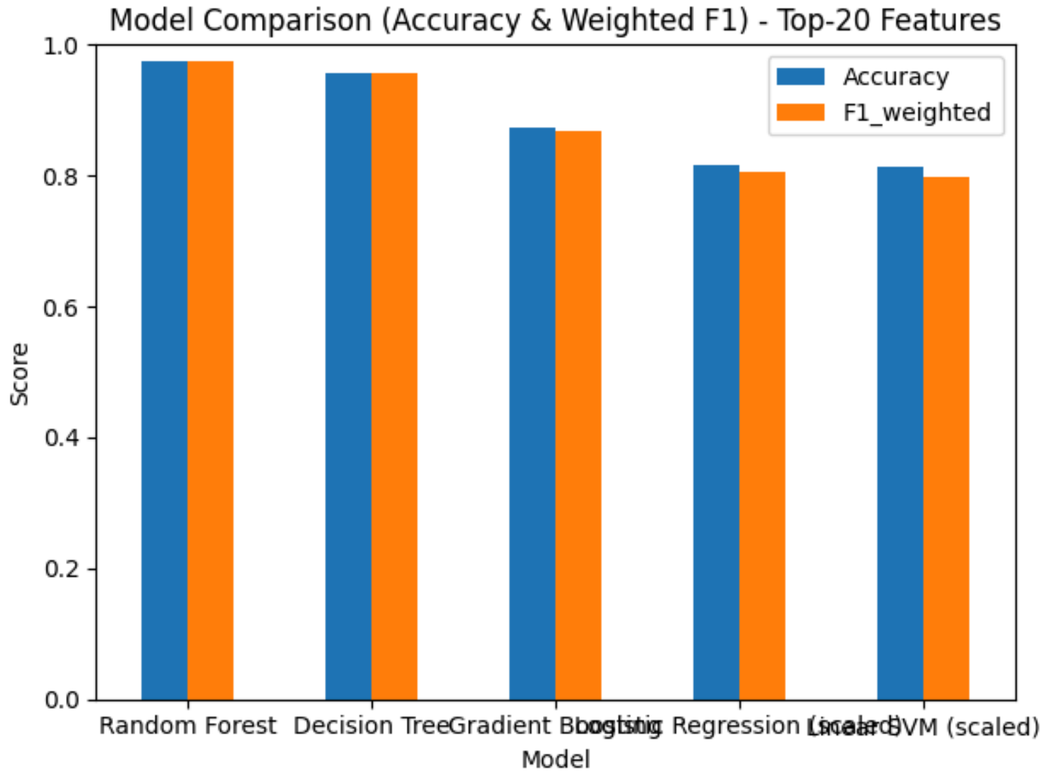


Model	Accuracy	F1-Score
Logistic Regression (scaled)	0.8258	0.8155
Decision Tree	0.9599	0.9598
Random Forest	0.9726	0.9725
Gradient Boosting	0.8748	0.8702
Linear SVM (scaled)	0.8247	0.8113

Results and Findings (Model)



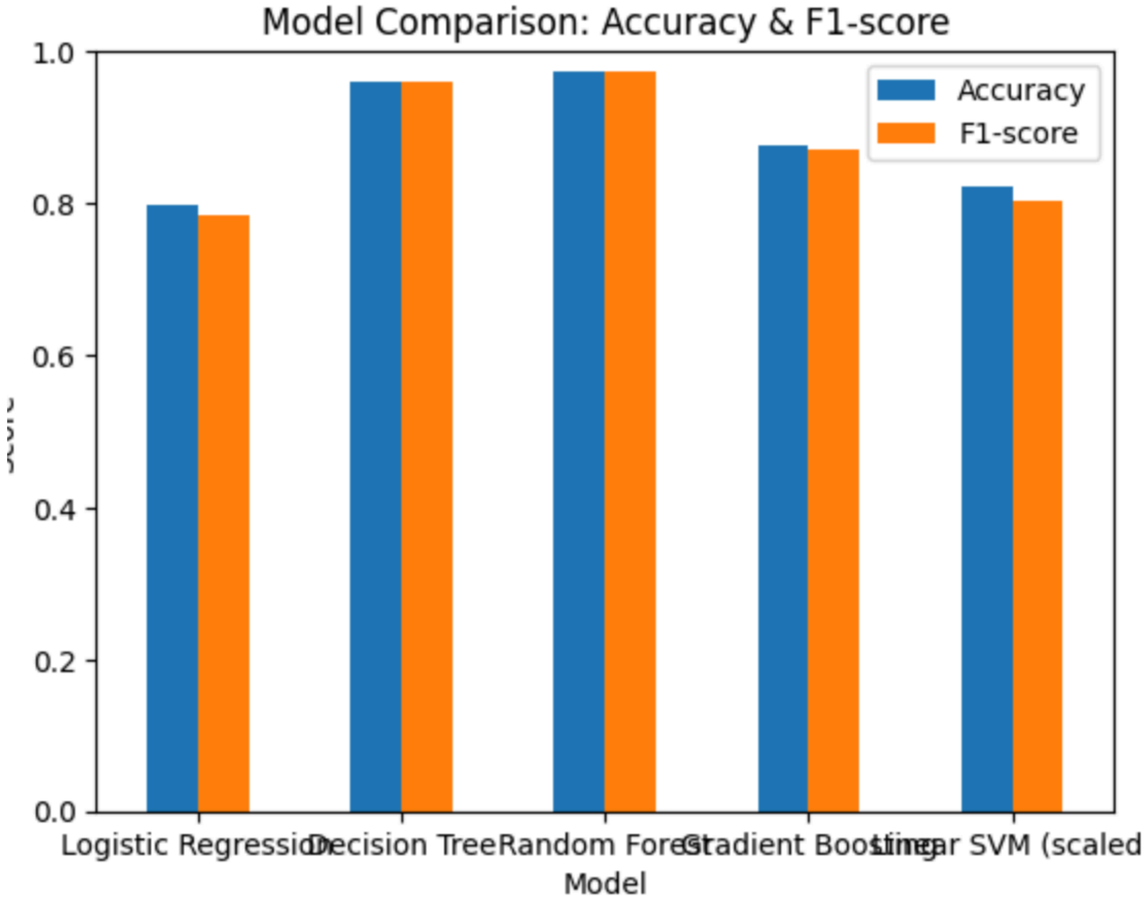
Model	Accuracy	F1-Score
Logistic Regression (scaled)	0.8258	0.8155
Decision Tree	0.9599	0.9598
Random Forest	0.9726	0.9725
Gradient Boosting	0.8748	0.8702
Linear SVM (scaled)	0.8247	0.8113



Model	Accuracy	F1-Score
Logistic Regression (scaled)	0.8153	0.8050
Decision Tree	0.9572	0.9572
Random Forest	0.9758	0.9757
Gradient Boosting	0.8727	0.8687
Linear SVM (scaled)	0.8247	0.8113

Final Evaluation Result

With The comparisons of the Top-20 feature, it increase very small point in the the accuracy and F1 scores. So, the overall feature selection with Random forest will be the final selected model with below metric.



Conclusion and Learning

Conclusion : Tree Base model have higher performance in this covertype dataset. In soil type, although some soil type are not correlated, some soil type and wilderness are correlated to the covertype.

Learning : Knowledge related to the Forest and it’s usage, data reading of 3 Classification model, tried log transform test,

Further Study

Cover type - 7 covertype will be include and tested. Will also try Smote and log transform.

Other Studies - By using the baseline of how the other forest data is predicted, further research related to the forest in Myanmar using the geographical information from the official release data from government or satellite image.


Thank You
3 Covertype Classification Report