Team Agni_astra - Fire-Ready Forests Data Challenge

Guruprasad Parasnis , Abhay Lal, Yash Vishe
National Data Platform : Forest Fire Challenge

Question 1 : Can you describe your full pipeline? Which data did you use? What kind of data preprocessing tasks did you perform? Can you visualize this pipeline through a flowchart?

The task involved three subdivisions, the first of them being the prediction of the Plant Functional Type (PFT), from the features gained by merging together two datasets : the REF_SPECIES and the CA_TREE datasets. The process of the approach we undertook is as follows :

1. Exploring the Target Variable (PFT)

We first examine the **distribution of the PFT** categories in the dataset by counting how many trees fall under each category. This helps to understand if the dataset is balanced or skewed towards certain types of trees.

2. Selecting and Checking Key Features

0.0.1 Distribution of Tree Diameter (DIA_cm) - Figure 1

- Histogram with KDE overlay showing tree diameter in centimeters.
- Positively skewed distribution with a peak near 20 cm.
- Long tail suggests presence of mature, older trees.

0.0.2 Distribution of Tree Height (HT) - Figure 2

- Distribution is right-skewed with a modal height between 30–50 ft.
- Long tail indicates a few extremely tall trees.

0.0.3 Distribution of Basal Area (BasalA) - Figure 3

- Highly skewed distribution with most trees under 5 sq ft of basal area.
- **Implications:** BasalA is a function of DIA and adds redundancy. Hence, we remove it to avoid multicollinearity.

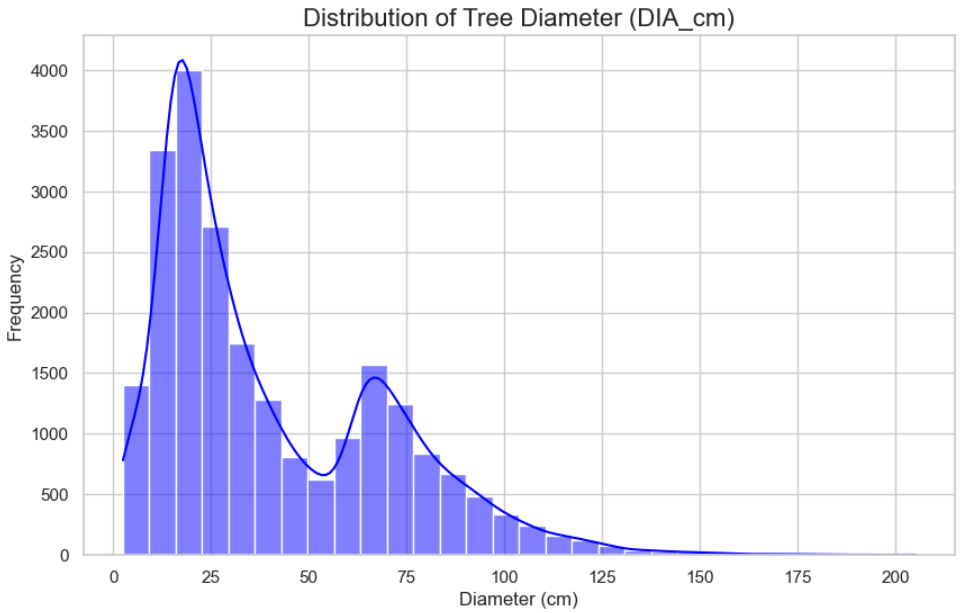

Figure 1: Distribution of Tree Diameter in FIA

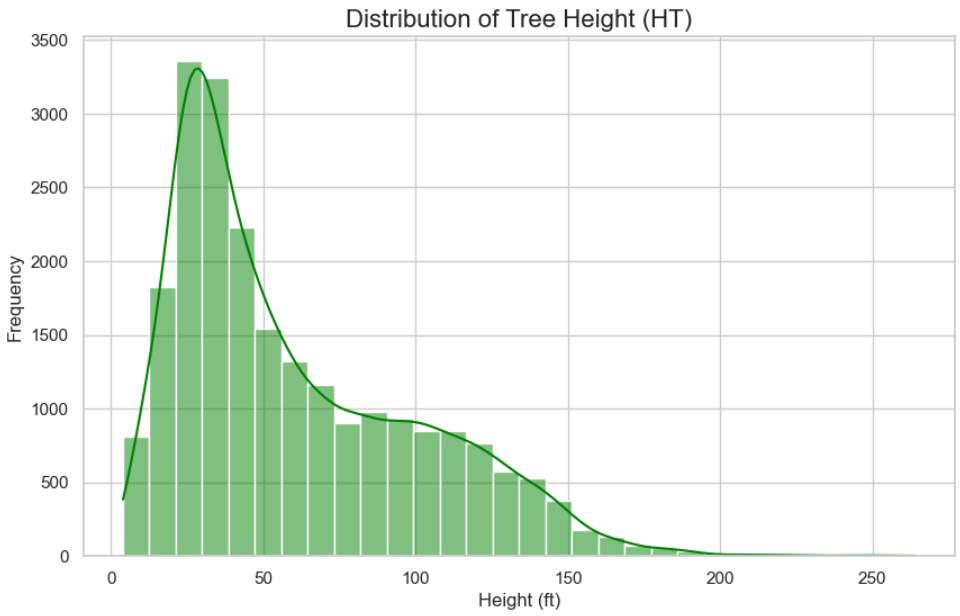

Figure 2: Distribution of Tree Height in FI

0.1 Categorical Distributions

0.1.1 Count of Trees by Plant Functional Type (PFT) - Figure 4

- Majority class: **Evergreen conifer**.
- Minority classes: **Evergreen broadleaf** and **Deciduous broadleaf**.

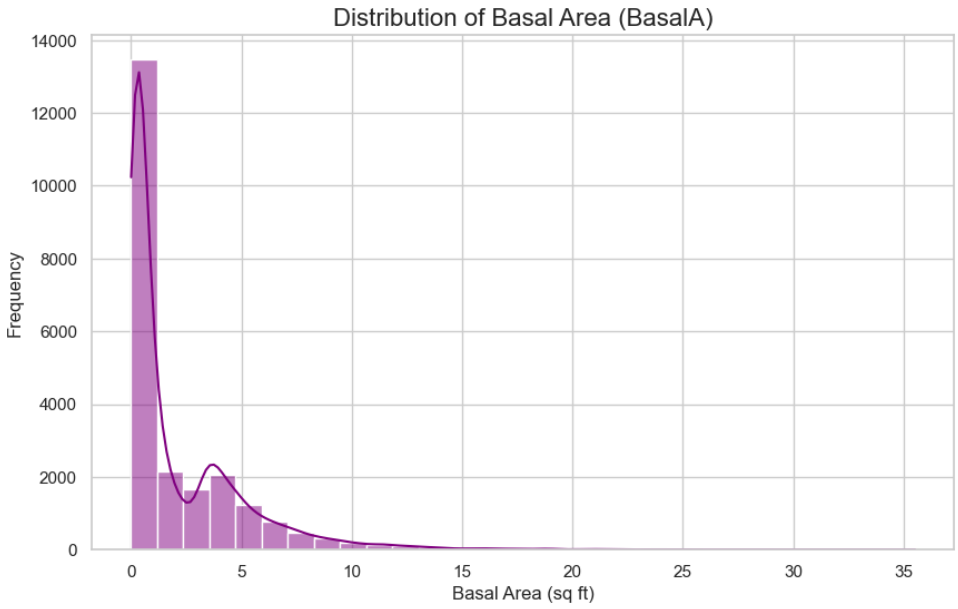

Figure 3: Distribution of Basal Area (BasalA) in FIA

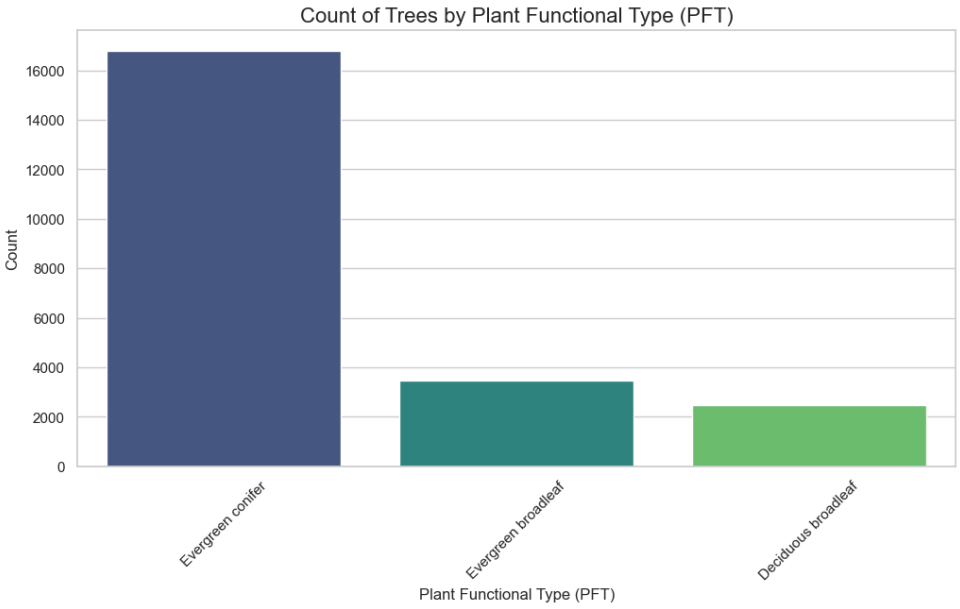

Figure 4: Tree Count by PFT in FIA

0.1.2 Top 10 Genera by Tree Count - Figure 5

- Dominated by *Pinus*, *Quercus*, *Calocedrus*, and *Abies*.
- Long-tail distribution: many genera have very few occurrences.
- **Implications:** Limited Genus Classification since many genera have limited occurences.

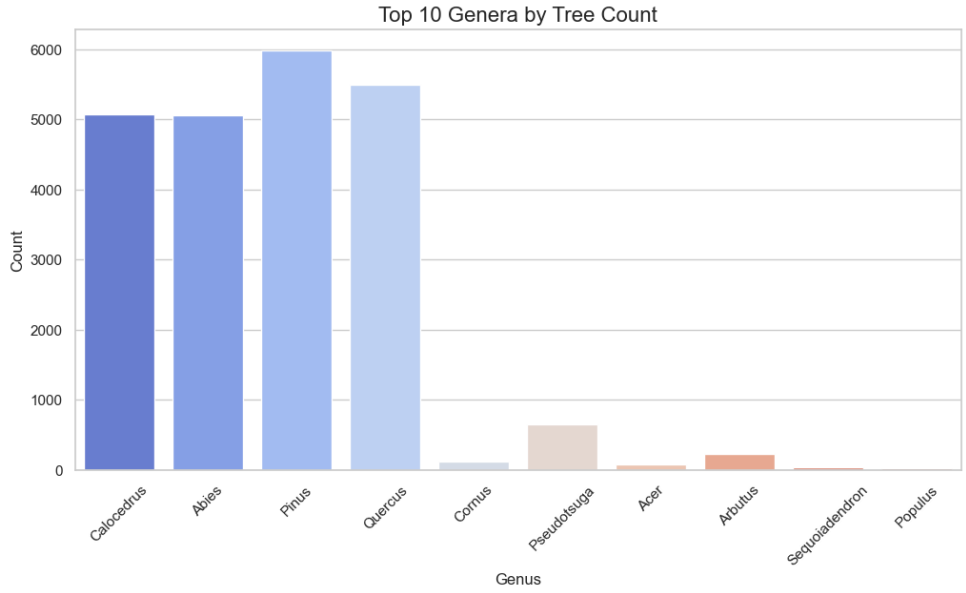

Figure 5: Count of top 10 genera by tree count in FIA.

0.2 Bivariate and Comparative Plots

0.2.1 Tree Diameter vs Height by PFT - Figure 6

- Strong positive relationship between DIA_cm and HT.
- Clusters are distinguishable by PFT; evergreen conifers tend to be larger and taller.
- **Implications:** Structural traits are good predictors for PFT classification.

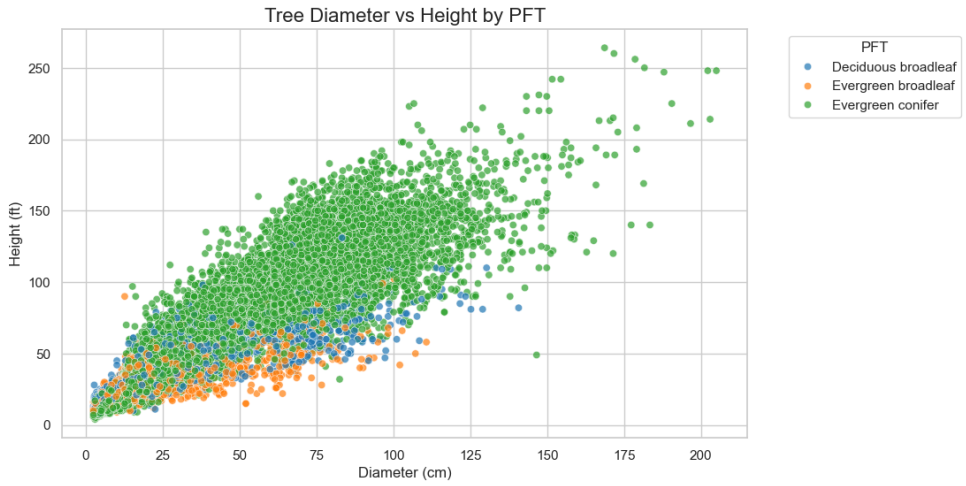

Figure 6: Tree Diameter vs Height by PFT

0.2.2 Boxplot of Tree Diameter by PFT - Figure 7

- Evergreen conifer shows the widest spread and highest median diameter.

- Evergreen broadleaf has narrower diameter range.
- **Implications:** Distinct statistical characteristics exist across PFTs; will prove useful for supervised learning tasks like tree-based models.

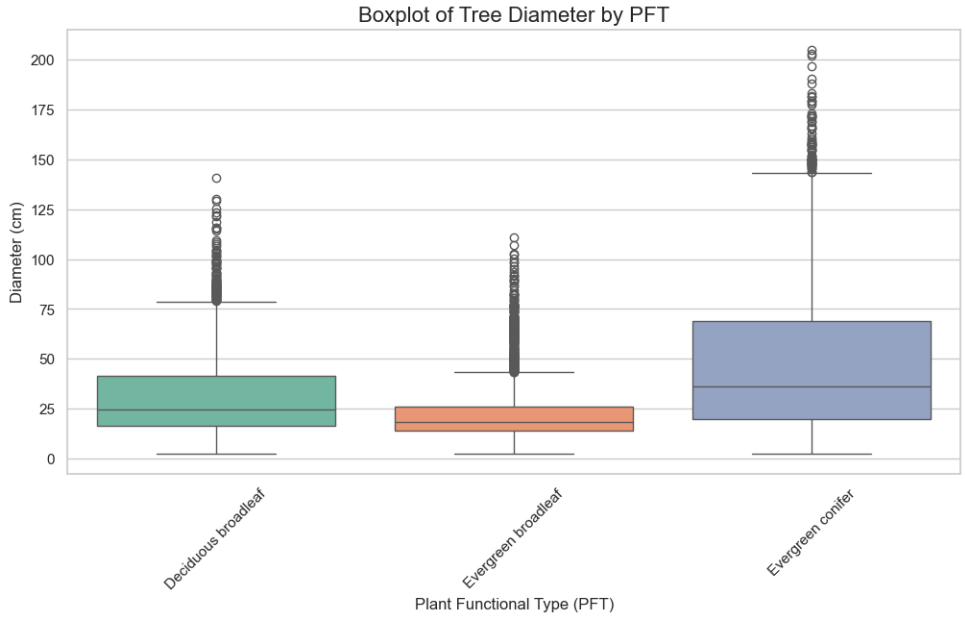

Figure 7: Boxplot of Tree Diameter by PFT

0.3 Correlation Analysis

0.3.1 Correlation Matrix of Numerical Features - Figure 8

- DIA_cm and BasalA: **0.95** (very high, due to mathematical relationship).
- DIA_cm and HT: **0.90**, HT and BasalA: **0.83**.
- **Implications:** Strong multicollinearity exists. Hence, we drop the BasalA feature.

0.4 Summary of Key Insights

Feature	Key Insight	Actionable Note
DIA_cm & HT	Right-skewed	Normalization done
BasalA	Derived from DIA	Dropped to reduce redundancy
PFT	Imbalanced classes	Applied SMOTE but no significant improvement
GENUS	Long-tail distribution	Took Top N classes
Correlation	High among structural traits	Extracted the important features from correlation values
Scatter/Boxplots	Clear separation by PFT	Used for classification models

We check for **missing values** as seen in Figure 9, to identify any incomplete records that might negatively impact the analysis.

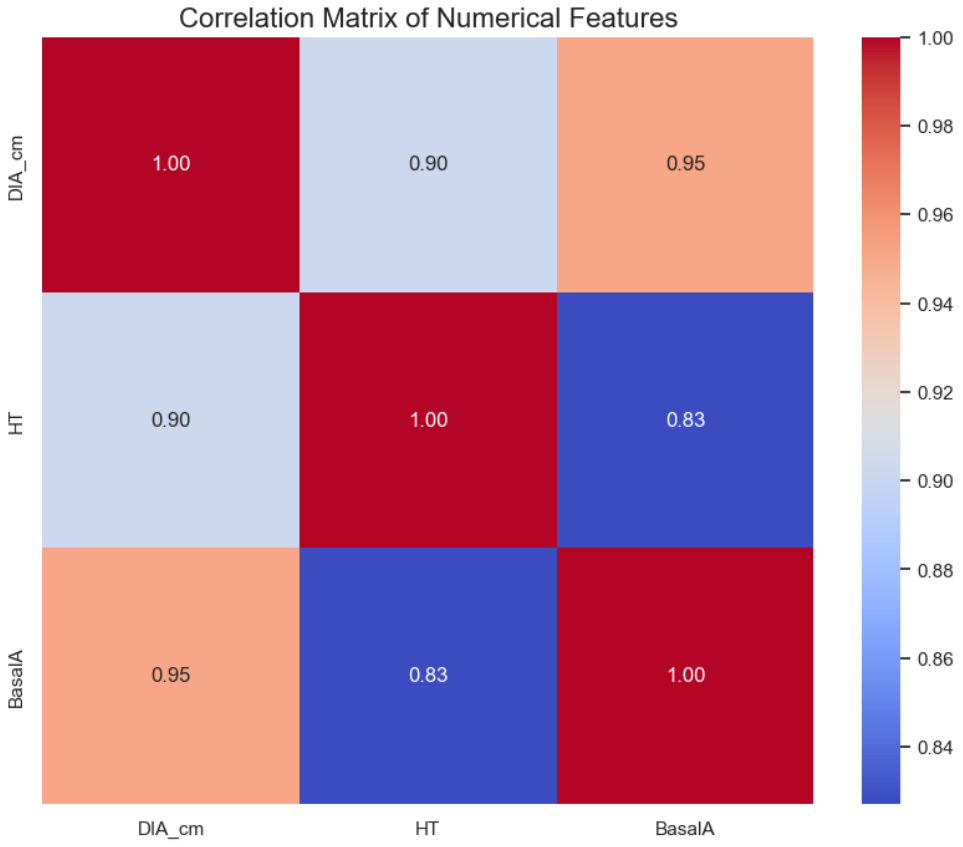

Figure 8: Correlation Matrix of Numerical Features

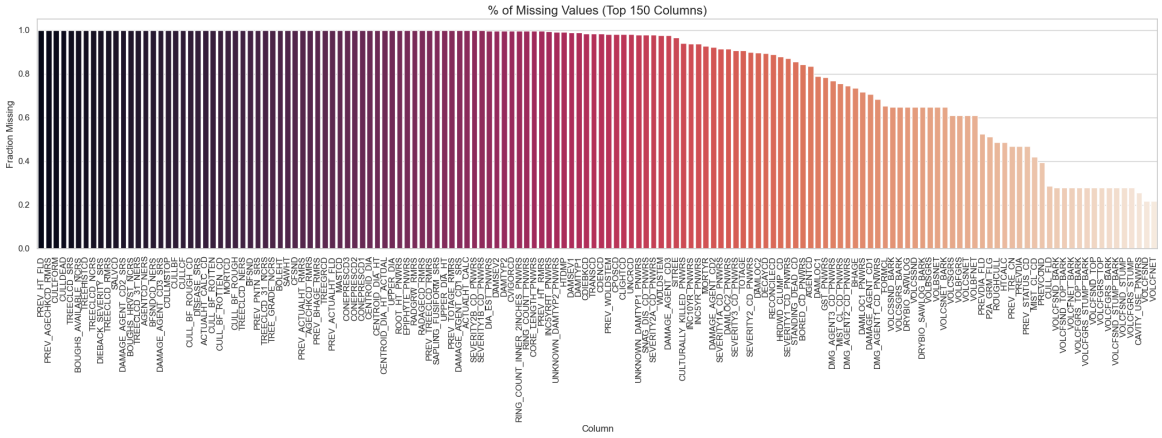

Figure 9: Missing values in FIA

As we saw, the PFT dataset contains more than 150 columns with significant null values, due to which their inclusion in the model training phase would negatively impact the performance since they cannot be good features to learn from.

0.5 Tree Height vs Diameter by Genus

The following scatter plot visualizes the structural relationship between tree height and diameter, grouped by Genus. Each point represents a single tree, with the x-axis denoting diameter (in inches) and the y-axis representing height (in feet).

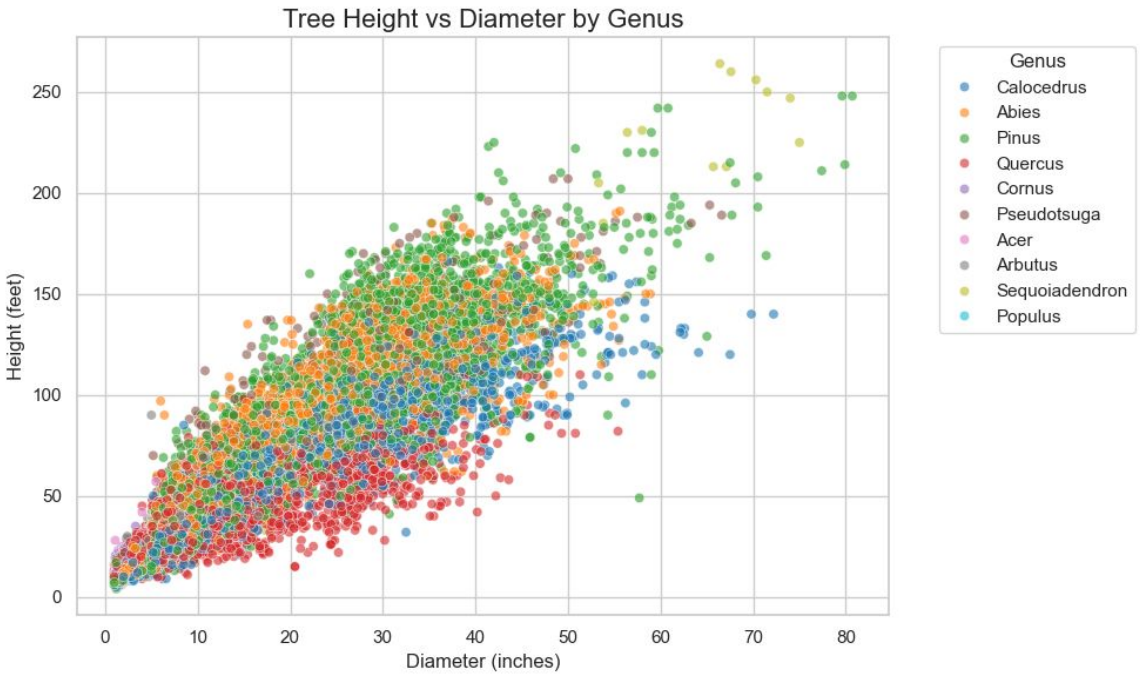

Figure 10: Tree Height vs Diameter grouped by Genus

A strong positive correlation between diameter and height is evident across all genera. However, different genera display unique growth tendencies:

- **Pinus** and **Pseudotsuga** species often achieve both high diameters and tall heights.
- **Quercus** trees tend to be broader but shorter, as they cluster lower in height for comparable diameters.
- **Sequoiadendron** exhibits extreme values, consistent with its known status as one of the tallest and thickest trees.

These structural signatures suggest that morphological features such as height and diameter, when combined with genus information, can significantly aid in taxonomic classification

0.6 Spatial Distribution of Tree Density

To better understand the geographic distribution of tree observations across the study region, a two-dimensional kernel density estimation (KDE) plot was generated using the latitude and longitude coordinates of all trees in the dataset.
The KDE visualization reveals distinct spatial patterns in tree density across California. Areas with higher density estimates are shown in bright yellow and green, indicating regions where a large number of tree observations were recorded. These clusters appear prominently in the central to northern parts of the state, which is consistent with forested mountain ranges such as the Sierra Nevada.

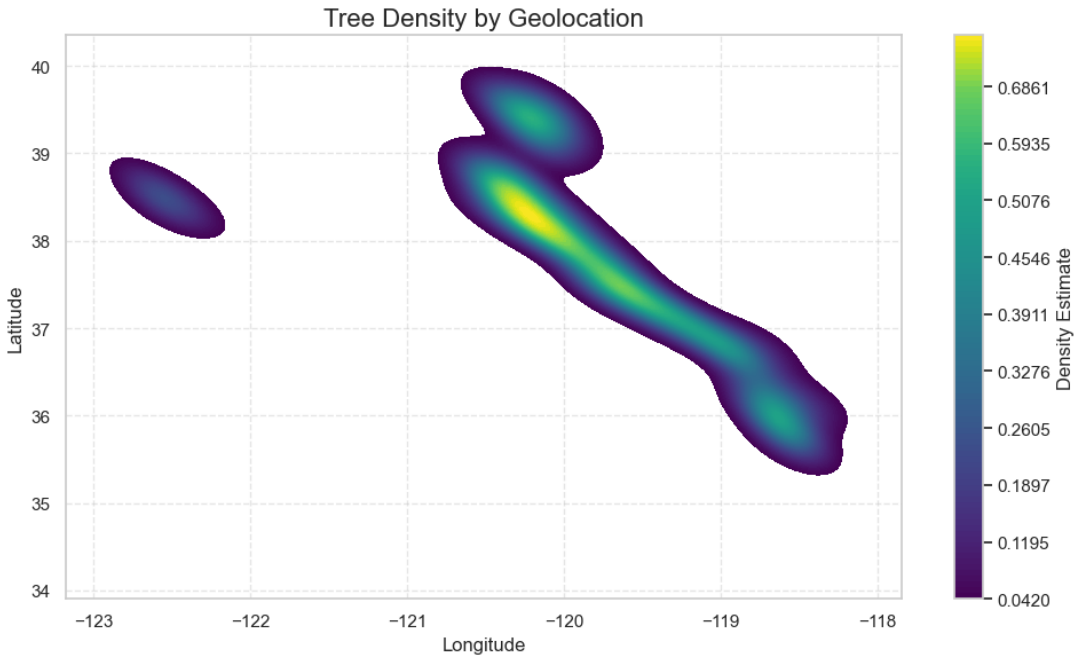

Figure 11: Tree Density by Geolocation

In contrast, regions shaded in darker blue and purple correspond to areas with lower observation density. These may be urban, agricultural, or desert areas where fewer trees are present or less sampling was conducted.
This plot highlights the spatial bias in the dataset viz. tree data are not uniformly distributed across the landscape. Consequently, when incorporating spatial features such as latitude and longitude into predictive models (e.g., for species or PFT classification), it is important to account for this imbalance. Otherwise, models may overfit to dense regions and underperform in sparsely sampled areas. To address this, spatial cross-validation or stratified sampling techniques should be considered during model development to ensure geographic generalizability.

After this thorough analysis of which features are important for influencing the PFT , we then select the below focused set of important columns:

- **Diameter (**`DIA`**)**
- **Height (**`HT`**)**
- **Basal Area (**`BasalA`**)**
- **Latitude (**`LAT`**)**
- **Longitude (**`LON`**)**
- **Species code (**`SPCD`**)**
- **Ecological sub-region code (**`ECOSUBCD`**)**
- **Plant Functional Type (**`PFT`**)**

3. Cleaning the Data by Removing Missing Values

To ensure reliability, any rows with missing values in the chosen columns are completely **removed from the dataset**. This step significantly reduces errors or bias in the analysis later on.

4. Converting Units for Diameter

Next, the **diameter measurements are converted from inches to centimeters** (multiplying by 2.54). This conversion ensures consistent measurement units throughout the dataset, making interpretation easier and more standardized, especially if sharing or comparing this data with international datasets or standards.

5. Filtering by Ecological Zones

The below figure showcases the count of trees for all the important ecological zones. The dataset is filtered to only include trees from these specific ecological zones:

- `M261Ep`, `261Ba`, `M261Em`, `263Am`, and `M261Ej`.

These zones are so chosen because they have enough data to be useful for modeling. Moroever, as shown above, we found that other ecological zones do not contribute much from a modelling perspective. This filtering simplifies the data, makeing it easier to analyze, and ensures that each ecological region has sufficient data for the model to learn effectively.

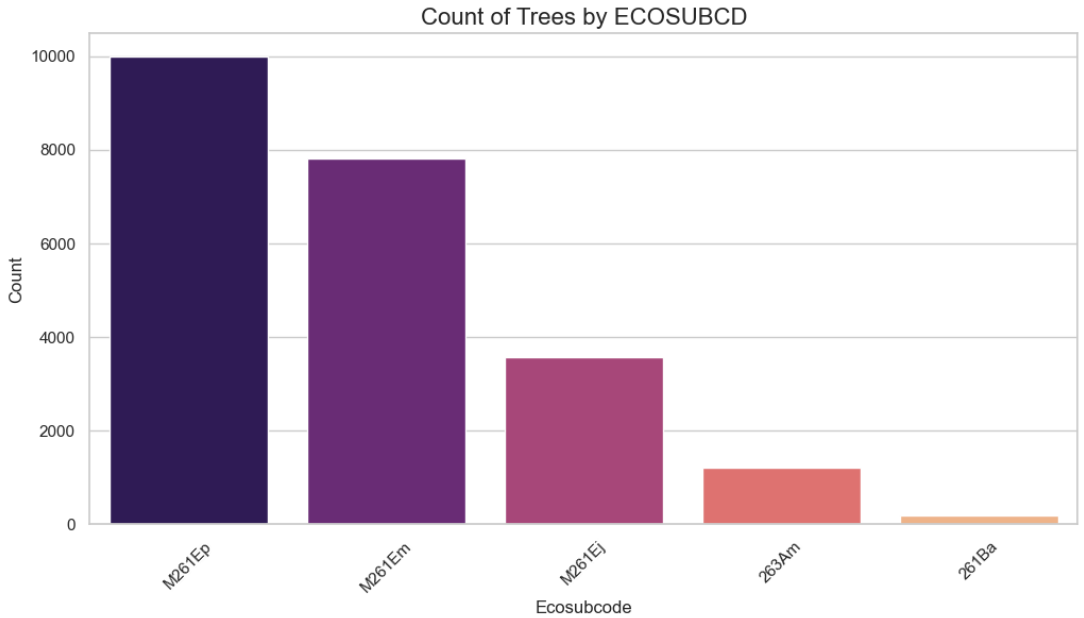

Figure 12: Tree Counts per Ecological Zone chosen

6. Shuffling and Encoding Ecological Zones

The filtered dataset is randomly shuffled, ensuring the order of data points does not bias the model during training. The ecological zone (`ECOSUBCD`) column, originally categorical, is converted to multiple numeric columns through **one-hot encoding**. Each ecological zone gets

its own column, filled with `0` or `1` values which explicitly tells the model which ecological zone each tree belongs to. One ecological zone (`ECOSUBCD_M261Ep`) is removed after encoding to avoid redundancy (multicollinearity), as its presence can be inferred from the absence of other zones.

7. Adjusting the Target Variable

A specific category label (`Deciduous`) is renamed to `Deciduous broadleaf` to improve clarity or correct inconsistencies in labeling. The `Broadleaf` category is entirely removed from the dataset, due to inadequate representation , simplifying the modeling and ensuring better accuracy.

8. Further Data Exploration and Group Analysis

We examine how tree types (**PFTs**) vary across different ecological regions. This helps understand which plant types are dominant or rare within specific regions.

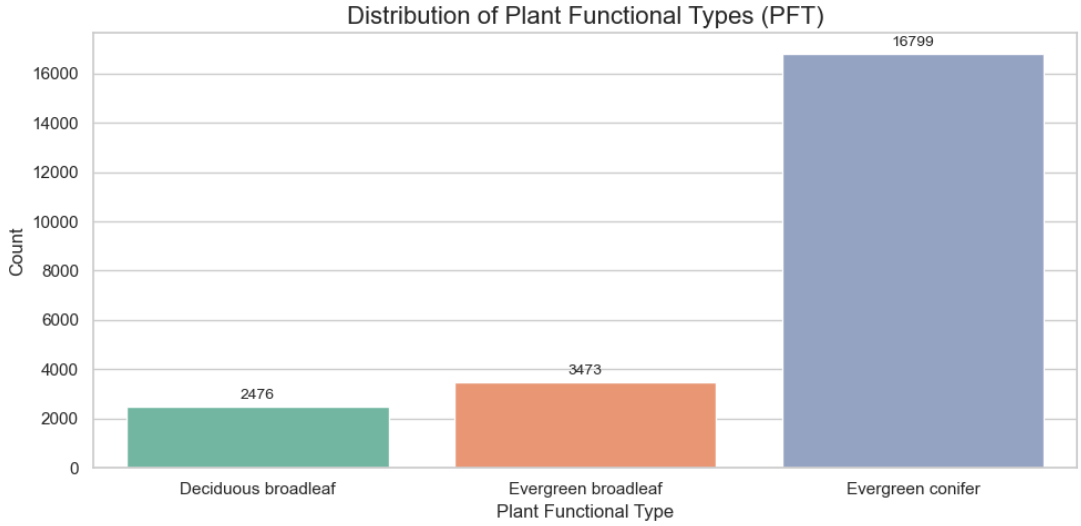

Figure 13: Distribution of Trees by ECOSUBCD in the PFT dataset

To explore interdependencies between key structural features, a pairplot was generated for diameter (DIA_cm), height (HT), and basal area (BasalA). This plot visualizes both individual feature distributions and bivariate relationships.
Diagonal plots: KDE plots indicate that all three variables are right-skewed, with BasalA showing the strongest skew due to its quadratic derivation from diameter.
Scatterplots:

- **DIA_cm vs HT:** Strong linear correlation that is taller trees generally have thicker trunks.
- **DIA_cm vs BasalA:** Non-linear quadratic relationship, derived from the geometric formula of basal area.
- **HT vs BasalA:** Also shows strong correlation, but more scatter than DIA-BasalA.

This visualization highlights substantial multicollinearity between the features. Since BasalA is functionally derived from DIA_cm, using both in predictive models may lead to redundancy. Hence, we drop it from the training process.

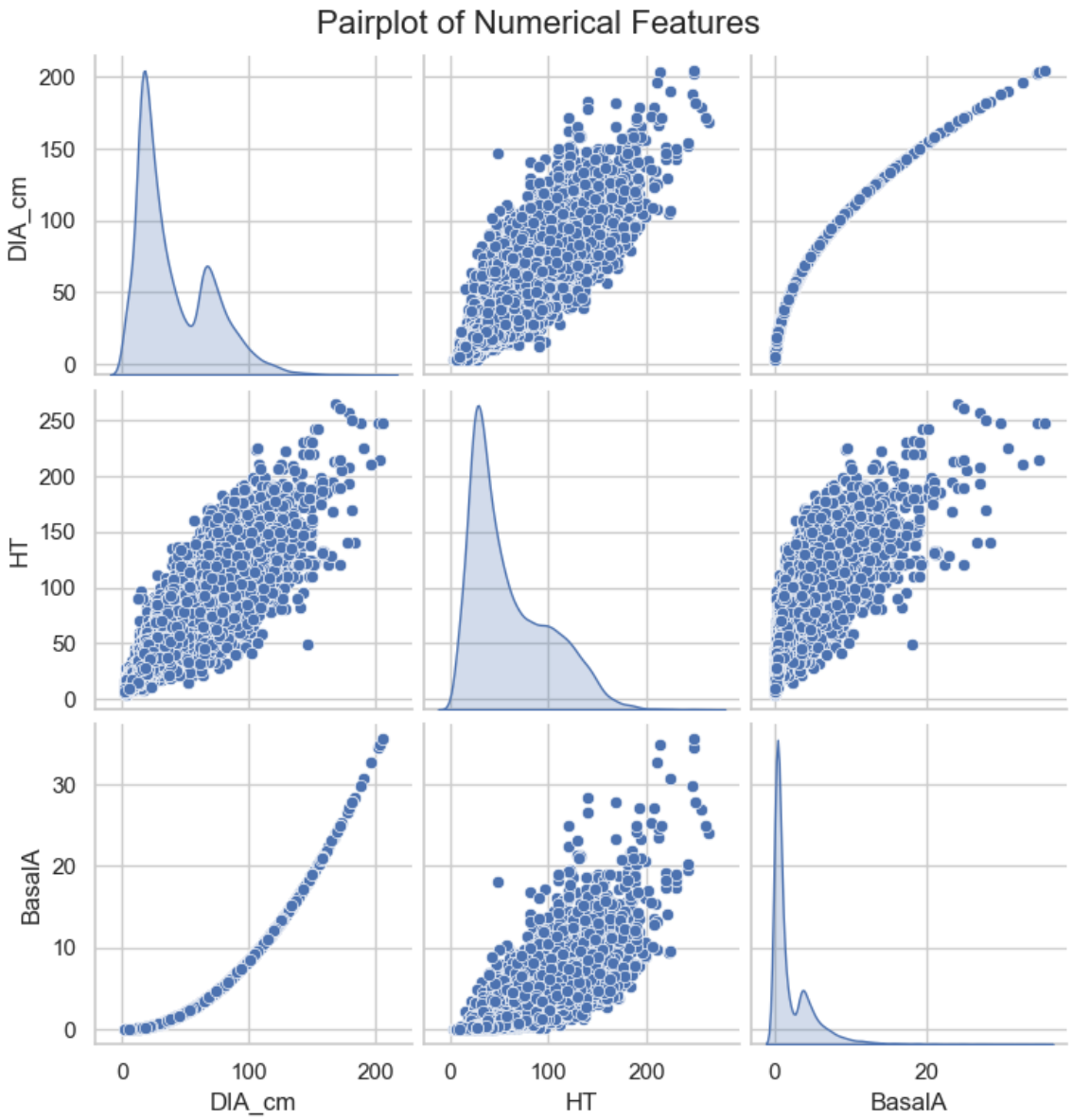

Figure 14: Pairplot of Numerical Features (DIA_cm, HT, BasalA)

9. Preparing the Final Dataset for Modeling

A final selection of variables is made:

- **Numerical features**: diameter (`DIA_cm`), height (`HT`), latitude (`LAT`), longitude (`LON`).
- **Categorical features**: selected ecological zones (`ECOSUBCD`).

These become the **input features (X)** used to predict the **target variable (y)**, which is the plant type .

10. Encoding the Target Variable

Since machine learning algorithms prefer numeric targets, the `PFT` labels are converted into numeric form using a simple encoding system like the (`LabelEncoder`). Each plant type gets a

unique number.

11. Splitting and Scaling the Data

The data is split into two sets:

- **Training set (75%)**: used to teach the model.
- **Test set (25%)**: held aside to evaluate the model later.

To maintain balance, the split preserves the proportions of different plant types. Numerical features are scaled using **standardization** (`StandardScaler`).

12. Training the Machine Learning Model

0.6.1 Objective - Model PFT

The goal of this modeling task is to classify individual trees into their corresponding PFTs, using features that describe the tree's structure, geographic location, and ecological region.

0.6.2 Features Used

Independent Variables:

- Numerical: `DIA_cm`, `HT`, `BasalA`, `LAT`, `LON`
- One-hot encoded categorical: `ECOSUBCD_261Ba`, `ECOSUBCD_263Am`, `ECOSUBCD_M261Ej`, `ECOSUBCD_M261Em`

Target Variable: `PFT`, a multi-class categorical variable with 3 classes.

0.6.3 Preprocessing Steps

- The dataset is split into train and test sets with stratification to maintain class proportions.
- Numerical features are scaled using `StandardScaler`.
- One-hot encoded features are appended post-scaling.

0.6.4 Model Architecture

A stacking ensemble classifier was constructed with four base learners and an XGBoost meta-learner. The following images illustrates the architecture along with the processing performed in the same.

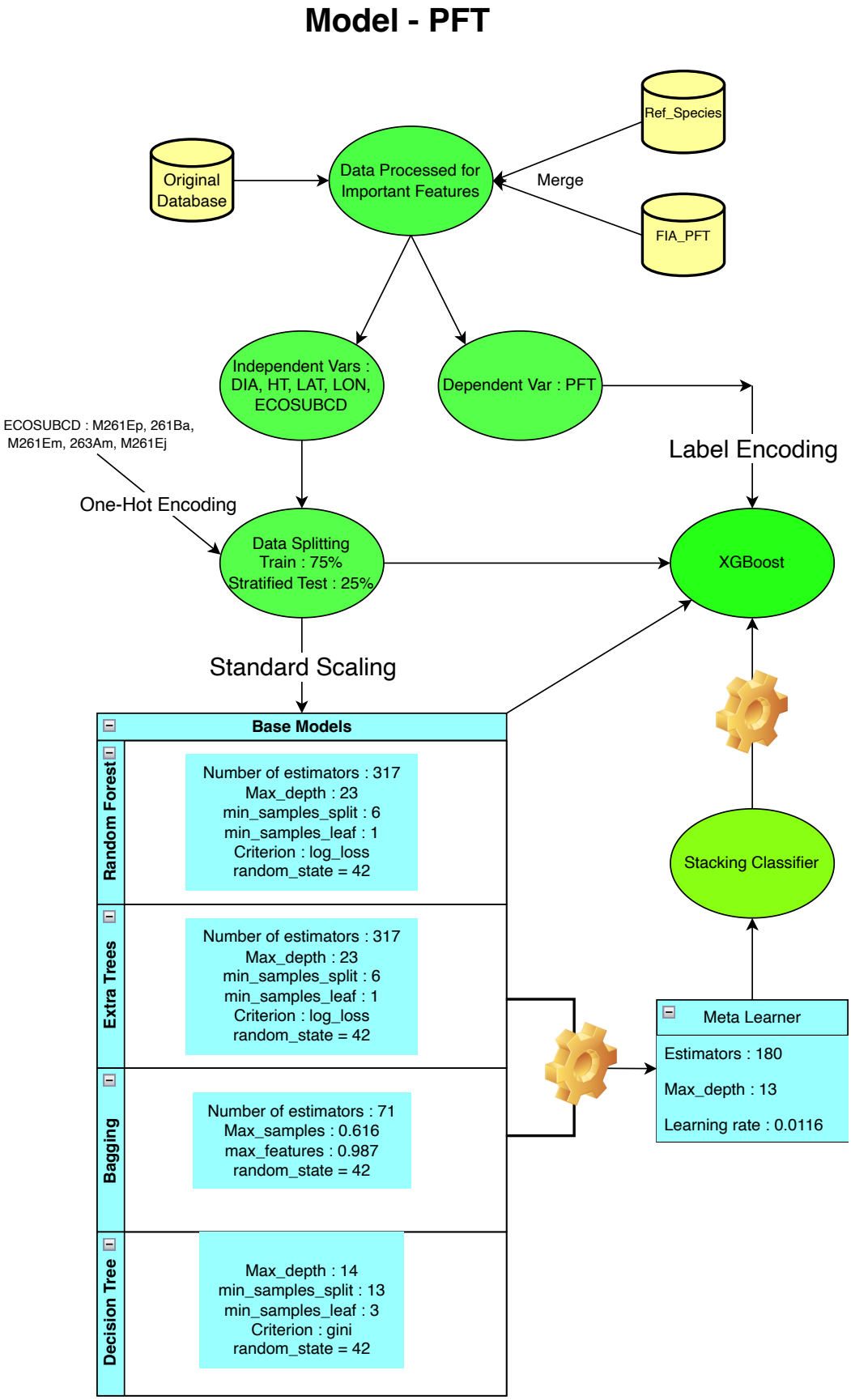

Figure 15: Model Architecture for PFT prediction

Base Models:

- **RandomForestClassifier:** 317 trees, `max_depth=23`, `criterion=log_loss`, `class_weight=balanced`
- **ExtraTreesClassifier:** Similar parameters as RF but without bootstrapping.
- **DecisionTreeClassifier:** `max_depth=14`, `min_samples_split=13`, `min_samples_leaf=3`
- **BaggingClassifier:** `n_estimators=71`, `max_samples=0.616`, no bootstrapping.

Meta-Learner: XGBoostClassifier (fine-tuned parameters):

- `n_estimators=180`, `max_depth=13`, `learning_rate=0.0116`
- `gamma=0.994`, `reg_alpha=0.711`, `reg_lambda=0.790`
- `subsample=0.803`, `colsample_bytree=0.886`
- `eval_metric=mlogloss`

0.6.5 Feature Importance

- Feature importances were derived from the final ensemble.
- A vertical bar chart was plotted to visualize which features contributed most to classification.

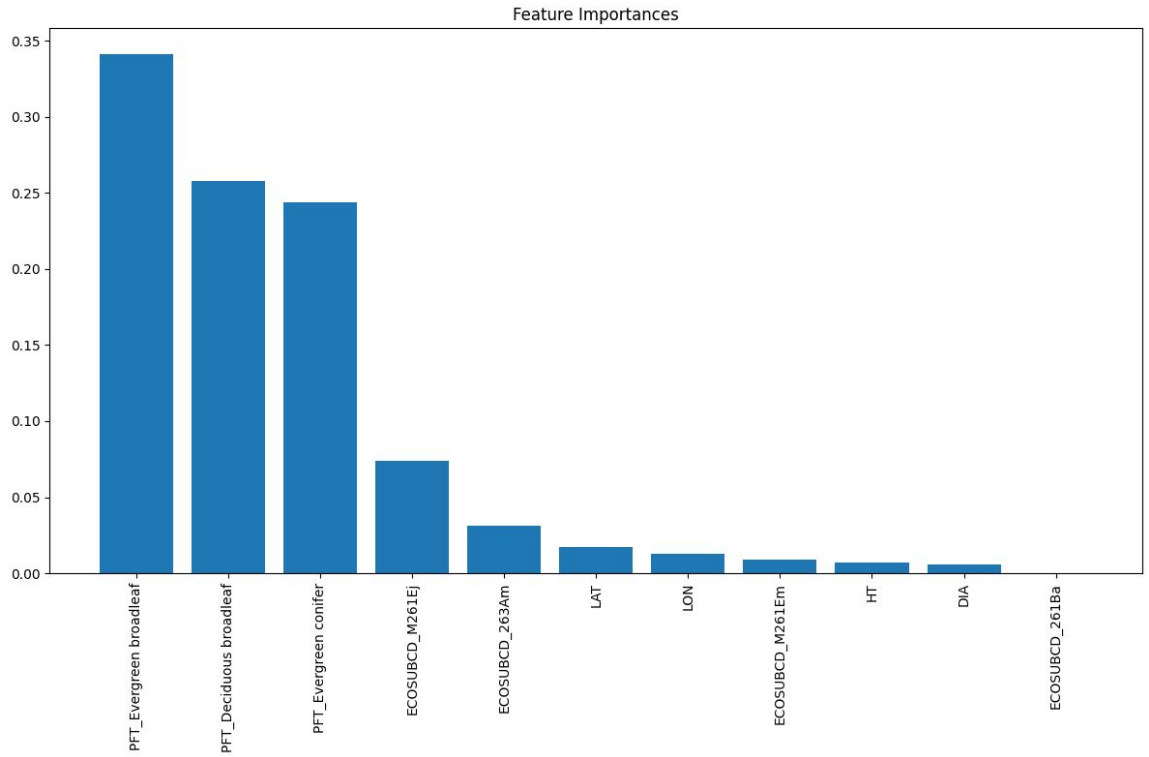

Figure 16: Feature Importance for the PFT model

13. Integrating the TLS Data

Another task in this is to predict the instances on the TLS data, containing precise tree locations and plot identifiers.

14. Updating the TLS Dataset

We add global coordinates (latitude and longitude) to the TLS dataset, enhancing geographic analysis capability, since these features are not available in the dataset.

0.7 Hierarchical Stacking Pipeline for Tree Classification

0.7.1 Objective

This new proposed pipeline aims to carry forward the predictions done on PFT and use the to classify the genus and species of the trees. Hence, we use a series of hierarchical models. This design enhances interpretability and predictive performance than other experiments we performed.

0.7.2 Model Hierarchy for next tasks

- **Model 2: Genus Prediction**
  Predicts the Genus using the **same set of features as Model 1**, augmented by the **predicted PFT as an additional input**. This leverages the ecological-genetic correlation between plant types and genus-level identity.
- **Model 3: Species Prediction**
  The final model predicts the tree species using all previously available features, **along with the predicted PFT and predicted Genus**. This hierarchical inclusion allows the model to focus on species-level variation within the ecological and genetic context.

0.7.3 Data Processing Pipeline

- **Feature Merging:** We process the available datasets by merging them to use common features wherever possible for better performance .
- **One-Hot Encoding:** Categorical variables like ECOSUBCD are transformed into binary columns.
- **Standard Scaling:** Applied to numerical variables to standardize feature ranges.
- **Stratified Splitting:** 75% of data is used for training and 25% for testing, with stratification to maintain class distributions.

0.7.4 Model Architectures

Each model in the pipeline is a stacking ensemble, composed of the following base learners:

- Random Forest
- Extra Trees
- Decision Tree

These are tuned using `RandomizedSearchCV` across parameters such as `n_estimators`, `max_depth`, `min_samples_split`, and `criterion`. The output of these base models is passed to a meta-learner:
Meta-Learner: XGBoost, with the following configuration:

- `n_estimators = 200`
- `max_depth = 5`
- `learning_rate = 0.01`
- `gamma = 0.1`, `min_child_weight = 2`, `scale_pos_weight = 2`

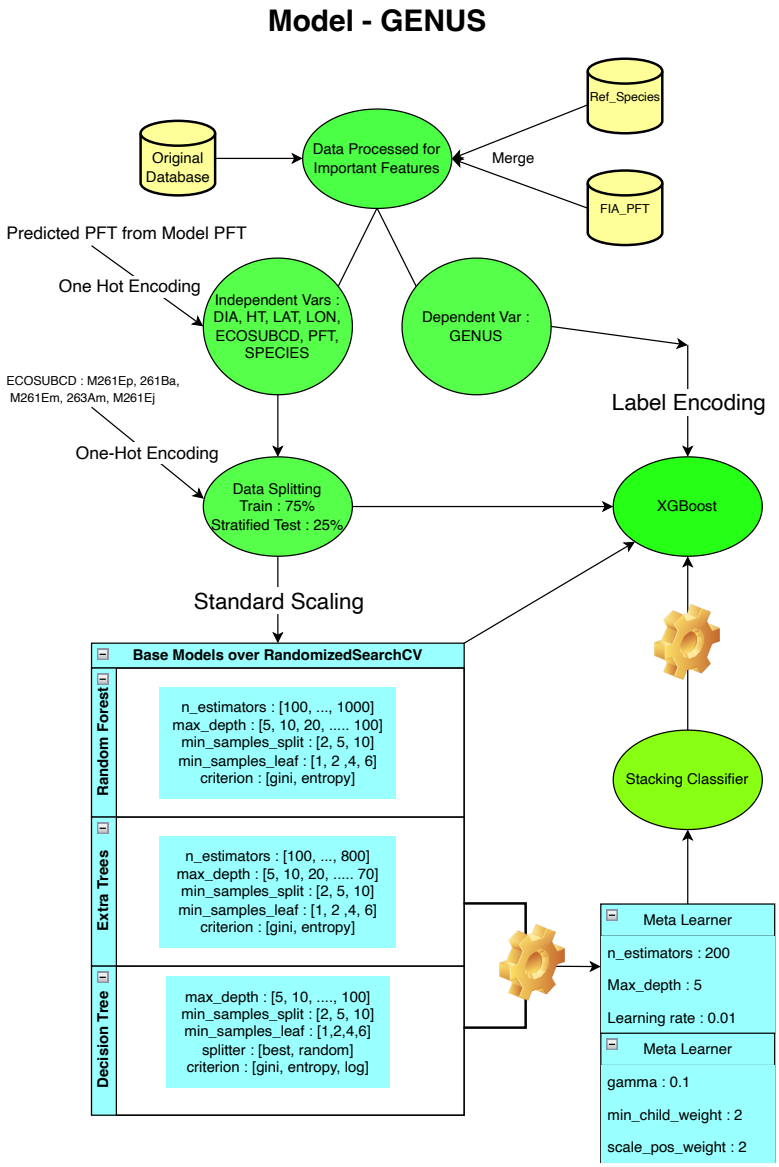

Figure 17: Model Architecture for Genus prediction

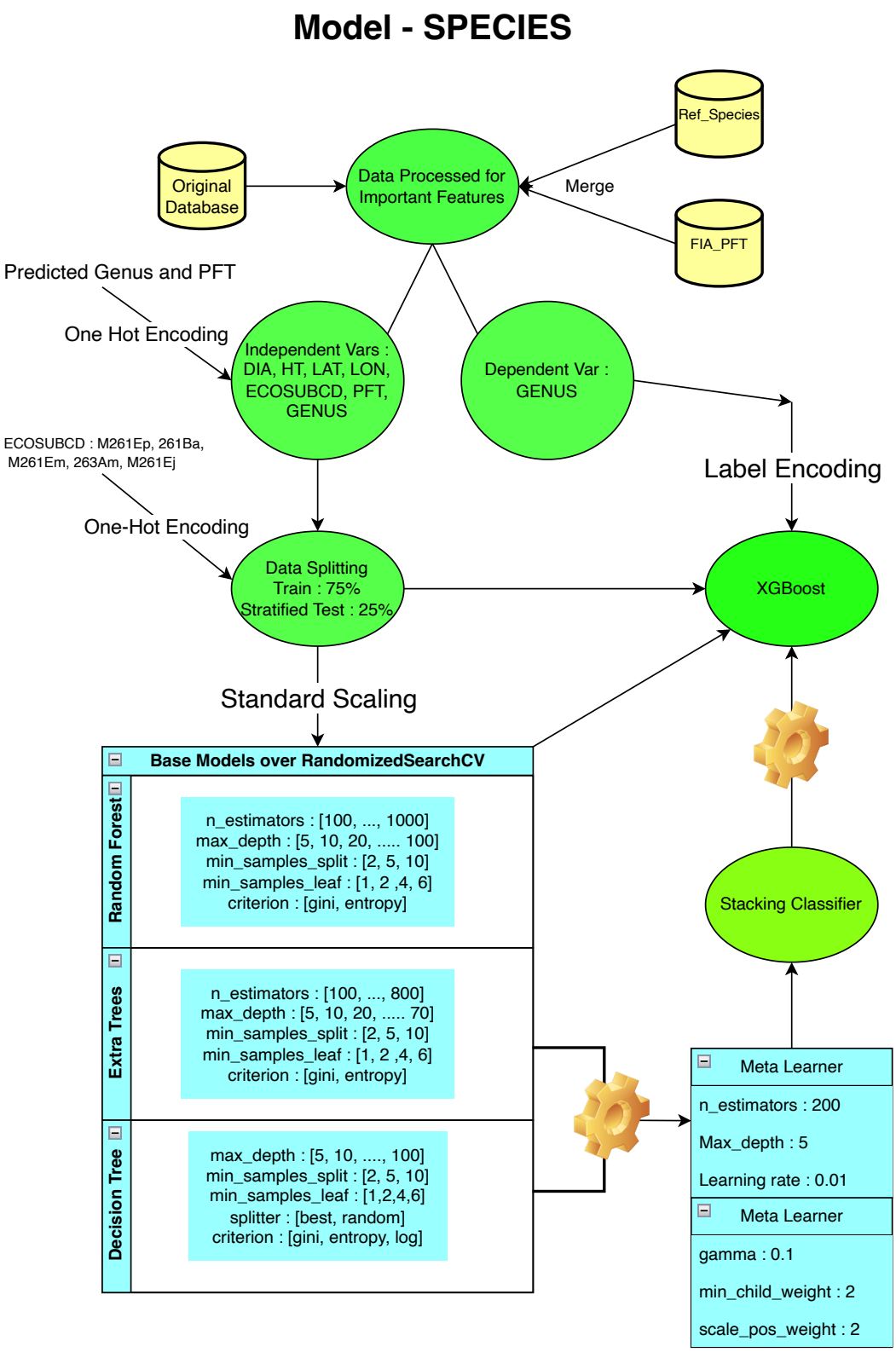

Figure 18: Model Architecture for Species prediction

0.7.5 Benefits of the Proposed Approach

- **Improved Accuracy:** The use of stacked models and merged common features along with precise preprocessing helps rectify fine-grained species.
- **Biological Interpretability:** Hierarchical prediction mirrors natural plant taxonomy.
- **Flexible Design:** Each model stage can be tuned independently and extended if additional labels or data sources are introduced.

Summary of the entire pipeline :
The pipeline developed for this task integrates ecological, morphological, and taxonomic data into a robust, **hierarchical machine learning architecture** aimed at classifying trees across multiple biological levels, beginning with Plant Functional Type (PFT), progressing through Genus, and ending in Species prediction. Initially, the two foundational datasets (`FIATree_PFT` and `REF_SPECIES`) were merged to enrich the tree-level observations with standardized species codes and ecological classifications. A rigorous preprocessing framework was implemented, including EDA techniques, outlier handling, one-hot encoding of ecological zones, standardization of numerical features, and stratified data splitting to preserve class balance. Visualizations such as distribution plots, pairplots, and KDE maps provided deep insight into feature distributions, multicollinearity, and geographic sampling bias. For modeling, a stacking ensemble framework was employed, combining `Random Forest`, `Extra Trees`, `Decision Tree`, and `Bagging` as base learners with a finely tuned `XGBoost` meta-learner. This ensemble was trained and validated using stratified cross-validation. Feature selection was guided by domain knowledge and statistical correlations, resulting in a focused feature set comprising structural traits (`DIA_cm`, `HT`), spatial attributes (`LAT`, `LON`), and ecological indicators (`ECOSUBCD`). The output of each stage (e.g., predicted PFT) was integrated into subsequent models to enhance classification granularity and biological realism. The final models, particularly those predicting Genus and Species, benefited from this layered pipeline, yielding improved accuracy and interpretability.

Question 2 : How well did your model perform across these classification tasks?

1 Model Performance on PFT Classification Task

1.1 Overall Classification Metrics

The model was evaluated on a test set containing 5,931 samples across five PFTs. The overall classification accuracy was **89.72%**. The weighted average F1-score of **0.89** indicates strong predictive performance across the dataset.

Table 1: Classification Report on Test Set

Class	Precision	Recall	F1-Score	Support
Deciduous broadleaf	0.75	0.64	0.69	718
Evergreen	0.68	0.75	0.71	106
Evergreen broadleaf	0.85	0.81	0.83	869
Evergreen conifer	0.93	0.96	0.95	4234
Flowering plants	1.00	0.50	0.67	4
Accuracy		**0.8972**		
Macro Avg	0.84	0.73	0.77	5931
Weighted Avg	0.89	0.90	0.89	5931

The model performs best on *Evergreen conifer*, which has the highest support and F1-score. Moderate performance is observed for *Evergreen broadleaf* and *Deciduous broadleaf*, while performance is unreliable for rare classes like *Flowering plants* due to limited data (only 4 samples).

1.2 Predicted vs. Ground Truth Proportions (Overall)

To assess ecological realism, we compare the predicted and actual distributions of PFTs across the entire dataset:

Table 2: Overall PFT Proportions: Ground Truth vs. Predictions

PFT	Ground Truth (%)	Predicted (%)	Difference (%)
Evergreen conifer	69.82	78.23	+8.41
Evergreen broadleaf	21.43	19.98	-1.45
Deciduous broadleaf	8.71	1.79	-6.92
Tree (unclassified)	0.05	–	–

The model overestimates *Evergreen conifer* and underestimates *Deciduous broadleaf*, consistent with lower recall for this class. No predictions were made for the very rare *Tree* category.

1.3 Site-Level Analysis

To evaluate model performance spatially, we compare predicted and observed PFT proportions at various ECOSUBCD sites:

Table 3: Predicted PFT Proportions by Site

Site (ECOSUBCD)	Evergreen conifer	Evergreen broadleaf	Deciduous broadleaf
261Ba	11.36	88.64	0.00
263Am	61.36	36.93	1.70
M261Ej	96.51	3.49	0.00
M261Em	87.67	9.08	3.25
M261Ep	74.16	25.36	0.48

Table 4: Ground Truth PFT Proportions by Site

Site (ECOSUBCD)	Evergreen conifer	Evergreen broadleaf	Deciduous broadleaf	Tree
261Ba	0.00	84.07	15.93	0.00
263Am	40.68	50.95	8.37	0.00
M261Ej	93.58	0.00	6.42	0.00
M261Em	89.31	2.46	8.11	0.12
M261Ep	94.83	0.00	5.17	0.00

Observations:

- **M261Ej, M261Em, M261Ep:** The model closely matches conifer-dominated sites, showing high accuracy where Evergreen conifer is prevalent.
- **263Am:** Mixed PFTs lead to slight underprediction of Deciduous broadleaf.
- **261Ba:** The model completely misses Deciduous broadleaf, which constitutes 16% of the site in the field data.

1.4 Summary of Model Strengths and Weaknesses

Strengths:

- High classification accuracy (89.72%).
- Strong performance on dominant classes, particularly Evergreen conifer (F1 = 0.95).
- Good site-level match in conifer-heavy ecosystems.

Weaknesses:

- Underprediction of Deciduous broadleaf, both in classification metrics (recall = 0.64) and proportion (-6.92%).
- Inability to capture minority or rare PFTs (e.g., Flowering plants and Tree).

1.5 Conclusion

The model is well-suited for classifying dominant PFTs and performs reliably in coniferous ecosystems. However, it requires improvement in representing less common functional types, such as Deciduous broadleaf. Future work may consider incorporating class balancing techniques or domain-specific augmentation to address these limitations.

2 Model Performance on Genus Classification Task

In this section, we evaluate the model's performance on the task of predicting the genus of tree species. A stacked ensemble model utilizing a tuned XGBoost classifier was employed to capture both feature-level diversity and class-specific nuances. The model was trained on ecologically relevant data procured earlier and evaluated on a held-out test set.

2.1 Overall Classification Metrics

The model achieved an overall accuracy of **83.26%** on the test set, with a weighted average F1-score of **0.83**. This indicates strong overall generalization performance across genera, although per-class metrics reveal variability based on class representation.

Table 5: Classification Report on Test Set (Genus)

Genus (Label)	Precision	Recall	F1-Score	Support
0 (Abies)	0.77	0.78	0.78	1477
1 (Acer)	0.81	0.95	0.88	22
2 (Arbutus)	0.85	0.83	0.84	60
3 (Calocedrus)	0.78	0.83	0.80	1451
4 (Cornus)	0.89	0.94	0.92	35
5 (Pinus)	0.81	0.77	0.79	1730
6 (Notholicarpos)	1.00	0.88	0.93	8
7 (Populus)	0.80	0.71	0.75	188
8 (Pseudotsuga)	0.99	0.99	0.99	1302
9 (Quercus)	1.00	0.80	0.89	10
Accuracy		**0.8326**		
Macro Avg	0.87	0.85	0.86	6283
Weighted Avg	0.83	0.83	0.83	6283

The model performs consistently well across both common and rare genera. In particular, it achieves near-perfect performance on *Pseudotsuga* and *Cornus*. For very small classes like *Notholicarpos* and *Acer*, performance remains strong despite limited sample sizes.

2.2 Predicted vs. Field Genus Distribution

Beyond classification accuracy, it is critical to assess how well the model reproduces the overall genus composition observed in the field. The following table compares the proportion of each genus in the predicted labels with the observed (ground truth) proportions from field data:

Table 6: Overall Genus Distribution: Ground Truth vs. Model Predictions

Genus	Field Observed (%)	Predicted by Model (%)
Abies	23.80	18.74
Calocedrus	22.26	18.35
Quercus	21.47	19.44
Pinus	18.26	19.60
Pseudotsuga	5.26	21.38
Populus	1.07	1.71
Acer	0.09	0.54
Sequoiadendron	0.23	0.23
Arbutus	3.21	–
Notholicarpos	2.98	–
Salix	0.79	–
Cornus	0.56	–

Key Observations:

- The model captures the major genera reasonably well (e.g., *Pinus*, *Quercus*, *Calocedrus*).
- *Pseudotsuga* is significantly overpredicted (5.3% in field vs. 21.4% in predictions).
- Several rare genera including *Arbutus*, *Notholicarpos*, *Salix*, and *Cornus* are completely missing from the model predictions due to no representation in the training data.

2.3 Impact of Training Data Distribution

The distribution of genera in the training data significantly influences model behavior. Table below presents the training set proportions:

Table 7: Training Data Genus Distribution

Genus	Proportion (%)
Pinus	27.54
Abies	23.51
Calocedrus	23.10
Quercus	20.72
Pseudotsuga	2.99
Arbutus	0.95
Cornus	0.55
Acer	0.35
Sequoiadendron	0.16
Populus	0.13

Insights:

- The top 4 genera (*Pinus*, *Abies*, *Calocedrus*, *Quercus*) dominate the training set, accounting for over 94% of the data.

- Rare classes like *Populus*, *Cornus*, and *Sequoiadendron* have minimal representation, which likely contributes to their low recall or complete absence in model predictions.
- Interestingly, *Pseudotsuga* is overpredicted despite having low training proportion (2.99%). This could be due to feature similarities with other dominant genera or overfitting to its distinctive traits.

This signifies the need for class balancing strategies such as:

- Reweighting class losses during training
- Oversampling underrepresented genera
- Aggregating rare genera under a hierarchical grouping

2.4 Why Site-Level Analysis Was Not Performed

Unlike the PFT task, genus-level predictions were not analyzed at the ECOSUBCD (site) level. The primary reasons for this are:

- Genus diversity within individual sites is often low, limiting the statistical robustness of comparisons.
- Many rare genera are sparsely distributed across regions, reducing confidence in any spatial pattern detection.
- The primary goal of this model was to evaluate genus classification globally across the entire dataset, rather than spatially at fine granularity.

Site-level genus analysis may be considered in future work with expanded datasets and richer spatial annotations.

2.5 Summary of Strengths and Weaknesses

Strengths:

- Achieves high accuracy (83.26%) and F1-score (macro avg = 0.86) on a multiclass classification task.
- Accurately captures most common genera and maintains good recall even for moderately sized classes like *Cornus* and *Populus*.
- Handles some low-frequency genera effectively (e.g., *Notholicarpos*, *Acer*).

Weaknesses:

- Overprediction of *Pseudotsuga* suggests possible confusion with other similar conifers.
- Complete omission of some rare genera in predictions (e.g., *Salix*, *Arbutus*) points to imbalance-driven underfitting.
- No ecological spatial validation due to lack of robust site-level genus distribution.

2.6 Conclusion

The genus classification model demonstrates strong generalization for dominant and moderately represented classes. However, class imbalance in the training data limits its ability to detect rare genera, and some genera are systematically underrepresented or missed in predictions. Addressing these issues will be crucial for improving ecological applicability and model robustness in future iterations.

3 Model Performance on Species Classification Task

This section analyzes the performance of the species classification task using a stacking model with a tuned XGBoost classifier. The goal was to accurately predict tree species using site and structural features.

3.1 Overall Classification Metrics

The model achieved a high test accuracy of **96.91%**, with training completed in only 6.64 seconds. The macro-averaged F1-score was **0.92**, and the weighted average was **0.97**, showing strong and consistent performance across species classes.

Table 8: Classification Report on Test Set (Species)

Species (Label)	Precision	Recall	F1-Score	Support
0	0.98	0.98	0.98	95
1	1.00	1.00	1.00	515
2	1.00	1.00	1.00	1329
3	1.00	1.00	1.00	860
4	0.91	0.88	0.90	34
5	1.00	1.00	1.00	10
6	0.95	0.97	0.96	291
7	0.99	0.99	0.99	301
8	0.85	0.72	0.78	291
9	0.50	0.33	0.40	3
10	1.00	1.00	1.00	20
11	1.00	1.00	1.00	238
12	0.90	0.94	0.92	822
13	1.00	1.00	1.00	8
Accuracy			**0.9691**	
Macro Avg	0.93	0.92	0.92	4817
Weighted Avg	0.97	0.97	0.97	4817

The model demonstrates nearly perfect performance on most species. Precision and recall remain high even for smaller classes such as *macrophyllum* and *douglasii*, with some drop observed for very rare classes like label 9.

3.2 Predicted vs. Ground Truth Species Distribution

The following table compares the species proportions from the model predictions and the field-observed (ground truth) values:

Table 9: Species Distribution Comparison (Predicted vs. Field Observed)

Species	**Ground Truth (%)**	**Predicted (%)**
decurrens	22.26	22.08
concolor	20.91	20.92
lambertiana	1.63	19.28
agrifolia	14.30	11.43
ponderosa	2.75	9.56
menziesii	8.48	9.41
kelloggii	3.49	3.11
jeffreyi	5.59	1.94
macrophyllum	0.05	1.32
lobata	0.28	0.93

Key Observations:

- Species such as *decurrens* and *concolor* were predicted with proportions closely matching ground truth.
- The model significantly **overpredicts lambertiana** (1.63% $\rightarrow$ 19.28%), likely due to feature similarities with dominant species.
- *macrophyllum*, a very rare species in the field (0.05%), is overrepresented in predictions (1.32%).
- Other rare species like *tremuloides*, *giganteum*, and *scouleriana* were not predicted and are omitted from the table.

3.3 Impact of Training Data Distribution

The training data distribution shows clear class imbalance, which likely influences model behavior. The most common species include *decurrens* and *concolor*, while species like *lobata* and *giganteum* are heavily underrepresented.

Table 10: Species Distribution in Training Data

Species	Proportion (%)
decurrens	26.76
concolor	24.56
ponderosa	15.31
chrysolepis	10.26
kelloggii	5.93
lambertiana	5.36
jeffreyi	5.00
menziesii	3.58
agrifolia	1.74
douglasii	0.76
macrophyllum	0.40
giganteum	0.18
tremuloides	0.11
lobata	0.05

Interpretation:

- The close match between training and prediction distributions for dominant species validates the model's generalization ability.
- Overprediction of some rare species like *lambertiana* may indicate class confusion or bias from correlated features.
- The underprediction or exclusion of ultra-rare species (e.g., *giganteum*) highlights the challenges of data imbalance.

3.4 Why Site-Level Evaluation Was Not Conducted

Site-level evaluation (e.g., ECOSUBCD-wise distribution) was not included for species prediction due to the following limitations:

- Species-level diversity per site is often too narrow to allow robust statistical comparison.
- Many species appear only sparsely or in isolated locations, complicating site-level aggregation.
- The current focus was to assess global predictive accuracy and distributional alignment rather than spatial stratification.

3.5 Summary of Strengths and Weaknesses

Strengths:

- Achieves excellent accuracy (96.91%) and high F1-scores across most species.
- Predicts dominant and moderately rare species accurately.
- Robust performance even in presence of training imbalance for mid-frequency species.

Weaknesses:

- Overprediction of *lambertiana* and *macrophyllum* raises concerns about class confusion.
- Several rare species are not predicted, reflecting underfitting or poor generalization in low-support classes.
- Lack of spatial validation limits ecological interpretation.

3.6 Conclusion

The species classification model demonstrates high accuracy and strong class-wise consistency, particularly for well-represented species. However, addressing the imbalance-driven prediction gaps for rare species remains an important direction for future improvement. Incorporating additional features, class-aware loss functions, or multi-stage hierarchical classifiers could help mitigate these limitations.

Question 3: Were the predicted distributions representative of the actual field data?

To evaluate the ecological realism of the models, we compared predicted class distributions against observed proportions from field data across all three classification levels: PFT, Genus, and Species. We also visualized the ground truth distributions in Figures 19 to 21.

PFT

The predicted PFT distribution showed strong alignment with the field data, particularly for the dominant class *Evergreen conifer*. However, the model underpredicted minority classes:

- **Evergreen conifer** was well predicted: 78.23% predicted vs. 69.82% ground truth.
- **Evergreen broadleaf** was closely matched: 19.98% predicted vs. 21.43% ground truth.
- **Deciduous broadleaf** was significantly underpredicted: 1.79% predicted vs. 8.71% ground truth.

The model appears biased toward dominant coniferous classes, consistent with the training set's imbalance. Nevertheless, the overall predicted PFT proportions are reasonably representative of the actual field distribution.

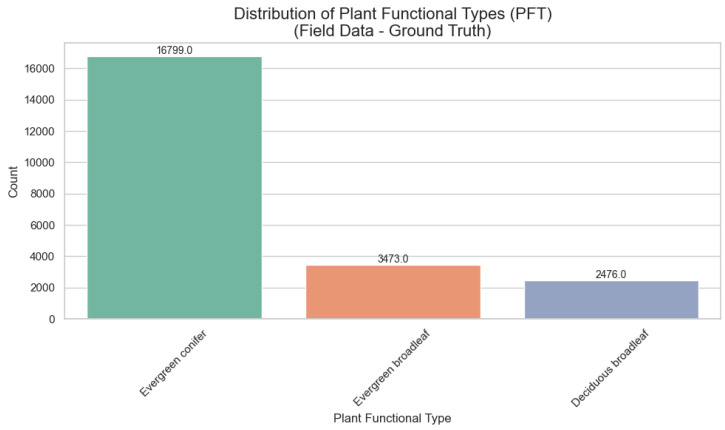

Figure 19: Distribution of Plant Functional Types (PFT) in Field Data

Genus

The genus classification model displayed stronger predictive accuracy, but the predicted genus proportions deviated more from the field observations:

- **Pseudotsuga** was heavily overpredicted: 21.38% predicted vs. 5.26% in the field.
- **Pinus**, **Quercus**, and **Calocedrus** were fairly close, with underprediction margins under 3%.
- Rare genera like **Arbutus**, **Cornus**, and **Populus** were completely missed.

These results suggest that although the model classified dominant genera well, it struggled with ecological diversity and underrepresented genera due to significant training imbalance.

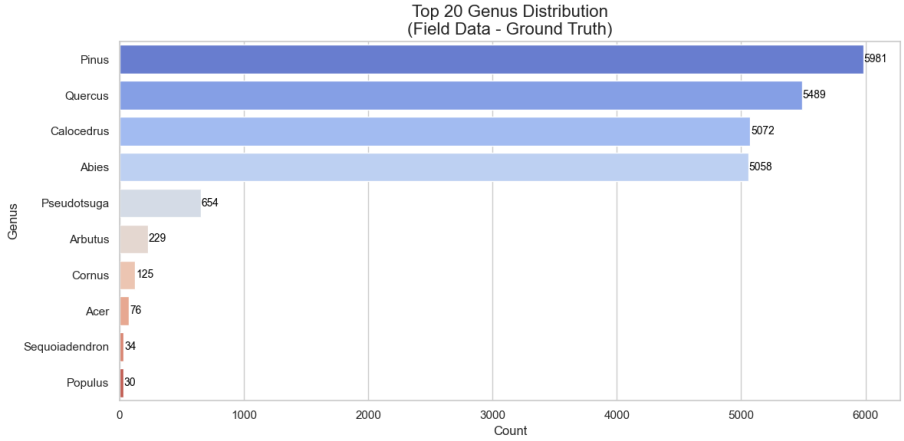

Figure 20: Top 20 Genus Distribution in Field Data

Species

The species model achieved the highest classification accuracy (96.91%) among all tasks. Its predicted distributions also showed reasonable alignment with field data for dominant species:

- **Decurrens** and **Concolor** predictions matched field data almost exactly (within 0.2%).
- **Lambertiana** was severely overpredicted: 19.28% predicted vs. 1.63% observed.
- Many rare species, such as *giganteum*, *macrohyllum*, and *lobata*, were absent or underestimated in predictions.

Despite high predictive accuracy, species-level prediction exhibits similar distributional challenges as genus classification: overfitting to common classes and under-representation of rare ones.

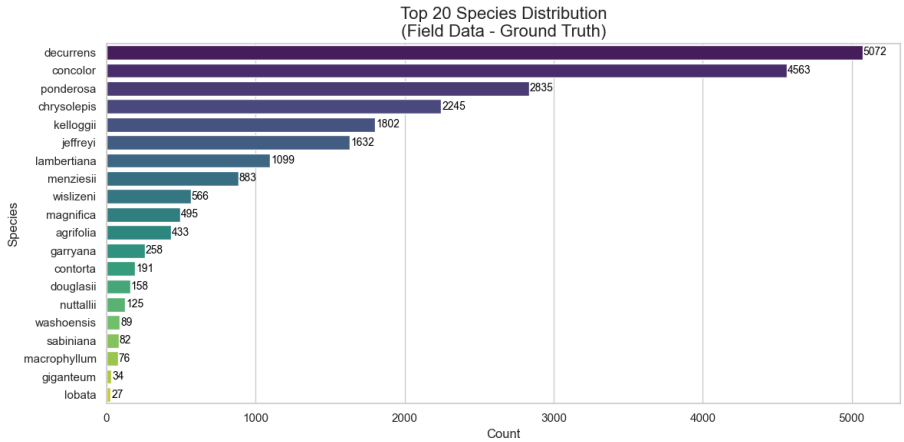

Figure 21: Top 20 Species Distribution in Field Data

Summary

- The models accurately reflect overall field distributions for dominant classes across all three tasks.
- However, the models tend to **overpredict dominant classes** and **underpredict rare or ecologically unique taxa**.
- These deviations are primarily attributed to class imbalance in the training data, as shown in earlier sections.
- Incorporating balanced training strategies, class-aware losses, or hierarchical taxonomic modeling could improve ecological fidelity.

Hence, while predicted distributions are broadly representative of the field data at a high level, finer ecological diversity is not fully captured indicating the need for more inclusive modeling approaches and better represented data in future work.

Question 4: Do you think that the field-collected data was representative of the surrounding site?

To assess the representativeness of the field-collected data, we compared the spatial distribution of ECOSUBCD sites across three datasets:

- **FIA dataset** (used for full-area model predictions)
- **TLS dataset** (LiDAR-based data)
- **Field dataset** (ground truth species/genus/PFT annotations)

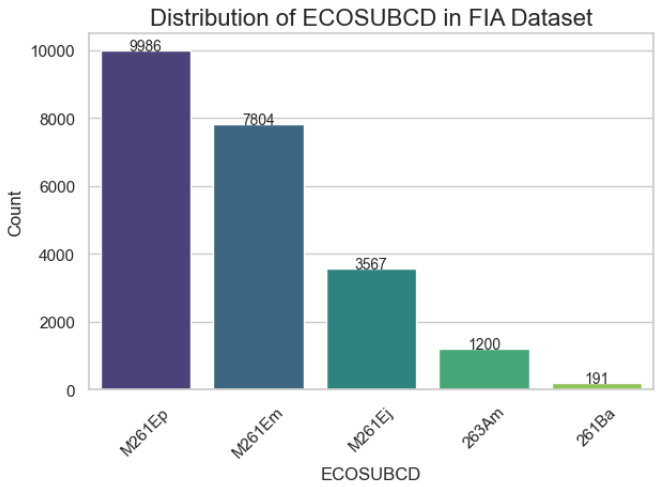

Figure 22: Distribution of ECOSUBCD in FIA Dataset (Model Input)

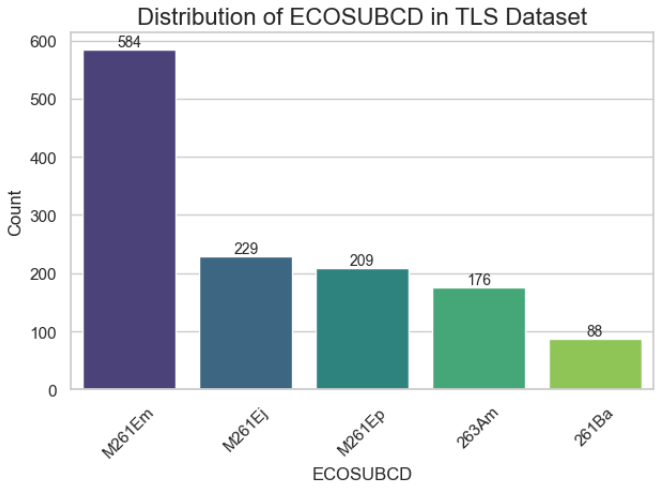

Figure 23: Distribution of ECOSUBCD in TLS Dataset

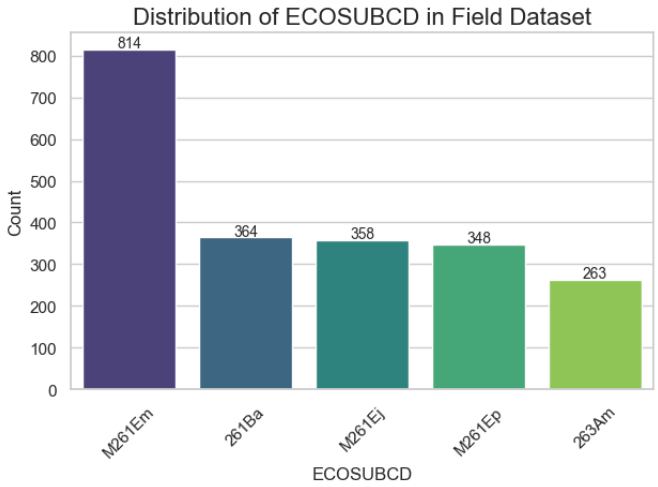

Figure 24: Distribution of ECOSUBCD in Field Dataset

Distributional Comparison Across Datasets

While the field dataset spans all major ECOSUBCDs present in the larger FIA dataset, their relative proportions differ significantly:

- **M261Ep** accounts for nearly 10,000 samples in FIA but is underrepresented in the field dataset (only 348 trees).
- **M261Em** is overrepresented in the field (814 trees) compared to its presence in the FIA dataset.
- Sites such as **263Am** and **261Ba** are sparsely represented across all datasets, but still included in the field sampling.

Thus, while all ECOSUBCDs are sampled, the field dataset is not proportionally distributed relative to the surrounding site vegetation as inferred from the FIA dataset.

Representativeness for PFT-Level Evaluation

We specifically evaluated model predictions and field data at the PFT level across ECOSUBCDs. This site-level comparison revealed several trends:

- For conifer-dominated sites such as **M261Ej**, **M261Ep**, and **M261Em**, the field-collected PFTs closely matched the model's site-wide predictions.
- In contrast, sites with mixed PFT compositions (e.g., **261Ba**, **263Am**) showed discrepancies. For instance, *Deciduous broadleaf* was present in field data but underrepresented in model predictions likely due to low sample size or spatial clustering.
- Overall, field PFT proportions captured the dominant ecological signals at most sites, despite sampling imbalances.

Why Genus and Species Site-Level Evaluation Was Not Performed

As noted previously, genus and species-level predictions were evaluated only globally and not at the ECOSUBCD level. This decision was made due to the following constraints:

- **Sparse per-site genus/species diversity**: Many sites had very few occurrences of rare species, making robust statistical comparisons difficult.
- **Sampling imbalance**: The field dataset is not evenly stratified across ECOSUBCDs for all taxa, reducing interpretability.
- **Model goal**: The genus and species classification tasks aimed to assess overall model generalization rather than local ecological fidelity.

Conclusion

In conclusion, while the field-collected dataset provides sufficient taxonomic and ecological diversity across the study area, it is not fully representative of the surrounding site distributions in the FIA dataset. Some sites are over-sampled while others are underrepresented, introducing potential bias. Nevertheless, for PFT-level evaluation, the field data is reasonably representative of local ecological composition, especially in dominant vegetation zones.
This insight is essential for interpreting model evaluation results and understanding the generalization performance of classifiers in real-world ecological settings.

Question 5: What worked well, and what were the limitations?

What Worked Well

- **High Accuracy Across Tasks:** The hierarchical ensemble models demonstrated excellent performance across all levels of prediction, achieving test accuracies of 89.7% for PFT, 83.3% for Genus, and 96.9% for Species classification. Moreover, the F1 scores of every class are equally good, showcasing efficient performance across classes. These results validate the effectiveness of the overall design and the robustness of the chosen model architecture.
- **Stacked Ensemble Strategy:** The use of diverse base models (Random Forest, Extra Trees, Decision Tree, and Bagging) combined with a finely-tuned XGBoost meta-learner significantly enhanced generalization and reduced overfitting, ensuring stable predictive performance.
- **Hierarchical Modeling Approach:** The layered pipeline wherein PFT predictions informed Genus, and Genus informed Species closely aligned with natural biological hierarchies and proved highly effective in refining class granularity and interpretability.
- **Thoughtful Feature Engineering:** Key preprocessing choices, such as dropping highly correlated variables, unit conversion, one-hot encoding of ecological regions, and filtering by data-rich zones, helped construct a clean, balanced, and interpretable feature set for downstream modeling.
- **Ecologically Coherent Predictions:** PFT predictions exhibited spatial coherence with known ecological distributions, especially in dominant zones like coniferous forests, adding confidence to the field applicability of the results.
- **Insightful Visual Analysis:** Exploratory plots such as pairplots, spatial KDEs, and feature importance charts offered deep insights into the structure of the data and the rationale behind key modeling decisions.

Limitations and Scope Considerations

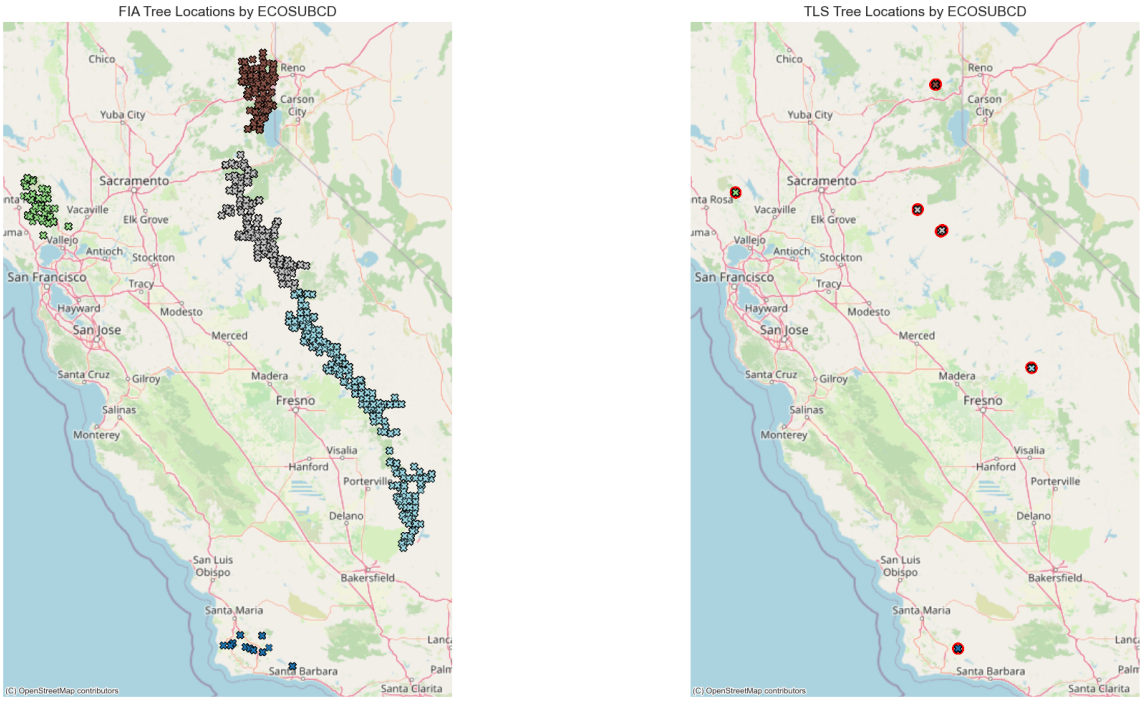

Figure 25: Comparison of tree sampling locations between FIA training data (left) and TLS test data (right). Each point is colored by its ECOSUBCD (ecological subregion).

Table 11: Comparison of FIA (Training) and TLS (Test) Tree Datasets

Aspect	**FIA (Train)**	**TLS (Test)**
Spatial Coverage	Broad, across Sierra Nevada, northern coast, southern CA	Sparse, localized to a few sites across California
Point Density	Thousands of data points	A few isolated locations
ECOSUBCD Variety	Diverse, covering many ecological zones	Limited, only a few zones represented
Sample Redundancy	High redundancy within regions	Very limited per ECOSUBCD
Suitability	Ideal for model training and pattern learning	Limited for generalization testing or validation

Explanation: The visual and tabular comparison highlights the disparity between the training and testing datasets. While the FIA data offers comprehensive spatial and ecological representation suitable for robust model training, the TLS data is sparse and geographically limited. As such, TLS is better suited for targeted validation rather than holistic evaluation. Its limited ECOSUBCD diversity and low density reduce its effectiveness as a true test set for broad generalization.

- **Class Distribution Challenges in Real-World Data:** While the models performed well overall, certain underrepresented classes such as *Deciduous broadleaf* or rare species

were less accurately predicted. This reflects the natural skew in ecological datasets and is more a product of biological rarity and field sampling distribution than a shortcoming of the modeling process.

- **Strong Predictions for Dominant Classes, Less Resolution for Rare Ones:** The model's high accuracy in dominant labels such as *Evergreen conifer* or *Pinus* aligns with their ecological and statistical prominence. Conversely, rare or niche taxa were predicted with lower resolution, which is an expected outcome in biodiversity modeling without oversampling or external augmentation.
- **Overprediction Reflects Ecological Dominance:** Instances of overprediction for species such as *lambertiana* or genus *Pseudotsuga* likely stem from their distinctive structural signatures and ecological abundance. These predictions are consistent with their dominance in the landscape and dataset, rather than model bias.
- **Site-Level Evaluation Limited to PFT:** Regional validation was conducted only for PFT predictions, due to availability of ecological overlays and manageable class sizes. Extending such evaluations to genus and species levels would require additional expert annotation or localized ecological metadata, which were outside the scope of this project.
- **Rare Class Generalization Beyond Current Scope:** Generalizing effectively to rare taxa typically requires advanced techniques such as hierarchical smoothing or class-aware loss functions. While promising, these were not implemented in the current study but represent exciting directions for future research.
- **Sampling Constraints from Ground Truth Data:** Some ecological regions, such as M261Ep, had limited representation in the field-collected ground truth data. While this did not adversely affect training, it did limit post-hoc analysis and evaluation for those specific zones.

Question 6 : What strategies or techniques could improve your current pipeline (e.g., feature engineering, additional data sources, advanced models)?

While the current pipeline demonstrates high predictive accuracy across classification levels, several strategies could further enhance its robustness, ecological generalizability, and rare-class sensitivity:

- **Advanced Class Balancing Techniques:** Despite multiple attempts like SMOTE, class imbalance remains a challenge particularly for rare genera and species. Inspite of trying undersampling and oversampling tehcniques the vast difference in distribution of FIA and TLS would require more sophisticated balancing methods such as focal loss , generative oversampling. These could be explored to better capture underrepresented classes.
- **Dimensionality Reduction:** While correlated features like `BasalA` were manually removed, implementing automated techniques such as PCA (Principal Component Analysis) could help uncover latent feature structures and reduce noise if we had more features in the TLS collected dataset. Since the FIA dataset had a lot of features but could not use those due to lack of those in TLS for model prediction.
- **Incorporation of Temporal Features:** Features capturing temporal patterns can improve the model's ability to distinguish between seasonal functional types like deciduous and evergreen species.
- **Geospatial Modeling Enhancements:** Rather than relying solely on latitude and longitude, integrating geospatial encoding techniques such as geohash, spatial embeddings, or proximity to ecological landmarks could allow models to better contextualize regional differences.
- **Continual Learning Frameworks:** To adapt the model as more field data is collected, exploring online learning or fine-tuning strategies can help the pipeline remain up-to-date without retraining from scratch.

These improvements would not only raise model performance, especially for underrepresented taxa, but also increase its reliability in ecological forecasting and conservation planning.

Question 7 : What new types of data could be included to enhance your model's predictive power?

To push the model's accuracy and better represent distribution across sites and ecological regions further, we could possibly explore incorporating additional data sources that capture other dimensions of ecosystem variation:

- **Multi-spectral and Hyperspectral Imagery:**
  Full-spectrum satellite data, including hyperspectral imaging, can enhance functional type discrimination by capturing subtle differences in canopy chemistry, water content, and structural traits. These datasets, when available, may be used directly or through reduced dimensions (e.g., PCA components) to improve model inputs.
- **Time-Series Remote Sensing:**
  Vegetation indices such as NDVI and EVI, when analyzed over time, provide valuable phenological metrics (e.g., green-up date, peak greenness, duration of season) that are closely tied to functional traits. These temporal signals can improve model generalizability and accuracy, particularly when distinguishing between phenologically distinct vegetation types. Platforms like Google Earth Engine and datasets such as MODIS and Sentinel-2 offer accessible sources for generating these features.
- **Soil and Geology Data:**
  High-resolution digital soil maps (e.g., soil texture, organic matter, pH) and geological data can inform species-environment relationships. Certain functional types exhibit strong affinities to specific soil or parent material types. Including these attributes as input features can enhance model interpretability and ecological relevance.
- **Disturbance and Land-Use History:**
  Incorporating historical land-use and disturbance records (e.g., fire, logging, storm events) can contextualize current vegetation states. Derived features like time since last disturbance or previous land cover class help distinguish successional stages and avoid misclassification of transitional communities.
- **Land Cover and Ecoregion Classifications:**
  Existing vegetation or land cover maps can provide coarse but informative prior knowledge that narrows the range of likely functional types. Similarly, ecoregion labels incorporate broad-scale climatic and biogeographic patterns, helping models generalize across regions and reducing the likelihood of ecologically implausible predictions.