FOREST COVER TYPE PREDICTION
Chaos-Aware Machine Learning Arquitecture

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Nicolás Martínes Pineda
20241020098
Universidad Distrital Francisco José de Caldas

Anderson Danilo Martínez
20241020107
Universidad Distrital Francisco José de Caldas

Gabriel Esteban Gutiérrez
20221020003
Universidad Distrital Francisco José de Caldas

Jean Paul Contreras
20242020131
Universidad Distrital Francisco José de Caldas

Environment & Objectives

Looking forward a tool to predict how seven types of trees should naturally distribute in a comprehend cell of 30x30 meters, we plan on build a reliable machine learning model that learns how variable ecology influents.

We'll build a prediction system that adapts the large data resource around 56 features that combines height, slope, aspect and soil labels.

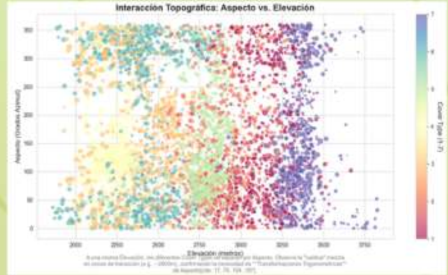
In summary, our main objectives rely on:

- Chaos-aware validation: elevation bands at 2,400 / 2,800 / 3,200 m.
- Uncertainty: aleatoric (entropy) + epistemic (model variance) with threshold amplification.
- Weighted ensemble (RF + XGBoost + LightGBM) for robust, calibrated predictions.

Approach & Planning

Classical ecological linear models fail to capture abrupt regime shifts in montane ecosystems. We adopt ensemble tree methods (RF, XGBoost, LightGBM) + domain feature engineering for robustness and interpretability.

Elevation × aspect coupling produces phase-transition zones (mid-elevation most chaotic). We encode aspect trigonometrically and stratify evaluation by elevation bands (2,400 / 2,800 / 3,200 m).



Data: Roosevelt NF – 15,120 samples, 56 cartographic features at 30 m resolution. Results are region-bound; multi-region validation reserved for future work.

Winning a competition ≠ deployable system, so we must design for reliability, observability and operational recovery. Operational focus: modular pipeline, experiment tracking (MLflow), low-latency serving (FastAPI/Kubernetes), and continuous drift monitoring (PSI, KL).

Gap proposal: prior work rarely treats regime-specific chaos as a design requirement. We look forward to contribute with chaos-aware pipeline + dual uncertainty + elevation-banded validation to manage sensitivity.

Methodology & Progress

<u>Requirements Specification</u>:

Functional requirements include: multi-class prediction (7 classes) with full probability vector; domain-aware features (sin/cos aspect, elevation bands, soil consolidation); uncertainty quantification (aleatoric, epistemic, combined); monitoring & drift detection; dual serving modes (real-time REST + batch GeoTIFF export). Non-functional requirements emphasize modularity, loose-coupling, high cohesion, interpretability, reproducibility (MLflow), availability and MTTR targets.

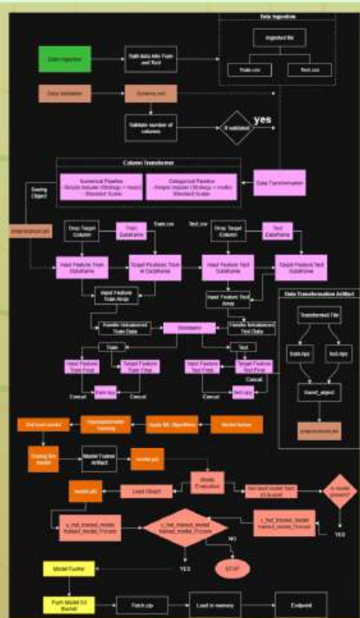
<u>Systems Analysis Summary</u>:

The data mix continuous topo inputs and many rare soil categories – mountains create fragile zones where tiny changes can flip labels, 56 features (topography, distances, 40 soil one-hots with 73% sparsity, wilderness binaries). Vulnerabilities: soil sparsity, threshold chaos (±50m around 2,400/2,800/3,200 m), geographic/time brittleness, aspect-elevation nonlinear coupling

<u>High-Level Architecture Design</u>:

The architecture follows separation of concerns, loose coupling (data contracts), high cohesion, scalability and maintainability. A modular, auditable pipeline from raw GIS to predictions, designed so any layer can be replaced without breaking the system, these layers are:

- 1) Data Ingestion
- 2) Validation
- 3)Feature engineering {sin/cos aspect, elevation bins, soil consolidation}
- 4) Model training {spatially blocked 5-fold CV, Optuna, weighted ensemble}
- 5) Prediction & uncertainty {aleatoric entropy + epistemic variance, threshold amplification}
- 6) Monitoring {PSI, KL, band-wise metrics}
- 7) Serving {FastAPI/Kubernetes, real-time & batch}



FOREST COVER TYPE PREDICTION
Chaos-Aware Machine Learning Arquitecture

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Nicolás Martínes Pineda
20241020098
Universidad Distrital Francisco José de Caldas

Anderson Danilo Martínez
20241020107
Universidad Distrital Francisco José de Caldas

Gabriel Esteban Gutiérrez
20221020003
Universidad Distrital Francisco José de Caldas

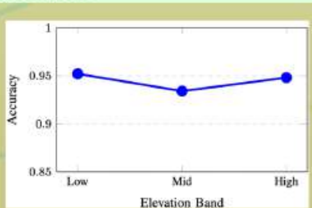
Jean Paul Contreras
20242020131
Universidad Distrital Francisco José de Caldas

Validation and Experimentation Plan:
We validated not only overall accuracy but also resilience: blocked CV prevents leakage across elevation regimes, planning on stress-test the model with realistic noise/drift scenarios, and we will quantify uncertainty so the system can defer to humans where predictions are unreliable

Results

Baseline and Ensemble Performance: Weighted ensemble (RF + XGBoost + LightGBM) – Accuracy 95.2%; ensemble improves calibration vs single models

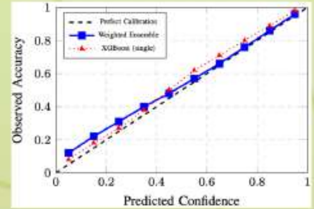
Band-Wise Performance Analysis: shows mid-elevation (2,400–2,800 m) as the most error-prone regime – motivates banded validation and uncertainty amplification.



Threshold Proximity Effects: proximity policy: ×2 uncertainty amplification in ±50 m windows → 19.6% of patches flagged for manual review.

Robustness Under Perturbations: perturbation tests (±25 m elevation, ±5° aspect, 5% noise) produce ⩽3 pp accuracy degradation – meets acceptance criteria.

Uncertainty Calibration: weighted ensemble yields near-diagonal reliability – probabilities align with observed accuracy



Overall Findings & Interpretations: Chaos-aware ensemble delivers high accuracy (95.2%), reliable calibration and perturbation robustness. ~19.6% of area flagged for manual review – supports a human-in-the-loop deployment model

Conclusions & Future Work

Mid-elevation transition zones drive most errors; uncertainty amplification and banded validation reduce silent failures and enable an auditable manual-review workflow.

Combining ecological domain knowledge with systems engineering produced interpretable, maintainable predictions ready for operational validation.

A chaos-aware, production-oriented pipeline (domain features + weighted ensemble + uncertainty & monitoring) achieves 95.2% accuracy while providing auditable decision support in transition zones.

Next steps: test the approach in other regions, improve uncertainty modeling, and integrate live sensors and retraining to handle seasonal and climate-driven shifts