Systems Analysis of Kaggle's Forest Cover Type Prediction: Elements, Relationships, Sensitivity, and Chaos

Nicolás Martínez Pineda*
20241020098
Universidad Distrital Francisco José de Caldas
Anderson Danilo Martínez Bonilla†
20241020107
Universidad Distrital Francisco José de Caldas
Gabriel Esteban Gutiérrez Calderón‡
20221020003
Universidad Distrital Francisco José de Caldas
Jean Paul Contreras Talero§
20242020131
Universidad Distrital Francisco José de Caldas

***Abstract*—The prediction of forest cover type is a fundamental problem in environmental informatics, as accurate classification of ecological regions supports sustainable land management and conservation strategies. In this workshop, we propose the design and implementation of a machine learning model that leverages the Kaggle dataset *Forest Cover Type Prediction*, applying systems engineering principles such as sensitivity analysis, complexity mapping, and architectural modeling to address the competition challenge. The results of our analysis reveal both the predictive capacity of the proposed model and the systemic insights gained from exploring inter-element relationships, variability, and chaotic behaviors, which together provide a stronger foundation for decision-making in real-world applications.**

I. Introduction

Machine learning competitions such as Kaggle's *Forest Cover Type Prediction* offer a practical context for addressing real-world classification problems through data-driven approaches. The dataset consists of cartographic variables (elevation, slope, aspect, soil type, and wilderness area) collected from the Roosevelt National Forest in Colorado, United States. The primary task is to predict the type of forest cover from seven possible classes, based on these environmental features. This context is particularly relevant for systems analysis, as the problem encapsulates both structured data and ecological complexity, providing a testbed for systemic modeling, sensitivity evaluation, and predictive performance. In practice, Kaggle competitions formalize the problem via clear data schemas, submission formats, and evaluation metrics, which directly shape model design and validation protocols [2]. Moreover, modern experimentation workflows commonly rely on cloud object storage (e.g., S3) for scalable, durable datasets and artifact management throughout the lifecycle of training and evaluation [3].

From a systems perspective, several challenges arise when attempting to model this problem. First, the dataset is highly dimensional, combining continuous and categorical variables, which increases the risk of overfitting and complicates feature interactions. To curb the curse of dimensionality and expose useful interactions, wrapper-based feature selection can iteratively search subsets guided by model performance, improving generalization and interpretability [1]. Second, the natural environment introduces nonlinear dynamics and feedback loops, aligning with the principles of chaos theory, where small changes in soil or slope can generate disproportionately large impacts on vegetation type. Third, competition constraints such as accuracy benchmarks and computational limits push us to consider efficiency and robustness in the design of our solution; these constraints are further reinforced by the standardized pipelines and evaluation rules defined by the platform [2]. Together, these challenges highlight the importance of systems thinking in structuring the problem, identifying dependencies, and mapping flows of information within the dataset.

Prior work in ecological modeling and environmental prediction has often relied on traditional statistical methods such as logistic regression and decision trees. However, these methods may struggle to capture the nonlinearities and complex feature interactions present in large-scale environmental data. Recent research, supported by the Kaggle community, has shown the effectiveness of ensemble methods, such as Random Forests and Gradient Boosting Machines, in improving predictive accuracy while mitigating overfitting. Within such pipelines, wrapper methods complement feature-importance analyses by refining the input space used by ensembles and sensitivity studies [1]. In addition, advances in sensitivity analysis provide tools to quantify the impact of input variability on model outcomes, making these approaches particularly suitable for ecological applications; cloud-backed storage patterns (e.g., S3) further facilitate reproducible cross-validation and large-scale ablations by decoupling compute from data [3]. By incorporating these prior developments into a systems engineering framework—and adhering to competition-defined protocols for data handling and evaluation [2]—this workshop aims to go beyond pure prediction and provide a holistic analysis of the forest cover classification problem, linking technical performance to systemic insights.

II. Systems Analysis Report

A. Systemic Analysis

Core Objective: To develop and validate a robust supervised machine learning model for the multi-class classification of forest cover type across the Roosevelt National Forest. The primary output is a high-resolution predictive map at a 30m × 30m spatial resolution, leveraging an extensive suite of cartographic and derived topographical features as input variables.

Data ecosystem structure:

Input Layer: 54 cartographic features divided into numerical (10), binary wilderness (4) and soil type category (40)

Numerical Topographic Features (10): Elevation (1859 – 3858 meters): Master ecological driver. Aspect (0 – 60 degrees): Slope orientation and solar exposure. Slope (0–40 degrees): Terrain steepness affecting drainage and stability. Horizontal_Distance_To_Hydrology: Distance to the nearest water source. Vertical_Distance_To_Hidrology: Elevation difference relative to water. Horizontal_Distance_To_Roadways: Distance to the nearest road. Horizontal_Distance_To_Fire_Points: Distance to historical fire locations. Hillshade_9am: Solar illumination index at 9 AM. Hillshade_Noon: Solar illumination index at noon. Hillshade_3pm: Solar illumination index at 3 PM.

Binary Wilderness Area Indicators (4 features)
Wilderness_Area1–4: One-hot encoded administrative zones (mutually exclusive)

Binary Soil Type Indicators (40 features)
Spil_Type1–40: One hot encoded soil classification, characterized by extreme sparsity (most of the categories have zero or very few occurrences)

Processing Layer: It is designed to transform raw cartographic inputs into predictive insights through supervised learning techniques. As a primary function, it needs to be applied to a multi-class classification algorithm that can accurately map the 54 input features to one of the seven forest cover types in the dataset. The key characteristics of this layer are: the task definition (the multi-class classification in the 7 target categories), the input generated by the 54 features previously described, the output (the prediction cover type label for each 30m x 30m forest patch), the challenge (to manage the feature heterogeneity between continuous vs sparse categorical) and the algorithmics approaches.

Output Layer: It delivers the final prediction generated by the classification pipeline. For each 30m x 30m forest patch in the Roosevelt National Forest, the system assigns species classification corresponding to one of the seven cover types defined in the dataset: Spruce/Fir (Type 1) – High-elevation conifers; Lodgepole Pine (Type 2) – Fire-adapted mid-elevation species; Ponderosa Pine (Type 3) – Dry montane forests; Cottonwood/Willow (Type 4) – Riparian, water-dependent species; Aspen (Type 5) – Deciduous species in montane regions; Douglas-fir (Type 6) – Mixed conifer zones at higher elevations; Krummholz (Type 7) – Alpine tree line formations. This layer also is going to translate the abstract patterns identified in the processing layer into ecological outcomes (species labels) providing a direct ecological interpretation that aligns with the real-world forest management practices.

Element Relationships and Interdependencies:

Hierarchical Variable Architecture: The predictive system is structured by a hierarchical organization of variables, reflecting both ecological principles and cartographic feature design. This architecture ensures that enviromental drivers are being represented at multiple levels of abstraction, from fundamental topographic forces to a categorical classification.

Master Enviromental Variables: These variables represent the primary ecological drivers of species distribution in mountainous ecosystems. They also establish the foundation upon which all other interactions occur. Elevation: As the dominant factor shaping vegetation patterns through altitudinal zonation. Determines temperature regimes, precipitation levels, and ecological transitions from foothill to alpine zones. Aspect: The directional orientation of slopes, modulating solar radiation and influencing moisture retention, snowmelt timing and microclimate gradients. Slope: Degree of terrain steepness, controlling drainage pattern, soil stability and habitat accessibility.

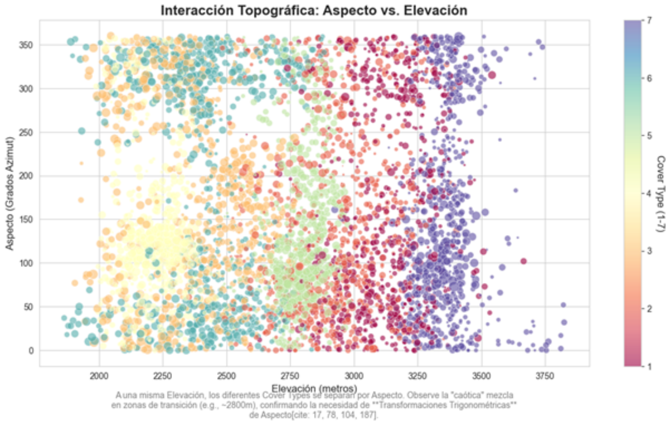
Derived Environmental Indicators: These are secondary variables, derived from the landscape structure, which capture more localized or dynamic environmental conditions.

Hillshade Variables (9 am, Noon, 3pm): Measures of solar illumination at different times of the day. Provide insight into diurnal energy input and shading, which directly affect vegetation growth and competition.

Distance Metrics: Horizontal/Vertical Distance to Hydrology: Represents accessibility to water resources, crucial for riparian species. Distance to roadways: Proxy for human accessibility and potential disturbances. Distance to fire points: Encodes historical disturbance regimes, which influence successions dynamics and fire-adapted species distribution.

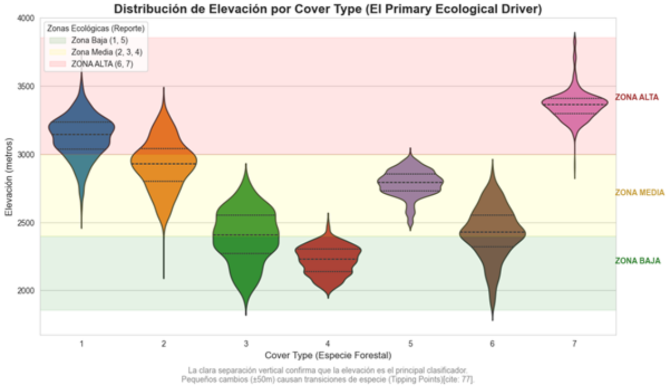
Categorical classification systems: These variables encode administrative and pedological constraints, represented as sparse binary indicators. Wilderness Areas (4 types): Mutually exclusive administrative zones, reflecting differences in land use restrictions and disturbance management. Soil Types (40 categories): An extremely sparse distribution, where most categories have limited or zero presence in the dataset. Soil composition imposes critical constraints on root systems, water retention, and nutrient availability.

Ecological Relationship Patterns: Figure 1 and Figure 2 shows a coupling between elevation and species distribution, reflecting the altitudinal zonation typical of mountain ecosys-



A una misma Elevación, los diferentes Cover Types se separan por Aspecto. Observe la "caótica" mezcla en zonas de transición (e.g., ~2800m), confirmando la necesidad de **Transformaciones Trigonométricas** de Aspecto[cite: 17, 78, 104, 187].

Fig. 1: Topographic interaction: Aspect vs Elevation with species distribution.



La clara separación vertical confirma que la elevación es el principal clasificador.
Pequeños cambios (±50m) causan transiciones de especie (Tipping Points)[cite: 77].

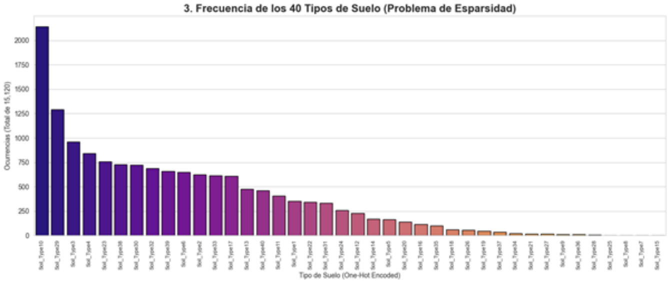
Fig. 2: Elevation vs Cover Type classified by zone.

tems where different types of species coexist at heights that function as critical points, such as 2,800 meters.

Figure 1: Topographic interaction: Aspect vs Elevation with species distribution.

Figure 2: Elevation vs Cover Type classified by zone. Low elevation zone (1859 – 2400m): Dominated by Spruce/Fir (Type 1) and Aspen (Type 5) Aspect influences moisture balance—north-facing slopes retain cooler, wetter microclimates favoring conifers, while south-facing slopes expose Aspens to more sunlight and drier conditions. Mid elevation zone (2400 – 3000m): Hosting Lodgepole Pine (Type 2), Ponderosa Pine (Type 3), and Cottonwood/Willow (Type 4). Here, aspect creates fine-grained ecological mosaics: Lodgepole Pine tolerates multiple orientations but thrives in fire-prone areas. Ponderosa Pine dominates sun-exposed, drier slopes. Cottonwood/Willow remain confined to riparian aspects with better hydrology access. High elevation zone (3000 – 3858m): Encompassing Douglas-fir (Type 6) and Krummholz (Type 7). Aspect here modulates survival under harsh alpine conditions—east/west orientations receive balanced solar input, while extreme north/south exposures determine growth limits at the tree line.

Aspect-Elevation Interactions: The Figure 1 illustrates that at the same elevation, cover types separate along the aspects dimension: As mentioned above, the 2800m, transitional "chaotic" mixing occurs, where aspect becomes the decisive factor for species identity. This confirms the necessity of trigonometric transformations (sin/cos) for aspect, since the circular nature of the variable (0° = 360°) encodes critical ecological transitions.



Spatial Network Effects: In addition to elevation and aspect, the distance-based metrics still refine distribution: Hydrology proximity aligns with riparian species (Type 4). Road accessibility relates to disturbance-adapted pines (Types 2, 3). Fire point distance modulates fire-adapted succession patterns (notably Lodgepole Pine).

B. Complexity & Sensitivity

For this point we focus mainly on identifying and explaining the risks, limitations and areas of uncertainty of the system that will influence the results of our model. We will not describe the data; we will diagnose your vulnerabilities.

Data Architecture Constraint Analysis

Temporal Brittleness: Being the single point time dataset, it eliminates capture of dynamic succession, variations by seasons and how it adapts to climate change. The model will have a measure of annual variations in precipitation and temperature.

Geographic Overfitting Risk: The dataset is limited to the Roosevelt National Forest. This creates the risk that the model is too close to the specific ecology of the Colorado Frontward and does not generalize to other forest systems.

Complexity of the Feature Space

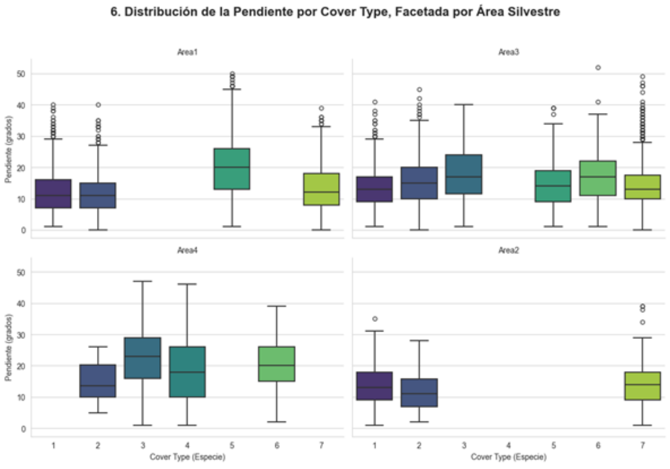
Sparsity Problem of Soil Type: The most important point. The 40 binary soil type variables show extreme scarcity, indicating that many occurrences less than or equal to zero. This has the potential to introduce noise into the model and make it more susceptible to noise in limited areas.

Curse of Dimensionality: The mixture of 56 features (numeric, binary, etc.) complicates the selection of algorithms and increases computational complexity.

Ecological and Algorithmic Sensitivity

Hyperparameter sensitivity: In such a complex system, the algorithms will be fragile. Small adjustments in regularization or learning rate can lead to large differences in performance, especially when trying not to over-adjust to artificial balance.

Ecological Transition Sensitivity: The system is chaotic in peripheral or transitional areas, such as the mid-mountain. It is necessary to indicate that small changes of 50 meters in elevation can cause a total transformation of species.



C. *Chaos and Randomness*

Non-Linear System Dynamics: The forest ecosystem represented in the dataset reveals chaotic behaviors and non-linear dependencies, consisten with complex adaptive systems. These dynamics manifest as tipping points, feedback loops, and emergent patterns that amplify the difficulty of prediction while enriching ecological realism.

Elevation Threshold Effects:

Sharp Transitions: Cover types shift abruptly at specific altitudinal thresholds (e.g., 2400m and 3000m), producing ecological “phase transitions”.

Hysteresis Effects: At identical elevations, species composition differs by slope orientation—north-facing slopes (cooler, moister) support conifers, while south-facing slopes (drier) support deciduous species.

Critical Transition Zones: Around 2800m, small fluctuations in elevation (± 50m) can trigger complete community replacement, illustrating sensitive dependence on initial conditions.

Aspect-Elevation Interaction Chaos:

Non-Linear Coupling: Aspect magnifies elevation effects, producing chaotic species mixing in transition zones where deterministic rules blur.

Butterfly Effect Manifestations: Minute differences in aspect or slope angle generate disproportionately large shifts in species distribution, reflecting chaotic sensitivity.

Emergent Zonation Patterns: Complex three-dimensional ecological boundaries arise from relatively simple topographic drivers (elevation, aspect, slope), defying linear predictability.

Soil-Vegetation Feedback Loops

The soil type features, while sparse, may encode recursive ecological processes: Pedogenesis-Vegetation Coupling: Three species actively shape soil chemistry, influencing nutrient cycling and erosion. Recursive Habitat Modification: Once established, vegetation alters microclimatic and soil conditions, reinforcing its own dominance. Path-Dependent Succession: Disturbance histories (e.g., fire events) leave irreversible soil-vegetation imprints, producing legacy effects that drive divergence even under similar topographic conditions.

Unforeseen Interaction Effects

Feature Interaction Emergences

1. Distance Variable Coupling Chaos

The interaction between horizontal and vertical distances to hydrology illustrates highly non-linear and chaotic behaviors: Multiplicative Effects: Cover type boundaries emerge not along single variables but through joint combinations. As shown in Figure 5, species distributions follow diagonal gradients (e.g., small vertical + medium horizontal distances) rather than clean thresholds, confirming multiplicative coupling. Accessibility Paradoxes: In low elevation zones, proximity to water strongly constrains species, while in higher elevations, both very near and very distant points to hydrology harbor overlapping species. This paradox mirrors how road proximity + hydrology distance may generate unexpected habitat niches, especially for disturbance-adapted species. Fire–Hydrology Interactions: The dispersion clouds in mid/high zones suggest historical disturbance overlays—species appear in configurations where fire history interacts chaotically with water availability, producing distributions irreducible to linear models. Tipping Points: Critical thresholds (e.g., vertical hydrology distance ∼0–100m, horizontal ∼200–400m) exhibit sharp mixing of cover types. Small environmental changes (±50m) may flip classifications entirely, a hallmark of chaotic phase transitions. Fractal Boundaries: The fragmented clustering patterns across all zones imply scale-dependent classification: patterns appear stable at broad resolution but dissolve into recursive complexity upon closer inspection.

2. Hillshade Temporal Complexity

Diurnal Energy Balance Effects: Morning, noon, and evening illumination values create microclimate gradients that non-linearly interact with slope and aspect. Seasonal Amplification: Fixed daily snapshots risk underrepresenting seasonal growth patterns, amplifying randomness in classification. Aspect–Illumination Feedback: Orientation-driven solar exposure reinforces aspect-driven species sorting, introducing feedback loops akin to chaotic oscillators.

Stochastic Elements and Random Components

1. Sampling Randomness Propagation

Artificial Balancing Effects: The perfectly stratified dataset introduces non-natural attractors, creating distortions in ecological realism. Spatial Sampling Patterns: Grid-based sampling may inflate autocorrelation, while random alternatives would yield greater noise. Measurement Error Cascades: Small hydrology/elevation measurement errors propagate non-linearly, especially near tipping zones, amplifying ecological classification uncertainty.

2. Model Training Chaos

Initialization Sensitivity: Neural networks may converge to drastically different boundaries depending on initial weights. Bootstrap Aggregation Variability: Ensemble variance clouds emerge from stochastic resampling, producing shifting decision boundaries. Hyperparameter Optimization Chaos: Non-convex search spaces yield multiple optima, with models converging to divergent but equally valid ecological interpretations.

Competition-Specific Randomness

Leaderboard Dynamics and System Feedback

Public–Private Split Chaos: Overfitting to the public leaderboard inflates ranking volatility, leading to surprise reversals on private data. Collective Intelligence Chaos: Participant behavior introduces emergent optimization strategies, with knowledge sharing accelerating path-dependent trajectories.

Feature Engineering Path Dependencies

Innovation Cascades: Early feature transformations (e.g., trigonometry aspect, binning elevation) bias the trajectory of later optimization. Knowledge Transfer Effects: Public kernels create convergent solutions, reducing diversity but accelerating local optima discovery. Ensemble Complexity Evolution: As feature engineering progresses, model stacking increases sophistication but reduces interpretability, creating a chaotic balance between performance and explainability.

Chaos Theory Applications

Strange Attractors in Feature Space: Artificial balance may create non-natural convergence points, skewing model learning toward artificial attractors rather than ecological truths. Fractal Boundary Conditions: Ecological tipping zones (e.g., hydrology distance thresholds, elevation transitions) exhibit fractal properties, with self-similar patterns repeating at multiple resolutions. Classification accuracy therefore becomes scale-dependent, with recursive boundary complexity intensifying at finer scales.

D. Conclusions

Elevation is the primary driver of the system, establishing a hierarchical architecture by creating three different climatic zones. This zoning is confirmed by the clear segregation of species distributions, altitude acting as a determining threshold for secondary environmental variables.

The Aspect-Elevation interaction is highly chaotic and nonlinear. The circular nature of the Aspect (where 359 is adjacent to 1) cannot be interpreted by models linearly, leading to a loss of critical information about solar orientation (North vs. South) on slopes.

The 40 variables coded for Soil Type with one-hot show extreme sparsity. A high proportion of these features shows very low or non-existent occurrence counts, which causes dimensionality to increase without providing a reliable predictive signal.

An important vulnerability for operational implementation is the artificial balance of classes and geographical limitation to a single area (Roosevelt National Forest). The model has the potential to over-conform to the particular circumstances of the Colorado Front Range, which could cause it to lose its ability to generalize.

The flowchart outlines a structured machine learning pipeline tailored for solving the Forest Cover Type Prediction challenge on Kaggle. This competition involves predicting the type of forest cover based on cartographic variables derived from US Forest Service data. The dataset includes both numerical and categorical features, such as elevation, slope, soil type, and wilderness area, with the target variable being a categorical label from 1 to 7 representing different forest types.

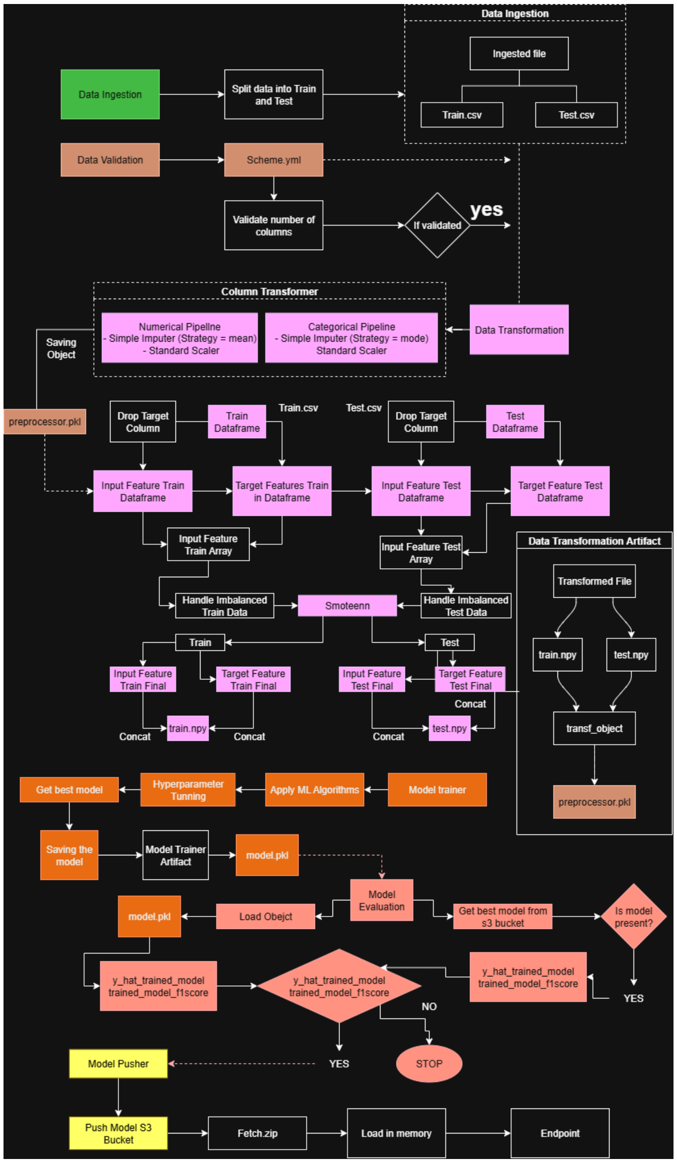


Fig. 3: Flow chart of the Kaggle's competition.

The first stage in the flowchart is data acquisition, which refers to downloading and loading the dataset provided by Kaggle. This includes both the training and test sets. The training set contains labeled data, which is essential for supervised learning, while the test set is used for evaluating model performance on unseen data. This step ensures that the raw data is available for further processing and analysis.

Next comes data preprocessing, a crucial phase where the raw data is cleaned and transformed. Although the competition dataset is relatively clean, preprocessing may still involve checking for anomalies, encoding categorical variables like `Soil_Type` and `Wilderness_Area` using one-hot encoding, and scaling numerical features such as elevation and distances. Feature engineering might also be applied here to create new informative features or interactions that could improve model performance.

Following preprocessing, the flowchart moves into exploratory data analysis (EDA). This step involves visualizing feature distributions, examining correlations between variables, and identifying patterns that could inform model selection. For instance, elevation might show strong separation between forest types, while certain soil types may be exclusive to specific covers. EDA also helps detect class imbalance, which is important for choosing appropriate evaluation metrics and resampling techniques.

The model selection phase is where different algorithms are considered. Given the nature of the dataset, tree-based models like Random Forests and Gradient Boosting (e.g., XGBoost, LightGBM) are particularly effective due to their ability to handle mixed data types and capture non-linear relationships. These models also offer feature importance metrics, which can guide further feature selection and interpretation. Simpler models like k-Nearest Neighbors or more complex ones like neural networks may also be explored depending on performance trade-offs.

Once a model is chosen, the flowchart proceeds to model training and validation. This involves splitting the training data into training and validation sets, tuning hyperparameters, and evaluating performance using metrics like accuracy or F1-score. Cross-validation may be employed to ensure robustness and prevent overfitting. The goal here is to find a model that generalizes well to unseen data.

Finally, the pipeline concludes with prediction and submission, where the trained model is applied to the test set to generate predictions. These predictions are formatted according to Kaggle's submission requirements and uploaded to the competition platform for scoring. This step marks the culmination of the workflow, translating all prior efforts into a measurable outcome.

References

[1] V. Verma, "A comprehensive guide to Feature Selection using Wrapper methods in Python," *Analytics Vidhya*, Oct. 15, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/. [Accessed: Sep. 27, 2025].

[2] Kaggle, "Competitions Setup Documentation." [Online]. Available: https://www.kaggle.com/docs/competitions-setup. [Accessed: Sep. 27, 2025].

[3] H. Golas, "S3 Storage: How It Works, Use Cases and Tutorial," *Cloudian Blog*, Apr. 12, 2021. [Online]. Available: https://cloudian.com/blog/s3-storage-behind-the-scenes/. [Accessed: Sep. 27, 2025].