

Milestone -2

Data Preprocessing & ML Training

1. Introduction

The objective of the **ExoHabitAI** project is to predict the potential habitability of exoplanets using data from the NASA TESS mission. This document outlines the theoretical basis for the data preprocessing decisions and the Machine Learning (ML) strategies employed to classify planets as "Potentially Habitable" or "Non-Habitable."

2. Theoretical Framework: Data Preprocessing

Raw astronomical data is inherently noisy, sparse, and scientifically complex. Preprocessing transforms this raw data into a structured format suitable for statistical learning.

2.1 Handling Missing Data (Imputation Theory)

- **The Challenge:** TESS detects planets via the *Transit Method* (measuring radius), but mass measurements require *Radial Velocity* follow-up, which is resource-intensive. Consequently, ~50% of the dataset lacks mass values.
- **Methodology:** We utilized **Median Imputation** for numerical features.
- **Theoretical Justification:** Astronomical distributions (like planetary mass) are often non-Gaussian and heavy-tailed (e.g., many small planets, few massive Jupiters). The mean is highly sensitive to outliers, whereas the median provides a robust estimator of central tendency for skewed distributions.

2.2 Feature Engineering (Domain-Driven Transformation)

ML models require features that correlate with the target variable. We engineered three key synthetic features based on astrobiological principles:

1. Habitability Score (Similarity Index):

- *Theory:* Based on the *Earth Similarity Index (ESI)* concept. It calculates the Euclidean distance of a planet's properties (Radius, Flux, Temperature) from Earth's values in a normalized vector space.
- *Mathematical Formulation:* $H = \frac{1}{n} \sum e^{-|x_i - x_{\text{earth}}|}$. This provides a continuous gradient of "Earth-likeness" rather than a binary label.

2. Stellar Compatibility:

- *Theory:* Spectral class determines the habitability window. G-type stars (like the Sun) offer stable luminosity. M-dwarfs are prone to flares, and O/B stars have short lifespans. We modeled this as a Gaussian function peaking at $T_{\text{eff}} \approx 5778 \text{ K}$.

3. Orbital Stability Score:

- *Theory:* Derived from Kepler's Third Law This feature captures the relationship between orbital period and distance, helping the model distinguish between stable orbits and eccentric/unstable ones.

2.3 Handling Class Imbalance

- **The Problem:** Habitable planets are rare anomalies (<0.5% of the dataset). A standard model maximizing accuracy would simply predict "Non-Habitable" for every instance, achieving 99.5% accuracy but 0% Recall.
- **Solution:** We employed **Cost-Sensitive Learning**.
 - **Class Weights:** We assigned a higher penalty to misclassifying the minority class (Habitable).
 - *Weight Formula:* $W_{\text{pos}} = \frac{N_{\text{total}}}{2 \times N_{\text{pos}}}$. This forces the algorithm to treat missing a habitable planet as a significant error.

3. Theoretical Framework: Machine Learning Models

3.1 Problem Formulation

We framed this as a **Binary Classification** problem:

- **Class 1 (Positive):** Candidate Habitable (Radius $\leq 2.5 R_{\oplus}$ AND In Habitable Zone).
- **Class 0 (Negative):** Non-Habitable (Gas Giants, Too Hot, Too Cold).

3.2 Baseline Model: Logistic Regression

- **Role:** Establishes a performance baseline.
- **Theory:** A linear classifier that models the probability of a class $P(Y=1|X)$ using the sigmoid function
- **Limitation:** Assumes a linear decision boundary, which is likely insufficient for the complex, non-linear boundaries of the Habitable Zone (e.g., the "Goldilocks" zone is a band, not a line).

3.3 Primary Model 1: Random Forest Classifier

- **Theory:** An ensemble learning method constructing multiple Decision Trees.
- **Why Selected:**
 - **Non-Linearity:** Can model complex boundaries (like the specific temperature range required for liquid water).
 - **Robustness:** Averaging multiple trees reduces variance and the risk of overfitting to noisy astronomical data.
 - **Feature Importance:** Provides interpretability by measuring the decrease in impurity (Gini importance) for each feature.

3.4 Primary Model 2: XGBoost (eXtreme Gradient Boosting)

- **Theory:** A boosting algorithm that builds trees sequentially. Each new tree corrects the errors of the previous ones.
- **Why Selected:**
 - **Handling Imbalance:** The `scale_pos_weight` parameter allows specific tuning for rare event detection.
 - **Regularization:** Includes L1/L2 regularization to prevent overfitting on small datasets.

4. Evaluation Methodology

4.1 Metrics Selection

Accuracy is a misleading metric for this project due to the imbalance. We prioritized:

- **Recall (Sensitivity):** The fraction of *actual* habitable planets correctly identified.
 - *Justification:* In exoplanet science, a False Positive (flagging a barren rock as habitable) is acceptable, but a False Negative (missing Earth 2.0) is a critical failure.
- **F1-Score:** The harmonic mean of Precision and Recall, ensuring the model doesn't just predict "Habitable" for everything to get high Recall.
- **ROC-AUC:** Measures the model's ability to distinguish between classes across all probability thresholds.

4.2 Feature Importance Analysis

To ensure the model learns physics and not noise, we analyzed feature contributions.

- **Expected Outcome:** The model should identify **Equilibrium Temperature** and **Insolation Flux** as top predictors, confirming it has learned the definition of the Habitable Zone.

5. Conclusion

The combination of domain-specific feature engineering (Habitability Score) and cost-sensitive learning (XGBoost/Random Forest) provides a robust framework for identifying rare habitable worlds in the TESS dataset. This approach moves beyond simple filtering, allowing for probabilistic ranking of candidates based on their physical similarity to Earth.