

ExoHabitAI: Milestone2-Documentation

A Multi-Stage Framework for Predicting Exoplanetary Habitability

Refer for more detailed Documentation – [Link](#)

Part 1: The Data Engineering Pipeline (Milestones 1.1 – 1.6)

Milestone 1.1: Data Architecture & Gold-Standard Curation

- **Objective:** Transition from raw observational noise to a reliable planetary catalog by applying a "Trust Hierarchy".
- **Scientific Logic:** Filters entries for `default_flag == 1` (primary records), `soltype == 'Published Confirmed'` (peer-reviewed), and `pl_controv_flag == 0` (uncontroversial existence).
- **Outcome:** Reduces 39,000 rows to ~4,355 "Gold-Standard" records, ensuring the AI learns from absolute certainties rather than disputed noise.

Milestone 1.2: Intelligent Null Recovery & Reference Alignment

- **Objective:** Perform non-destructive "Scientific Data Recovery" to fill sparse rows.
- **Scientific Logic:** Uses a targeted update method (`overwrite=False`) to cross-reference multiple NASA sub-tables.
- **Outcome:** Recovers missing physical parameters from secondary peer-reviewed sources without overwriting primary "Gold-Standard" observations.

Milestone 1.3: Physics-Based Imputation & Empirical Synthesis

- **Objective:** Achieve data completeness by applying Astrophysical Laws rather than statistical guesses.
- **Scientific Logic:**
 - **Kepler's Third Law.**
 - **Mass-Radius Power Law.**
- **Outcome:** A "Physics-Complete" dataset where every planet obeys the fundamental laws of celestial mechanics.

Milestone 1.4: Statistical Consolidation & Final ML-Ready Synthesis

- **Objective:** Reach absolute zero-null density across all 263+ features.
- **Scientific Logic:** Applies **Median Imputation** as a final statistical fallback. Median is chosen over Mean to prevent outliers (like extreme Super-Jupiters) from shifting the dataset's physical distribution.

- **Outcome:** A dense feature matrix with **Exactly 0 Nulls**, ready for matrix multiplications in high-performance algorithms.

Milestone 1.5: Physical Sanitization & Statistical Standardization

- **Objective:** Validate physical feasibility and remove anomalies.
- **Scientific Logic:**
 - **Constraint Filtering:** Removes "Impossible Worlds" (e.g., negative mass, eccentricity).
 - **Outlier Capping (IQR):** Clips extreme values using IQR to prevent "Super-Giant" stars from distorting model gradients.
- **Outcome:** A sanitized dataset where every row adheres to thermodynamics and orbital mechanics.

Milestone 1.6: Habitability Indexing & Feature Engineering

- **Objective:** Synthesize a continuous "Habitability Score" using a 4-Pillar Physics Model.
- **Scientific Logic:** Uses a **Geometric Mean** of Gaussian scores for Radius, Flux, and Temperature.

Outcome: Creates the supervised learning target. Planets with a score > 0.7 are identified as the "True Gems" for training.

Part 2: The Machine Learning Pipeline (Milestones 2.1 – 2.4)

Milestone 2.1: Advanced Balancing & Feature Engineering

- **Objective:** Solve the "Needle in a Haystack" problem (Class Imbalance).
- **Scientific Logic:**
 - **Leakage Control:** Drops pillars like `pl_rade` to force the AI to use indirect stellar clues.
 - **SMOTE-ENN:** Generates synthetic "Gems" and removes noisy overlaps to create a clear decision boundary.
- **Outcome:** A perfectly balanced dataset (Class 0: ~4,300, Class 1: ~4,300).

Milestone 2.2: Baseline Benchmarking & Reference Performance

- **Objective:** Establish a "Null Hypothesis" using Logistic Regression and Shallow Decision Trees.
- **Scientific Logic:** Establishes the "floor" for performance, initially showing a **Recall of 0.50** (missing 50% of habitable worlds).
- **Outcome:** Justifies the necessity for more complex, non-linear ensemble models.

Milestone 2.3: Primary Model Implementation & High-Recall Optimization

- **Objective:** Deploy **Random Forest** and **XGBoost** to capture multi-dimensional physics patterns.
- **Scientific Logic:** Optimizes for the **F2-Score**, prioritizing the discovery of habitable planets (Recall) over avoiding false alarms (Precision).
- **Outcome:** Significant increase in detection sensitivity and the first set of serialized production models (.pkl).

Milestone 2.4: Meta-Ensemble Synthesis & Threshold Optimization

- **Objective:** Achieve "Extreme Sensitivity" using a **Stacking Classifier** and **Dynamic Threshold Optimization**.
- **Scientific Logic:**
 - **Meta-Learner:** A Logistic Regression "Referee" learns when to trust XGBoost vs. Random Forest.
 - **Threshold Shift:** Moves the decision boundary to **0.0034** to capture 100% of targets.
- **Outcome:** **Final Metrics: 100% Recall and 95% Overall Accuracy.**

Final Performance Summary

Metric	Category 0 (Non-Habitable)	Category 1 (Habitable)	Total Model Performance
Recall	0.95 (95%)	1.00 (100%)	No Habitable Planets Missed
Precision	1.00	0.10	Focused Discovery List
Accuracy	—	—	95% Final Accuracy