ExoHabitAI

Milestone 1 - Data Preprocessing & Feature Engineering

Name: Satya Sri Dheeraj M
Date: 18 January 2026

1. Objective

The objective of this milestone is to transform raw NASA TESS exoplanet data into a **clean, machine-learning–ready dataset** for habitability analysis. The preprocessing pipeline focuses on data validation, physics-aware imputation, and feature engineering while maintaining scientific interpretability.

---

2. Selected Core Features

Only features directly relevant to habitability assessment were retained.

Planetary Features

- `pl_orbper` – Orbital period (days)
- `pl_orbsmax` – Semi-major axis (AU)
- `pl_rade` – Planet radius (Earth radii)
- `pl_bmasse` – Planet mass (Earth masses)
- `pl_orbeccen` – Orbital eccentricity
- `pl_insol` – Insolation flux (Earth flux)
- `pl_eqt` – Equilibrium temperature (K)

Stellar Features

- `st_teff` – Stellar effective temperature (K)
- `st_rad` – Stellar radius (Solar radii)
- `st_mass` – Stellar mass (Solar masses)
- `st_met` – Stellar metallicity [Fe/H]
- `st_logg` – Stellar surface gravity

Identifiers

- `pl_name` – Planet identifier
- `disc_year` – Discovery year

---

3. Data Cleaning Pipeline

3.1 Critical Feature Filtering

Observations missing any of the following essential parameters were removed:

- Planet radius
- Insolation flux
- Equilibrium temperature
- Stellar temperature

These parameters are required for any meaningful habitability evaluation.

3.2 Physical Validation

All retained values were validated against physically realistic ranges:

- Planet radius: 0.1–30 Earth radii
- Planet mass: 0.01–13,000 Earth masses
- Orbital eccentricity: 0–1
- Insolation flux: > 0
- Stellar temperature: 2,000–10,000 K

Rows violating physical constraints were removed.

---

4. Physics-Based Imputation

Missing values were handled using astrophysical relationships rather than purely statistical methods.

4.1 Semi-Major Axis Estimation

For planets missing orbital distance, Kepler's Third Law was applied:

$$a = (M_\star \cdot P^2)^{1/3}$$

where:

- a = semi-major axis (AU)
- M⋆ = stellar mass (Solar masses)
- P = orbital period (years)

Remaining gaps were filled using period-stratified median values.

4.2 Planetary Mass Estimation

Planetary mass was estimated from radius using empirically derived mass–radius relationships:

- Rocky planets: $M \propto R^{3.7}$
- Neptune-like planets: $M \propto R^{2.06}$
- Gas giants: scaled relative to Jupiter

4.3 Statistical Imputation

Lower-impact features were filled using median imputation:

- Orbital eccentricity
- Stellar metallicity
- Stellar surface gravity

---

5. Feature Engineering

5.1 Derived Physical Features

The following features were computed:

- **Planet density**: M / R^3
- **Escape velocity proxy**: $\sqrt{(M / R)}$
- **Tidal heating indicator**: e / a^3
- **Flux variation**: Insolation × eccentricity (or) pl_insol×pl_orbeccen

These features enhance physical interpretability and model learning.

5.2 Categorical Features

Stellar type: M, K, G, F, Other (based on temperature)

Planet type: rocky, super-Earth, Neptune-like, Jupiter-like (based on radius)

All categorical features were one-hot encoded.

---

6. Habitability Labels

Three habitability definitions were implemented:

Zone	Purpose
Conservative	Near-Earth analogs
Standard	**ML target label**
Optimistic	Exploratory analysis

The **Standard Habitable Zone** label (`habitable_candidate`) is used for supervised learning.

---

7. Final Dataset Summary

Metric	Value
Initial rows	39,212
Final rows	12,894
Retention rate	32.88%
Total features	34
Data completeness	99.99%
Standard HZ positives	92 (0.71%)

---

8. Scope Boundary

This milestone does **not** include Feature scaling, Class balancing, Model training and Explainability analysis