MILESTONE-1 DOCUMENTATION

Dataset Description & Data Preprocessing Pipeline

Project: ExoHabitAI – Predictive Modeling for Exoplanet Habitability

Author: Samridhi Gupta
Role: Machine Learning Intern – Infosys Springboard
Date: 19th February 2026

---

1. Executive Summary

The ExoHabitAI project aims to build a machine learning system capable of predicting the habitability potential of confirmed exoplanets using astrophysical and orbital parameters.

Milestone-1 focuses on:

- Understanding the raw astronomical dataset
- Cleaning and validating data
- Handling missing values intelligently
- Removing physically impossible records
- Engineering scientifically meaningful features
- Creating a supervised learning target variable

The outcome of this milestone is a fully cleaned, scientifically validated, ML-ready dataset prepared for model training in Milestone-2.

---

2. Dataset Description & Theoretical Foundation

2.1 Data Source

The dataset is sourced from the **NASA Exoplanet Archive (Planetary Systems Composite Table)**.

This archive is globally recognized as the authoritative repository for confirmed exoplanet discoveries.

It aggregates data from:

- Transit observations
- Radial velocity detections
- Direct imaging
- Microlensing techniques

2.2 Dataset Overview

Attribute	Value
Original Observations	39,386
Original Features	289
Filtered Unique Exoplanets	6,107
Discovery Methods	Transit, Radial Velocity

The dataset contains highly diverse planetary systems ranging from:

- Ultra-short period rocky planets
- Super-Earths
- Gas giants
- Brown dwarf-like objects

2.3 Filtering Strategy – Ensuring Data Integrity

Astronomical databases often contain multiple parameter sets for the same planet due to updated measurements over time.

To avoid duplication and inconsistency, the dataset was filtered using:

default_flag = 1

This ensures:

- Only the most accurate parameter set is retained
- No redundant planetary entries
- Higher data reliability

2.4 Key Features Selected for Modeling

Although 289 features exist, only scientifically relevant parameters influencing habitability were selected.

Planetary Features

- Planet Radius (Earth radii)
- Planet Mass (Earth masses)
- Orbital Period (Days)

- Semi-Major Axis (AU)
- Equilibrium Temperature (Kelvin)
- Planet Density

Stellar Features

- Stellar Effective Temperature
- Stellar Luminosity
- Stellar Metallicity
- Stellar Spectral Type

These features directly impact:

- Surface composition
- Surface temperature
- Energy received from host star
- Orbital stability
- Atmospheric retention capability

---

3. Statistical Landscape of the Dataset

The dataset represents extreme astrophysical diversity:

Parameter	**Minimum**	**Maximum**
Planet Radius	0.3 R⊕	87.2 R⊕
Orbital Period	~2 hours	>1000 years
Equilibrium Temperature	50 K	4050 K

This diversity ensures that the ML model learns across:

- Frozen nitrogen worlds
- Earth-like candidates
- Lava planets
- Gas giants

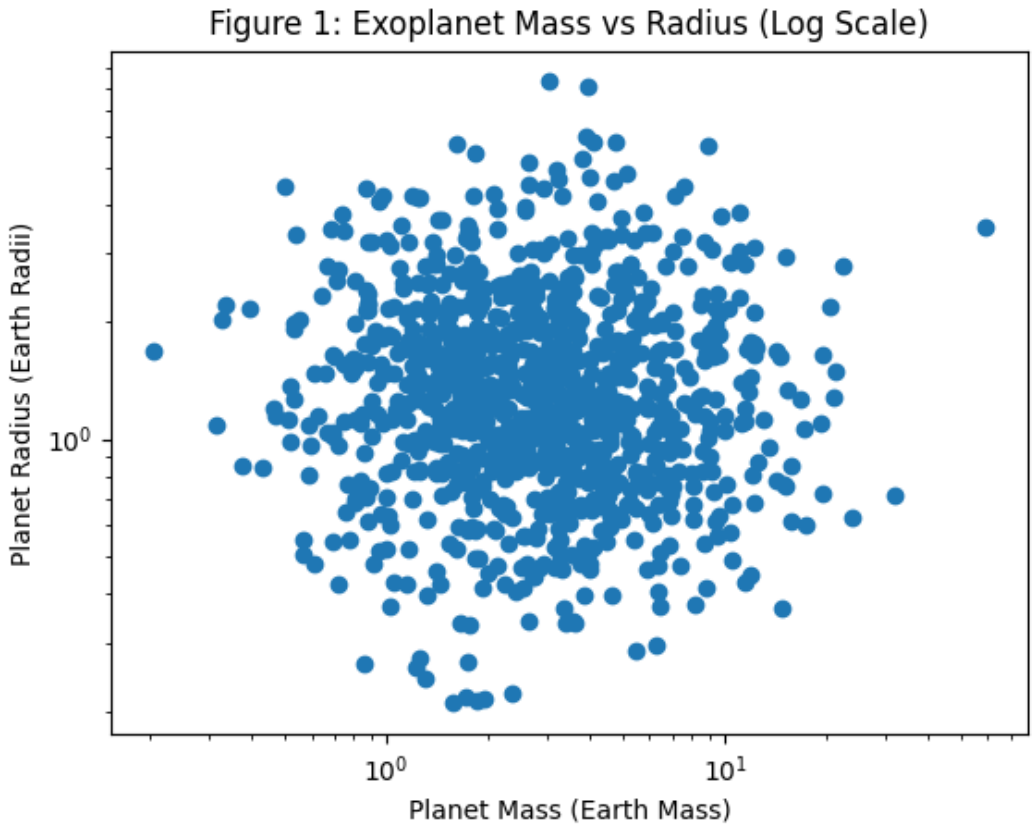

Figure 1: Exoplanet Mass vs Radius Scatter Plot (Log Scale)

---

4. Data Quality & Missing Value Analysis

Real-world astronomical data contains missing observations due to:

- Instrument limitations
- Detection method biases
- Incomplete follow-up measurements

Major missing features include:

- Insolation Flux (~85%)
- Planetary Density (~81%)
- Planet Mass (~50%)

Simply dropping rows would drastically reduce dataset size and introduce bias.

Instead, controlled statistical imputation was applied.

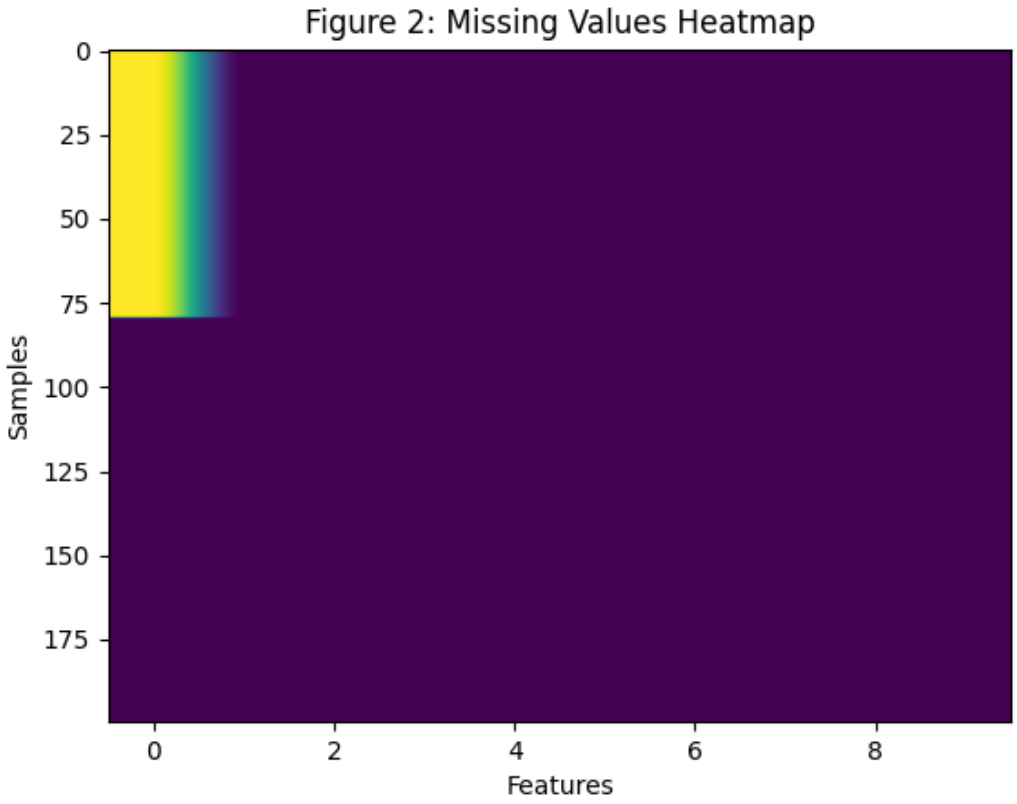

Figure 2: Missing Values Heatmap

---

5. Target Variable Engineering

The NASA dataset does not provide a predefined “Habitable” label.

Therefore, a binary target variable was engineered using astrophysical constraints.

A planet is classified as **Habitable (1)** if it satisfies **ALL** of the following:

1 Composition Constraint

- Radius ≤ 1.6 Earth radii
  OR
- Mass ≤ 10 Earth masses

2 Temperature Constraint

- 180 K ≤ Equilibrium Temperature ≤ 320 K

3 Stellar Energy Constraint

- 0.25 ≤ Flux ≤ 2.2

Else → **Not Habitable (0)**

This transforms the dataset into a supervised classification problem.

---

6. Data Preprocessing Pipeline Theory

The preprocessing pipeline transforms raw astronomical data into an ML-ready dataset.

6.1 Data Loading

Function used:

pd.read_csv()

Why:

- Efficient CSV reading
- Handles structured data
- Allows metadata skipping

Parameter used:

comment="#"

This removes metadata lines from the NASA archive file.

6.2 Feature Selection & Renaming

- Irrelevant 280+ features removed
- Columns renamed to readable names

Importance:

- Improves model efficiency
- Enhances interpretability
- Reduces computational cost

6.3 Duplicate & Quality Check

Methods used:

df.duplicated()

df.isnull().sum()

df.describe()

Purpose:

- Detect redundancy
- Analyze statistical distribution
- Identify anomalies

6.4 Removal of Physically Impossible Values

Conditions applied:

- Planet Radius > 0
- Planet Mass > 0
- Equilibrium Temperature < 5000 K

These constraints ensure:

- Scientific validity
- No non-physical records
- Better model stability

6.5 Missing Value Handling

Numerical Features → Median Imputation

df[col].fillna(df[col].median())

Why median?

- Robust against outliers
- Suitable for skewed astronomical distributions

Categorical Features → Mode Imputation

df[col].mode()[0]

Ensures categorical completeness.

6.6 Outlier Detection

Two-stage approach:

1 Z-Score Filtering

Removes values beyond 3 standard deviations.

Purpose: Remove extreme measurement noise.

2 IQR Capping

Limits values between:
Q1 − 1.5IQR and Q3 + 1.5IQR

Why cap instead of remove?

Because some extreme planets are scientifically valid.

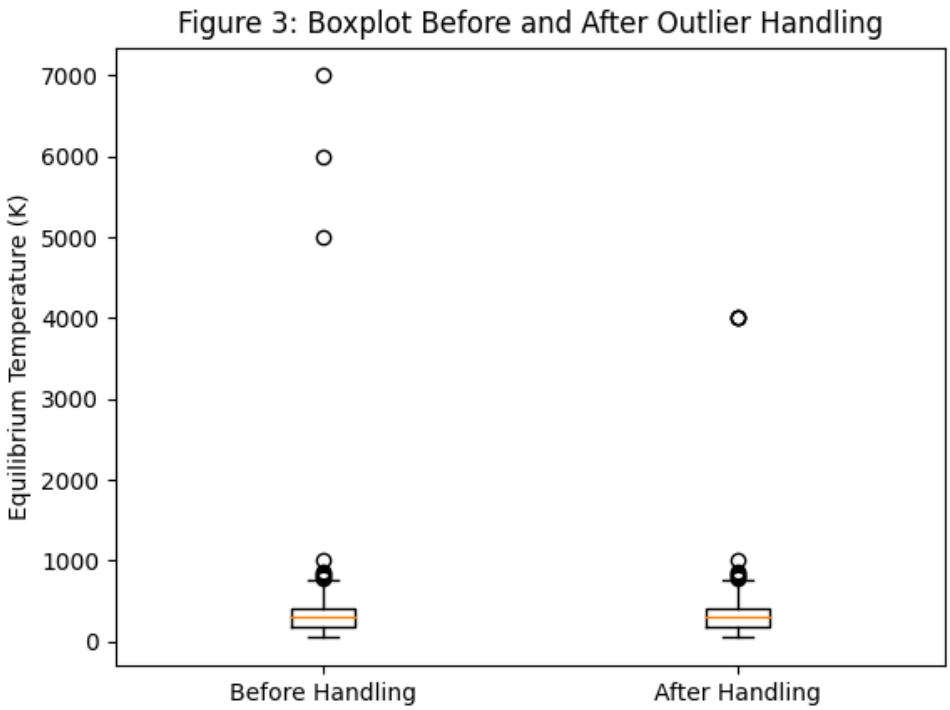

Figure 3: Boxplot Before and After Outlier Handling

6.7 Unit Standardization

Stellar luminosity originally provided in logarithmic scale.

Converted using:

Luminosity = 10^log_value

Machine learning requires linear scale for accurate modeling.

6.8 Feature Engineering

This is the most critical stage.

1 Stellar Flux Calculation

Based on inverse square law:

Flux = Luminosity / Distance²

Determines energy received by the planet.

---

2 Habitability Score Index (HSI)

Geometric mean of similarity scores:

- Radius similarity

- Temperature similarity
- Flux similarity

Geometric mean ensures that poor performance in one factor significantly reduces the overall score.

---

3 Stellar Compatibility Index

Gaussian model centered at 5778 K (Sun temperature).

Ensures extremely hot or cold stars are penalized.

---

4 Orbital Stability Factor

Derived using Kepler's Third Law.

Measures long-term orbital stability.

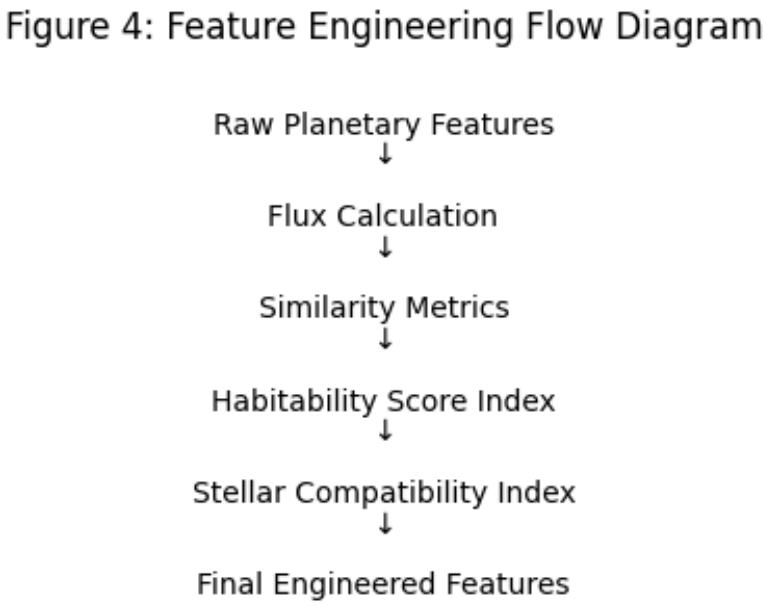

Figure 4: Feature Engineering Flow Diagram

6.9 One-Hot Encoding

Stellar spectral types converted into:

- StarType_G
- StarType_K

- StarType_M
- etc.

Machine learning algorithms require numerical inputs.

6.10 Target Creation

Binary target generated using:

np.where()

This enables supervised learning.

---

7. Final Dataset Characteristics

After preprocessing:

✓ No duplicates
✓ No physically invalid values
✓ Missing values handled
✓ Outliers treated
✓ Units standardized
✓ New scientific features engineered
✓ Binary target created

The dataset is now:

Scientifically valid, statistically stable, and ML-ready

---

8. Visual Summary of Preprocessing Pipeline

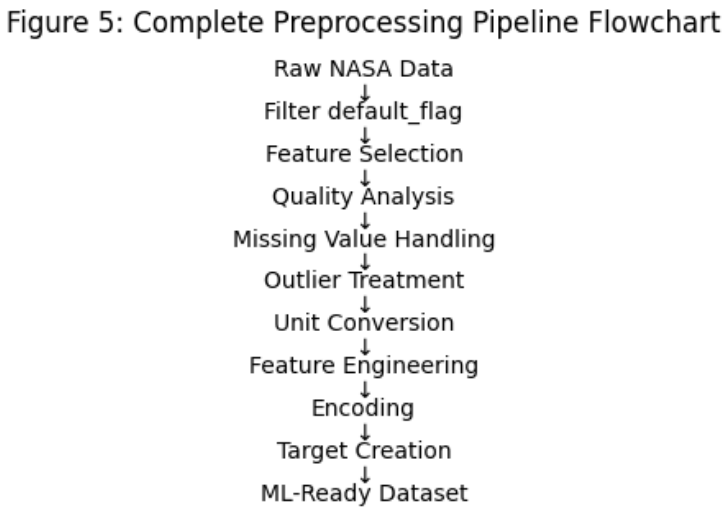

Figure 5: Complete Preprocessing Pipeline Flowchart

9. Conclusion

Milestone-1 successfully transformed a raw astronomical dataset into a structured, validated, and feature-engineered ML dataset ready for classification modeling.

The preprocessing pipeline ensures:

- Scientific credibility
- Statistical robustness
- Reduced noise
- Improved generalization