MILESTONE – 1 DOCUMENTATION

Dataset Description & Data Preprocessing framework

Project: Predicting the Habitability of Exoplanets Using Machine Learning

1. Project Overview

The objective of **ExoHabitAI** is to design a machine learning framework capable of assessing whether a confirmed exoplanet has the potential to support life.

Milestone-1 concentrates on preparing a scientifically reliable dataset by:

- Studying raw astronomical records
- Eliminating redundant and invalid observations
- Treating missing values without bias
- Removing non-physical entries
- Engineering astrophysically meaningful features
- Creating a supervised binary target variable

---

2. Dataset Background & Scientific Context

2.1 Data Source

The dataset originates from the globally recognized
NASA Exoplanet Archive
specifically the *Planetary Systems Composite Table*

The archive compiles confirmed exoplanet discoveries using methods such as:

- Transit Photometry
- Radial Velocity
- Direct Imaging
- Microlensing

2.2 Dataset Snapshot

Metric	Value
Raw Observations	39,386
Total Attributes	289
Unique Confirmed Planets (Filtered)	6,107
Major Detection Methods	Transit, Radial Velocity

The dataset includes diverse planetary systems ranging from:

- Compact rocky planets
- Super-Earths
- Gas giants
- Brown dwarf–like objects

2.3 Data Filtering Strategy

Astronomical datasets often store multiple parameter revisions for the same planet.
To ensure consistency, filtering was performed using:

```
Default_flag = 1
```

This guarantees:

- Retention of the most reliable measurement
- Elimination of redundant entries
- Increased scientific consistency

---

3. Selected Modeling Features

Out of 289 attributes, only parameters directly influencing habitability were retained.

Planetary Parameters

- Planet Radius (R⊕)
- Planet Mass (M⊕)
- Orbital Period
- Semi-Major Axis
- Equilibrium Temperature
- Planetary Density

Stellar Parameters

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

These factors influence:

- Surface conditions
- Atmospheric retention
- Energy exposure
- Orbital mechanics
- Chemical composition

---

4. Statistical Characteristics of the Data

The dataset reflects extreme astrophysical variability.

Parameter	Minimum	Maximum
Radius	0.3 R⊕	87.2 R⊕
Orbital Period	~2 hrs	>1000 yrs
Equilibrium Temp	50 K	4050 K

This broad range ensures the ML model learns from:

- Ice worlds
- Earth analogs
- Lava planets
- Massive gas giants

---

5. Data Quality Assessment

Astronomical data often contains incomplete records due to:

- Observational limitations
- Instrument sensitivity
- Detection biases

High missing percentages were observed in:

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

Dropping such rows would reduce dataset reliability.
Hence, structured imputation methods were applied.

---

6. Preprocessing Methodology

The transformation pipeline converts raw astrophysical measurements into structured ML-ready data.

6.1 Data Loading

```
pd.read_csv()
```

Parameter used:

```
comment="#"
```

Reason:
Removes metadata comments present in NASA files.

6.2 Feature Reduction & Renaming

- Removed ~280 irrelevant columns
- Renamed selected attributes for clarity

Benefits:

- Reduced dimensionality
- Improved readability
- Lower computational overhead

6.3 Duplicate & Integrity Checks

Functions used:

```
df.duplicated()
df.isnull().sum()
df.describe()
```

Purpose:

- Detect repeated planets
- Evaluate missing distributions
- Identify extreme anomalies

6.4 Removal of Non-Physical Records

Applied logical constraints:

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

This prevents unphysical or corrupted measurements from affecting training.

6.5 Missing Value Treatment

Numerical Columns → Median Imputation

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

Why median?

- Handles skewed astronomical data
- Less sensitive to outliers

Categorical Columns → Mode Imputation

```
df[col].mode()[0]
```

Ensures complete categorical representation.

6.6 Outlier Management

Two-step strategy:

Z-Score Filtering

Removes extreme noise beyond $\pm 3\sigma$.

IQR Capping

Bounds values within:

```
Q1 − 1.5IQR  to  Q3 + 1.5IQR
```

Capping preserves rare but scientifically valid planets.

6.7 Unit Normalization

Stellar luminosity originally stored in logarithmic scale.

Converted using:

```
Luminosity = 10^log_luminosity
```

Machine learning models require linear relationships for accurate feature interaction.

---

7. Feature Engineering

This phase introduces scientifically informed derived features.

7.1 Stellar Flux

Using inverse square law:

```
Flux = Luminosity / (Distance²)
```

Determines stellar energy reaching the planet.

7.2 Habitability Similarity Index (HSI)

Computed as geometric mean of similarity metrics:

- Radius similarity
- Temperature similarity
- Flux similarity

Geometric mean ensures poor performance in one metric strongly lowers the overall score.

7.3 Stellar Compatibility Index

Modeled as Gaussian centered at 5778 K (solar temperature).

Penalizes stars that are excessively hot or cold.

7.4 Orbital Stability Indicator

Derived from Kepler's Third Law to measure long-term orbital sustainability.

---

8. Categorical Encoding

Stellar spectral types transformed via One-Hot Encoding:

- StarType_G
- StarType_K
- StarType_M
- etc.

Ensures numerical compatibility for ML algorithms.

---

9. Target Variable Construction

Since no predefined habitability label exists, a binary variable was created.

A planet is labeled **Habitable (1)** if:

- Radius ≤ 1.6 $R_\oplus$ OR Mass ≤ 10 $M_\oplus$
- 180 K ≤ Temperature ≤ 320 K
- 0.25 ≤ Flux ≤ 2.2

Else → **Not Habitable (0)**

Implemented using:

```
np.where()
```

This converts the problem into a supervised classification task.

---

10. Final Dataset State

After preprocessing:

✔ Duplicates removed
✔ Missing values handled
✔ Outliers treated
✔ Units standardized
✔ Engineered scientific indicators added
✔ Binary classification target created

The dataset is now:

Scientifically consistent, statistically balanced, and ready for ML model development.

---

11. Conclusion

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

The structured preprocessing ensures:

- Physical realism
- Statistical robustness
- Reduced observational noise
- Improved predictive generalization