

Milestone 1: Exoplanet Dataset Understanding and Analysis

Dataset Overview

- **Source:** NASA Exoplanet Archive
- **Format:** CSV
- **Nature of Data:** Astronomical observational data
- **Records:** Each row represents a confirmed exoplanet
- **Scope:** Includes planetary, orbital, and stellar parameters

The dataset is research-oriented and not preprocessed for machine learning tasks. It contains raw observational values, uncertainty ranges, metadata, and missing entries, which require significant preprocessing.

Structure of the Dataset

The dataset consists of a large number of columns that can be broadly categorized into the following groups:

Planetary Physical Parameters

These attributes describe the physical nature of the exoplanet and are critical for habitability analysis.

Examples include:

- Planet radius (relative to Earth)
- Planet mass (relative to Earth)
- Planet density
- Equilibrium temperature
- Orbital period

Planet size and temperature are primary indicators of whether a planet could retain an atmosphere and support liquid water.

Orbital and Environmental Parameters

These features define how the planet interacts with its host star.

Examples include:

- Distance from the host star

- Semi-major axis
- Stellar energy received (insolation)
- Orbital eccentricity

Even Earth-sized planets may be uninhabitable if they orbit too close or too far from their star. Orbital parameters strongly influence surface temperature.

Host Star Characteristics

These attributes describe the star around which the planet orbits.

Examples include:

- Star temperature
- Star mass and radius
- Luminosity
- Metallicity
- Spectral type

Habitability depends not only on the planet but also on the star. Cooler stars (such as K and M-type stars) often have habitable zones closer to the star.

Discovery and Metadata Fields

These fields provide contextual information.

Examples include:

- Discovery method
- Discovery year
- Telescope or mission used
- Data quality flags

These features are useful for analysis and filtering but do not directly influence habitability prediction and are not suitable for ML model input.

Data Quality Observations

While exploring the dataset, the following issues were identified:

- **Missing Values:**
Many key planetary parameters such as mass and density are missing for a large number of planets.
- **Inconsistent Coverage:**
Not all planets have complete stellar or orbital data.

- **Duplicate Records:**
Some host stars have multiple planets, and some planets may appear multiple times due to updated measurements.
- **Non-ML-Friendly Format:**
The dataset includes commented headers, uncertainty bounds, and categorical values that require encoding.

Absence of Habitability Labels

An important observation is that the dataset does not contain a predefined habitability label.

- No column directly states whether a planet is habitable.
- Habitability must be derived using scientific assumptions.

This project does not predict confirmed habitability but rather **habitability potential** based on observable proxy variables such as temperature, size, and stellar conditions.

Relevance to Machine Learning

The dataset is suitable for machine learning after preprocessing because:

- It contains quantitative features correlated with habitability.
- Relationships between stellar and planetary features can be learned by models.
- Feature importance analysis can reveal key habitability drivers.

The dataset offers rich scientific information but requires careful cleaning, feature selection, and domain-driven interpretation. The insights gained from this milestone lay the foundation for data preprocessing, habitability score engineering, and machine learning model development in subsequent phases.