

Data Description Report

NASA Exoplanet Archive Dataset

File Name: PS_2026.02.07_05.49.41.csv

1. Introduction

This report presents a detailed description of the NASA Exoplanet Archive dataset used in the ExoHabitAI project. The dataset contains confirmed exoplanet observations along with their corresponding stellar and discovery-related parameters. It serves as the foundational data source for analyzing and predicting exoplanet habitability potential.

2. Dataset Overview

- Source: NASA Exoplanet Archive
- File Format: CSV (Comma-Separated Values)
- Total Records: Approximately 39,315 rows
- Total Features: 289 columns
- Domain: Astrophysics / Exoplanetary Science

Each row represents a reported observation of a confirmed exoplanet, including planetary characteristics, host star properties, and discovery metadata. Multiple entries may exist for the same planet due to values reported by different scientific publications.

3. Feature Categories

3.1 Planetary Parameters

- pl_name – Planet name
- pl_rade – Planet radius (Earth radii)
- pl_bmasse – Planet mass (Earth masses)
- pl_dens – Planet density
- pl_orbper – Orbital period (days)
- pl_orbsmax – Semi-major axis (AU)
- pl_eqt – Equilibrium temperature (Kelvin)
- pl_insol – Insolation flux

3.2 Stellar Parameters

- st_teff – Stellar effective temperature (Kelvin)
- st_mass – Stellar mass (Solar masses)
- st_rad – Stellar radius (Solar radii)

- st_lum – Stellar luminosity

3.3 Discovery Metadata

- disc_method – Discovery method
- disc_year – Year of discovery
- disc_facility – Discovery facility

4. Data Characteristics

- Multiple records per planet due to different literature sources.
- Presence of missing values (NaN) in several physical parameters.
- High dimensionality (289 features) requiring feature selection.
- Includes uncertainty bounds (_err1, _err2) for scientific precision.

5. Relevance to ExoHabitAI Project

The most relevant features for habitability prediction include planetary radius, planetary mass, orbital distance, equilibrium temperature, insolation flux, stellar temperature, and stellar mass. These parameters help determine whether a planet lies within the habitable zone and may support Earth-like conditions.

6. Conclusion

The NASA Exoplanet Archive dataset is a comprehensive astrophysical catalog of confirmed exoplanets. Due to its high dimensionality, duplicate measurements, and missing values, preprocessing steps such as filtering default entries, feature selection, and data cleaning are required before applying machine learning techniques.