DATASET DESCRIPTION DOCUMENT

1. Dataset Name

Planetary Systems (PS) – Confirmed Exoplanet Dataset

2. Dataset Source

The dataset was obtained from the NASA Exoplanet Archive, a publicly accessible and peer-reviewed astronomical database maintained by the California Institute of Technology (Caltech).

Specifically, the data was downloaded from the

Planetary Systems (PS) Table, which contains confirmed exoplanets along with detailed planetary and stellar parameters.

Source Platform: NASA Exoplanet Archive

Table Used: Planetary Systems (PS)

File Name: PS_2026.02.07_06.09.53.csv

Data Type: Confirmed exoplanets only

Access Type: Publicly available scientific dataset

This archive compiles data from peer-reviewed scientific publications and space missions such as Kepler, TESS, and other international observatories.

3. Dataset Overview

This dataset contains detailed information about confirmed exoplanets and their host stars. The attributes include planetary properties (mass, radius, orbital characteristics), stellar properties (temperature, luminosity), and discovery-related parameters.

The dataset is designed to support:

Astronomical research

Statistical analysis of planetary systems

Machine learning–based habitability prediction

Comparative planetology studies

For this project, the dataset is used to analyze exoplanet characteristics and develop a predictive model to estimate their potential habitability.

4. File Format

Format: CSV (Comma-Separated Values)

Encoding: Standard UTF-8

Structure: Tabular format

Each row represents one confirmed exoplanet.

Each column represents a planetary or stellar attribute.

The CSV format enables easy integration with tools such as Python (Pandas), R, Excel, and machine learning frameworks.

5. Dataset Size

Number of Rows: Varies depending on archive version (auto-generated by NASA archive at download time)

Number of Columns: Multiple planetary, orbital, stellar, and detection-related attributes

Since the archive is regularly updated, the dataset size may increase over time as new exoplanets are discovered and confirmed.

6. Important Attributes / Features

The dataset includes numerous attributes. The most relevant features for habitability analysis are categorized below:

A. Planetary Parameters

Planet Name: Official name of the exoplanet

Planet Mass: Mass of the planet (usually in Jupiter or Earth masses)

Planet Radius: Radius of the planet

Orbital Period: Time taken to complete one orbit around its host star

Equilibrium Temperature: Estimated surface temperature assuming black-body radiation

Semi-Major Axis: Average distance between planet and host star

These parameters are crucial for determining whether a planet lies within the habitable zone.

B. Stellar Parameters

Host Star Name: Name of the parent star

Stellar Effective Temperature: Surface temperature of the host star

Stellar Luminosity: Energy output of the star

Stellar Radius: Size of the host star

Stellar characteristics directly influence planetary temperature and potential habitability.-

C. Positional & Distance Parameters

Distance from Earth: Measured in parsecs or light-years

Right Ascension & Declination: Astronomical coordinates

These parameters are useful for observational astronomy and comparative studies.

D. Discovery Parameters

Discovery Method: Technique used to detect the exoplanet (e.g., Transit, Radial Velocity, Imaging)

Discovery Year: Year in which the planet was confirmed

These features help analyze trends in detection technology and dataset bias.

7. Target Variable

The dataset does not contain a predefined “habitability” label.

Instead, habitability will be derived using feature engineering techniques based on:

Planet mass and radius

Equilibrium temperature

Orbital distance

Stellar temperature and luminosity

A habitability score or classification label (Habitable / Potentially Habitable / Non-Habitable) will be generated during preprocessing and modeling.

8. Data Preprocessing Plan

The following preprocessing steps will be applied:

Handling missing values (imputation or removal)

Removing duplicate entries

Feature selection based on scientific relevance

Unit normalization and scaling

Encoding categorical variables (e.g., discovery method)

Creating derived features such as Habitable Zone indicator

These steps ensure the dataset is suitable for machine learning model training.

9. Use of Dataset in the Project

The dataset will be used for:

Exploratory Data Analysis (EDA)

Feature engineering

Training machine learning models

Model evaluation and validation

Ranking exoplanets based on predicted habitability potential

The ultimate goal is to develop an AI-based predictive system that assists in identifying promising candidates for future astronomical study.

10. Limitations of the Dataset

Missing values in some planetary attributes

Observational bias toward larger planets

Measurement uncertainties

No direct habitability ground-truth label

Dataset updates over time may change model reproducibility

These limitations will be addressed during preprocessing and model evaluation.

11. Conclusion

The Planetary Systems (PS) dataset from the NASA Exoplanet Archive provides comprehensive and scientifically validated information about confirmed exoplanets and their host stars.

It forms a reliable foundation for developing a machine learning system aimed at predicting exoplanet habitability and contributing to data-driven astronomical research.