Title: - NASA Exoplanet Dataset Description

Topic: - Predicting the Habitability of Exoplanets Using Machine Learning

1. Introduction

This document outlines the structure and content of a comprehensive dataset consolidating verified information on exoplanets and their stellar hosts. Compiled from the NASA Exoplanet Archive, the dataset offers a rich collection of vetted scientific measurements suitable for in-depth analysis. Its primary applications include supporting astronomical research, statistical modeling, and the development of machine learning algorithms for planetary science.

2. Data Provenance

- **Origin:** NASA Exoplanet Archive
- **Access Point:** `https://exoplanetarchive.ipac.caltech.edu`
- **Extraction Date:** February 13, 2026
- **Core Sample:** Planetary systems identified predominantly by the TESS mission, forming the default archive view.

3. Dataset Dimensions

- **Instances (Rows):** 39,386
- **Features (Columns):** 289

Each row corresponds to a unique planetary object and its association with a specific stellar host.

4. File Specifications

- **Format:** Comma-Separated Values (CSV)

- **Encoding:** UTF-8
- **Structure:** Two-dimensional tabular data with a header row.
- **Metadata:** Lines prefixed with the '#' character provide additional context, constraints, and data type information.

5. Feature Categories

The dataset's attributes are organized into several thematic groups, detailed below.

A. Primary Identifiers
These fields provide unique keys for each planetary system entry.

Field Name	Description
`rowid`	A unique integer identifier for the record.
`pl_name`	The primary designation of the exoplanet.
`hostname`	The commonly used name of the central star.
`pl_letter`	The alphabetical suffix distinguishing the planet (e.g., b, c).
`hd_name`	Identifier from the Henry Draper catalog.
`hip_name`	Identifier from the Hipparcos catalog.
`tic_id`	Identifier from the TESS Input Catalog.

B. Stellar Host Characteristics
These attributes describe the fundamental properties of the stars in the systems.

Field Name	Description
`st_mass`	Mass of the star, measured in solar units.

Field Name	Description
`st_rad`	Radius of the star, measured in solar units.
`st_teff`	Effective temperature of the photosphere, in Kelvin.
`st_lum`	Bolometric luminosity relative to the Sun.
`st_age`	Estimated age of the star, in billions of years.
`st_met`	Metallicity of the star, typically in dex.

C. Planetary Physical Attributes

These fields detail the intrinsic physical state and composition of the exoplanets.

Field Name	Description
`pl_massj`	Planetary mass expressed in Jupiter masses.
`pl_radj`	Planetary radius expressed in Jupiter radii.
`pl_dens`	Bulk density of the planet.
`pl_eqtemp`	Estimated equilibrium temperature.
`pl_grav`	Surface gravity, often given as log10(cm/s^2).

D. Orbital Architecture

These parameters define the geometric and dynamic characteristics of the planetary orbits.

Field Name	Description
`pl_orbper`	Time required to complete one orbit, in days.

Field Name	Description
`pl_orbsmax`	Semi-major axis of the orbit, in Astronomical Units (AU).
`pl_orbeccen`	Eccentricity of the orbit (unitless).
`pl_orbincl`	Inclination of the orbital plane, in degrees.
`pl_orblper`	Argument of periastron (longitude of periastron), in degrees.

E. Discovery Context

This group contains metadata pertaining to the detection and announcement of the planets.

Field Name	Description
`discoverymethod`	The primary technique used to detect the planet (e.g., Transit, R.V.).
`disc_year`	The calendar year of the discovery announcement.
`disc_facility`	The observatory, survey, or instrument responsible for the discovery.
`pl_pubdate`	Publication date of the discovery paper.
`releasedate`	Date the data was released in the archive.

F. Record Reliability Indicators

Fields that help assess the confidence and completeness of a given entry.

Field Name	Description
`default_flag`	A binary flag (1 = Yes) indicating the most reliable representation for a planet.
`pl_nnotes`	The count of informational notes or annotations for this record.

Field Name	Description
`rowupdate`	The timestamp of the most recent modification to the database row.

G. Observational Metadata

Quantitative information on the observations supporting the stellar and planetary parameters.

Field Name	Description
`st_nphot`	Total number of photometric observations archived.
`st_nrvc`	Number of radial velocity measurements archived.
`st_nspec`	Number of spectroscopic observations archived.
`pl_ntranspec`	Count of transmission spectroscopy observations for the planet.

6. Data Completeness

The dataset is not fully dense; fields may contain null (`NaN`) values. These gaps are inherent to astronomical data and occur when specific measurements are unattainable, not yet performed, or have not been published for a given planet or star. Appropriate handling of these missing values is a necessary step in preprocessing.

7. Core Characteristics

- **Typology:** Structured, quantitative scientific data.
- **Origin:** Primarily observational, with derived and modeled parameters.
- **Temporal Coverage:** Encompasses discoveries from the first confirmed exoplanets through early 2026.

- **Maintenance:** The source archive is updated periodically with new discoveries and revisions.
- **Data Quality:** Subject to rigorous peer-review and validation processes before ingestion.

8. Potential Use Cases

- Automated classification of exoplanet types.
- Assessment of planetary habitability potential.
- Predictive modeling of undiscovered planetary properties.
- Large-scale statistical surveys of planetary system architectures.
- Informing observational target selection for future missions.
- Analyzing discovery trends and biases over time.

9. Known Constraints & Considerations

- A subset of parameters are model-dependent estimates rather than direct measurements.
- The sample is subject to observational biases (e.g., detection methods favor large planets close to their stars).
- Data completeness varies significantly across different parameters.
- The archive may contain multiple entries for the same physical planet, reflecting different analysis sources or model assumptions, with the `default_flag` indicating the preferred solution.

10. Summary

This dataset represents a substantial and authoritative compilation of known exoplanets and their host stars, containing over 39,000 entries and nearly 300 distinct attributes. Its breadth and depth make it an invaluable resource for sophisticated scientific inquiry, enabling detailed exploration of exoplanet demographics, stellar-planetary interactions, and the development of advanced predictive models in the field of astronomy.