NASA EXOPLANET ARCHIVE

1. Data Source Overview:

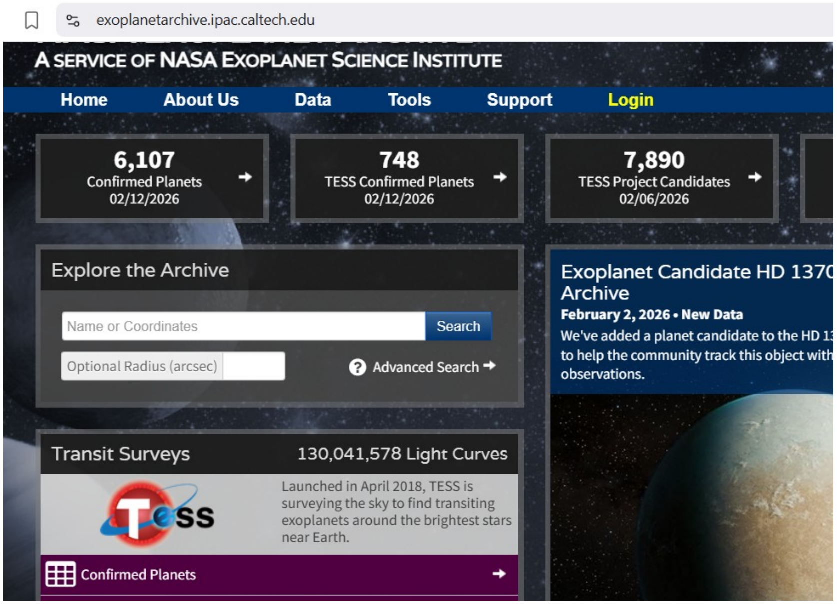
The dataset used in this project was obtained from the **NASA Exoplanet Archive**, an official website maintained by NASA's Exoplanet Science Institute.

The archive provides publicly accessible, continuously updated data on confirmed exoplanets, planetary systems, and detection methods.

From the homepage :

- **6,107 confirmed planets (as of 02/12/2026)**
- **748 TESS confirmed planets**
- **7,890 TESS project candidates**

This confirms that the archive is dynamic and regularly updated.



Exoplanet Candidate HD 137(
Archive

February 2, 2026 • New Data

We've added a planet candidate to the HD 1
to help the community track this object with
observations.

Launched in April 2018, TESS is surveying the sky to find transiting exoplanets around the brightest stars near Earth.

2. Dataset Download and Structure:

The dataset was downloaded from the Planetary Systems table in CSV format.

- **Data records (rows): 39,386**
- **Attributes (columns): 43**

- **Total data cells: 39,386 × 43 = 1,693,598 values**
- **File size: 35 MB**
- **Metadata rows: 295**
- **Header row: 296**

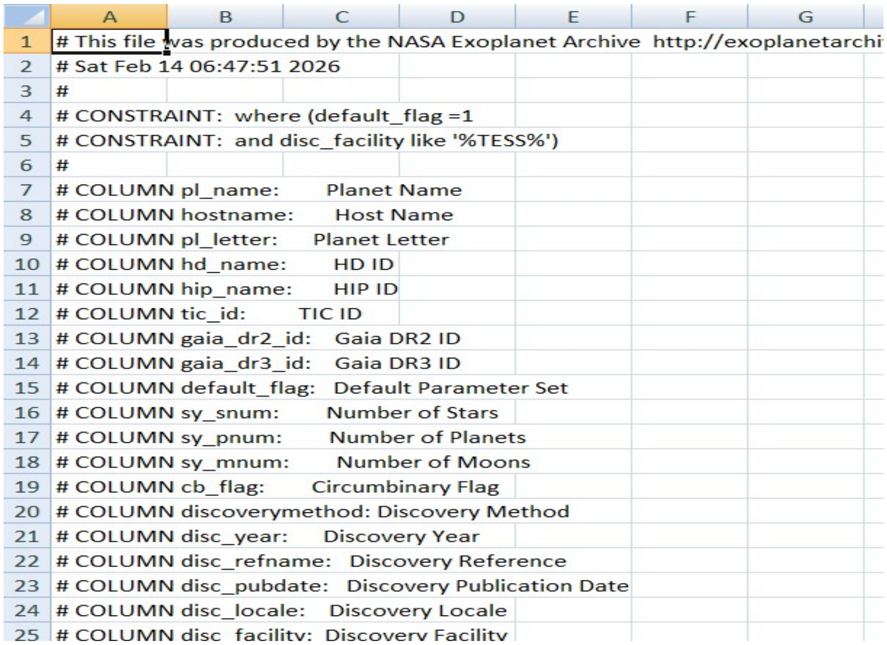
That is complete dataset size information.

Metadata Section (Rows 1–295):

This section includes:

- Download timestamp
- Query constraints
- Column definitions
- Source URL

Each metadata row is prefixed with #.



	A	B	C	D	E	F	G
1	# This file was produced by the NASA Exoplanet Archive http://exoplanetarchi						
2	# Sat Feb 14 06:47:51 2026						
3	#						
4	# CONSTRAINT: where (default_flag =1						
5	# CONSTRAINT: and disc_facility like '%TESS%')						
6	#						
7	# COLUMN pl_name: Planet Name						
8	# COLUMN hostname: Host Name						
9	# COLUMN pl_letter: Planet Letter						
10	# COLUMN hd_name: HD ID						
11	# COLUMN hip_name: HIP ID						
12	# COLUMN tic_id: TIC ID						
13	# COLUMN gaia_dr2_id: Gaia DR2 ID						
14	# COLUMN gaia_dr3_id: Gaia DR3 ID						
15	# COLUMN default_flag: Default Parameter Set						
16	# COLUMN sy_snum: Number of Stars						
17	# COLUMN sy_pnum: Number of Planets						
18	# COLUMN sy_mnum: Number of Moons						
19	# COLUMN cb_flag: Circumbinary Flag						
20	# COLUMN discoverymethod: Discovery Method						
21	# COLUMN disc_year: Discovery Year						
22	# COLUMN disc_refname: Discovery Reference						
23	# COLUMN disc_pubdate: Discovery Publication Date						
24	# COLUMN disc_locale: Discovery Locale						
25	# COLUMN disc_facility: Discovery Facility						

3. Feature Categorization:

Identification Features:

- pl_name
- hostname
- hd_name

- hip_name
- gaia_dr2_id

The below **ten features** were selected based on astrophysical relevance to planetary habitability. Identification fields, observational metadata, and uncertainty parameters were excluded to reduce dimensionality and focus on physically meaningful predictors.

Category	Column Name	Data Type	Unit	Simple Description	Why It Matters for Habitability
Planet	pl_rade	Float	Earth Radii	Planet radius	Determines rocky vs gas planet
Planet	pl_bmasse	Float	Earth Masses	Planet mass	Indicates gravity & atmosphere retention
Planet	pl_dens	Float	g/cm³	Planet density	Identifies rocky composition
Planet	pl_eqt	Float	Kelvin	Equilibrium temperature	Key surface temperature indicator
Planet	pl_orbsmax	Float	AU	Distance from star	Determines habitable zone position
Star	st_teff	Float	Kelvin	Star surface temperature	Affects radiation exposure
Star	st_lum	Float	Solar Luminosity	Star brightness	Defines habitable zone range
Star	st_mass	Float	Solar Masses	Star mass	Influences star lifespan
Star	st_rad	Float	Solar Radii	Star radius	Stellar classification factor
Star	st_met	Float	dex	Stellar metallicity	Indicates rocky planet formation probability

4.Data Quality Assessment:

Advantages of the Dataset:

1. Large Dataset Size

- 39,386 planetary records
- Over 1.69 million data values

Advantage 1:

Sufficient data volume for supervised machine learning.

Advantage 2:

Reduces overfitting risk compared to small datasets.

2. Scientifically Verified Source

- Data maintained by NASA
- Peer-reviewed discovery references included

Advantage 1:

High reliability and authenticity.

Advantage 2:

Suitable for academic and research-level modeling.

3. Structured and Standardized Format

- CSV format
- Clearly defined columns
- Standard astronomical units (Kelvin, AU, Earth Masses)

Advantage 1:

Easy to load into Python (pandas).

Advantage 2:

No unit conversion inconsistencies.

4. Physically Meaningful Features

Includes:

- Planet radius
- Planet mass
- Equilibrium temperature
- Orbital distance
- Stellar temperature and luminosity

Advantage 1:

Features directly relate to habitability physics.

Advantage 2:

Strong foundation for feature engineering.

5. Controlled Parameter Selection

Constraint applied:

```
default_flag = 1
```

Advantage 1:

Removes duplicate parameter solutions.

Advantage 2:

Improves dataset consistency.

Limitations of Dataset :

1. Missing Values

Common missing columns:

- pl_bmasse
- pl_dens
- st_met

Disadvantage 1:

Incomplete mass values prevent density calculation.

Disadvantage 2:

Reduces usable dataset size after cleaning.

2. Observational Bias

Filtered for TESS-discovered planets.

Disadvantage 1:

Transit method dominates dataset.

Disadvantage 2:

Small, close-in planets are overrepresented.

This may bias habitability predictions.

3. Measurement Uncertainty

Many columns include:

- Upper uncertainty (_err1)
- Lower uncertainty (_err2)
- Limit flags (_lim)

Disadvantage 1:

Measurements contain scientific uncertainty.

Disadvantage 2:

Ignoring uncertainty may reduce prediction accuracy.

4. No Direct Habitability Label

Dataset does NOT contain:

- Habitable / Not Habitable column
- Habitability score

Disadvantage 1:

Target variable must be engineered manually.

Disadvantage 2:

Model output depends on defined assumptions.

5. Impact on Machine Learning:

The dataset is large and scientifically reliable, which is good for building machine learning models.

Because it contains 39,386 records and important physical features like radius, mass, temperature, and distance from star, it provides enough information for predicting planetary habitability.

However, there are some challenges:

1. Some important values like planet mass and density are missing in many rows.
   This may reduce the usable dataset after cleaning.
2. The dataset mainly includes TESS-detected planets.
   This creates detection bias, meaning the model may learn patterns specific to transit-detected planets rather than true habitability.
3. There is no direct habitability label in the dataset.
   So the target variable must be created manually using defined conditions.
4. Some columns are unnecessary for modeling (IDs, references, uncertainty values).
   These must be removed to avoid noise.

6.CONCLUSION:

While the dataset demonstrates strong scientific reliability and structural integrity, preprocessing and bias acknowledgment are essential before applying machine learning models for habitability prediction. Overall, the dataset is strong and suitable for machine learning, but model performance will depend on proper cleaning, feature selection, and clear definition of habitability criteria.

---Documentation by:

Anusree .Challa