ExoHabitAI Dataset Description

1. Introduction

- The ExoHabitAI dataset is taken from NASA Exoplanet Archieve.
- The dataset consists of thousands of confirmed exopalnets discovered by TESS (Transiting Exoplanet Survey Satellite) which is NASA's planet hunter.
- Each row represents an individual exoplanet and it contains 39212 rows.
- Each column represents planetary, orbital, steller characteristsics and it contains 289 columns.
- The physical and orbital attributes that directly influence planetary habitability such as radius, mass, equilibrium temperature, orbital distance, and steller luminosity.
- The data is scientific, real and authoritative.

2. Why ExoHabitAI dataset is suitable

- By using this dataset we can predict the habitability potential of exoplanet.
- The dataset contains all the necessary scientific parameters required to decide habitability :
  - Planet size
  - Planet mass
  - Temperature
  - Distance from star
  - Star properties

3. Understanding Attributes (Columns)

- Identification of Attributes :

Column	**Meaning**
pl_name	Planet name
hostname	Host star name
pl_letter	Planet letter (b, c, d…)
hd_name, hip_name	Catalog IDs
tic_id, gaia_dr2_id	Space telescope IDs

- Discovery Information :

Column	**Meaning**
disc_year	Discovery year
disc_method	Transit, Radial Velocity
disc_facility	TESS, Kepler

- Orbital Parameters :

Column	**Meaning**
pl_orbper	Orbital period (days)
pl_orbsmax	Semi-major axis (AU)
pl_orbeccen	Orbital eccentricity
pl_orbincl	Orbital inclination

- Planet Physical Properties :

Column	**Meaning**
pl_rade	Planet radius (Earth radii)
pl_bmasse	Planet mass (Earth mass)
pl_dens	Density
pl_eqt	Equilibrium temperature

- Steller Properties (Host Star) :

Column	**Meaning**
st_teff	Star temperature
st_rad	Star radius
st_mass	Star mass
st_lum	Star luminosity

- Habitability Indicators (Target Features) :

Column	Meaning
pl_insol	Stellar flux received
pl_eqt	Planet equilibrium temp
pl_hzflag	In habitable zone (if available)

4. Nature of the Data

- **Data Type :**
  - It is numerical and continuous data.
  - There are some categorical and missing values.

- **Data Quality :**
  - It is having real scientific measurements.
  - There are some Missing values.

5. Before Using the Dataset :

- Remove identification columns
- Handle missing values
- Select habitability-related features
- Normalize numerical values