

MILESTONE-1 DATA PREPROCESSED DOCUMENT

PART 1: DATA DESCRIPTION

1. Introduction to Dataset

The dataset used in this project is obtained from the NASA Exoplanet Archive. It is a publicly accessible scientific repository that contains validated information about planets discovered outside our solar system.

The archive is maintained by the Infrared Processing and Analysis Center (IPAC) at Caltech under NASA's guidance. The data is curated from peer-reviewed scientific publications and space missions.

This dataset contains information about:

- Confirmed exoplanets
- Orbital properties
- Physical characteristics
- Stellar properties
- Detection methods
- Observational records

2. Dataset Organization

According to the documentation

NASA Exoplanet Archive

, the dataset is structured into multiple logical groups:

A) Identification Information

- Planet name
- Host star name
- Alternative catalog IDs

B) Discovery Information

Column Name Description

discoverymethod Method used to detect the planet

disc_year Year of discovery

disc_facility Observatory or mission used

pl_pubdate Publication date

releasedate Archive release date

C) Planetary Properties

Feature Description Unit

pl_rade Planet Radius Earth Radii

pl_bmasses Planet Mass Earth Masses

pl_orbper Orbital Period Days

pl_orbsmax Semi-Major Axis AU

pl_eqt Equilibrium Temperature Kelvin

pl_dens Planet Density g/cm³

These parameters help classify planets as rocky, gas giants, or habitable candidates.

D) Stellar Characteristics

Feature Description

st_teff Star Effective Temperature

st_lum Stellar Luminosity

st_met Stellar Metallicity

Feature Description

st_spectype Star Spectral Type

Stellar characteristics help determine the habitable zone and radiation environment.

E) Observational Data

As mentioned in the documentation tables

NASA Exoplanet Archive:

Column Meaning

st_nphot Number of photometric time series

st_nrvc Radial velocity curves

pl_ntranspec Transmission spectroscopy measurements

pl_nespec Eclipse spectroscopy

st_nspec Stellar spectra measurements

pl_ndispec Direct imaging spectroscopy

3. Benefits and Limitations

Benefits

- Scientifically validated
- Standardized format
- Frequently updated
- Suitable for ML models

Limitations

- Missing values
- Measurement uncertainties
- Different unit systems
- Incomplete planetary parameters

PART 2: DATA PREPROCESSING THEORY

(Reference: ExoHabitAI Preprocessing Guidelines
ExoHabitAI Data Preprocessing
)

4. Expected Features for Model

The selected features for habitability prediction:

1. Planet Radius
2. Planet Mass
3. Orbital Period
4. Semi-major Axis
5. Equilibrium Temperature
6. Planet Density
7. Host Star Temperature
8. Stellar Luminosity
9. Stellar Metallicity
10. Star Type

5. Data Quality Assessment

Identify:

- Missing values
- Null values
- Duplicate rows
- Inconsistent units (km vs Earth radii)

Generated:

- Summary statistics
- Missing value heatmap

Example Summary Statistics Table

Feature	Mean	Median	Std Dev	Min	Max
Radius	2.3	1.8	1.5	0.3	18
Mass	8.5	5.2	12	0.1	320
Temp	900K	750K	600	200	3000

6. Handling Missing Data

Feature Type Method Used

Planetary Physical Values Mean/Median Imputation

Star Temperature Median

Star Type Mode

Completely Empty Rows Removed

7. Outlier Detection

Methods Used:

- Z-Score
- IQR (Interquartile Range)

Removed Examples:

- Negative radius
- Surface temperature < -300°C
- Physically impossible density

8. Unit Standardization

All units converted to standard astronomical units:

Parameter Converted To

Radius Earth Radii

Mass Earth Mass

Distance AU

Temperature Kelvin

9. Feature Engineering

1) Habitability Score Index

Computed using:

- Temperature closeness to habitable range
- Earth-like radius similarity
- Distance from star
- Stellar luminosity

Output: Numerical score (0–1)

2) Stellar Compatibility Index

Based on:

- Star temperature
- Star size
- Radiation stability

3) Orbital Stability Factor

Based on:

- Orbital period
- Semi-major axis

Stable orbit = Higher habitability chance

10. Categorical Encoding

Star types (G, K, M, F) converted using:

→ One-Hot Encoding

Example:

Star Type G K M F

G 1 0 0 0

M 0 0 1 0

11. Feature Scaling

Applied:

- StandardScaler
- MinMaxScaler

Reason:

Machine learning models perform better when features are normalized.

12. Target Variable Creation

Created:

Binary Classification:

Condition	Label
------------------	--------------

Habitable	1
-----------	---

Non-Habitable	0
---------------	---

FINAL DATASET

After preprocessing:

- Missing values handled
- Outliers removed
- Units standardized
- New engineered features added
- Categorical encoded
- Features scaled

Final dataset saved as:

preprocessed.csv

Uploaded to:

data/preprocessed/

As required in preprocessing guidelines

ExoHabitAI Data Preprocessing

CONCLUSION

The NASA Exoplanet dataset provides structured planetary and stellar data suitable for scientific and machine learning applications

NASA Exoplanet Archive.

Through systematic preprocessing:

- Data quality was improved
- Features were standardized
- Meaningful indices were engineered
- Dataset prepared for habitability prediction

This processed dataset is now ready for training machine learning models in ExoHabitAI.