Milestone–1: Dataset Description and Data Preprocessing

---

1. Dataset Description

1.1 Overview of the Dataset

The dataset used in this project is obtained from the **NASA Exoplanet Archive**, which provides scientifically verified information about confirmed exoplanets and their host star systems. It contains detailed records related to planetary discovery, orbital parameters, physical properties, and stellar characteristics.

This dataset is widely used for astronomical research, habitability analysis, and machine learning applications related to space science and exoplanet studies.

1.2 Source of Data

- **Provider:** NASA Exoplanet Archive
- **Website:** https://exoplanetarchive.ipac.caltech.edu
- **Generated Date:** February 13, 2026
- **Selection Criteria:** Planets detected mainly using the TESS mission

---

1.3 Dataset Size

- **Number of Records (Rows):** 39,386
- **Number of Attributes (Columns):** 289

Each row represents an individual exoplanet entry associated with a host star system.

1.4 Data Format

- **File Type:** CSV (Comma-Separated Values)
- **Encoding:** Text-based
- **Structure:** Tabular format with headers
- **Comments:** Metadata lines start with "#"

1.5 Main Components of the Dataset

The dataset is organized into several logical categories:

A. Planet Identification Information

Includes attributes such as:

- Planet name
- Host star name
- Catalog IDs
- Designation letters

B. Stellar (Host Star) Properties

Includes:

- Star mass
- Star radius
- Effective temperature
- Luminosity
- Metallicity
- Age

C. Planetary Physical Characteristics

Includes:

- Planet mass
- Planet radius
- Density
- Equilibrium temperature
- Surface gravity

D. Orbital Parameters

Includes:

- Orbital period
- Semi-major axis
- Orbital eccentricity
- Inclination

E. Discovery and Observation Data

Includes:

- Discovery method
- Discovery year
- Facility
- Publication date

F. Data Quality and Validation

Includes:

- Default flag
- Update dates
- Validation notes

G. Observation Count Information

Includes:

- Photometric observations
- Spectral measurements
- Radial velocity records

1.6 Missing Values

Some columns contain missing (NaN) values due to:

- Incomplete observations
- Unavailable measurements
- Instrument limitations

Handling missing values is an important step in data preprocessing.

1.7 Data Characteristics

- **Type:** Structured scientific dataset
- **Nature:** Observational and experimental
- **Time Span:** Multi-year data (up to 2026)
- **Update Frequency:** Periodic
- **Reliability:** Peer-reviewed and verified

1.8 Applications of the Dataset

The dataset can be used for:

- Exoplanet classification
- Habitability prediction
- Machine learning modeling
- Statistical analysis
- Discovery trend analysis
- Astronomical research

1.9 Limitations of the Dataset

- Some parameters are estimated
- Observational bias may exist
- Incomplete records for some planets
- Multiple records for the same planet

1.10 Conclusion of Dataset Description

The NASA Exoplanet Archive dataset is a comprehensive and reliable source of exoplanetary data. With more than 39,000 records and 289 attributes, it provides valuable information for habitability analysis and predictive modeling in astronomy.

---

2. Data Preprocessing

2.1 Introduction to Data Preprocessing

Data preprocessing is the process of cleaning, transforming, and organizing raw data into a suitable format for analysis and machine learning. In this project, preprocessing ensures that astronomical data is consistent, reliable, and ready for predictive modeling.

The preprocessing workflow includes data cleaning, handling missing values, outlier detection, unit standardization, feature engineering, and data scaling.

2.2 Expected Features

The following important features are selected for analysis:

- Planet radius
- Planet mass
- Orbital period
- Semi-major axis
- Equilibrium temperature
- Planet density
- Host star temperature
- Star luminosity
- Star metallicity
- Star type

2.3 Data Quality Assessment

Before preprocessing, data quality is evaluated by identifying:

- Missing values
- Null values
- Duplicate records
- Inconsistent measurement units

Additionally, the following analyses are performed:

- Generation of summary statistics
- Visualization of missing values using heatmaps

2.4 Handling Missing Data

Different methods are applied based on feature type:

Feature Type	Method
Planetary physical values	Mean / Median Imputation
Star temperature	Median Imputation
Categorical features	Mode Imputation
Completely missing rows	Removal

This approach helps maintain dataset integrity while minimizing data loss.

2.5 Outlier Detection

Outliers are extreme values that can negatively affect model performance.

Methods Used:

- Z-Score Method
- Inter-Quartile Range (IQR) Method

Actions Taken:

- Remove physically impossible values
- Cap extreme but realistic values

Examples:

- Negative planet radius
- Surface temperature below –300°C

2.6 Unit Standardization

To maintain consistency, all features are converted into standard astronomical units:

Feature	Standard Unit
Radius	Earth radii
Mass	Earth masses
Distance	Astronomical Units (AU)
Temperature	Kelvin

This ensures uniform representation of data.

2.7 Feature Engineering

Feature engineering converts raw physical data into meaningful indicators for machine learning.

A. Habitability Score Index

This score is calculated using:

- Temperature proximity to habitable range

- Planet radius similarity to Earth
- Distance from star
- Stellar luminosity

It represents the potential for supporting life.

B. Stellar Compatibility Index

This index measures:

- Host star temperature
- Star size
- Radiation stability

Higher values indicate better life-supporting conditions.

C. Orbital Stability Factor

This factor uses:

- Orbital period
- Semi-major axis

Stable orbits increase long-term habitability.

2.8 Categorical Encoding

Categorical attributes such as star types (G, K, M, F, etc.) are converted into numerical format using:

- One-Hot Encoding

This enables machine learning models to process categorical data effectively.

2.9 Feature Scaling

Numerical features are scaled to a common range using techniques such as:

- Min-Max Scaling
- Standardization (Z-score normalization)

Scaling improves model convergence and performance.

2.10 Target Variable Creation

Based on project objectives, target variables are created as:

- Binary Classification (Habitable / Non-habitable)
- Multi-Class Classification (Low, Medium, High Habitability)

This enables predictive modeling.

2.11 Final Dataset Storage

After preprocessing:

- The final dataset is named **preprocessed.csv**
- It is uploaded to the GitHub repository under:
  data/preprocessed/

This ensures proper version control and accessibility.

2.12 Conclusion of Data Preprocessing

Data preprocessing plays a crucial role in improving data quality and model performance. By handling missing values, detecting outliers, standardizing units, and engineering meaningful features, the dataset becomes suitable for accurate and reliable machine learning analysis.