Milestone – 1:Data Collection, Understanding and Preprocessing Strategy

Project Title: ExoHabitAI – Habitability Prediction of Exoplanets using Machine Learning

1. Introduction

The discovery of exoplanets has increased significantly in the last two decades due to advanced space missions and ground-based telescopes. However, identifying potentially habitable planets from thousands of confirmed exoplanets requires systematic data analysis and preprocessing.

The objective of this project is to prepare a clean and scientifically meaningful dataset from the NASA Exoplanet Archive that can be used for machine learning models to predict planetary habitability.

This milestone focuses on:

• Data collection

• Understanding the dataset

• Designing a preprocessing pipeline

2. Data Collection

2.1 Data Source:

The dataset used in this project is taken from the NASA Exoplanet Archive – Planetary Systems Table, a publicly available astronomical database containing physical, orbital and stellar parameters of confirmed exoplanets.

2.2 Dataset Characteristics:

• File format: CSV

• Contains multiple entries for the same planet

• Large number of features (200+ columns)

• Includes planetary, stellar, discovery and observational parameters

3. Dataset Understanding

3.1 Planetary Parameters

Orbital period, Semi-major axis, Planet radius, Planet mass, Planet density, Insolation flux, Equilibrium temperature.

These describe the physical and orbital characteristics of the planet:

- Orbital period
- Semi-major axis
- Planet radius
- Planet mass
- Planet density
- Insolation flux
- Equilibrium temperature

These parameters are directly related to planetary environment and habitability.

3.2 Stellar Parameters

Stellar effective temperature, Stellar mass, Stellar radius, Stellar luminosity, Stellar metallicity, Spectral type.

These describe the host star:

- Stellar effective temperature
- Stellar mass
- Stellar radius
- Stellar luminosity
- Stellar metallicity
- Spectral type

Since planetary habitability depends strongly on the host star, these features are important for model training.

3.3 Discovery and Observational Parameters

Discovery method, Discovery year, Telescope and instrument used.

These include:

- Discovery method
- Discovery year
- Telescope and instrument used

These features help in exploratory analysis and understanding detection trends.

4. Data Preprocessing Strategy

4.1 Selection of Best Record for Each Planet

The dataset contains a column named **default_flag**, where:

- default_flag = 1 → most reliable and recommended planetary parameters
- default_flag = 0 → alternative measurements

To avoid data duplication and data leakage, only rows with:

default_flag = 1

will be retained.

This ensures:

- one row per planet
- most accurate scientific values
- reduced dataset size
- better generalization for machine learning

4.2 Removal of Irrelevant Columns

Several columns do not contribute to habitability prediction and will be removed:

Identifier columns

- rowid
- hd_name
- hip_name
- tic_id
- gaia_dr2_id

- gaia_dr3_id

Reference and HTML text columns

- pl_refname
- disc_refname
- st_refname
- sy_refname

Sky coordinate columns

- ra, dec, glon, glat, elon, elat

These columns increase dimensionality but do not add useful learning information.

4.3 Handling Multiple Unit Representations

Some planetary parameters are available in both:

- Earth units
- Jupiter units

To maintain consistency:

- Earth-based units will be retained
- Jupiter-based columns will be removed

This ensures uniform scaling and easier interpretation.

4.4 Removal of Error and Limit Columns

Columns representing uncertainties such as:

- err1
- err2
- lim

will be removed because:

- they represent measurement bounds
- they are not primary physical properties
- they introduce unnecessary sparsity

Only the central measured values will be used for model training.

4.5 Handling Missing Values

Missing data will be handled using:

- Dropping columns with very high missing percentage
- Median imputation for numerical features
- Mode imputation for categorical features

Median is preferred because astronomical data is highly skewed and contains extreme values.

4.6 Removal of Physically Invalid Values

Rows containing physically impossible values will be removed:

- Planet radius ≤ 0
- Planet mass ≤ 0
- Stellar temperature ≤ 0

This ensures the dataset remains scientifically valid.

4.7 Categorical Encoding

Categorical features such as:

- Discovery method
- Stellar spectral type

will be transformed using **One-Hot Encoding** so that they can be used in machine learning models.

4.8 Feature Scaling

Numerical features will be scaled using:

- StandardScaler / RobustScaler

This step ensures:

- equal contribution of all features
- faster convergence of ML algorithms
- improved model performance

5. Target Variable Creation

A new habitability label will be created based on scientific constraints such as suitable planet radius and equilibrium temperature.

Since the dataset does not contain a direct habitability label, a new target variable will be created based on scientific constraints such as:

- Planet radius in the rocky planet range
- Equilibrium temperature suitable for liquid water

This converts the problem into a supervised classification task.

6. Expected Outcome

After completing the preprocessing steps:

- Dataset will contain one row per planet
- Only relevant scientific features will remain
- Missing values will be handled
- Categorical data will be encoded
- Numerical data will be scaled
- Habitability label will be created

The final dataset will be ready for:

- Exploratory Data Analysis
- Feature importance analysis
- Machine learning model development

7. Conclusion

This milestone established a structured preprocessing pipeline for transforming raw astronomical data into a clean and machine learning–ready format. By selecting the most reliable planetary records using the default_flag, removing irrelevant attributes and handling missing values appropriately, the dataset becomes suitable for building a robust habitability prediction model.