Milestone 1: Data Collection, Description, and Preprocessing Strategy

Dataset Description

The dataset is obtained from the NASA Exoplanet Archive and contains confirmed exoplanet observations along with planetary and stellar characteristics used to assess habitability.

Selected Features (Habitability-Relevant)

Feature	Description	Importance
Planet Radius	Size of planet relative to Earth	Determines rocky vs gaseous composition
Planet Mass	Mass relative to Earth	Affects gravity and atmosphere retention
Orbital Period	Time to orbit host star	Indicates seasonal cycles
Semi-major Axis	Average distance from star	Defines habitable zone location
Equilibrium Temperature	Estimated surface temperature	Determines liquid water possibility
Planet Density	Mass per unit volume	Indicates internal composition
Host Star Temperature	Surface temperature of star	Defines habitable zone limits
Star Luminosity	Energy output of star	Controls planetary climate
Star Metallicity	Chemical composition of star	Relates to planet formation
Star Type	Spectral classification	Affects radiation stability

1. Data Quality Assessment

• Identify missing values and null entries.

• Detect duplicate planet records.

• Check inconsistent measurement units.

• Generate summary statistics for all numerical features.

• Visualize missing data patterns using heatmaps.

2. Handling Missing Data

• Numeric planetary features are imputed using median values to reduce skew impact.

• Star temperature is imputed using median for robustness.

• Categorical star types are filled using mode or labeled as 'Unknown'.

• Rows with excessive missing values are removed to maintain data integrity.

3. Duplicate Removal

• Remove multiple entries for the same planet using planet identifiers.

• Retain the row with the most complete data.

• Ensures one reliable observation per planet.

• Prevents data leakage and bias in machine learning.

4. Outlier Detection

• Z-score and IQR methods detect extreme values.

• Negative or physically impossible values are removed.

• Extreme but realistic values are capped.

• Improves model robustness by reducing noise.

5. Unit Standardization

• Planet radius standardized to Earth radii.

• Planet mass standardized to Earth masses.

• Orbital distance standardized to Astronomical Units (AU).

• Temperature standardized to Kelvin.

• Ensures scientific consistency across observations.

6. Feature Engineering

• Habitability Score combines temperature, radius, distance, and luminosity.

- Stellar Compatibility Index evaluates host star suitability.
- Orbital Stability Factor assesses long-term orbital stability.
- Transforms astrophysical data into ML-interpretable signals.

7. Categorical Encoding

- Star types encoded using One-Hot Encoding.
- Prevents ordinal bias in machine learning models.
- Allows algorithms to interpret stellar classifications effectively.

8. Feature Scaling

- StandardScaler applied to numerical features.
- Ensures uniform feature contribution.
- Prevents dominance of features with large magnitudes.
- Improves convergence of machine learning algorithms.

9. Target Variable Creation

- Binary classification: habitable vs non-habitable.
- Based on temperature range and Earth-like size.
- Provides supervised learning target variable.