Project Report: Exoplanet Habitability Data Preprocessing

1. Objective of the Data Preprocessing

The objective of this preprocessing pipeline is to convert raw exoplanet observational data into a clean, structured, and machine-learning-ready dataset. Astronomical datasets often contain missing values, noise, and physically inconsistent measurements. Therefore, this pipeline ensures that all observations are scientifically valid, statistically consistent, and suitable for predicting planetary habitability.

2. Data Quality Assessment

The raw dataset was first examined for:

- Missing values across planetary and stellar features
- Duplicate records
- Entirely empty rows
- Structural inconsistencies

Duplicate rows were removed and fully empty rows were dropped to ensure dataset integrity.
A heatmap visualization was generated to understand the distribution of missing values and guide imputation strategies.

3. Feature Selection

Only scientifically meaningful parameters relevant to habitability modeling were retained:

Planetary Features

- Planet radius
- Planet mass
- Orbital period
- Semi-major axis
- Equilibrium temperature

Stellar Features

- Host star temperature
- Stellar radius
- Stellar mass
- Stellar metallicity
- Spectral type

This selection ensures the model focuses on astrophysically meaningful predictors.

4. Missing Data Handling

Different imputation strategies were applied depending on feature type:

- Numerical astrophysical values were imputed using the median to preserve distribution robustness
- Stellar temperature was also filled using the median
- Categorical spectral type was imputed using the mode

This approach prevents statistical bias while maintaining physical realism.

5. Removal of Physically Impossible Values

To maintain scientific validity, observations violating physical laws were removed.
Examples include:

- Negative planetary radius or mass
- Zero or negative orbital parameters
- Non-physical stellar measurements

This filtering step ensures the model learns only from astronomically plausible systems.

6. Outlier Detection Using IQR

The Interquartile Range (IQR) method was applied across all numerical parameters. Values outside the acceptable statistical range were removed to eliminate extreme observational noise while preserving meaningful astrophysical variation.

This step improves model stability and prevents skewed learning.

7. Feature Engineering

Several domain-inspired indices were created to convert astrophysical properties into machine-learning-interpretable signals.

7.1 Habitability Score

A composite index combining:

- Temperature proximity to Earth-like conditions
- Radius similarity to Earth
- Mass similarity to Earth

This normalized score represents how Earth-like a planet is.

7.2 Stellar Compatibility Index

This index measures how suitable the host star is for supporting habitable planets.

It incorporates:

- Stellar temperature similarity to the Sun
- Stellar mass similarity to the Sun

A higher score indicates a star more likely to host stable habitable environments.

7.3 Orbital Stability Encoding

Spectral type was simplified into the first stellar class letter (G, K, M, etc.).
One-hot encoding was then applied to convert star categories into machine-learning-friendly format.

8. Target Variable Construction

A binary classification target was created:

- Planets with habitability score > 0.6 → labeled **Habitable (1)**
- Others → labeled **Non-Habitable (0)**

This transforms the dataset into a supervised learning problem.

9. Feature Scaling:- Min-Max normalization was applied to numerical columns so that all values fall between 0 and 1.

This ensures:

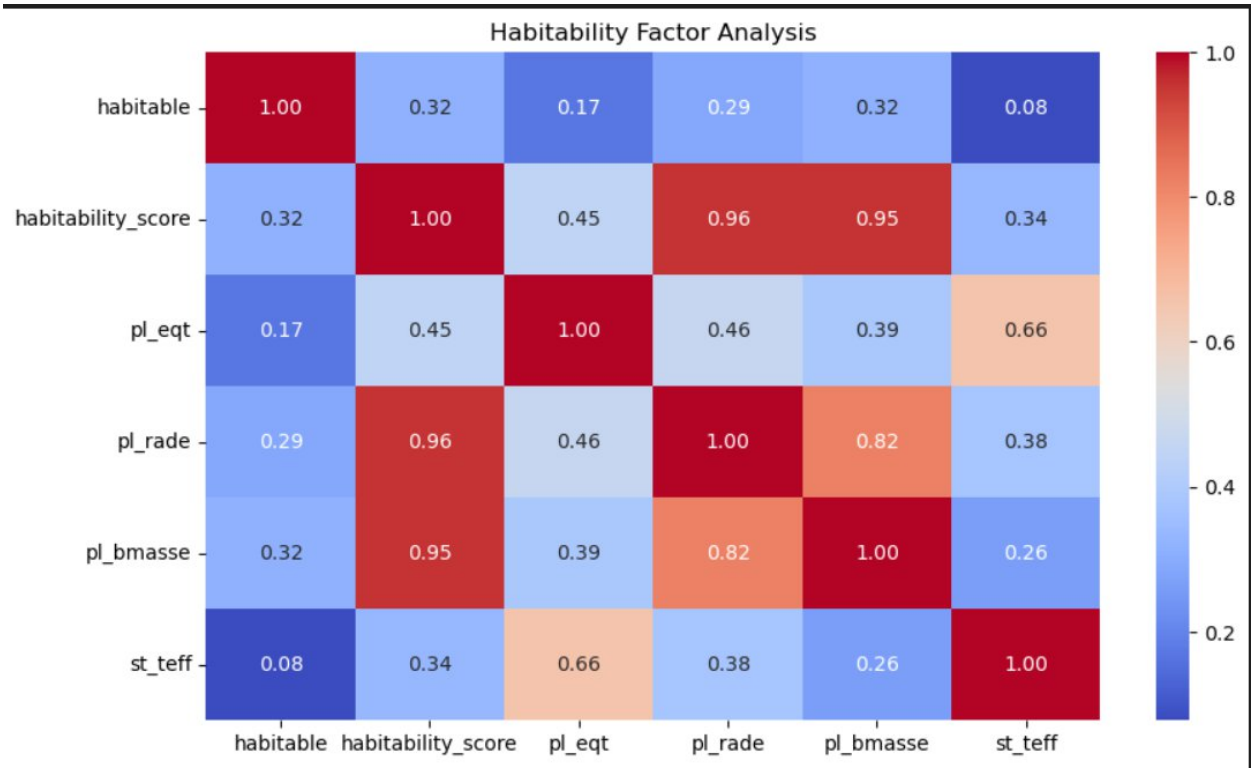
- Stable gradient learning
- Fair contribution of all features
- Faster convergence for ML algorithms

10. Correlation Analysis

A heatmap visualization was generated to examine relationships between:

- Habitability score
- Stellar parameters
- Planetary temperature and mass

This step validates whether engineered features meaningfully relate to the prediction target.



Final Conclusion

The preprocessing pipeline successfully transforms raw exoplanet observations into a scientifically reliable machine-learning dataset.
By combining astrophysical validation, statistical cleaning, and domain-informed feature engineering, the final dataset captures both planetary properties and stellar environmental conditions that influence habitability.