Data Description Report:

Predicting the Habitability of Exoplanets Using Machine Learning

1. Introduction

The primary objective of this report is to provide a detailed technical description of the dataset used in the **ExoHabitAI** project. This project utilizes Machine Learning to predict the habitability potential of exoplanets by analyzing physical, orbital, and stellar parameters. The data is sourced from the official **NASA Exoplanet Archive**.

2. Dataset Summary

- **Source:** NASA Exoplanet Archive (.https://exoplanetarchive.ipac.caltech.edu/)
- **Data Format:** CSV (Comma-Separated Values).
- **Total Records:** ~39,315 observations.
- **Total Features:** 289 columns.

3. Core Features & Data Dictionary

This table describes the most important parameters used by the **ExoHabitAI** model to detect and rank habitable exoplanets.

Feature Name	Category	Unit	Description
pl_name	Metadata	Name	Unique identifier for the exoplanet.
pl_rade	Planetary	Earth Radii	The size of the planet compared to Earth. Crucial for identifying rocky worlds.
pl_bmasse	Planetary	Earth Mass	The weight of the planet. Helps distinguish between Earth-like and Gas Giants.
pl_orbper	Orbital	Days	The time taken for the planet to complete one orbit around its star.
pl_orbsmax	Orbital	AU	The average distance between the planet and its star (Semi-major axis).
pl_eqt	Thermal	Kelvin	Estimated surface temperature. Essential for the "Habitable Zone" check.
st_teff	Stellar	Kelvin	The surface temperature of the host star (Effective Temperature).
st_rad	Stellar	Solar Radii	The size of the host star compared to our Sun.
st_met	Stellar	dex	Metallicity of the star, indicating the presence of heavy elements.
sy_dist	System	Parsecs	The distance of the planetary system from Earth.
default_flag	Quality	Binary (0/1)	A flag indicating the most accurate, peer-reviewed measurement (1 = Best)

Extended Data Features for ExoHabitAI

Feature Name	Category	Scientific Importance
pl_dens	Planetary	**Density:** Helps detect if the planet is made of Rock, Water, or Gas.
pl_insol	Energy	**Insolation Flux:** Measures how much energy the planet gets from its sun compared to Earth.
st_lum	Stellar	**Luminosity:** The total brightness of the star, which determines the size of the Habitable Zone.
pl_orbeccen	Orbital	**Eccentricity:** How circular or oval the orbit is. High eccentricity can lead to extreme weather.
st_age	Stellar	**Star Age:** Older stars are generally more stable, which is better for life to evolve.
disc_year	Metadata	**Discovery Year:** Helps in understanding how detection technology has improved over tim

4. Data Cleaning & Preparation

- **Removing Duplicates:** We used the default_flag to keep only the best, most accurate version of each planet. This removed scientific "noise."
- **Cleaning Missing Data:** We identified columns with too many empty spaces (Null values) and removed them. This ensures the AI doesn't learn from "blank" information.
- **Dropping Unnecessary Columns:** We removed columns like URLs and internal IDs that have nothing to do with science. This helps the AI focus only on important data like size and heat.
- **Unit Consistency**: We made sure all measurements are in the same units (for example, all temperatures in Kelvin) so the model doesn't get confused.
- **Outlier Removal:** We removed planets with "impossible" numbers (like zero mass or extreme errors) to keep the dataset scientifically correct.

5. Scientific Integrity

To ensure our research is accurate and trustworthy, we followed these three rules:

- **Peer-Review Only:** We only used data that has been double-checked and confirmed by independent scientists to avoid "rumors" or unproven planets.
- **Physics Check:** We removed all "impossible" data (like planets with negative mass or temperatures hotter than their stars) to stay realistic.
- **Data Precision:** We filtered out measurements with high error margins, focusing only on the most accurate and sharp data available from NASA.

6. Feature Engineering

Raw numbers can be confusing, so we combined different data points to create "Smart Indicators." These help us see clearly if a planet can support life:

- **The Goldilocks Test:** By combining the **Star's Temperature** and the **Planet's Distance**, we find the "Habitable Zone." This is the perfect area where it is not too hot and not too cold for liquid water.
- **Surface Composition Profile:** We used the planet's **Mass** and **Radius** to calculate its **Density**. This helps us separate "Rocky" planets like Earth from "Gas Giants" like Jupiter.
- **Energy Balance:** We measured the **Insolation Flux**, which is the total heat and light a planet gets from its sun. Getting the same amount of energy as Earth is a key sign of habitability.
- **Climate Stability:** We analyzed the **Orbital Shape**. A circular path means the weather stays stable all year, while an oval path causes dangerous and extreme temperature changes.

7. Feature Importance (Key Life Indicators)

After analyzing the data, we identified the most important factors that determine if a planet is habitable. These are the primary "Life Indicators" that our research focuses on:

1. Energy Intake (Insolation Flux): This is the most critical factor. It measures how much heat and light a planet gets from its star. If a planet receives a similar amount of energy as Earth, it has the best chance of maintaining a stable temperature for liquid water.

2. Surface Type (Planet Density): The density of a planet tells us what it is made of. We prioritize planets with a high density, which indicates a **solid, rocky surface**. This is essential because life (as we know it) needs a ground to stand on, unlike gas giants like Jupiter.

3. Distance and Heat (The Goldilocks Zone): The distance from the star (**`pl_orbsmax`**) combined with the star's heat (**`st_teff`**) is a major indicator. A planet must be in the "Sweet Spot"—not too close to be burned and not too far to be frozen.

4. Climate Stability (Orbital Shape): We look at how circular the planet's path is. A perfect circle means the planet's climate stays steady. If the path is a long oval (high eccentricity), the planet will experience extreme weather shifts that make life very difficult.

5. Host Star Quality (Stellar Luminosity): The brightness and radiation of the star are very important. Some stars are too violent and can strip away a planet's atmosphere. We look for stable stars that provide a safe environment for their planets.

Conclusion

The NASA Exoplanet Archive is a massive database, but its raw form is cluttered with too many columns, duplicates, and missing values. To fix this, we performed "Data Preprocessing" by filtering out errors and selecting only the most important scientific features. By cleaning and organizing this information, we have created a reliable, high-quality dataset. This refined foundation is now ready for AI models to accurately identify planets that could support life, bringing us one step closer to finding a "Second Earth."