Exoplanet Habitability Prediction Project

1. Introduction

Exoplanets are planets that exist outside our solar system and orbit stars other than the Sun. With advancements in space exploration, large amounts of exoplanet data are now available. However, this raw data cannot be directly used for machine learning applications. This project focuses on preparing exoplanet data for habitability prediction. Habitability refers to whether a planet has conditions suitable for life, such as proper temperature and Earth-like size. By preprocessing the dataset, this project ensures that the data is clean, consistent, and ready for further analysis and prediction tasks.

2. Objective

The objective of this project is to preprocess exoplanet data to make it suitable for machine learning models. Raw astronomical data often contains missing values, inconsistent formats, and categorical variables. This project aims to clean column names, handle missing data, encode categorical features, and normalize numerical attributes. Another goal is to create a habitability label based on scientific thresholds such as temperature and orbital distance. The final outcome is a structured dataset that can be directly used for training and evaluating habitability prediction models.

3. Dataset Description

The dataset used in this project contains information about exoplanets and their host stars. It includes planetary features such as radius, mass, orbital period, density, and equilibrium temperature. Stellar properties like star temperature, luminosity, metallicity, and star type are also included. These features are important for determining whether a planet falls within the habitable zone of its star. The dataset is inspired by NASA-style exoplanet archives and represents realistic astronomical observations.

Dataset Overview

Column	Category	Description
planet_radius	Planet	Size of planet compared to Earth
planet_mass	Planet	Mass of the planet
orbital_period	Orbit	Time taken for one revolution

4. Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. In this project, preprocessing includes cleaning column names to ensure consistency, handling missing values using median imputation, and encoding categorical features such as star type into numerical form. Feature scaling is applied using standardization so that all numerical values are on a similar scale.
Additionally, a habitability label is created using astrophysical rules. These steps ensure reliable and accurate input data for future machine learning models.

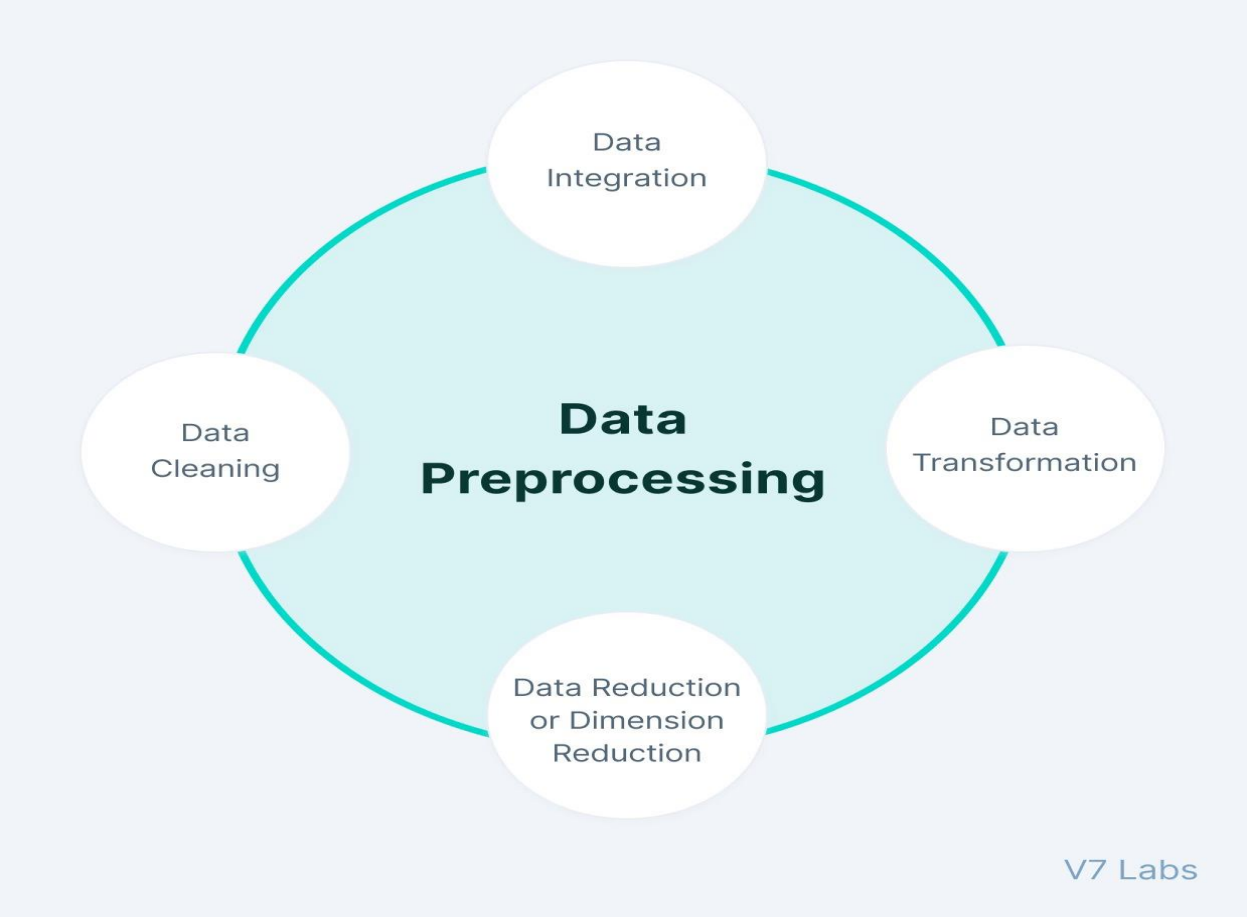

Fig: 1

Planetary Features

Feature	Purpose	Description
planet_density	Composition	Identifies rocky or gaseous planet
equilibrium_temp	Temperature	Estimates surface temperature

Stellar Features

Feature	Purpose	Description
star_temp	Star Property	Controls planetary climate
luminosity	Radiation	Energy emitted by star
metallicity	Formation	Probability of planet formation

5. Habitability Criteria

Habitability in this project is determined using scientific conditions commonly used in astronomy. A planet is considered potentially habitable if its equilibrium temperature allows liquid water to exist Planet size is also considered to ensure Earth-like conditions, avoiding gas giants. The distance of the planet from its host star is checked to ensure it lies within the habitable zone. By combining these criteria, a binary habitability label is generated, which serves as the target variable for machine learning tasks.

Habitability Logic

Condition	Threshold	Reason
Temperature	200–350 K	Allows liquid water
Planet Radius	0.5–2 Earth	Earth-like size
Orbit Distance	0.3–1.5 AU	Within habitable zone

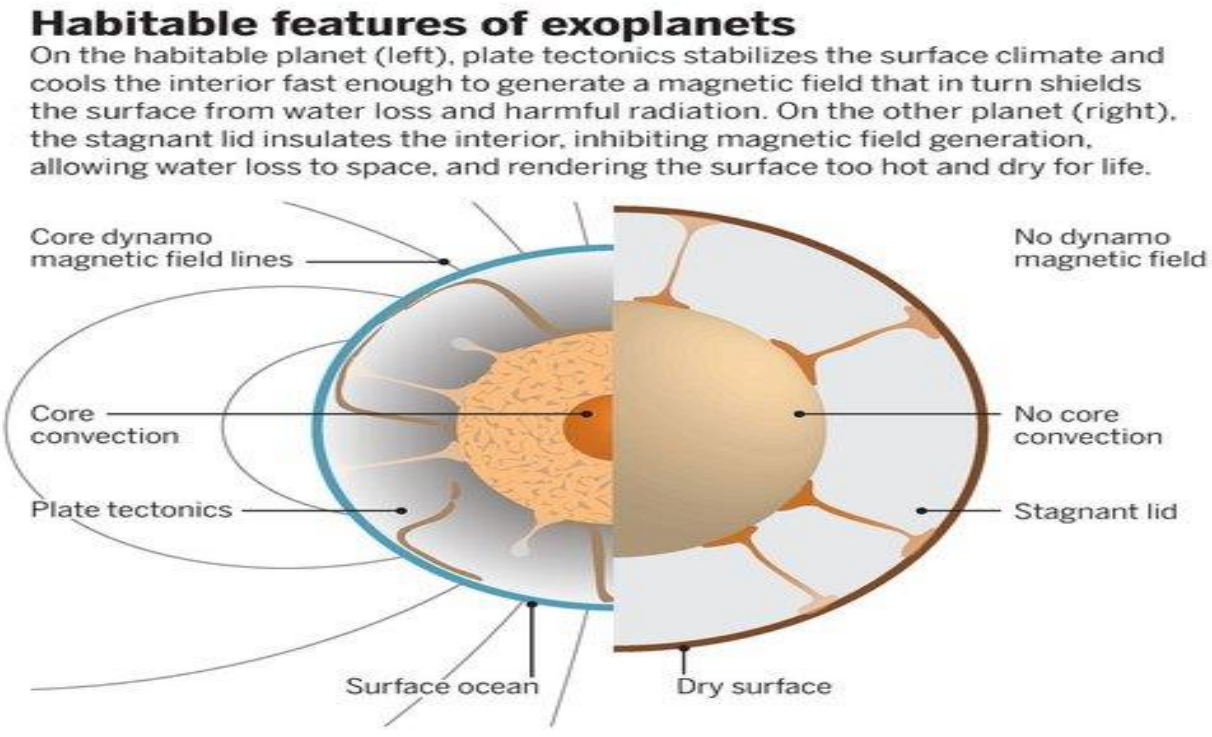

Fig: 2

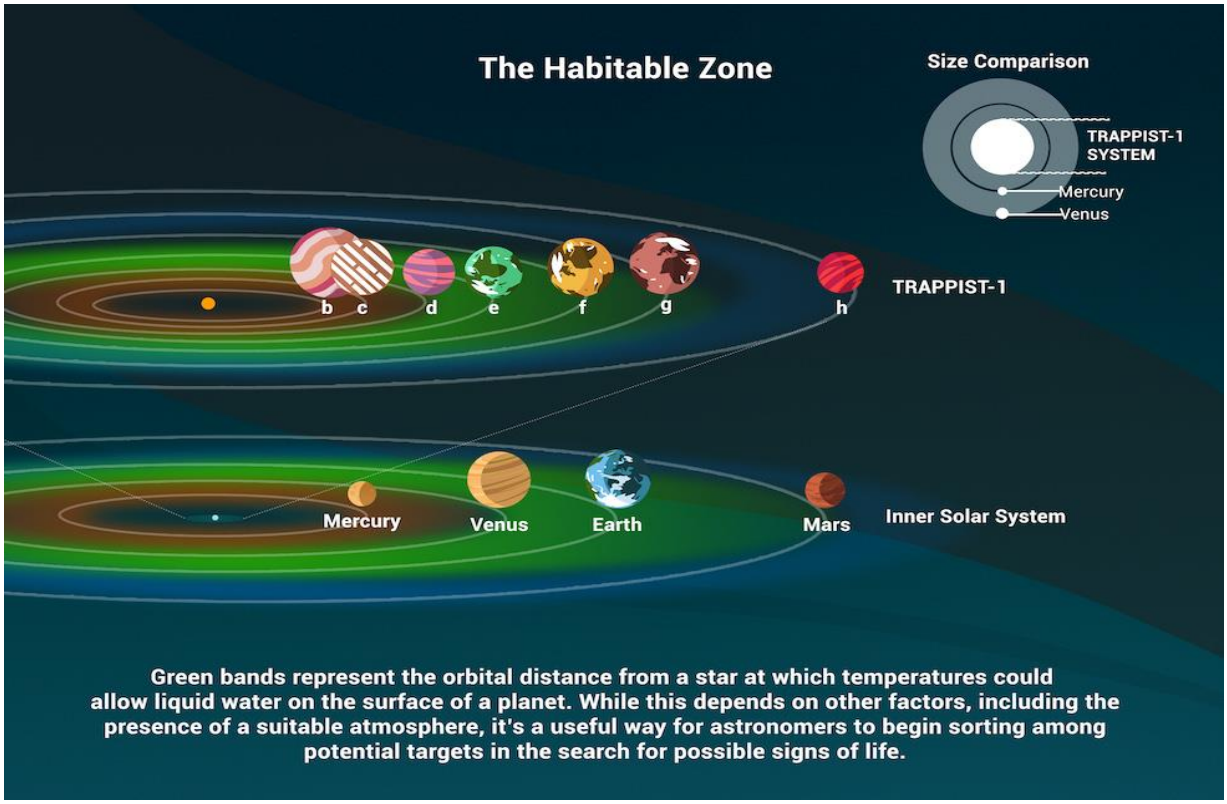

Fig: 3

6. Output

The final output of this project is a fully preprocessed dataset saved in CSV format. This dataset contains cleaned, scaled, and encoded features along with a habitability label. The processed data is suitable for training machine learning classification models. This output demonstrates the successful transformation of raw astronomical data into a structured format that supports predictive modeling. It also serves as a reusable dataset for future research and academic projects.

Output	Type	Meaning
habitable	Binary	1 = Habitable, 0 = Non-habitable

7. Applications

This project has multiple applications in both academic and research domains. It can be used as a foundation for machine learning-based exoplanet habitability prediction. Students can use the processed dataset for experimentation with different algorithms. Researchers can extend the project to include deep learning or advanced feature selection techniques. The project is also suitable for internships, data science learning, and astronomy-related studies.

8. Conclusion

This project demonstrates a complete and professional approach to exoplanet data preprocessing. By applying systematic cleaning, transformation, and feature engineering techniques, the raw dataset is converted into a machine learning-ready format. The project highlights the importance of preprocessing in achieving accurate predictions. It provides a strong base for future work such as model training, performance evaluation, and deployment of habitability prediction systems.