

1. Introduction

In recent years, the discovery of exoplanets has increased rapidly due to advanced space missions such as Kepler, TESS, and ground-based observations. Among thousands of confirmed exoplanets, only a small fraction may have conditions suitable for life. Identifying such potentially habitable planets manually is time-consuming and complex due to the large volume of data.

In this project, I focus on predicting the habitability of exoplanets using machine learning techniques. By using planetary and stellar parameters from reliable astronomical databases, machine learning models can learn patterns associated with habitable conditions and help classify exoplanets as potentially habitable or non-habitable.

2. Data Sources

2.1 NASA Exoplanet Archive

The primary dataset used in this project is obtained from the NASA Exoplanet Archive, which is a publicly available and reliable database maintained by NASA and Caltech. It provides detailed information about confirmed exoplanets and their host stars.

The archive contains important parameters such as:

- 1) Planet radius and mass
- 2) Orbital period and semi-major axis
- 3) Equilibrium temperature
- 4) Stellar effective temperature
- 5) Stellar mass, radius, and luminosity

These parameters are essential for habitability analysis and machine learning model training.

2.2 Habitable Exoplanets Catalog (HEC)

To label planets for supervised learning, I use the Habitable Exoplanets Catalog (HEC) provided by the Planetary Habitability Laboratory (PHL). This catalog identifies planets that lie within the habitable zone and satisfy basic physical criteria for habitability.

This catalog helps in:

Identifying known potentially habitable planets

Creating target labels (habitable / non-habitable)

Validating machine learning predictions

3. Understanding Planetary Habitability

3.1 Habitable Zone (HZ)

The habitable zone, also known as the Goldilocks Zone, is the region around a star where conditions may allow liquid water to exist on the surface of a planet. Since liquid water is essential for life as we know it, planets located within this zone are considered potential candidates for habitability.

However, being in the habitable zone does not guarantee habitability. Other factors such as atmosphere,

planetary composition, and stellar radiation also play significant roles.

3.2 Important Habitability Factors

In this project, habitability is assessed using the following factors:

Distance from host star

Planetary size (rocky vs gas planets)

Surface or equilibrium temperature

Stellar temperature and luminosity

These parameters are directly or indirectly related to the ability of a planet to maintain stable surface conditions.

4. Feature Selection for Machine Learning

The performance of machine learning models depends heavily on selecting relevant features. The main features used in this project include:

Planetary Features

Planet radius

Planet mass

Orbital period

Semi-major axis

Equilibrium temperature

Stellar Features

Stellar effective temperature

Stellar luminosity

Stellar mass

Stellar radius

These features are normalized and preprocessed before model training.

5. Machine Learning Approach

5.1 Data Preprocessing

The dataset contains missing and incomplete values. To handle this:

Missing values are removed or imputed

Numerical features are normalized

Class imbalance is handled using resampling techniques

Since habitable planets are very rare, the dataset is highly imbalanced.

5.2 Machine Learning Models

I experiment with different machine learning models to classify exoplanets:

Random Forest Classifier

Support Vector Machine (SVM)

K-Nearest Neighbors (KNN)

Gradient Boosting / XGBoost

These models are chosen because they handle non-linear relationships and complex feature interactions effectively.

5.3 Evaluation Metrics

Accuracy alone is not sufficient due to class imbalance. Therefore, I use:

Precision

Recall

F1-score

Confusion matrix

Special importance is given to recall, as correctly identifying potentially habitable planets is more critical than misclassifying non-habitable ones.

6. Examples of Potentially Habitable Exoplanets

Some well-known exoplanets often considered in habitability studies include:

Planet Name Reason for Habitability

TRAPPIST-1e Rocky planet in habitable zone

Kepler-442b Earth-like size, stable orbit

LHS 1140 b Dense rocky planet, HZ location

GJ 1002 b Earth-mass planet in habitable zone

These planets are used as reference points for validating machine learning predictions.

7. Challenges and Limitations

Limited number of confirmed habitable planets

Missing atmospheric and surface composition data

Class imbalance in datasets

Habitability definitions depend on assumptions

Despite these challenges, machine learning provides a scalable and efficient approach for habitability prediction.

8. Conclusion

In this project, I demonstrate how machine learning techniques can be used to predict the habitability of exoplanets using astronomical data. By leveraging datasets from the NASA Exoplanet Archive and Habitable Exoplanets Catalog, meaningful patterns related to habitability can be learned.

This approach helps astronomers prioritize promising exoplanets for future observation and research, contributing to the broader search for life beyond Earth.