# STELLAR ANALYTICS

## TECHNEX’25 IIT (BHU) VARANASI

NAME: Abhay Pratap Singh
TEAM NAME: The Exiled Spaceman
DATE OF SUBMISSION: 14/02/2025
Contact: psabhay2003@gmail.com

# PROBLEM STATEMENT

By the year 2224, Earth is no longer the vibrant cradle of life it once was. The world's remaining nations formed an unprecedented alliance — The Stellar Coalition. Together, they launched The Stellar Project. The primary objective of this Stellar Project is to identify which exoplanets can sustain life by creating a custom "Habitability Index" using features such as Earth Similarity Index (ESI), long-term stability and more.
Our goal is to classify exoplanets into three categories:

- Potentially Habitable
- Marginally Habitable
- Non-Habitable

# ASSUMPTIONS

The key assumptions made while performing the complete analysis were:

- It is assumed that the dataset provided in the problem statement represents current exoplanet characteristics and no new changes have been done in the dataset.
- All the features (such as Planet Mass, Planet Radius and others) are in their SI units.
- Features related to discovery of exoplanets such as P_DETECTION and P_DISCOVERY_FACILITY are not significant contributors to a planet's habitability and therefore, have been dropped during the data cleaning process.
- Features related to Host Star's name, its position in the sky are also irrelevant to a planet's habitability and therefore, have also been dropped during the data cleaning process.
- It is assumed that all these astrophysical features are spanned over several orders of magnitude. Therefore, normalization of this data is necessary.

# CUSTOM HABITABILITY INDEX

### 3.1. Definition/Formula of the Index

My Custom Habitability Index is derived by scoring each exoplanet based on several features derived in feature engineering:

- **ESI (Earth Similarity Index):**
  - Score 2 if ESI ≥ 0.35
  - Score 1 if 0.25 ≤ ESI < 0.35
- **Long-Term Stability:**
  - Score 2 if Stability ≥ 1.2
  - Score 1 if 1.0 ≤ Stability < 1.2
- **Habitability Zone Distance (HZD):**
  - Score 2 if -1.2 ≤ HZD ≤ -0.8
  - Score 1 if -1.5 ≤ HZD < -1.2 or -0.8 < HZD ≤ -0.5
- **Flux Ratio:**
  - Score 2 if |Flux Ratio - 12.57| ≤ 1
  - Score 1 if |Flux Ratio - 12.57| ≤ 2

- **Escape Velocity Ratio:**
    - Score 2 if Ratio ≥ 0.15
    - Score 1 if 0.12 ≤ Ratio < 0.15
- **Tidal Force Ratio:**
    - Score 2 if 0.8 ≤ Ratio ≤ 1.2
    - Score 1 if 0.5 ≤ Ratio < 0.8 or 1.2 < Ratio ≤ 1.5

The total score is the sum of points across these six features (maximum possible = 12).

### 3.2. Explanation of Feature Relevance and Weights

- **ESI:** The Earth Similarity Index (ESI) is a scale to physically compare other planets to Earth. The scale is between 0 (no similarity to Earth) and 1 (Earth-like). Planets with an ESI between 0.8 and 1.0 are more likely to be similar to Earth
  .
- **Long-Term Stability:** Long term stability highlights planets with minimal extreme climate variations, increasing the likelihood of sustaining life long-term. Higher Stability: Values > 0.5 indicate stable orbits, favourable for maintaining consistent climates.

- **Habitability Zone Distance (HZD):** Measures a planet's location relative to the habitable zone; planets too close or too far from their star may not retain liquid water.

- **Flux Ratio:** To measure how much energy a planet receives from its star. The amount of radiation received from the host star directly impacts habitability.

- **Escape Velocity Ratio:** To measure the ratio of escape velocity of planet relative to Earth's escape velocity. The planets with value close to 1 will have almost same Escape Velocity as of Earth.

- **Tidal Force Ratio:** The Tidal Force Ratio of Earth to Sun is 1:2 = 0.5, any value close to it, indicates strong correlation with Earth and similar habitability.

Each feature is weighted based on its expected impact on habitability. Higher scores contribute to a more favourable habitability classification.

## Approach

The key approach for this problem statement is as follows:

- **Data loading, exploring and cleaning:** The very first step is to load the csv file and explore its shape, info and other details. The data contains null values which must be filled and certain irrelevant columns which must be dropped before proceeding further.

- **Outlier Handling and Normalization:** The next step is to address outliers and normalize the data which is highly spanned over a wide spectrum of magnitude. This will help create a better model for data analysis.

- **Visualizing Correlation matrix and dropping highly correlated features:** The next step is to visualize a correlation matrix to help understand how different features of a dataset are correlated and dropping features with strong positive correlation because we can only keep one feature for further calculations as other correlated features will be directly proportional to it. It is cautious to note that we do not drop those features which will be used to derive other features in feature engineering.

- **Feature Engineering:** This part is one of the most important parts of entire code because feature engineering derives new features using existing features from the dataset which will be directly taken into account for defining a habitability class. In this part, I derived features such as Earth Similarity Index (ESI), Long-Term Stability, Star-Planet Energy Flux Ratio, Habitability Zone Distance (HZD), Escape Velocity Ratio, and Tidal Force Ratio.

- **Defining a custom "Habitability Index":** This section requires a lot of logical thinking because here, a custom "habitability index" is to be defined. The approach I used has already been discussed in the section 3.1.

- **Model Development:** The last part is to develop a machine learning model to classify the exoplanets as per the problem statement. I used Decision Tree classifier using Entropy criterion for measuring pure/impure split and random state = 42 for random sampling of data. I also set a max depth = 3 to prevent unnecessary splitting and an overfitting model.  This process is called pruning using the hyper parameters I used.

- **Model Evaluation:** The final step is to evaluate the model performance by checking the model accuracy, precision, recall, and F1 score to see how the model performs.

## FINDINGS

- **Accuracy: 96.02%**
  Out of 226 samples, about 96% were correctly classified.
- **Weighted Precision: 96.19%**
  When the model makes a prediction, about 96% of the time it is correct—this average takes into account the number of samples in each class.
- **Weighted Recall: 96.02%**
  Across all classes, the model correctly identifies about 96% of the actual samples.
- **Weighted F1 Score: 95.38%**
  This score, balancing precision and recall, indicates the model performs very well overall.

## NOTEBOOK LINK

The link for Colab Notebook is attached below:

https://colab.research.google.com/drive/13CRntOvLXsn1uBk9v9EjQL_w80R7x5Q8?usp=sharing

# Appendix

## Resources:

- https://en.wikipedia.org/wiki/Planetary_habitability

- https://exoplanets.nasa.gov/news/109/in-the-zone-how-scientists-search-for-habitable-planets/

- https://www.mdpi.com/2076-3263/8/8/280

- https://astrobiology.nasa.gov/quick-facts/earth-similarity-index/#:~:text=The%20Earth%20Similarity%20Index%20

- https://exoplanetarchive.ipac.caltech.edu/docs/poet_calculations.html

- https://pmc.ncbi.nlm.nih.gov/articles/PMC7414196/

- https://phl.upr.edu/projects/earth-similarity-index-esi

- https://arxiv.org/abs/1509.08922

- https://academic.oup.com/mnras/article/471/4/4628/4096396

- https://www.aanda.org/articles/aa/full_html/2019/10/aa35297-19/aa35297-19.html