april 5th, 2025

exoplanet research

stuff i did

- attended the washington state science and engineering fair (last weekend) with my derivative research – **An Analysis of Exoplanet Habitability and Most Influential Stellar and Planetary Parameters to Habitability through the Lens of Machine Learning**
  - i won first place in the physics and astronomy category!
- starting feature engineering for planetary system clustering ML modeling

- going to be busy with AP and exam studies in the coming two weeks so might not be able to present a lot next time! :(

An Analysis of Exoplanet Habitability and Most Influential Stellar and Planetary Parameters to Habitability through the Lens of Machine Learning

Christina Liu

INTRODUCTION

Are we alone in this universe? Are there any exoplanets other than Earth where humans are able to thrive? The search of potentially habitable exoplanets has been an active research field in astrophysics throughout the past decade.

As of January 28, 2025, there were **5,834** confirmed exoplanets documented in the **NASA Exoplanet Archive** dataset, each associated with hundreds of parameters. With advancements in the observational capabilities of satellite and telescope based techniques, the number of discovered exoplanets continues to grow.

To identify potential habitable exoplanets among such a large and ever-growing set of candidates, machine learning (ML) has been increasingly adopted to predict habitability. Furthermore, ML model feature importance analysis techniques such as **SHAP** (**SH**apley **A**dditive ex**P**lanations) provide unique opportunities for studying and identifying stellar and planetary parameters that impact the habitability and how they impact it, which is the focus of this research work.

RESEARCH OBJECTIVES

This research aims to study the influential stellar and planetary parameters to habitability through the lens of machine learning, with the following goals:

- Build high-quality ML models (**Random Forest**, **XGBoost**) to predict exoplanet habitability.
- Conduct feature important analysis via the **SHAP** technique to identify influential stellar and planetary parameters to habitability.
- Perform analysis through **SHAP** to understand how different stellar and planetary parameter values positively or negatively affect the exoplanet habitability.

DATA SOURCES

The primary data sources for this study:

- Planetary Systems Composite Data @ **NASA Exoplanet Archive**: **5,834** confirmed exoplanets as of January 28, 2025.
- Habitable World Catalog (**HWC**), PHL @ UPR Arecibo: **5,599** exoplanets as of January 28, 2025.

The experiment joined data from NASA Exoplanet Archive and HWC. The HWC dataset has a ***P_HABITABLE*** data field, which indicates exoplanet habitability and is used to label training data.

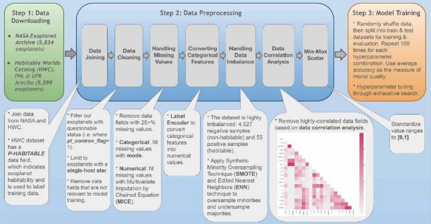
Table extracted from Google Colab notebook showcasing just some of the data fields from the combined NASA Exoplanet Archive + HWC dataset.

MODEL SELECTION

Tree-based machine learning models – specifically **Random Forest** and **XGBoost** – were chosen to build classifiers for predicting exoplanet habitability.

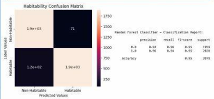
Research in machine learning shows that tree-based models still outperform deep learning models on the tabular dataset. The dataset for this study is entirely tabular based data and therefore tree-based models were well suited for this classification problem.

DATA PROCESSING & MODEL TRAINING

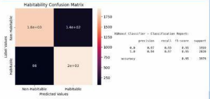


MODEL EVALUATION

Random Forest Evaluation

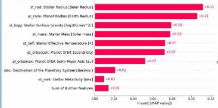


XGBoost Evaluation



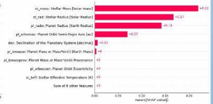
FEATURE IMPORTANCE ANALYSIS

Random Forest Feature Importance by SHAP



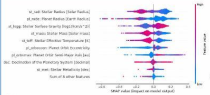
The most influential features of Random Forest classifier: (1) Stellar Radius, (2) Planet Radius, (3) Stellar Surface Gravity, (4) Stellar Mass, (5) Stellar Effective Temperature, (6) Planet Orbit Eccentricity, (7) Planet Orbit Semi-Major Axis.

XGBoost Feature Importance by SHAP



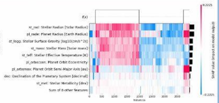
The most influential features of XGBoost classifier: (1) Stellar Mass, (2) Stellar Radius, (3) Planet Radius, (4) Planet Orbit Semi-Major Axis, (5) Declination of the Planetary System.

Deep Dive on Random Forest through SHAP

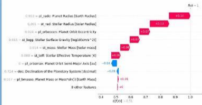


SHAP **beeswarm** plot shows that:

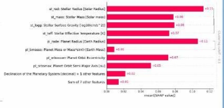
- Relatively higher values in stellar radius, planet radius, stellar mass, and stellar effective temperature lead towards non-habitable prediction, while relatively lower values in those parameters lead towards habitable prediction.
- Relatively higher planet orbit semi-major axis leads towards habitable prediction, while relatively lower value leads towards non-habitable prediction.



SHAP **heatmap** plot groups samples (exoplanets) that have the same model outputs for the same reasons together (e.g., the exoplanets that are predicted to be habitable due to stellar radius, etc.).



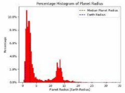
SHAP **waterfall** plot shows how the stellar and planetary parameter values influence habitability for a specific sample (exoplanet).



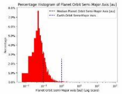
SHAP plot shows correlated features. Stellar radius is correlated with stellar mass. Same goes with planet radius and planet mass.

DISCUSSION

SHAP analysis indicates that relatively larger planet radius leads towards non-habitable. This indeed matches the reality. The below figure shows the percentage histogram of planet radius, with a black vertical line indicating the median and a blue line indicating Earth. As shown in the figure, Earth (habitable) is on the far left of the median, while the exoplanets on the far right are gas-giants and non-habitable.



Similarly, SHAP analysis indicates relatively higher planet orbit semi-major axis leads towards habitable. As shown in the figure, Earth (habitable) is on the far right of the median, while the exoplanets on the far left are too close to their host stars and thus non-habitable.



CONCLUSIONS

A **Random Forest** and **XGBoost** model were trained to predict exoplanet habitability with high accuracy at **0.95**.

Feature importance analysis through **SHAP** identified several influential stellar and planetary parameters to habitability, including stellar radius, stellar mass, stellar effective temperature, planet radius, and planet orbit semi-major axis.

Further analysis by **SHAP** showed that stellar radius, stellar mass, stellar effective temperature, and planet radius have different impacts on habitability than planet orbit semi-major axis. The relatively higher values in stellar radius, stellar mass, stellar effective temperature, and planet radius lead towards habitable exoplanet while the relatively higher value in planet orbit semi-major axis leads towards non-habitable exoplanet.

FUTURE WORK

- Train a **Neural Network** model for habitability prediction and feature importance analysis, and compare with tree-based models.
- Study **Planetary Systems** as a whole to understand what planetary systems might be more likely to host habitable planets.

KEY REFERENCE

- Seager, Sara. "Exoplanet Habitability." *Science* 340.6132 (2013): 577-581.
- Saha, Snehanshu, et al. "Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets." *Astronomy and computing* 23 (2018): 141-150.
- Basak, Suryoday, et al. "Habitability classification of exoplanets: a machine learning insight." *The European Physical Journal Special Topics* 230 (2021): 2221-2251.
- Lundberg, Scott. "A unified approach to interpreting model predictions." *arXiv preprint arXiv:1705.07874* (2017).
- Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." *Advances in neural information processing systems* 35 (2022): 507-520.

ideas | feature engineering

- looking at **single-host planetary systems**
    - analyze each planetary system with its host star and member exoplanets at the planetary system level.

ideas | feature engineering

- looking at **single-host planetary systems**
    - analyze each planetary system with its host star and member exoplanets at the planetary system level.
- determine stellar parameters

a certain planetary system

[host stellar parameters]

ideas | feature engineering

- looking at **single-host planetary systems**
  - analyze each planetary system with its host star and member exoplanets at the planetary system level.
- determine stellar parameters

a certain planetary system	[host stellar parameters]

- we can't analyze each individual planet in this planetary system analysis
  - we'll look at aggregated statistics about the planets that can offer us insight into their data without singling them out.

a certain planetary system	[host stellar parameters]	[a bunch of exoplanets in the system]

[parameters of the exoplanets within the planetary system on the aggregated system level]

model training

- we can then collect a bunch of these and build a **k-mean clustering model** that groups similar systems together
  - now we find patterns – what similarities do they have? can we derive a system of classification out of them?
- train a **random forest** and an **xgboost** (best for tabular data) to predict the probability of a planetary system to host a habitable exoplanet
  - and conduct **SHAP** analysis to find patterns - which features are most important? how they affect the probability?

a certain planetary system	[host stellar parameters]	[a bunch of exoplanets in the system]

[parameters of the exoplanets within the planetary system on the aggregated system level]

feature engineering – aggregated exoplanet parameters

a certain planetary system	[host stellar parameters]	[a bunch of exoplanets in the system]

[parameters of the exoplanets within the planetary system on the aggregated system level]

- %s of each exoplanet type
  - terrestrial, super-earth, neptune-like, gas giant
- distributions – maybe include min, 25%, 50%, 75%, and max for mass, radius, orbit semi-major axis, orbital period, orbit eccentricity, etc?
- "Framework for the architecture of exoplanetary systems" paper statistics – coefficients of similarity and variation on mass, radius, orbital eccentricity, etc.

feature engineering – aggregated exoplanet parameters

a certain planetary system

[host stellar parameters]	[a bunch of exoplanets in the system]

[parameters of the exoplanets within the planetary system on the aggregated system level]

2023 paper defining four different planetary systems

two ways to determine the planetary system architecture type:

1. coefficient of similarity – positive for ordered, negative for anti-ordered

$$C_s(q) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(\log \frac{q_{i+1}}{q_i} \right)$$

where q_i is some planetary quantity q (ex. mass, radius, orbital period, etc.) for the i^{th} planet in a system.

this is super helpful for distinguishing **ordered** and **anti-ordered** planetary systems

2023 paper defining four different planetary systems

two ways to determine the planetary system architecture type:

2. coefficient of variation – measure magnitude of variation in a set of numbers

$$C_v(q) = \frac{\sigma(q)}{\overline{q}}$$

"while similar systems will have a low value of the coefficient of variation, mixed systems will have a high value of coefficient of variation"

- %s of each exoplanet type
  - terrestrial, super-earth, neptune-like, gas giant
- distributions – maybe include min, 25%, 50%, 75%, and max for mass, radius, orbit semi-major axis, orbital period, orbit eccentricity, etc?
- "Framework for the architecture of exoplanetary systems" paper statistics – coefficients of similarity and variation on mass, radius, orbital eccentricity, etc.

that's all for this week. :)