Central Sound Regional Science & Engineering Fair (CSRSEF) 2025

Research Plan

Author: Christina Liu
Email: christinaliu2026@gmail.com
Date: January 2025

Project Title

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

Rationale

Provide a summary of your research. Highlight why it is important/interesting and describe any social/societal impact.

Are we alone in this universe? Are there any exoplanets out there other than Earth where humans can live? The search and discovery of potentially-habitable exoplanets beyond our solar system has been one of the most interesting active research fields in astrophysics throughout the past decade. A crucial part of the process is the research that goes into classifying these exoplanets so that they can be more easily identified and analyzed.

This research aims to apply machine learning (ML) techniques to exoplanet habitability classification, with a goal to build high quality machine learning models based on stellar and exoplanet data from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu) and the Habitable Worlds Catalogue (HWC), PHL @ UPR Arecibo (https://phl.upr.edu/hwc). The models built from this research are used to predict exoplanet habitability and identify the stellar and planetary parameters that most influence an exoplanet's habitability.

As of January 22, 2025, there are 5,830 confirmed exoplanets in the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu) dataset. With the enhanced observational capabilities of several ongoing satellite-based or telescope-based exoplanet discovery missions, e.g., the Transiting Exoplanet Survey Satellite (TESS, https://tess.mit.edu) and the James Webb Space Telescope (JWST, https://science.nasa.gov/mission/webb), the dataset of identified exoplanets will continue to grow. The machine learning methods from this research as well as the understanding of the most influential stellar and planetary parameters to exoplanet habitability can be applied to the newly discovered exoplanets to efficiently identify ones that are likely habitable for further study and exploration.

Research Question/Hypothesis(es)/Engineering Goal(s)/Expected Outcomes

Briefly explain how these relate to the rationale above.

This research aims to answer the following two questions:
(1) Given the data for exoplanets and their stellar hosts obtained from NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu), can we determine whether or not certain exoplanets are likely to be habitable?
(2) What are the stellar and planetary parameters that influence the exoplanet habitability the most?

I hypothesized that there exist some specific set of stellar and planetary parameters that, when combined, can largely determine the likelihood of the habitability for a given exoplanet. With careful data processing, feature engineering, and model training, a proper machine learning model should be able to identify the decision boundary between the habitable and non-habitable exoplanets among a large dataset with good precision and recall. I believed that the stellar and planetary parameters that influence exoplanet habitability the most could include stellar effective temperature, stellar radius, exoplanet orbit semi-major axis (i.e., the distance between the exoplanet and its stellar host), exoplanet radius, and exoplanet atmosphere composition.

The engineering goal of this research is to train machine learning models based on the data from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu) to predict exoplanet habitability with good precision and recall. Feature importance analysis is also conducted on the machine learning models to understand which stellar and planetary parameters play the most influential roles in the exoplanet habitability prediction. The expected outcomes include the machine learning models with high accuracy and the feature importance analysis results that largely match with my hypothesis of the most influential parameters to exoplanet habitability.

Procedures

Explain in some detail all procedures and experimental design created by you – do not include any work done by mentors or others associated with your project. Be sure to include information about your data collection methods.

The research involves the following major steps:
(1) **Data Fetching**: starts off by downloading exoplanet and stellar data from NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu), specifically the Planetary Systems Composite Data (https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PSCompPars), and the Habitable Worlds Catalog (HWC) dataset from PHL @ UPR Arecibo (https://phl.upr.edu/hwc/data), which contains a *P_HABITABLE* data field that can be relied on as the labels for the training data.

(2) **Data Processing**: the Habitable Worlds Catalog (HWC), PHL @ UPR Arecibo dataset contains a *P_HABITABLE* data field that identifies potential habitable exoplanets. *P_HABITABLE=1* indicates conservative habitable exoplanets (more likely to be rocky planets capable of surface liquid water), *P_HABITABLE=2* indicates optimistic habitable exoplanets (might include water worlds or mini-Neptunes, with less likelihoods of habitable conditions), and *P_HABITABLE=0* indicates non-habitable exoplanets. This dataset is joined with the Planetary Systems Composite Data from the NASA Exoplanet Archive and the *P_HABITABLE* data field is used to mark the labels for machine learning model training (*P_HABITABLE=1 or 2* as habitable while *P_HABITABLE=0* as non-habitable).

(3) **Feature Engineering**: feature engineering techniques are applied to pre-process the data to transform it into the feature dataset that is ready for machine learning model training, which includes but is not limited to: one-hot or label encoding for converting categorical features into numerical values, Multivariate Imputation by Chained Equation (MICE) technique for filling missing numerical values, Synthetic Minority Oversampling Technique (SMOTE) for oversampling minorities and Edited Nearest Neighbors (ENN) technique for downsampling majorities to balance training data, correlation analysis for identifying and removing highly correlated features, and MinMaxScalar for standardizing value ranges of all features to between 0 and 1.

(4) **Machine Learning Model Training**: machine learning models that are proper for this classification problem and dataset are selected. For each of the selected machine learning models, the feature dataset is randomly shuffled and then split it into a training set for model training and a test set for model evaluation. This process is repeated with multiple rounds and the mean accuracy is reported as representative of the model's performance. During training, hyperparameters are tuned and techniques are applied to minimize

the risk of overfitting.

(5) **ML Feature Importance Analysis**: after the machine learning models are trained and evaluated, feature importance analysis is conducted on the models to understand which features play the most influential roles in exoplanet habitability prediction. The results are then compared with the hypothesis.

(6) **ML Model Comparison**: the performance of machine learning models are compared to understand which models might be better suited to the given classification problem and dataset.

Risk and Safety

Briefly identify potential risks and the safety procedures taken to minimize those risks.

This research only involves downloading exoplanet and stellar data from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu) and the Habitable Worlds Catalogue (HWC), PHL @ UPR Arecibo (https://phl.upr.edu/hwc), which are widely open to public. As my research involves analyzing the data, and training machine learning models based on the data at home, there aren't any potential risks.

Data Analysis

Briefly describe the analysis performed on the data collected.

Exploratory Data Analysis (EDA) is applied on the dataset to identity patterns and trends, explore relationships between data fields, and uncover potential problems or anomalies. The EDA process includes but is not limited to:
(1) Descriptive statistics to summarize characteristics of the dataset, such as the count, min, max, mean, p25, p50, p75, p90, p95, p99, p99.9, and standard deviation for numerical data fields.
(2) Frequency tables to understand the distribution of categorical data fields.
(3) Bar charts to understand the distribution of categorical data fields.
(4) Histograms to visualize the distribution of numerical data fields.
(5) Scatter plots to understand relationships between two numerical data fields.

Correlation analysis is also applied on the dataset to identify highly-correlated data fields and remove them from the feature set, reducing data redundancy and collinearity for better preparation for the machine learning model training.

Bibliography

List major references such as books, journal articles, internet sites, etc. used in the preparing your project.

1. NASA Exoplanet Archive: Planetary Systems Composite Data. Retrieved from https://exoplanetarchive.ipac.caltech.edu/

2. The Habitable Worlds Catalog (HWC), PHL @ UPR Arecibo. Retrieved from https://phl.upr.edu/hwc

3. Seager, Sara. "Exoplanet Habitability." *Science* 340.6132 (2013): 577-581.

4. Kopparapu, Ravi Kumar, et al. "Habitable zones around main-sequence stars: new estimates." *The Astrophysical Journal* 765.2 (2013): 131.

5. 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.

6. Basak, Suryoday, et al. "Habitability classification of exoplanets: a machine learning insight." *The European Physical Journal Special Topics* 230 (2021): 2221-2251.

7. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

8. Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32.

9. Lundberg, Scott. "A unified approach to interpreting model predictions." *arXiv preprint arXiv:1705.07874* (2017).

10. Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. "SICE: an improved missing data imputation technique." *Journal of big Data* 7.1 (2020): 37.

11. 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.

12. Shwartz-Ziv, Ravid, and Amitai Armon. "Tabular data: Deep learning is not all you need." *Information Fusion* 81 (2022): 84-90.