Predicting Exoplanet Habitability from NASA using Machine Learning and Data Analysis

DS-450: Data Science Senior Capstone

Samia Mahmood

smahmood@bellarmine.edu

January 11, 2026

Executive Summary

With the discovery of exoplanets revolutionizing our understanding beyond our solar system, the ability to systematically analyze large-scale astronomical datasets has become increasingly important in characterizing exoplanets. This project will analyze data from NASA's Exoplanet Archive dataset which contains over one million data points and over 6,071 confirmed exoplanets discovered by space- and ground-based missions. This project aims to apply a structured data analytics framework and three machine learning models to conduct predictive analysis and classification.

This project proposes to predict the potential habitability of exoplanets based on key planetary and stellar features. The project will also identify the top 5 Earth-Like exoplanets based on how similar the key planetary and stellar features are with the features of Earth and Sun respectively. Using these key features, this project will also classify and characterize exoplanets into categories. This project will demonstrate the application of computational modeling in Astronomy and Astrophysics analyses to provide meaningful insights into exoplanet characteristics and accelerate the discovery process of exoplanets.

Project Idea

As of 2024, over 5,630 exoplanets have been discovered, and by 2026, this number has grown to over 6,071 confirmed exoplanets, primarily discovered through Transit and radial velocity surveys, with a majority of exoplanets identified by the Kepler, K2, and TESS missions [16]. While the number of exoplanet detections continue to rapidly increase, characterizing these planets and assessing their potential for harboring life has posed major challenges due to the variability, incompleteness, and high dimensionality of astronomical data. The focus of this project is to employ data-driven methodologies to analyze key features of the planets and their host stars by building upon previous work conducted by [28].

By leveraging unsupervised and supervised machine learning techniques and data-driven analysis, this project aims to explore relationships between stellar and planetary features, develop predictive models to determine the exoplanet's habitability potential, classify them by identifying their natural groupings, and determine the top Earth-analogous planets. The potential habitability of an Earth-Like exoplanet will be determined by whether it is a rocky Earth-sized or Super-Earth-sized planet orbiting Sun-like (G-type) Main Sequence stars within the star's

Habitable Zone (HZ) inside the Galactic Habitable Zone (GHZ) of the Milky Way Galaxy in accordance with established Astro biological models [10], [22]. This project will ensure computation results align with established astrophysical criteria. This project subsequently seeks to enhance our understanding of exoplanet formation, classification, and potential for habitability through data science and machine learning techniques.

Background

Within the past decade, significant improvements in high-precision space missions along with improved computational methodologies have enabled more advanced data analysis at a greater scale while fundamentally shaping our understanding of the Galaxy [1]. In 2009 NASA launched the Kepler mission to search for primarily Earth-Sized exoplanets orbiting the habitable zone (HZ) of Type-G (Sun-Like) stars [9]. Following its completion, Kepler's successors, K2 and TESS, extended the search to include exoplanets orbiting brighter and closer stars which helped improve occurrence-rate estimates. These missions transformed planetary science by increasing statistical analysis of high-precision data while leading to a rapid growth in data volume. This as a result, has necessitated the adoption of data-driven methodologies and machine learning techniques for efficient analysis and interpretation.

Early analysis of exoplanet data characterized planets based on their similarity to Earth and measured their habitability potential by developing habitability scoring frameworks. The Earth Similarity Index (ESI) developed by [18] quantitatively measures how Earth-like an exoplanet is based on four select features: planet surface temperature, planet density, planet mass, and escape velocity. Another habitability scoring metric developed by [18] is the Planetary Habitability Index (PHI) which is based on interior chemistry of the exoplanet to determine the planet's potential for sustaining life through the presence of liquid water. Another habitability index used is the Biological Complexity Index (BCI) proposed by [19] which focused on the biological and chemical composition, geophysical properties, age, and temperature of the planet to determine their potential for habitability. Similar to the Earth Similarity Index, the Cobb-Douglas Habitability Score (CDHS) was developed by [20], [21] and calculated using the parameters: planet surface temperature, planet density, planet mass, and escape velocity for planetary habitability assessments.

Prior studies utilized exploratory data analysis (EDA) with machine learning model comparison studies to characterize exoplanets, determine planet habitability, and explore parameter relationships [2], [3], [8], [11], [17].

Methods focusing on hybrid feature-selection to balance predictive performance and computational efficiency have been conducted by [7]. Applications of machine learning for astronomical data have leveraged supervised learning for classification and regression analyses. Research focused on exoplanet detection, false positive mitigation, and validation pipelines has utilized decision tree-based modeling for regression analysis along with K-Nearest Neighbor (KNN) and support vector machine-based modeling for transit classification [4], [23], [24]. Research conducted to estimate Earth similarity and perform habitability assessment has applied deep learning and regression modeling primarily through Decision Tree based modeling and Support Vector Machine learning algorithms [4], [5], [6]. In recent years, deep learning has transformed astrophysical and astronomical analysis. Convolutional Neural Networks (CNNs) paired with classification algorithms such as with K-Nearest Neighbor (KNN) have been used to classify transit detections and perform candidate validation [25]. CNN pipelines have been used to specifically detect weak shallow transit signals, further improving the efficiency of false-positive categorization [12]. Further research to improve sensitivity for Earth-Sized exoplanets and characterize noise sources using CNN architecture and TESS candidate vetting have been conducted by [13], [14], [15], [26]. While a majority of studies conducted on exoplanet data have utilized supervised learning, research conducted by [27] has recently introduced the application of unsupervised learning using K-Means clustering for exoplanet classification.

Current exoplanet analysis utilizes EDA, feature selection, and machine learning. A majority of machine learning techniques utilized for astrophysical analysis rely on supervised learning for planet classification and habitability predictions. However, few incorporate unsupervised learning. This study proposes to integrate unsupervised learning and supervised learning with structured reproducible EDA for habitability prediction, Earth-analogous prediction, and planetary classification. For unsupervised learning, this project will analyze the natural groupings of exoplanets for classification and determining Earth-Like candidates without a predefined habitability scoring metric.

Additionally, many exoplanet studies rely on established habitability scoring metrics such as the Cobb-Douglas Habitability Score (CDHS), Earth Similarity Index (ESI), Planetary Habitability Index (PHI), and Biological Complexity Index (BCI). These scoring metrics rely on a limited subset of features which may present limitations for high-dimensional data. Therefore, this project proposes to create a new habitability scoring metric using a comprehensive set of key planetary and stellar features and will serve as the target variable for supervised

learning to perform predictive analysis. To ensure robustness and comparability with established methods, the Earth Similarity Index (ESI) will be used as a benchmark target in cases where data completeness or feature availability limits the construction of the proposed metric.

The NASA Exoplanet Archive dataset contains 87 columns and 39,332 rows and provides a comprehensive open access collection of all exoplanets discovered thus far. This dataset contains key planetary and stellar features required for habitability analysis including planetary planet radius, planet mass, orbital period, orbital semi-major axis, planet equilibrium temperature, stellar type, stellar radius, and stellar effective temperature. This dataset enables large-scale statistical and comparative studies of exoplanet populations across diverse stellar environments. The dataset, however, presents several analytical challenges such as redundant attributes, duplicate rows for individual exoplanets, and presence of incomplete data. Addressing these challenges requires methodical preprocessing and feature selection to augment the data in order to apply statistical analysis and machine learning techniques.

Modeling

Using unsupervised learning, I will utilize K-Means clustering to classify exoplanets by identifying the natural groupings of exoplanets based on key planetary and stellar features. The K-Means clustering model will partition exoplanets into k clusters. This model will minimize the Euclidean distance between data points in standardized feature space. As a distance-based method, this approach groups exoplanets which exhibit similar feature properties, which will enable identification of Earth-analog candidates based on proximity to Earth-like features. This will support exploratory analysis and planetary classification without relying on predefined habitability labels.

Using supervised learning, I will employ Random Forest and Gradient Boosted Decision Tree (GBDT) modeling due to their robustness and ability to handle missing data, nonlinear relationships, and high dimensional features. Random Forest modeling will be used to predict habitability likelihood for exoplanets and estimate feature importance. Feature importance analysis will be used to interpret the astrophysical significance of predictive variables. The model will build an ensemble of decision trees which will train on bootstrapped samples. Each tree will then vote on classification based on whether the planet is habitable or non-habitable. Then a random feature selection is performed at splits to reduce overfitting. Gradient Boosted Decision Tree (GBDT) modeling will be used

to improve the predictive accuracy of Random Forest to determine the habitability potential of exoplanets and will allow for the ranking of Earth-Like exoplanets. The model will sequentially build trees where each tree corrects the errors of the previous one and will optimize a loss function related to habitability classification or scoring. The target variable used for predictive analysis for random forest and gradient boosted decision tree modeling will be the habitability score of the exoplanet.

Tools

This project will utilize Python due to its interpretable syntax and vast ecosystem of libraries used for machine learning and data science. For data handling and preprocessing, the Pandas and NumPy libraries will be utilized to allow for data manipulation, data cleaning, filtering, and handling missing data for high-dimensional datasets. These steps will be conducted to prepare the data for machine learning analysis.

Modeling will be conducted using the Scikit-learn library which enables the creation of machine learning workflows and data pipelines. For supervised learning, this library will enable classification using Random Forest modeling, predictive modeling using Gradient Booosted Decision Trees (GBDT), and unsupervised learning to identify exoplanet groupings using K-Means clustering. To explore more complex feature representations, the Tensorflow library may be utilized.

The Matplotlib and Seaborn libraries will be used for data visualizations to generate plots and display feature distributions, correlation matrices, and classification results. This will ensure data interpretability and convey results in a comprehensive manner understood by both technical and non-technical viewers. Jupyter Notebook will be utilized for the development environment to display analysis workflow and aid in reproducibility due to its ability to integrate code, visualizations, and documentation.

Conclusion

This project presents a data-driven approach for exploring exoplanet habitability by integrating exploratory data analysis with unsupervised and supervised machine learning. Using the NASA Exoplanet Archive, this project addresses the current challenges in efficiently analyzing large-scale high-dimensional astronomical data for planet characterization and habitability assessment. By leveraging K-Means clustering alongside Random Forest and Gradient Boosted Decision Tree modeling, this study will predict exoplanet habitability, determine the top 5 Earth-

like exoplanets, classify exoplanets through their natural groupings, and study feature relationships using key stellar and planetary features. Overall, this study will highlight the importance of machine learning applications in Astrophysics and Astronomy.

References

[1] S. Sen, S. Agarwal, P. Chakraborty, and K. Singh, “Astronomical big data processing using machine learning: A comprehensive review,” Experimental Astronomy, vol. 53, no. 2, pp. 333–368, 2022.

[2] J. H. Jiang, P. E. Rosen, C. X. Liu, Q. Wen, and Y. Chen, “Analysis of habitability and stellar habitable zones from observed exoplanets,” Galaxies, vol. 12, no. 6, p. 86, 2024.

[3] J.-V. Rodrıguez, I. Rodrıguez-Rodrıguez, and W. L. Woo, “On the application of machine learning in astronomy and astrophysics: A textmining- based scientometric analysis,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 12, no. 5, p. e1476, 2022.

[4] P. Pratyush and A. Gangrade, “Automation of transiting exoplanet detection, identification and habitability assessment using machine learning approaches,” in arXiv preprint arXiv:2112.03298, 2021.

[5] R. Jagtap, U. Inamdar, S. Dere, M. Fatima, and N. B. Shardoor, “Habitability of exoplanets using deep learning,” in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE, 2021, pp. 1–6.

[6] G. K. Jayan, J. Mathew, C. Stephen et al., “Artificial intelligence (AI) to predict earth similarity index (ESI): An analysis of regression models on esi data.” Grenze International Journal of Engineering & Technology (GIJET), vol. 10, 2024.

[7] G. Koumbis, “Examining deep learning artificial neural networks with hyperparameter optimization and ensemble learning in the search for habitable exoplanets: An analysis,” ProQuest Dissertations and Theses, 2024.

[8] M. S. Jakka, “Assessing exoplanet habitability through data-driven approaches: A comprehensive literature review,” arXiv preprint arXiv:2505.11204, 2025.

[9] W. Borucki, D. Koch, G. Basri et al., “The Kepler mission: Finding the sizes, orbits and frequencies of earth-size and larger extrasolar planets,” in Scientific Frontiers in Research on Extrasolar Planets, ser. ASP Conference Series, vol. 294, 2003.

[10] C. H. Lineweaver, Y. Fenner, and B. K. Gibson, “The galactic habitable zone and the age distribution of complex life in the milky way,” Science, 2004.

[11] S. Basak, A. Mathur, A. J. Theophilus, G. Deshpande, and J. Murthy, “Habitability classification of exoplanets: a machine learning insight,” The European Physical Journal Special Topics, vol. 230, no. 10, pp. 2221–2251, 2021.

[12] C. Shallue and A. Vanderburg, “Identifying exoplanets with deep learning: A five planet resonant chain around kepler-80 and an eighth planet around kepler-90,” The Astronomical Journal, 2018.

[13] S. Zucker and R. Giryes, “Shallow transits—deep learning. I. feasibility study of deep learning,” The Astronomical Journal, 2018.

[14] “A convolutional neural network based ensemble model for exoplanet detection,” 2020.

[15] L. Yu, A. Vanderburg et al., “Identifying exoplanets with deep learning III: Automated triage and vetting of tess candidates,” The Astronomical Journal, 2019.

[16] NASA Exoplanet Archive. NASA Exoplanet Archive: Planetary Systems Composite Data. 2026. Available online: https://exoplanetarchive.ipac.caltech.edu/ (accessed on 11 January 2026).

[17] E. Yılmaz, M. E. Artan, and A. B. Yanartaş. “EXOLIFE: Detection and Habitability Estimation of Exoplanets Using Machine Learning Techniques,” Turkish Journal of Remote Sensing, vol. 6, no. 2, pp. 85–96, 2024.

[18] D. Schulze-Makuch, A. Mendez, A.G. Fairen, et al., “A two-tiered approach to assessing the habitability of exoplanets. Astrobiology,” vol. 11, no. 10, 2011.

[19] L.N. Irwin, A Mendez, A.G. Fairen, D. Schulze-Makuch, “Assessing the Possibility of Biological Complexity on Other Worlds, with an Estimate of the Occurrence of Complex Life in the Milky Way Galaxy,” Challenges Journal for Planetary Health, vol. 5, no. 1, pp. 159-174, 2014.

[20] S. Agrawal, S. Basak, S. Saha, K. Bora, J. Murthy, “A Comparative Analysis of the Cobb-Douglas Habitability Score (CDHS) with the Earth Similarity Index (ESI),” arXiv:1804.11176v1, 2018.

[21] K. Bora, S. Sahab, S. Agrawal, M. Safonova, S. Routh, A. Narasimhamurthy, “CD-HPF: New Habitability Score Via Data Analytic Modeling,” arXiv:1604.01722v1, 2016.

[22] J. F. Kasting, R. Kopparapu, R. M. Ramirez, and C. E. Harman, “Remote Life Detection Criteria, Habitable Zone Boundaries, and the Frequency of Earthlike Planets around M and Late-K Stars.” Proceedings of the National Academy of Sciences (PNAS), vol. 111, no. 35, pp. 12641-12646, 2013.

[23] S. Saha, S. Basaka, K. Borab, M. Safonovc, S. Agrawala, P. Sarkara, J. Murthyd, “Theoretical Validation of Potential Habitability via Analytical and Boosted Tree Methods: An Optimistic Study on Recently Discovered Exoplanets,” arXiv:1712.01040v1, 2017.

[24] D. Chia-Tien Lo, et al, “Exoplanet Detection Using Machine Learning Models Trained on Synthetic Light Curves,” arXiv:2507.19520v1, 2025.

[25] N. Schanche, et al, “Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys” Royal Astronomical Society, vol. 483, no. 4, pp. 5534–5547, 2019.

[26] A. Chaushev, et al, “Classifying Exoplanet Candidates with Convolutional Neural Networks: Application to the Next Generation Transit Survey,” Royal Astronomical Society, vol. 483, no. 4, pp. 5232–5250, 2019.

[27] Y. Jin, L. Yang, C. Chiang, “Identifying Exoplanets with Machine Learning Methods a Preliminary Study,” International Journal on Cybernetics & Informatics (IJCI). vol. 11, no.1, 2022.

[28] S. Mahmood, I. Rahman, S. Sarkar, and A. Mahmood. (2026). Discovering Earth 2.0: A data-driven exploration of NASA’s Exoplanet Dataset. *Proceedings of the IEEE International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2026)*. (Accepted for publication).