Exploring Potential Habitats Beyond Earth through a Multilayered Statistical Analysis of NASA's Confirmed Exoplanets Dataset

Anumanchi Agastya Sai Ram Likhit[1], Chintada Lohind Sai Durga[2]

[1]Indian Institute of Science Education and Research Bhopal [2]Koneru Lakshmaiah Education Foundation

Abstract

This study introduces "ExoCluster," a composite model combining ensemble machine learning in multiple layers to identify potentially habitable exoplanets within NASA's Confirmed Exoplanet Dataset. The first phase employs ensemble machine learning to establish initial exoplanet clusters based on the most important parameters. The second phase further refines these clusters by incorporating additional important habitability parameters, analyzing the data, and doing clustering using ensemble machine learning. This dual-layer approach enhances precision in predicting potential habitable planets. ExoCluster effectively filters the extensive list of over 5,535 confirmed exoplanets, identifying promising candidates for in-depth spectroscopic exploration. This model demonstrates the significant potential of integrated statistical methods in advancing the field of exoplanetary research.

Introduction

One of the most interesting questions we've always had about the universe and our place in it is, "Are we alone?". Detecting and studying exoplanets is very important for helping us figure out where we fit in the universe and if there is life elsewhere.

With thousands of exoplanets already found, it's very crucial to figure out which ones might have life. NASA's Confirmed Exoplanet Dataset has a long list of these discoveries. The hard part is to analyze this data to find planets that might be like Earth in terms of being habitable.

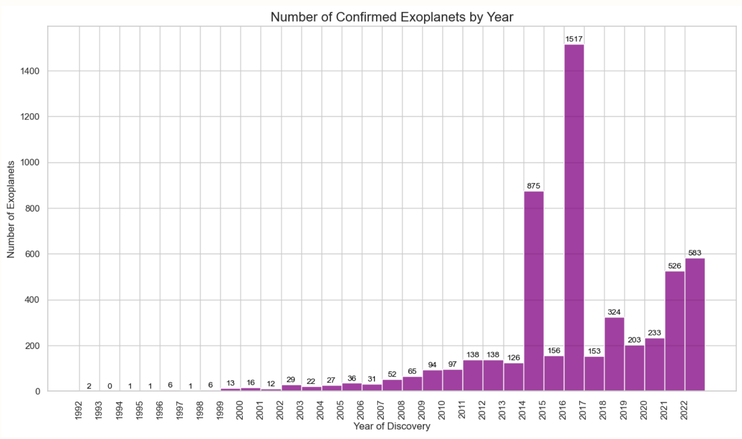


Figure 1: Annual Discovery of Confirmed Exoplanets

In this part of our study, we propose "ExoCluster" a model that combines the power of ensemble machine learning and other statistical tools to analyze this vast dataset. Using this, we aim to find potential habitats in the dataset.

ExoCluster model doesn't just randomly picks planets; it methodically categorizes them based on various parameters that are crucial for habitability.

Methodology

The NASA Exoplanet Data Set contains the details of 5,535 planets, which include many planetary and stellar properties. Here we have two phases. In each phase, we try to cluster the exoplanets and try to find the suitable cluster. In each phase, we combined three powerful clustering methods to understand our data better: 1) KMeans; 2) Hierarchical Clustering; 3)Gaussian Mixture Model (GMM)

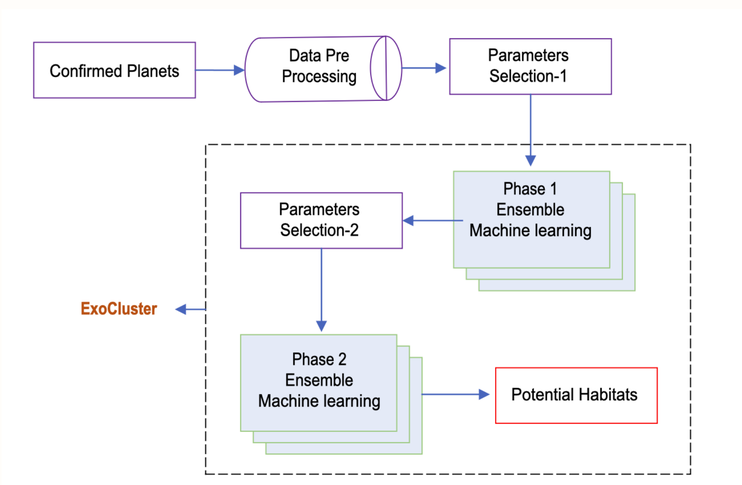


Figure 2: Data Flow

After applying these techniques separately, we created a unique map (co-association matrix) showing how often each pair of data points ended up in the same cluster across all methods. We then transformed this map into a new landscape, known as the Dissimilarity Matrix, which represents the differences between data points. Finally, we applied hierarchical clustering again to this new landscape to get our final grouping. This ensemble approach combines the strengths of each method, leading to a more accurate and reliable understanding of our data. In the first phase, we consider two parameters. 1. how far the planet is from us; this enables us to consider the possibilities as a next earth and to do observations; and 2. how far the planet is from its host star to see whether the planet is in a habitable zone or not and conduct the ensemble learning, which gives us a list of 205 exoplanets.

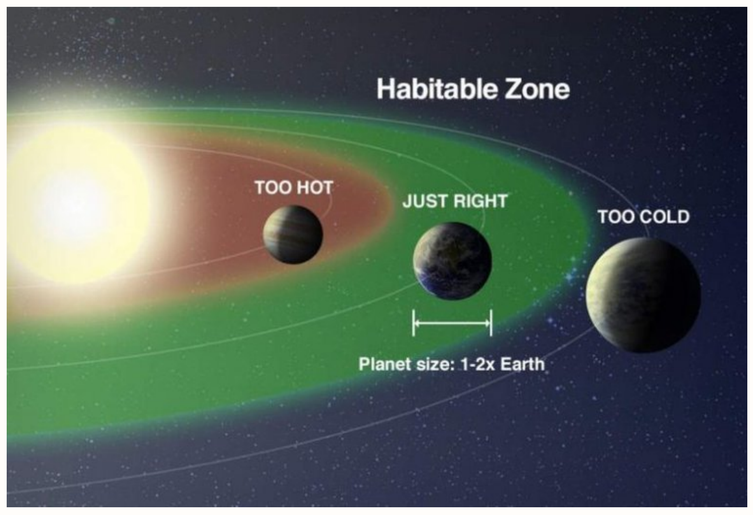
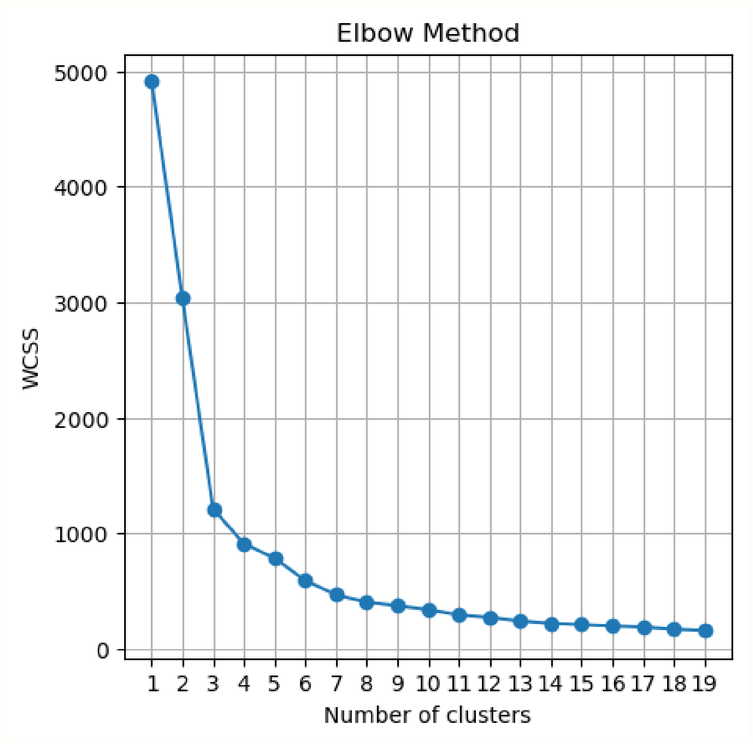


Figure 3: Habitable Zone

In the second phase, we conducted exploratory data analysis on the shortlisted candidates chosen from phase one, and we strategically chose the parameters of planet mass, host star mass, and planet orbit eccentricity. These parameters cover important features for habitability and help minimize the count of missing values in the dataset. We then conduct ensemble learning, make clusters, and try to find the suitable cluster that maps to earth.

Phase 1 Observations and Results

The decision on the number of clusters to form is crucial. By using the elbow method(the elbow method is a graphical method for finding the optimal K value in a k-means clustering algorithm), we determine a logical value for the optimal number of clusters, as shown in the below figure.



The elbow method plots the Within-Cluster Sum of Squares (WCSS) against the number of clusters. As clusters increase, WCSS decreases, but at the 'elbow' point, this decrease slows significantly. From this point, further increases in clusters do not significantly improve compactness, indicating the optimal number of clusters.

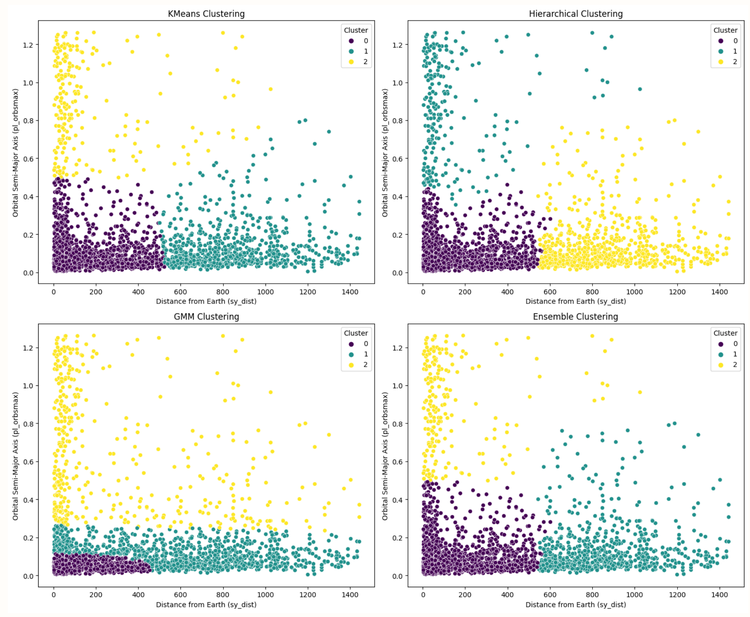


Figure 4: Clustering Phase 1

Model	Cluster1	Cluster 2	Cluster 3
KMeans	1499	218	739
Hierarchical	1515	229	712
GMM	982	1051	423
Ensemble	1539	712	205

Table 1: Cluster Distribution Across Different Techniques

Phase 2 Observations and Results

After careful analysis of different parameters and their relations with others, we have chosen the parameters for phase two, and using the elbow method, we got the optimal number of clusters for phase two at four

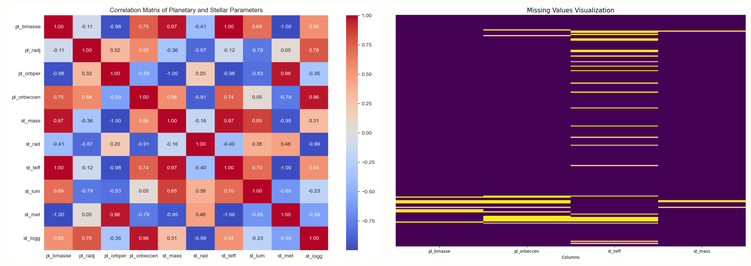


Figure 5: E.D.A on Shortlisted Candidates

Model	Cluster 1	Cluster 2	Cluster 3	Cluster 4
KMeans	91	60	37	17
Hierarchical	119	21	22	43
GMM	111	32	35	27
Ensemble	59	32	78	36

Table 2: Cluster Distribution Across Different Techniques

Cluster 3 is the final shortlisted cluster as earth lies in that cluster.

Conclusion

Using this model, we can focus on the shortlisted candidates as a priority for spectroscopic observations and do more observations on them.

References

[1]

42nd Meeting of the Astronomical Society of India 2024 (ASI - 2024), IISc Bangalore