Can the Correlation of Stellar Metallicity and Exoplanet Properties Determine Planet Habitability?

Sathvik Malla & Tony Rodriguez

Received June 09, 2024
Accepted August 20, 2024
Electronic access September 15, 2024

Stellar metallicity can help determine atmospheric composition. Typically, a high stellar metallicity means that the star is very metal-rich. The metals can indicate that the atmospheric composition of the star's planets contains life-supporting gases like CO2 and O2, and a metal-rich star indicates an abundance of those gases. To answer whether correlation of stellar metallicity and exoplanet properties determine planet habitability, we used a Gaussian clustering model to find the correlation between the stellar metallicity and the planet's properties like mass and radius. The clusters will help us figure out what types of masses and radii are within higher or lower stellar metallicities. The silhouette score, which represents how definite the clusters are, is around 0.67, indicating that the clusters are strong and could give us strong evidence for explanations of these correlations. From the results of the clustering model, we can see that the clusters for the mass and the stellar metallicity are clearly discrete compared to the clusters created for the radius and the stellar metallicity. Thus, we can see that the majority of the heavier masses of the exoplanets have larger stellar metallicities than their stellar hosts. Since the model shows that more massive planets tend to have host stars with higher metallicities, it could indicate that chemicals such as CH4, H2, and O2 are present in those exoplanet atmospheres.

Introduction

Every day, astronomers have been exploring and discovering exoplanets that could potentially be places that can support life since roughly the late 1990s, which means it is a relatively new field of study in astronomy. However, it is challenging to discover this when exoplanets are light-years away as they are so much dimmer compared to their host stars. Therefore, they use advanced techniques such as spectroscopy, which is based on principles of refraction and diffraction, to study chemicals in the atmosphere that determine what chemicals are in each exoplanet's atmosphere. Astronomers and scientists do this for two reasons. One reason is to answer the popular question: is there life out there in the universe? The second reason is to wonder whether humans can find another planet to thrive with good living conditions if Earth is no longer a safe place to stay. Astronomers go through hundreds of exoplanets every day through large datasets, filtering out which exoplanets are confirmed or not, and if they have proper living conditions for general life or humans. However, no confirmed result or exoplanet fits these conditions, and it is important to speed up the results and generate faster results so astronomers and scientists can achieve their goals.

How are exoplanets determined? A common method to identify exoplanets is by taking pictures of exoplanets through high-quality telescopes. Space agencies like NASA send powerful telescopes into space, capturing detailed pictures of planets, stars, neutron stars, and other unknowns of the universe. Astronomers interested in exoplanets take pictures of the newly discovered exoplanets from telescopes such as Kepler, Tess, and James Webb; however, their distance to Earth means it will take years to reach Earth. Astronomers also use transit photometry, which is an observation of a star's brightness or flux over time. If there is a periodic decrease in flux over time, this could mean that there is an exoplanet revolving around that star. There are many more techniques involved in exoplanet discovery, but the methods mentioned above are usually the most accurate and the least time-consuming.

We will be mainly focusing on exoplanets and their stellar hosts' properties. Exoplanet properties are crucial information in learning about them, but they will not provide much data to understand if planets are habitable. A straightforward and reliable way to determine planet habitability is to observe planets inside the habitable zone of stars. However, if we want to determine exoplanet habitability outside the habitable zone, the best way to understand planet habitability is to look at star flux. The star flux determines not just the chemicals in the atmosphere, but also the stellar metallicities and their ratios that determine if there is an abundant amount of those chemicals that can support life. Larger ratios mean higher metallicities and smaller ratios mean lower metallicities. In other words, the more metals present in a host star, the more likely it is that its planets originated from a similar chemical composition.

A star is mainly composed of hydrogen and helium with small percentages of other elements and metals such as oxygen and iron. To define all these quantities, we can represent the

hydrogen proportion as A, the helium proportion as B, and the other metals and elements proportion as C. Therefore, we can say that A+B+C = 1. This is the general way to define the chemical composition and metallicity of a star. However, in this research experiment, we will measure stellar metallicity based on the Sun's metallicity, as mentioned above, which is the Iron-to-Hydrogen ratio or the [Fe/H] ratio. To measure the [Fe/H] ratio of other stars relative to the Sun, we can use equation (1).

$$[\text{Fe/H}] = \log\left(\frac{(\text{Fe/H})_*}{(\text{Fe/H})_\odot}\right) = \log\left((\text{Fe/H})_*\right) - \log\left((\text{Fe/H})_\odot\right) \quad (1)$$

The abundance ratio [Fe/H] of the Sun is 0. The range of metallicities stars with the ratio [Fe/H] can have is between -4.5 and 1.0.[1].

In this paper, we will discover a way to find exoplanet habitability the easy way through correlations between exoplanet properties and stellar properties. Specifically, we will be looking through the mass and the radius of an exoplanet as they are great information to help us understand the density of the exoplanet, its size, and possibly its gravitational pull. The stellar properties we specifically investigate are stellar metallicities (specifically [Fe/H]), and the effective temperatures measured in Kelvin. We will use machine learning models, which have been used sparingly in previous analyses, to create our correlations and other data visualization methods of our solar system planets to understand the general scope of the data.

Background

People have attempted to find if chemical composition in exoplanet atmospheres correlates to exoplanet properties like exoplanet mass and radius. In the nature.com article (put link), the researchers have strived to discover if there is a possible correlation between exoplanet size and its stellar hosts' metallicity. However, their results suggest that there isn't any correlation between those two, and extraterrestrial planets do not correspond with metal abundance. These results cannot be conclusive as they probably need a larger sample size to produce accurate results, numbers, and correlations. More importantly, the study did not use machine learning to separate the two classes of exoplanets and identify specific trends in them.

Other researchers believe that stellar metallicities are a significant factor in determining planet formation and evolution of planetary systems. Not only were they able to identify the significance of stellar metallicities, but they were able to note the correlations between metallicities and their planet sizes. After accessing the ever-expanding dataset of exoplanets, they were able to discover "the average metallicity of the solar-like stars known to have giant planets is higher compared to F, G, and K dwarfs not known to host giant planets."[2]. However, the sample size of the outdated article influences the trends and correlations they have found as they have changed over time due to newly discovered exoplanets.

Researchers have noted how the abundance of methane, or CH4, is detected in F, G, or K-type stars (sun-like stars) and their Earth-like planets that are near their stellar hosts' habitable zones[3]. The study and the data have shown that methane is critical in understanding an exoplanet's habitability, receiving data from planet-to-star distance and orbit eccentricity. Stellar radiance and photodissociation are crucial factors of methane production, which is why the researchers observe the stellar metallicity of these F, G, or K-type stars. This will be helpful in our research when finding which metallicities and exoplanet sizes ensure habitability and methane presence in exoplanet atmospheres.

Dataset

We explored and modeled through the NASA exoplanet archive public dataset, which provides information not just on exoplanet properties, but with stellar properties that could provide more information about exoplanet habitability. The numerical dataset consists of 34,892 exoplanet detections (sometimes with various detections of the same planet) regardless of controversial flags and other factors that determine planet existence. The data has been compiled from telescopes on Earth and launched into space such as the James Webb Space Telescope and more. The types of data the archive hosts include exoplanet mass (M*sin(i)) and radius, stellar metallicities, its ratios, its effective temperature (K), and planet orbit eccentricity. In our clustering model, we will focus on the exoplanet mass and radius, and their stellar host's metallicity depending on its type of metallicity ratio. We believe that a planet's mass and radius are good enough to determine planet size, and we will compare these properties to Earth to scale them with the typical properties needed for life support. Stellar metallicities give us good insight into the atmosphere composition of exoplanets. The sun has a metallicity ratio of [Fe/H], which we observe as critical for life as it supports organisms on its very own planet, the Earth. We filtered out the data from 34,892 to 17,026 exoplanets to only observe non-controversial planets with stars of metallicity ratios [Fe/H] in our model. We will use data transformation and take the logarithm of planet mass and radius to effectively observe the data distributions.

We can clearly see the distributions of the planet masses and radii in our dataset, which seems fair since it displays a rough Gaussian distribution, which is what we expected and gives us the confidence to proceed to cluster our data. However, if the graph is extremely skewed left or right, we would need to modify our data more, such as taking the logarithm of the number of planets on the y-axis. With a large enough sample size of independent exoplanets, and relatively close Gaussian

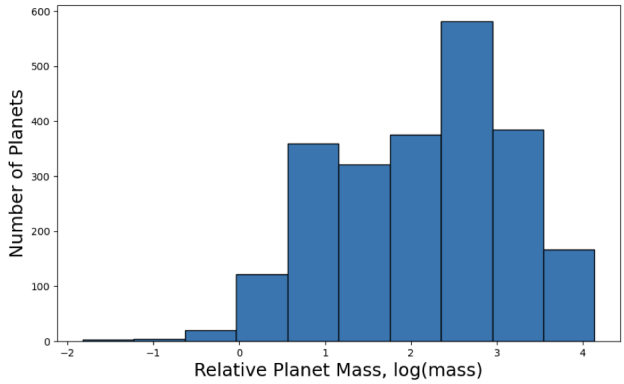

Fig. 1 The number of planets and the logarithm of planet masses (relative to Earth Mass).

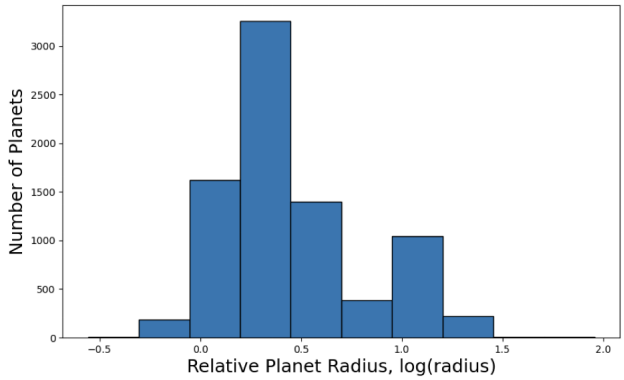

Fig. 2 The number of planets and the logarithm of planet radii.

distributions, we believe this is enough evidence to prove our correlation between planet properties and stellar metallicities.

Materials and Methods

In this research problem, we will investigate my research models through Google Colab's jupyter notebook. We chose to use clustering models as they best represented the types of planet sizes that correspond to high or low stellar metallicities. Using clustering prediction models, we can see the trends of large and small exoplanets and how their properties relate to metallicities. We used the Gaussian Mixture clustering model from Sci-kit Learn as the clusters created by the model were easy to observe and locate. We attempted other clustering models like Birch and KMeans and used the clustering models with different numbers of clusters to examine which clustering models suit this research investigation. We examined the silhouette scores of each graph, which determines how well the clusters are defined, and we examined them for both the logarithm of the exoplanet radius and logarithm of the exoplanet mass versus their star's corresponding stellar metallicity.

We observed the chart below of the most popular and useful clustering models based on Sci-Kit Learn's training and testing on sample datasets to provide important information on which clustering model could suit which types of data.

We observed which types of clusters our dataset will most likely have, and the clusters we most likely need to pay attention to which clusters fit with which ranges of stellar metallicities. Looking at the sixth row, which most likely can represent our data the best, we see that KMeans, BIRCH, and Gaussian Mixture are the most optimal for our data to test and see which ones can work. The bottom right corner shows the silhouette scores of each model, which is an important factor in determining the best models for our dataset while not using too much memory in the process.

Gaussian Mixture Model

The Gaussian Mixture model is a popular mixture clustering model that assumes that every data point is created from a mixture of many Gaussian distributions with different and unspecified parameters. These Gaussian distributions, which are also known as normal distributions, can be affected by sample size, mean, or standard deviation of the sample. To define the clusters, the Gaussian Mixture models compare similarities to each data point and fit the data through a process called maximum likelihood estimation, which is an effective way to assume probability distribution parameters. It is a flexible method but requires a lot of hyperparameter tuning that we need to make assumptions for covariance matrices. To represent the combined Gaussian mixture model, we can use equation (2).

$$N(\mu,\sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\sigma^{-1}(x-\mu)\right) \quad (2)$$

μ is the mean of the dataset, σ is the covariance matrix, d is the number of features, and x is the number of data-points.

KMeans Clustering Model

The KMeans clustering model is one of the simplest clustering models out there that uses unsupervised machine learning algorithms to cluster data points based on specific similarities. The "k" in KMeans refers to the number of defined clusters the model needs to look for when going through the dataset. The "means" in KMeans refers to averaging out the dataset to find specific centroids of those defined clusters. It selects random centroids of clusters, and depending on those parameters, uses calculations to find an optimized cluster with a stable centroid over a desirable number of iterations. Implementation requires user input on the number of clusters you want the model to define. One of the disadvantages of KMeans is that it is sensitive to outliers, and in a huge dataset like the

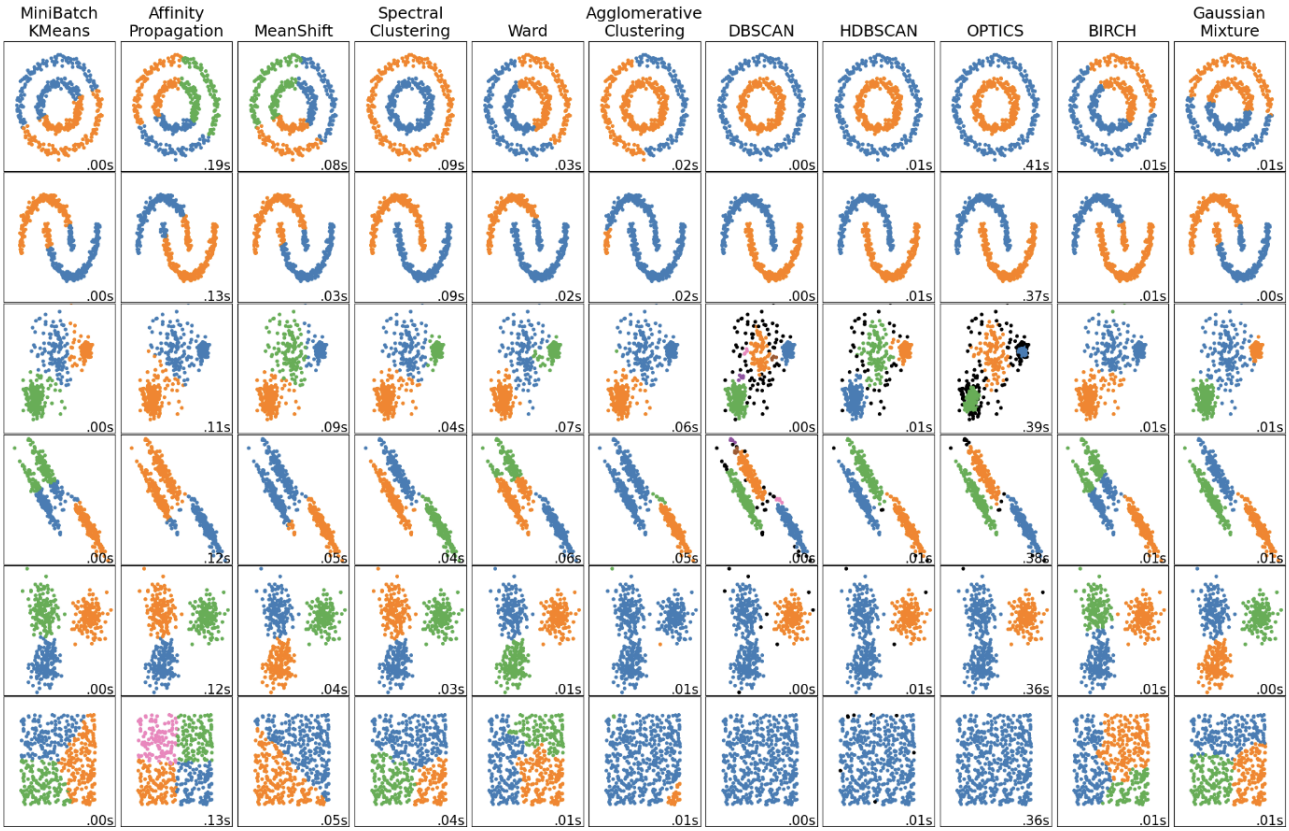

Fig. 3 A detailed chart showing different clustering models available in Sci-kit Learn with different data.[4]

NASA exoplanet archive, outliers will be common. Additionally, KMeans assumes that the dataset has clusters that look like spheres, which brings a huge limitation to clustering. We can use the objective function J (3) to represent the KMeans clustering model.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| \left(x_i^j - c_j \right) \right|^2 \tag{3}$$

The number of clusters is defined by k, n is the number of cases, i and j are two different cases, and c is the centroid of case j.

BIRCH Clustering Model

We will investigate one more clustering model that could also fit our clusters, which is the BIRCH clustering model. The BIRCH clustering model, also known as Balanced Iterative Reducing and Clustering using Hierarchies, is a memory-efficient model and great for large datasets to make defined clusters based on specific similarities between data points. Instead of directly making clusters from a dataset, it generalizes the data while preserving as much data as possible before making clusters on that generalization. One of the major limitations of BIRCH is that it only clusters based on metric attributes and not on categorical ones. The model consists of two stages: it first builds the clustering feature tree, and the second part is Global Clustering, hence why BIRCH is also known as Two Step Clustering. The data is loaded into the Clustering Feature tree, or the CF tree, condense the data with a smaller tree, and use global clustering and refining to obtain the best clusters achievable. With a large dataset, it is important to use some clustering models like BIRCH to handle the data, but it may not be ideal as it is important to make specific clusters based on specific data points and not generalizations of those data points. With the BIRCH model, we can derive three important statistics (4) (5) (6) of our cluster.

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{N} \tag{4}$$

$$R = \sqrt{\left(\frac{\sum_{i=1}^{N} (x_i - x_0)^2}{N} \right)} \tag{5}$$

$$D = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2}{N(N-1)}} \tag{6}$$

The centroid location is at x_0, R is the radius of the cluster, and D is the diameter of the cluster. There are also other formulas that cover other ways to measure and confirm distances between clusters.

Final Decisions

After observing all these clusters, the Gaussian Mixture clustering model can cluster our model the best as it is a well-known and capable clustering model for our large dataset, and since we need to pay attention to specific data points, the Gaussian Mixture model is effective at making good assumptions while not being super sensitive to outliers.

In our Gaussian Mixture clustering model, we decide to only define two clusters without having any randomness in our data. We fit our data in the model based on estimated model parameters and store the predictions of our model into an array. We then display our model's predictions with color-coded clusters using a scatter plot to locate specific data points of our dataset. We would use the same procedure for the other two models, which are shown below as further evidence in deciding that the Gaussian mixture clustering model will work the best for our research.

Results and Discussion

Before modeling my data with the Gaussian Mixture model, the graphs below will show how we made the decision ultimately to create the Gaussian Mixture model. We started by creating three stellar metallicities versus logarithm of the planet radius graphs and using each clustering model we chose above to fit the dataset into clusters in each graph. The yellow and purple dots represent two different clusters depending on planet size. The dots placed on the top of the graph represent the larger planets, and the dots placed on the bottom represent the smaller planets.

The silhouette scores of the Gaussian Mixture, KMeans, and BIRCH models are 0.64, 0.63, and 0.66, respectively. Usually, silhouette scores that are lower than 0.5 are not considered as they will not have clearly defined clusters. Next, we will observe our three models through another factor, which is the logarithm of the planet mass, as part of our evaluation process to choose the most effective model:

The silhouette scores for these three models are 0.67, 0.66 and 0.66 respectively. Therefore, since we observe that the Gaussian Mixture model outputs more defined clusters and overall better silhouette scores, we would choose this model for our data observations. Using the specific clusters, we will inspect how these exoplanet properties compare to our solar system's properties to scale and observe which conditions are ideal for supporting life.

We can observe here that most of the planets in this study have larger radii and masses, which we can see from the above

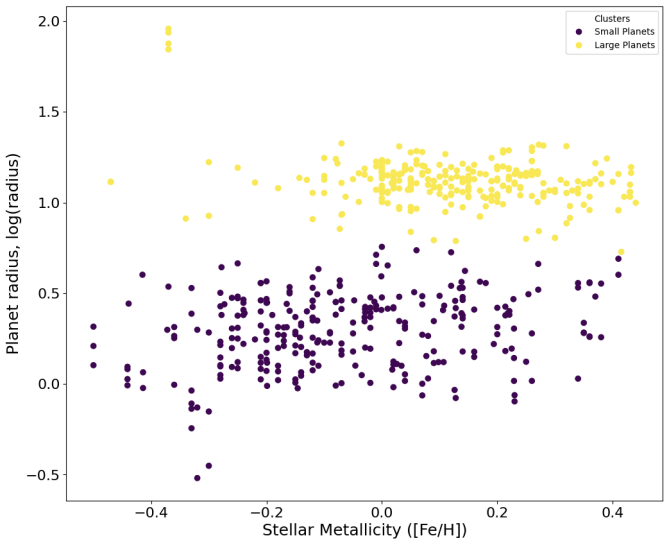

Fig. 4 Clusters modeled by Gaussian Mixture Model between the logarithm of planet radius and stellar metallicity ([Fe/H])

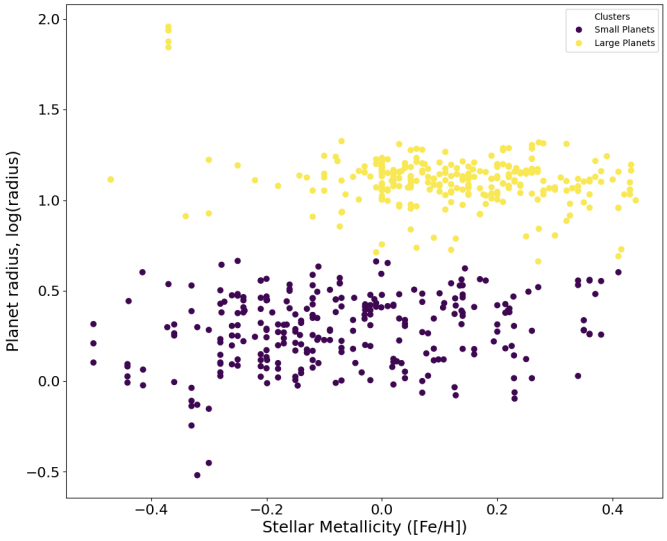

Fig. 5 Clusters modeled by KMeans Model between the logarithm of planet radius and stellar metallicity ([Fe/H]).

graphs could determine more defined clusters than the planets with smaller sizes. However, that leaves a very small percentage of our dataset that matches with Earth, which in the introduction explains why scientists and astronomers are struggling to find a habitable planet with the right size and conditions as Earth.

Let us look at a revised version of our Gaussian Mixture Model Fitting Logarithm of the Planet Mass (Earth Mass) versus Stellar Metallicity to pay close attention to the clusters to see specific trends:

We can see that the large cluster near the top right corner refers to the abundance of massive exoplanets with larger Earth

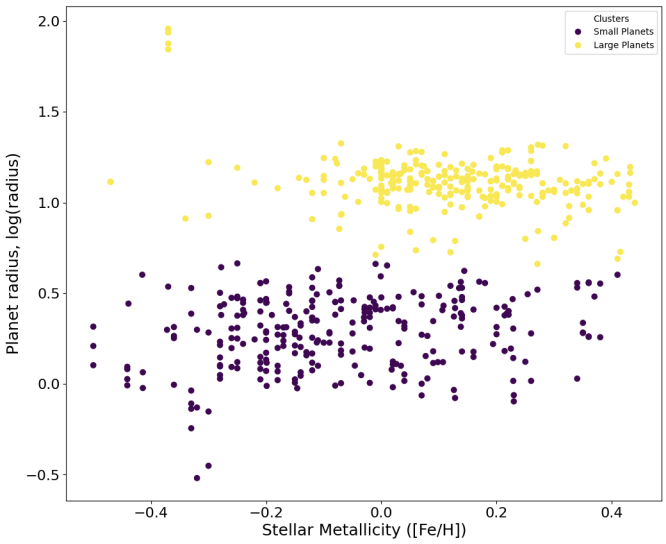

Fig. 6 Clusters modeled by BIRCH between the logarithm of planet radius and stellar metallicity ([Fe/H])

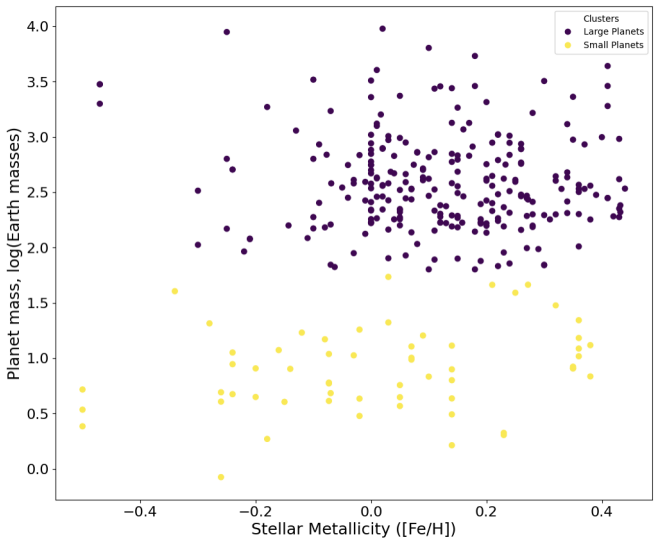

Fig. 8 Clusters modeled by KMeans Model between the logarithm of planet mass (Earth Mass) and stellar metallicity ([Fe/H]).

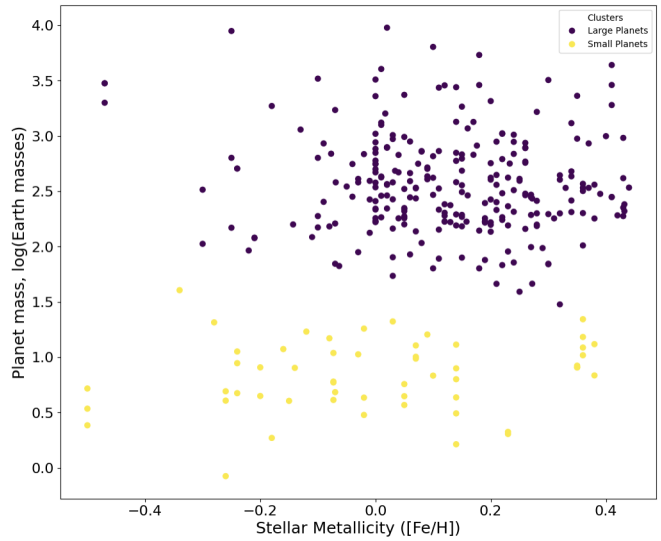

Fig. 7 Clusters modeled by Gaussian Mixture Model between the logarithm of planet mass (Earth Mass) and stellar metallicity ([Fe/H]).

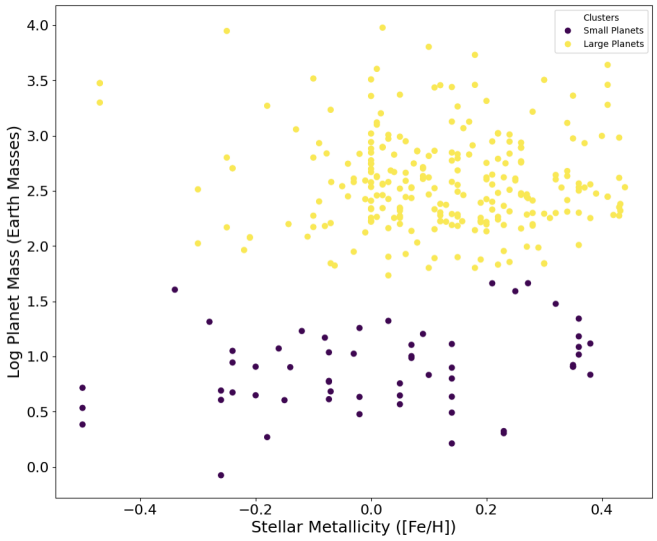

Fig. 9 Clusters modeled by BIRCH Model between the logarithm of planet mass (Earth Mass) and stellar metallicity ([Fe/H]).

masses and radii. From this clustering, we can observe that planets with larger Earth masses tend toward larger stellar metallicities with the exception of very few outliers. However, the cluster in the bottom left corner is not defined as clearly due to the limited statistics of Earth-sized planets discovered as mentioned previously. However, a significant number of the smaller exoplanets do tend to have lower and negative stellar metallicities.

Figure 13 shows that the same trend displayed in the Logarithm of the Planet Mass versus Stellar Metallicity graph is shown in the graph above. The planets with larger radii tend to have larger stellar metallicities, but due to limited statistics of Earth-sized planets or similar, the trend is unclear. However, we can observe that most planets have stellar hosts with lower and negative stellar metallicities.

Conclusion

The Gaussian Mixture model, which best represents the desired clusters, demonstrates that larger planets with larger masses and radii correlate to higher stellar metallicities, while smaller planets tend to correlate with smaller stellar metallicities. The

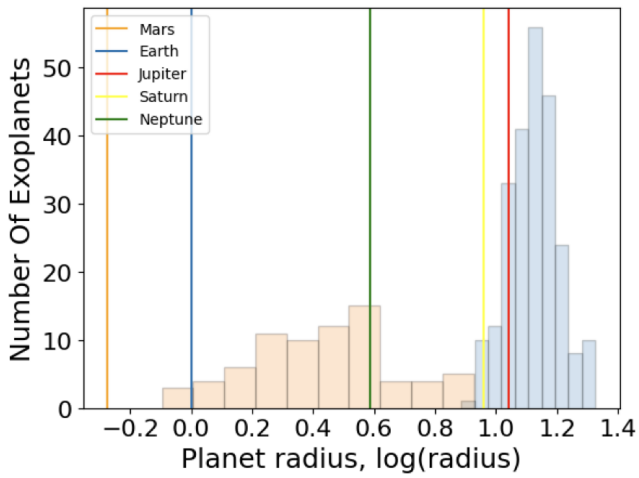

Fig. 10 The Number of Exoplanets and the Logarithm of Planet Radius with indications of planets within our solar system. The orange bars represent the smaller planets, and the blue bars represent the larger planets.

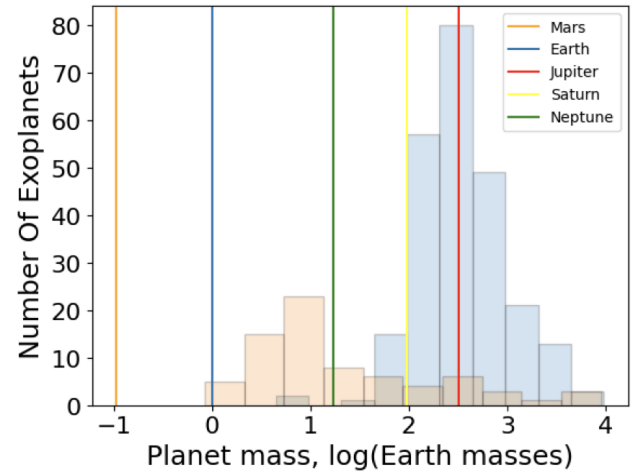

Fig. 11 Bar Graph showing the Number of Exoplanets and the Logarithm of Planet Mass (Earth Mass) with indications planets within our solar system. The orange bars represent the smaller planets, and the blue bars represent the larger planets.

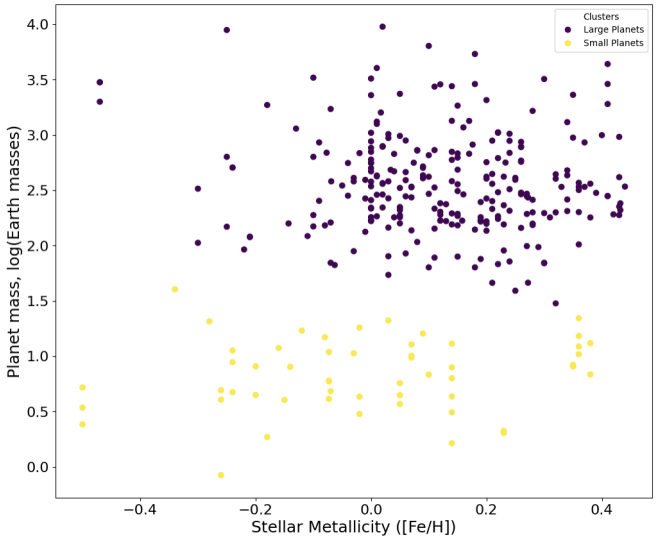

Fig. 12 Clusters modeled by a revised Gaussian Mixture Model between the logarithm of planet mass (Earth Mass) and stellar metallicity ([Fe/H]).

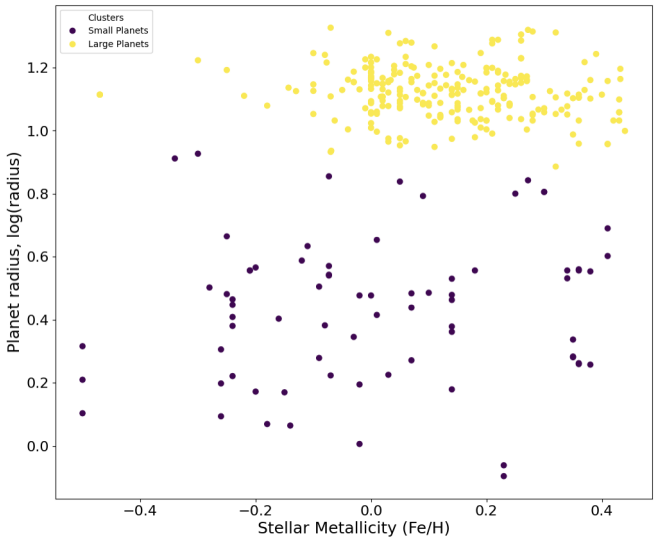

Fig. 13 Clusters modeled by a revised Gaussian Mixture Model between the logarithm of planet radius and stellar metallicity ([Fe/H]).

model did have a strongly defined cluster for larger planets, however it was not able to locate a good cluster for the smaller planets and what they correlated to since the cluster is spread throughout the x-axis (stellar metallicity). Even though our dataset has a limitation due to the small cluster size of the smaller planets, this model still best represents our data more than the others because, as mentioned before, it has the best-defined clusters out of all the models with high silhouette scores. The Gaussian Mixture model is also a well-known model that can represent our data without using too much memory, and it is not sensitive to outliers, unlike others we tried to experiment with such as the KMeans clustering model. In our dataset, we see that there is not a large enough sample size for the Gaussian mixture model to create a defined cluster for the smaller planets compared to the larger ones. To make stronger conclusions, we need larger sample sizes by recording exoplanet information on the influx of newly discovered exoplanets every year. We can also do more research into clustering models other than Gaussian that might have more optimizable parameters for our data. More importantly, what do the high and low metallicities tell us about the chemical composition of the exoplanets' atmospheres? As

mentioned in the introduction, different chemicals can change each exoplanet's stellar host's flux as the light passes through the exoplanet atmosphere and reaches the telescopes. As the iron-to-hydrogen metallicity ratio in stellar hosts increases, the abundance of essential chemical compounds of life such as H2, O2, CO, and CH4 increase in exoplanets. There is evidence that more massive planets tend to have a higher abundance of compounds like oxygen and hydrogen, but these compounds cannot bank on the existence of life or possible human habitability. Earth and the Sun hold as an outlier to their respective datasets due to limited statistics of exoplanets at these masses. However, there is still more data needed on Earth-sized planets to confirm the theory with strongly defined clusters. Since the clusters defined by the Gaussian Mixture model do not completely tell the whole story, it will still be an ongoing investigation as we find the most optimal and specific exoplanet and stellar host properties for future human habitability and possibly extraterrestrial life.

Acknowledgements

This research was supported by InspiritAI and its AI X Mentorship Program. The author wishes thanks to Inspirit AI, and his mentor Tony Rodriguez for supporting this research and giving necessary tools and knowledge to help make this research and investigation possible.

References

1 University of Maryland: Department of Astronomy, *Chapter 4: Galactic Chemical Evolution*, n.d., https://www.astro.umd.edu/~richard/ASTRO620/QM_chap4.pdf, Accessed: 2024-04-07.

2 L. Ghezzi, K. Cunha, V. V. Smith, F. X. De Araújo, S. C. Schuler and R. De La Reza, *The Astrophysical Journal*, 2010.

3 A. Akahori, Y. Watanabe and E. Tajika, *arXiv.org*, 2023.

4 Scikit-learn, *Comparing different clustering algorithms on toy datasets*, n.d., https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html, Accessed: 2024-09-06.