The Aligarh Journal of Statistics
Vol. 45 (2025), pp. 165-182
ISSN 0971-0388

A Study on Clustering and Classification of Kepler's Confirmed Habitable Rocky Exoplanets

Sushovon Jana[1], **Hrittik Banerjee**[2]
[1,2]Department of Applied Statistics, Maulana Abul Kalam Azad University of Technology, Kolkata, India
Corresponding author: Sushovon Jana, *E-mail: susostat@gmail.com*

AMS Subject Classification: 62H30, 62H10

Received: 20/10/2023 Accepted: 27/05/2025

Abstract
The term "exoplanet" holds significant interest in contemporary science. A planet orbiting a star beyond the solar system serves as a foundation for exploring concepts of alien life in the universe. This study utilizes data from Bryson et al. (2020), drawn from the Kepler DR25 planet candidate catalog, to focus on habitable rocky exoplanets with high inclusion probability. Intrinsic groupings among these exoplanets are analyzed based on their physical attributes. The optimal number of planetary groups is identified through the maximum number of clustering indices, and the exoplanets are grouped using the most appropriate clustering technique. The groups primarily differ based on the host star's temperature and the orbital period of the exoplanets, two critical physical attributes. Best-fitted multivariate probability distributions and their estimated parameters are employed to classify unknown habitable rocky exoplanets into the identified groups. A classification rule based on the maximum probability density value is proposed and applied, yielding satisfactory classification results.

Keywords: Habitable rocky planets, Cluster analysis, Clustering index, Multivariate probability distribution, Density-based classification

1 Introduction

The concept of exoplanets has captivated both scientific and public imagination, especially with advancements in detection techniques expanding our understanding of planetary systems beyond our solar system. Exoplanets, which orbit stars other than the Sun, exhibit diverse compositions and characteristics, ranging from gas giants and hot Jupiters to water worlds, super-Earths, and rocky planets. Among these, rocky exoplanets located within the habitable zone (HZ) are particularly significant due to their potential to host conditions conducive to life (Bryson et al., 2020). The HZ denotes the circumstellar region within which a terrestrial planet, characterized by an atmosphere similar to that of Earth, comprising chiefly CO_2, H_2O, and N_2, can support the presence of liquid water on its surface, a fundamental requirement for sustaining life as currently known(Konatham et al., 2020).
Liquid water is central to the development of life, serving as a medium for biochemical reactions. Moreover, biological activity on such planets could alter atmospheric compositions in detectable ways, providing potential biosignatures for remote sensing (Rogers, 2016). Understanding the classification and characteristics of habitable rocky exoplanets is thus essential for identifying potentially life-supporting environments and prioritizing targets for future exploration, such as lander missions.
The search for exoplanets began in earnest in the 1990s, facilitated by groundbreaking discoveries enabled by advanced observational methods. By November 2022, over 5,246 confirmed exoplanets were identified across 3,875 planetary systems, with 842 systems hosting multiple planets. The Kepler Space Telescope has played a pivotal role in these discoveries, using the transit method to detect planets cross-

ing in front of their host stars (Borucki et al., 2010; Koch et al., 2010). This method provides critical insights into the size, orbital period, and other key properties of exoplanets.

The extensive data generated by missions like Kepler have made statistical and machine learning techniques indispensable for analyzing and classifying exoplanets based on their habitability potential (Jiang et al., 2024). Machine learning methods have been applied to group exoplanets based on attributes like radius, orbital period, and host star temperature, enhancing our understanding of exoplanet distribution and habitability (Basak et al., 2021). Advanced statistical models have further refined our understanding of the HZ by incorporating stellar luminosity and planetary atmospheres (Kasting et al., 2014). Bayesian methods have estimated the probability of liquid water under varying conditions (Damiano et al., 2024), while multivariate statistical techniques have modeled the complex interdependencies among exoplanetary characteristics (Fisher et al., 2022).

This study leverages data from the Kepler DR25 planet candidate catalog (Thompson & Kepler Team, 2018) to focus on rocky exoplanets in the habitable zone. The transit method provides a robust framework for analyzing planetary characteristics by detecting minute variations in starlight. Focusing on rocky exoplanets within the HZ is motivated by their higher habitability potential and relevance for future exploratory missions. Unlike gaseous or liquid planets, rocky exoplanets provide solid surfaces, which are critical for many life forms and practical for robotic or human exploration missions.

The methodological framework includes clustering and classification techniques tailored to the properties of exoplanets. Clustering reveals intrinsic groupings based on physical attributes, offering insights into the diversity and patterns among habitable rocky exoplanets. Optimal cluster determination using multiple clustering indices ensures robustness and reduces subjective biases in group selection (Rousseeuw, 1987; Tibshirani et al., 2001). To refine classification, multivariate probability distributions are employed to model exoplanetary characteristics, enabling the assignment of unclassified exoplanets to existing groups. A classification rule based on the maximum probability density ensures statistically sound decisions.

The primary aim of this study is to identify intrinsic groupings of Kepler-confirmed habitable rocky exoplanets by analyzing key physical attributes, such as host star temperature and orbital period, which are critical determinants of habitability. By employing advanced clustering and probabilistic classification methods, the research establishes a robust statistical framework for categorizing exoplanets. This approach enhances understanding of exoplanet diversity, systematically refines the classification process, and provides insights into their potential to sustain liquid water and support life. The study's novel methodology contributes to developing a framework that can guide future astronomical research and support exploratory missions to habitable rocky exoplanets.

2 Data Description

The data is provided by Bryson et al. (2020) using the Kepler DR25 planet candidate catalog from Vizier. Bryson et al. (2020) estimated the occurrence rates of rocky exoplanets ($0.5R_{\oplus} \leq r \leq 1.5R_{\oplus}$) located within the habitable zones of various main-sequence dwarf stars, utilizing data from the Kepler DR25 planet candidate catalog in conjunction with Gaia-derived stellar properties. Their analysis employed differential population models that account for variations in planetary radius, instellation flux, and the effective temperature of the host stars.

The provided data contain 117 observations, representing 117 exoplanets with inclusion probability (occurrence rate). The dataset included 17 features, but only five features were considered for this analysis: `Rad.Rgeo.`, `Per.d.`, `Instel`, `Teff.K.`, and `IncProb`.

- **Rad.Rgeo.**: This column provides the planetary radius, i.e., the ratio of the exoplanet's radius in comparison to Earth's radius.
- **Per.d.**: Orbital period of the exoplanet in days.
- **Instel**: Instellation in Earth units. Instellation of an exoplanet is the rate at which energy from its host star illuminates it. The instellation depends on both the power output (luminosity) of the star

and the distance of the planet's orbit from the star. A planet orbiting a dim M star would need to be closer in than a planet orbiting a brighter G star to receive the same instellation. If instellation is measured in units of Earth's average instellation, a planet with unit instellation would have a surface temperature similar to Earth's. A planet with an instellation of 10 Earth units would be a very hot place, and one with an instellation of a tenth of an Earth unit would be a very cold place.

- **Teff.K.**: Effective temperature of the host star of the exoplanet in Kelvin.
- **IncProb**: This shows the probability for an exoplanet of being rocky and in the habitable zone.

For this work, only exoplanets with a high inclusion probability (≥ 0.5) of being rocky habitable exoplanets were considered. Additionally, four main physical properties were analyzed: radius ratio, orbital period, instellation, and host star temperature, as described in the specified columns.

3 Methodology

3.1 Diagrammatic Presentation

First, basic visualizations of the columns of the dataset are examined to gain an initial understanding of the data distribution and relationships between the features. Some visualization methods discussed below help in exploring these aspects:

- **Histogram:**
  A histogram is a graphical representation of the distribution of data. It is displayed as a set of adjacent rectangles (bars), where each bar represents a group or class of data. The height of each bar shows the frequency of observations in that class. Histograms help in determining the overall distribution and shape of the data in a rough sense.
- **Scatter Plot:**
  Scatter plots graphically depict the relationship between two variables in a dataset. Each point on the plot represents a single observation, positioned based on the values of two variables on the Cartesian plane. The independent variable is plotted along the X-axis, while the dependent variable is plotted along the Y-axis. These plots can reveal the presence and direction of any correlation between the two variables and are especially useful in regression and correlation analysis.
- **Heat Map:**
  A heat map is a visual representation of data in matrix form where values are encoded with colors. When used with a correlation matrix, it visually highlights the strength and direction of linear relationships among numerical features. Heat maps are particularly helpful in feature selection during the preprocessing phase of Machine Learning, as they assist in identifying highly correlated or independent variables for model training.

3.2 Cluster Analysis

Clustering is an unsupervised machine learning technique used for discovering the intrinsic grouping in data based on different features.

3.2.1 Estimation of Optimum Number of Clusters

Determining the optimum number of clusters is a key challenge in cluster analysis, as different clustering algorithms often produce varying numbers of data clusters. To address this issue, the method proposed by Jana & Pal (2019) has been considered, which provides a systematic approach for determining the optimal number of clusters in a dataset. This method employs 30 indices (Table 1), as described in Charrad et al. (2014), to evaluate and recommend the best clustering scheme.

It selects the most appropriate clustering structure based on the results obtained by varying all combinations of the number of clusters, distance measures, and clustering methods. The table below lists the clustering validity indices used in this approach, along with their respective criteria for determining the optimal number of clusters.

Name of the Index	**Criterion for Optimal Cluster Number**
Ch	Index value should be maximum
Duda	Fewest clusters with index $>$ critical value
Pseudot2	Fewest clusters with index $<$ critical value
Cindex	Index value should be minimum
Gamma	Index value should be maximum
Beale	Choose clusters where critical value $\geq \alpha$
Cubic Clustering Criterion	Index value should be maximum
Ptbiserial	Index value should be maximum
Gplus	Index value should be minimum
DB	Index value should be minimum
Frey	Last level before index drops below 1.00
Hartigan	Max. difference between hierarchy levels
Tau	Index value should be maximum
Ratkowsky	Index value should be maximum
Scott	Max. difference between hierarchy levels
Marriot	Max. second difference across levels
Ball	Max. difference between hierarchy levels
Trcovw	Max. difference between hierarchy levels
Tracew	Max. second difference across levels
Friedman	Max. difference between hierarchy levels
McClain	Index value should be minimum
Rubin	Min. second difference across levels
KL	Index value should be maximum
Silhouette	Index value should be maximum
Gap	Fewest clusters with critical value ≥ 0
Dindex	Use graphical interpretation
Dunn	Index value should be maximum
Hubert	Use graphical interpretation
SDindex	Index value should be minimum
SDbw	Index value should be minimum

Table 1: Clustering Validity Indices

3.2.2 Selection of Appropriate Clustering Technique

After getting the optimal number of clusters, it is very important to choose an appropriate clustering technique.

Partitional Clustering

In partitional approach, number of clusters and centroid values are usually specified and a set of objects which are closer to a particular centroid than other centroids can form a cluster. K-means is one of the well-known simplest unsupervised learning algorithms (MacQueen, 1967) of partitional clustering. Another partitional clustering is Partitioning around Medoids (PAM) algorithm (Kaufman & Rousseeuw, 1990).

K-means Clustering: K-means clustering is a commonly employed partitional clustering technique that organizes data into K clusters. The process begins by selecting K initial centroids, which serve as representative points for the clusters. Data points are subsequently assigned to the nearest centroid, with proximity typically measured using a distance metric such as Euclidean distance. Once all data points are assigned to clusters, the centroids are recalculated as the mean positions of the points within each cluster.

This procedure alternates between reassigning points to the nearest centroids and updating the centroids until a stopping criterion is met. Common convergence criteria include the stabilization of centroids, where no significant changes occur between iterations, or the attainment of a predefined number of iterations. Although K-means is a straightforward and computationally efficient algorithm, it operates greedily, aiming to minimize an objective function (e.g., within-cluster sum of squares). Despite its simplicity and speed, the algorithm is known to converge only to a local minimum, and the overall optimization problem is classified as NP-Hard. The Sum of Squared Errors (SSE), also known as the within-cluster sum of squares, quantifies the compactness of the clusters formed by K-means clustering. For a data set $D = \{x_1, x_2, \ldots, x_N\}$, consisting of N points and the clustering result $C = \{C_1, C_2, \ldots, C_K\}$, where C_k represents the set of points in the k-th cluster, the SSE is defined as:

$$\text{SSE}(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \tag{1}$$

where $c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}$ is the centroid of cluster C_k. The goal of K-means clustering is to determine a partitioning of the dataset that minimizes the Sum of Squared Errors (SSE), as defined in equation (1). This optimization is achieved through an iterative process involving two key steps: assigning data points to the nearest centroid and updating the centroids based on the mean position of the assigned points. These steps are repeated until the SSE score converges to a stable value, reflecting an optimal or near-optimal configuration of the clusters for the given centroids.

Fuzzy C-means Clustering: Fuzzy C-means (FCM) is an approach to clustering that allows a single piece of data to belong to two or more clusters (Dunn, 1973). In data sets with overlapping clusters, strict or hard assignments of points to clusters may not effectively capture the underlying structure. To address this limitation, fuzzy clustering techniques are employed. The fuzzy C-means (FCM) clustering algorithm allows for partial membership of data points in multiple clusters, with membership values ranging continuously between 0 and 1. This approach provides greater flexibility in identifying overlapping cluster structures. The objective function for FCM, designed to minimize the weighted within-cluster dispersion, is expressed as follows:

$$\text{SSE}(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} w_{xik}^{\beta} \|x_i - c_k\|^2 \tag{2}$$

In fuzzy C-means clustering, the membership weight w_{xik} for a data point x_i belonging to cluster C_k is computed as:

$$w_{xik} = \left[\sum_{j=1}^{K} \left(\frac{\|x_i - c_k\|}{\|x_i - c_j\|} \right)^{\frac{2}{\beta - 1}} \right]^{-1}$$

This weight determines the degree of membership of x_i to C_k and is updated during each iteration of the algorithm. Using these weights, the weighted centroid for cluster C_k, denoted as c_k, is calculated as:

$$c_k = \frac{\sum\limits_{x_i \in C_k} w_{xik}^{\beta} x_i}{\sum\limits_{x_i \in C_k} w_{xik}^{\beta}}$$

The fuzzy C-means algorithm follows a process similar to K-means clustering, where the algorithm iteratively minimizes the sum of squared errors (SSE). This involves alternately updating the membership weights w_{xik} and the centroids c_k until convergence is achieved, typically when the centroids stabilize. Fuzzy clustering encompasses various methods, including *standard*, *polynomial*, the *GK* (Gustafson–Kessel) algorithm (Gustafson & Kessel, 1979), the *GKB* (Gustafson–Kessel–Babuska) variant (Babuška, van der Veen & Kaymak, 2002), and *medoids* (Rdusseeun & Kaufman, 1987). Each variant adapts the clustering process to different data characteristics and optimization criteria.

Hierarchical Clustering

Hierarchical and Partitional are two main conventional approaches in cluster analysis. Hierarchical clustering algorithms are either agglomerative algorithms or divisive algorithms and merging or splitting process is done on the basis of a distance measure like Euclidean distance, Manhattan distance, etc. (Johnson, 1967). Different methods have been suggested to compute distance between two clusters in agglomerative algorithms. Single Linkage Clustering, Complete Linkage Clustering and Average Linkage Clustering are most common among them. Ward's minimum variance method (Ward, 1963) is another common technique where the clusters are merged by minimizing the increase in the sum of intra-cluster summed squared distances.

3.2.3 Multivariate Distribution fit on estimated clusters

Although several multivariate distributions exist in the literature, only five of the most general distributions were considered in this study: Normal, Skew-Normal, t, Skew-t, and Skew-slash.

Multivariate Normal

The multivariate normal distribution's probability density function is presented by

$$\phi(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma) = \frac{e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu})}}{\sqrt{(2\pi)^p |\Sigma|}} \tag{3}$$

The mean vector $\underset{\sim}{\mu}$ and covariance matrix Σ are the two main parameters of this distribution.

Multivariate Skew-Normal

The probability density function of the multivariate skew-normal distribution is given by

$$p(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma, \underset{\sim}{\delta}) = 2\, \phi_p(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma)\; \Phi_1\Big(\underset{\sim}{\delta}^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}) \mid 0,\; 1 - \underset{\sim}{\delta}^T \Sigma^{-1} \underset{\sim}{\delta}\Big) \tag{4}$$

where $\underset{\sim}{\mu}$ is the location vector, Σ is the scale (covariance) matrix, and $\underset{\sim}{\delta}$ is the skewness vector (Azzalini & Valle, 1996).

Here, $\phi_p(\cdot \mid \underset{\sim}{\mu}, \Sigma)$ denotes the density function of a p-variate normal distribution with mean vector $\underset{\sim}{\mu}$ and covariance matrix Σ, while $\Phi_1(\cdot \mid 0, \sigma^2)$ represents the cumulative distribution function of a univariate normal distribution with mean 0 and variance σ^2. The factor of 2 accounts for the skewness adjustment in the distribution.

Multivariate t-Distribution

A p-dimensional t-distribution with parameter set $\theta = (\underset{\sim}{\mu}, \Sigma, \nu)$ has the probability density function (McLachlan & Peel, 2000) given by

$$p(\underset{\sim}{x} \mid \theta) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{p/2}|\Sigma|^{1/2}} \left(1 + \frac{\delta_\Sigma(\underset{\sim}{x}, \underset{\sim}{\mu})}{\nu}\right)^{-\frac{\nu+p}{2}} \tag{5}$$

where $\Gamma(\cdot)$ denotes the Gamma function, $\nu > 0$ is the degrees of freedom parameter, and $\delta_\Sigma(\underset{\sim}{x}, \underset{\sim}{\mu})$ is the squared Mahalanobis distance defined as

$$\delta_\Sigma(\underset{\sim}{x}, \underset{\sim}{\mu}) = (\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}).$$

In this context, $\underset{\sim}{\mu}$ defines the location vector, Σ denotes the positive-definite scale matrix, and ν serves as the shape parameter controlling tail behavior. For small ν, the distribution has heavier tails than the multivariate normal, providing robustness to outliers, while as $\nu \to \infty$, the distribution converges to the multivariate normal distribution.

Multivariate Skew-t Distribution

The probability density function of a p-dimensional unrestricted multivariate skew-t distribution (Sahu et al., 2003) is expressed as

$$p(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma, \underset{\sim}{\delta}, \nu) = 2^p \, t_{p,\nu}(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Omega) \, T_{p,\nu}(\underset{\sim}{x}^* \mid \underset{\sim}{0}, \Lambda), \tag{6}$$

where

$$\Delta = \operatorname{diag}(\underset{\sim}{\delta}), \qquad \Omega = \Sigma + \Delta,$$

$$\underset{\sim}{x}^* = \underset{\sim}{q}\sqrt{\frac{\nu + p}{\nu + d(\underset{\sim}{x})}}, \qquad \underset{\sim}{q} = \Delta\Omega^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}),$$

$$d(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\mu})^T \Omega^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}), \qquad \Lambda = I_p - \Delta\Omega^{-1}\Delta.$$

The function $t_{p,\nu}(\cdot \mid \underset{\sim}{\mu}, \Omega)$ denotes the density of the p-variate t distribution with location $\underset{\sim}{\mu}$, scale matrix Ω, and ν degrees of freedom. The function $T_{p,\nu}(\cdot \mid \underset{\sim}{0}, \Lambda)$ represents the corresponding cumulative distribution function (CDF). The degrees of freedom parameter ν controls the heaviness of the distribution tails. For small values of ν, the distribution exhibits heavy tails, providing robustness against outliers. As $\nu \to \infty$, the multivariate skew-t distribution converges to the multivariate skew-normal distribution, analogous to how the classical t distribution converges to the normal distribution.

Multivariate Skew-Slash Distribution

The multivariate skew-slash distribution, originally formulated by Tian et al. (2017), extends the classical slash family by incorporating skewness into the multivariate setting. A random vector $\underset{\sim}{X} \in \mathbb{R}^p$ is said to follow a multivariate skew-slash distribution with location vector $\underset{\sim}{\mu} \in \mathbb{R}^p$, positive definite dispersion matrix $\Sigma \in \mathbb{R}^{p\times p}$, and skewness vector $\underset{\sim}{\alpha} \in \mathbb{R}^p$, denoted as

$$\underset{\sim}{X} \sim \text{MSS}_p(\underset{\sim}{\mu}, \Sigma, \underset{\sim}{\alpha}),$$

if its probability density function (pdf) is given by

$$f(\underset{\sim}{x}) = 2\,\phi_p(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma) \left[\Phi\Big(\underset{\sim}{\alpha}^\top \Sigma^{-1/2}(\underset{\sim}{x} - \underset{\sim}{\mu})\Big) - \frac{\phi(0) - \phi\Big(\underset{\sim}{\alpha}^\top \Sigma^{-1/2}(\underset{\sim}{x} - \underset{\sim}{\mu})\Big)}{\underset{\sim}{\alpha}^\top \Sigma^{-1/2}(\underset{\sim}{x} - \underset{\sim}{\mu})}\right], \quad \underset{\sim}{x} \neq \underset{\sim}{\mu}. \tag{7}$$

In this formulation, $\phi_p(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma)$ denotes the pdf of a p-variate normal distribution with mean $\underset{\sim}{\mu}$ and covariance Σ. $\Phi(\cdot)$ and $\phi(\cdot)$ are the univariate standard normal cumulative distribution function (CDF) and probability density function (PDF), respectively. The skewness component, controlled by $\underset{\sim}{\alpha}$, modulates departure from symmetry, while the slash construction introduces heavy-tailed behavior, thereby enhancing robustness against outliers. This makes the multivariate skew-slash distribution particularly useful in modeling asymmetric and heavy-tailed multivariate data.

3.3 Density-based classification model

After fitting the appropriate distribution to the data, the respective parameters are obtained for each cluster. Let there are k subpopulations i.e., k clusters and each cluster have its estimated parameters $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ for k subpopulations respectively, where $x_1, \ldots, x_n$ are the p-variate independent observations of the data x. Let $f_1(x \mid \hat{\theta}_1), \ldots, f_k(x \mid \hat{\theta}_k)$ be the best fitted probability densities with corresponding estimated (obtained by the maximum likelihood method) parameters. A classification rule is proposed for an object based on the conditional probability. The conditional probability of the observation y coming from subpopulation C_i given the values of the components of the vector y is

$$\frac{f_i(y \mid \hat{\theta}_i)}{\sum_{j=1}^{k} f_j(y \mid \hat{\theta}_j)} \tag{8}$$

A new object is classified as belonging to C_i if the conditional probability with respect to C_i exceeds that of every other subpopulation, i.e.,

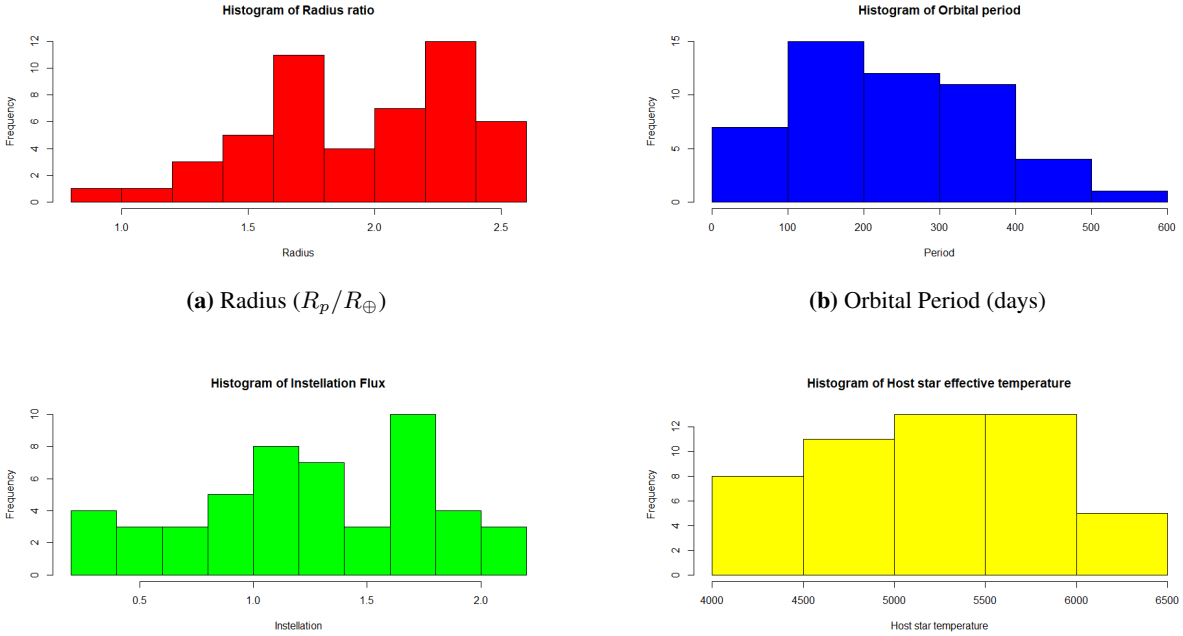
$$\frac{f_i(y \mid \hat{\theta}_i)}{\sum_{j=1}^{k} f_j(y \mid \hat{\theta}_j)} > \frac{f_m(y \mid \hat{\theta}_m)}{\sum_{j=1}^{k} f_j(y \mid \hat{\theta}_j)}, \quad \forall\, m \in \{1, 2, \ldots, k\},\ m \neq i. \tag{9}$$

Analytical Results

After filtering the data to include only observations with an inclusion probability of 0.5 or higher, a dataset of 53 exoplanets is selected. From this, four key physical attributes are considered: radius, orbital period, instellation, and the temperature of the host star. To facilitate model training and evaluation, 50 observations are used as the training dataset, while the remaining three observations are reserved for testing the classification model. The frequency distributions of the training dataset are examined through histograms for each of the four attributes.

The histograms reveal distinct patterns in the distributions (Figure 1). Both the radius and instellation exhibit non-symmetric bimodal distributions, suggesting the presence of two dominant clusters or groupings within the dataset for these attributes. In contrast, the orbital period shows a unimodal distribution with positive skewness, indicating that most exoplanets have shorter orbital periods, with fewer instances of longer-period planets. Meanwhile, the temperature of the host star displays a unimodal distribution with negative skewness, implying that cooler stars are more frequent in the dataset, with a gradual decline in the number of hotter stars. These distributional characteristics provide critical insights into the variability and clustering tendencies of the chosen attributes, forming the basis for subsequent analyses and classification.

The scatter plot matrix (Figure 2) provides a comprehensive visual representation of all pairwise scatter plots for the four attributes under investigation: radius, orbital period, instellation, and host star temperature. This matrix effectively highlights the relationships between the variables, revealing notable patterns and trends. A particularly strong positive association is observed between the orbital period of exoplanets and the temperature of their host stars, suggesting that longer orbital periods tend to correspond to stars with higher temperatures. Other pairwise relationships appear more moderate in strength. For instance, the orbital period and instellation exhibit a negative relationship, indicating that as the orbital period



(a) Radius ($R_p/R_\oplus$)

(b) Orbital Period (days)

(c) Instellation (Earth units)

(d) Host Star Temperature (K)

Figure 1: Histograms of four physical attributes with their respective units.

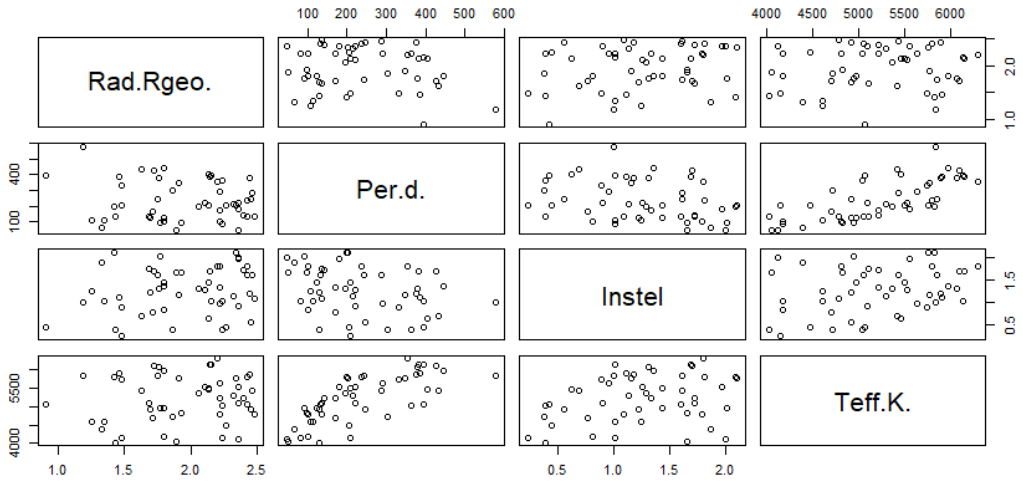


Figure 2: Pair-wise scatter plot of four attributes.

increases, the received stellar radiation generally decreases.

To further quantify these associations, a heat map of correlation (Figure 3) is presented, offering a detailed numerical perspective. The correlation coefficient between orbital period and host star temperature is calculated to be 0.73, confirming a strong positive relationship. In contrast, the correlation between orbital period and instellation is -0.31, reflecting a moderate negative association.

To determine the optimal number of clusters in the dataset, analyses were conducted using four distance metrics: Euclidean, Maximum, Manhattan, and Minkowski. Each distance metric was evaluated across various clustering methods, including single linkage, complete linkage, average linkage, Ward.D, Ward.D2, McQuitty, and K-means. This comprehensive approach ensured that the clustering results were robust and unbiased by the choice of method or distance metric.

The outcomes revealed consistent patterns across the majority of methods, with over 70% of them sug-

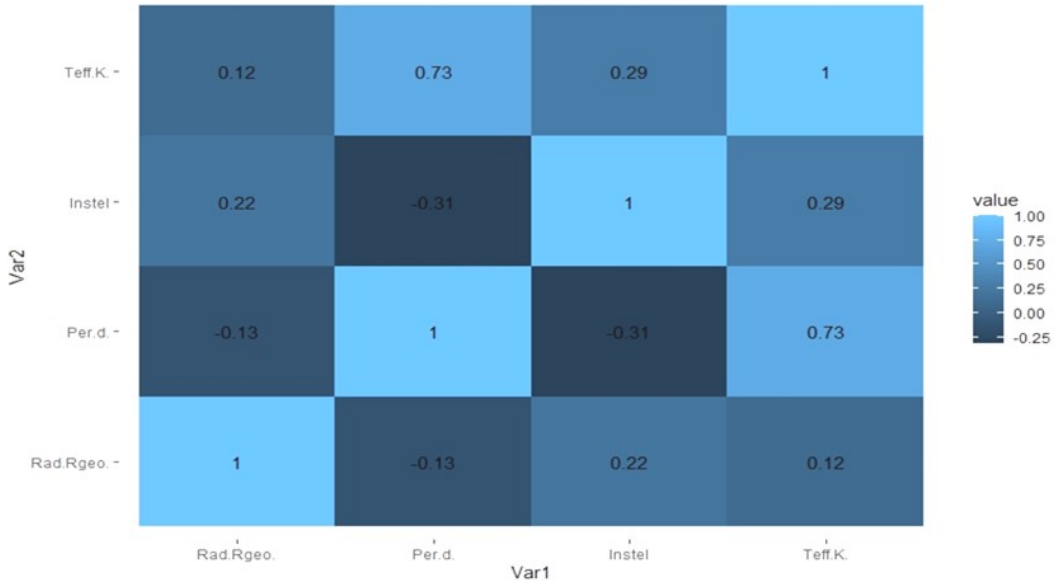


Figure 3: Correlation heat map of the data.

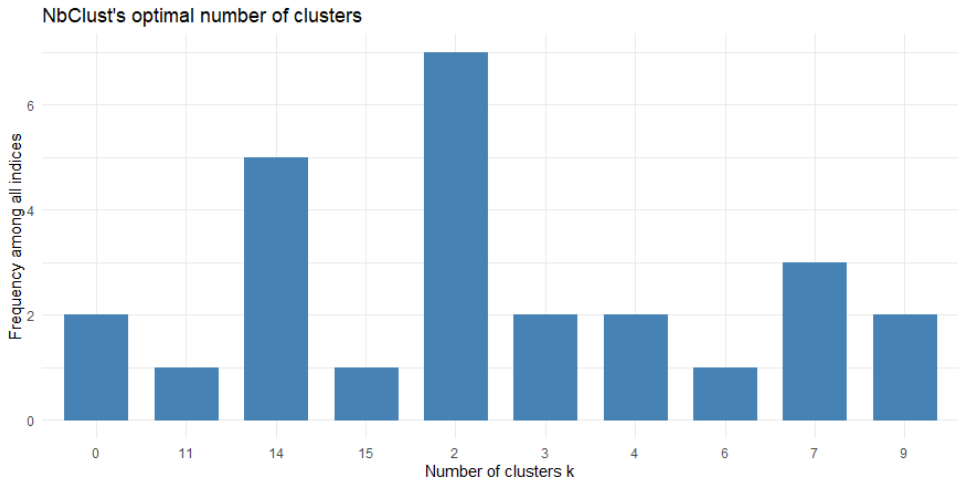


Figure 4: Bar diagram for different number of clusters.

gesting the presence of two distinct clusters in the data (Figure 4). This high level of agreement among different techniques reinforces the validity of the identified clustering structure and highlights the suitability of dividing the exoplanets into two primary groups based on their physical attributes. The consensus across diverse linkage and distance methods provides a strong basis for proceeding with further analyses using the two-cluster configuration.

Fuzzy clustering was initially employed to evaluate whether the clusters in the dataset exhibited overlap. The analysis revealed that the polynomial-type fuzzy clustering model provided the best fit for the data. The resulting cluster plot (Figure 5) demonstrates that the clusters are well-separated, with minimal to no overlap between them. This clear separation provided confidence in the clustering structure and justified the subsequent adoption of a partition-based clustering approach.

K-means clustering was chosen as the partition-based method for its simplicity and effectiveness in handling non-overlapping clusters. The clustering results from K-means mirrored those obtained using fuzzy clustering, which was expected given the clear separations observed. To further validate the clustering quality, the silhouette coefficient for K-means clustering with respect to the Euclidean distance metric was calculated, yielding a value of 0.5124. This value indicates a moderate level of cluster cohesion

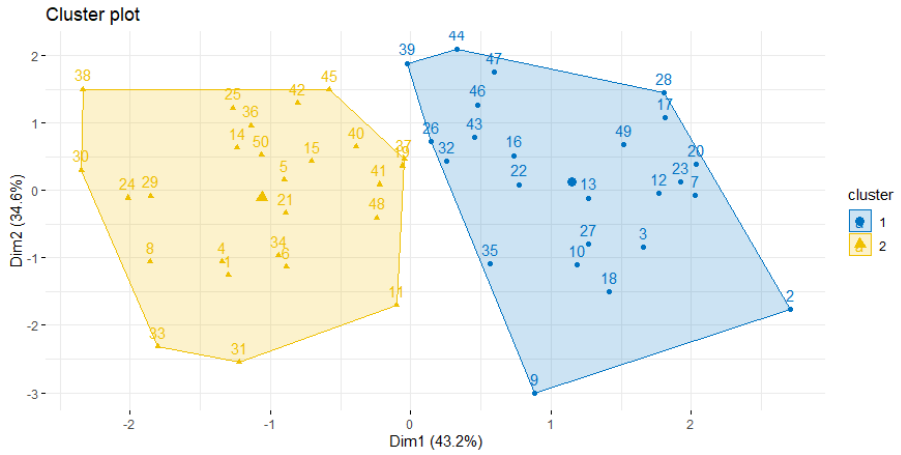


Figure 5: Cluster plot for Fuzzy clustering.

and separation, confirming the adequacy of K-means clustering in capturing the intrinsic grouping of the data.

To determine the most suitable clustering method for the data, hierarchical clustering was applied using

Distance	Method	Silhouette Index
Euclidean	Single	0.4584
Euclidean	Complete	0.5381
Euclidean	Average	0.5685
Euclidean	Ward	0.4960
Euclidean	Ward.D2	0.5381
Manhattan	Single	0.05445
Manhattan	Complete	0.4929
Manhattan	Average	0.4322
Manhattan	Ward	0.55583
Manhattan	Ward.D2	0.5276
Minkowski	Single	0.4584
Minkowski	Complete	0.5381
Minkowski	Average	0.5685
Minkowski	Ward	0.4960
Minkowski	Ward.D2	0.5381
Maximum	Single	0.4689
Maximum	Complete	0.5221
Maximum	Average	0.58412
Maximum	Ward	0.58412
Maximum	Ward.D2	0.58412

Table 2: The silhouette index values correspond to various hierarchical clustering models.

various distance metrics (Euclidean, Maximum, Manhattan, and Minkowski) and different linkage methods (single linkage, complete linkage, average linkage, Ward.D, Ward.D2, McQuitty). The silhouette index values for each combination of distance metric and linkage method were then computed to evaluate the quality of the clustering results (Table 2).

Higher silhouette values indicate better-defined clusters. By comparing the silhouette values across the different hierarchical clustering configurations, it became possible to identify the best-performing clustering method for the dataset. This comparison serves as a crucial step in determining the most effective approach for segmenting the exoplanet data into meaningful clusters.

Hierarchical clustering using the Maximum distance metric and the Ward.D2 linkage method was identified as the most appropriate clustering technique for the dataset. This choice was based on its superior

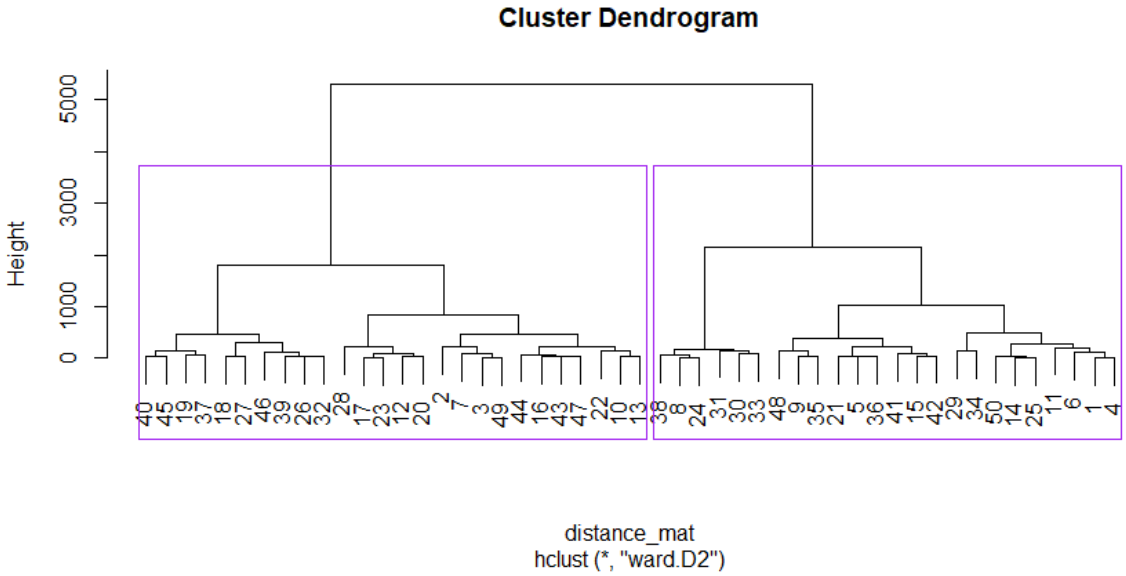


Figure 6: The dendrogram corresponds to the most appropriate clustering technique.

performance, as indicated by the silhouette index, which demonstrated the highest cohesion within clusters and separation between clusters compared to other methods. The clustering dendrogram, provided in Figure 6, visually represents the hierarchical structure of the data.

The clustering analysis resulted in two distinct clusters, as illustrated in Figure 7. Cluster 1 contains

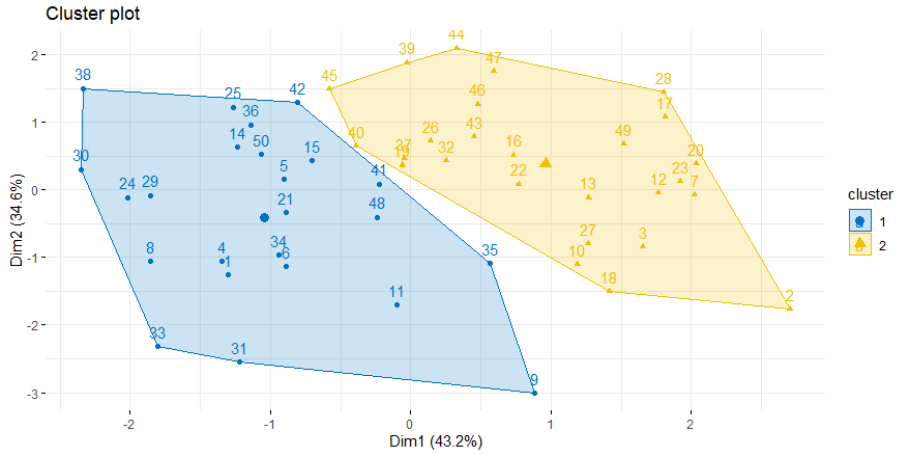


Figure 7: The cluster plot corresponds to the most appropriate clustering technique.

24 observations, while Cluster 2 consists of 26 observations. Summaries of the key attributes for the exoplanets in each cluster are presented in Tables 3 and 4, respectively. The summaries reveal significant differences between the two clusters. Exoplanets in Cluster 1 tend to be smaller in size compared to those in Cluster 2. Furthermore, exoplanets in Cluster 2 exhibit longer orbital periods and are associated with host stars having higher temperatures than those in Cluster 1.

To further investigate the clustering results, the patterns of association among the four attributes—radius, orbital period, instellation, and host star temperature—were examined separately for each cluster. Correlation heat maps for both clusters are provided below (Figure 8), highlighting the variations in relationships between the attributes within the different groups.

The correlation analyses reveal intriguing variations in attribute relationships when the dataset is divided into clusters. In Cluster 1, the correlation between orbital period and instellation has increased signifi-

Attribute	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Radius	0.910	1.630	1.830	1.871	2.245	2.480
Period	46.83	98.42	129.59	156.01	205.85	395.13
Instellation	0.2400	0.5325	1.0550	1.1133	1.6600	2.0100
Host Star Temperature	4023	4335	4744	4652	4950	5098

Table 3: Summary of first cluster of exoplanets.

Attribute	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Radius	1.190	1.752	2.130	2.008	2.295	2.460
Period	142.5	210.1	310.7	308.7	384.3	578.9
Instellation	0.630	1.117	1.325	1.380	1.698	2.100
Host Star Temperature	5213	5465	5764	5715	5896	6290

Table 4: Summary of second cluster of exoplanets.

cantly to -0.74, indicating a stronger negative relationship compared to the entire dataset. Conversely, the correlation between orbital period and host star temperature has decreased to 0.45, suggesting a weaker positive association.

In Cluster 2, the attribute relationships exhibit moderate correlations. The correlation between orbital period and host star temperature is 0.56, while the correlation between orbital period and instellation is -0.51. Additionally, the radius of exoplanets in Cluster 2 shows improved associations with the other three attributes, particularly with orbital period, where the correlation is -0.53. These variations highlight the nuanced differences in the physical properties and associations of exoplanets across the clusters. Exoplanets in Cluster 1 are likely to be smaller, rocky planets or Super-Earths with moderate levels of stellar radiation, situated in the habitable zones of cooler stars, and potentially capable of supporting life. In contrast, exoplanets in Cluster 2 are likely to be larger, possibly Mini-Neptunes or more massive Super-Earths, orbiting hotter stars and receiving higher levels of radiation. Although these planets may still reside within the habitable zone, their potential for supporting life may be less Earth-like due to the higher radiation environment.

To facilitate classification of exoplanets into these clusters, multivariate probability distributions were

Method	AIC	BIC	EDC	Log-Likelihood
Normal	690.6437	707.1364	676.3608	-331.3218
Skew-Normal	679.9964	701.2014	661.6328	-321.9982
Skew-Slash	682.0592	704.4422	662.6753	-322.0296
Skew-t	682.3032	704.6862	662.9193	-322.1516
t	692.7909	710.4617	677.4879	-331.3955

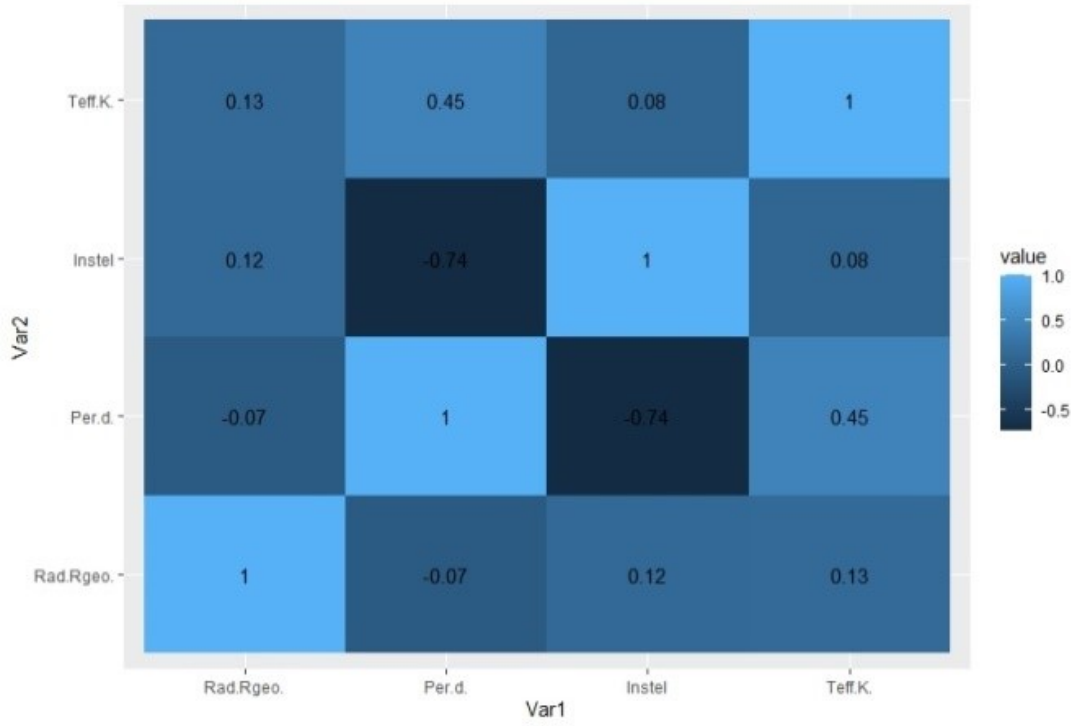
Table 5: Information criterion values for Cluster 1.

fitted separately to each subpopulation. The best-fit distribution was selected using four information criteria: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Efficient Determination Criterion (EDC), and Log-Likelihood values. Table 5 provides the criterion values for various probability distributions applied to Cluster 1, offering a quantitative basis for identifying the optimal distribution for accurate classification.

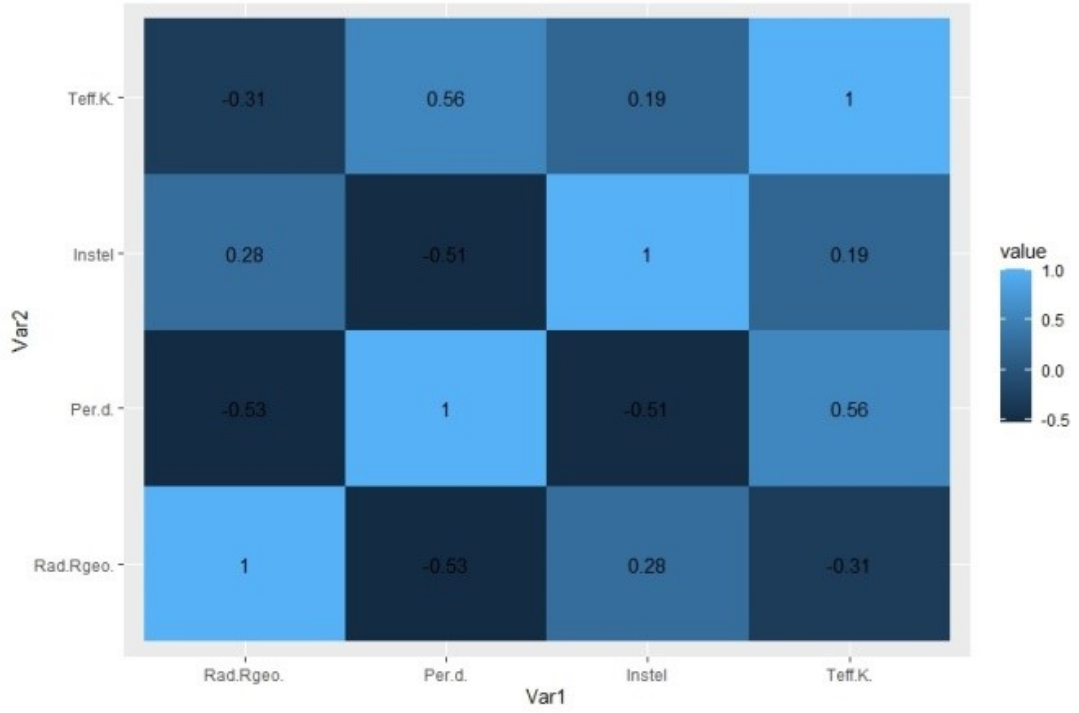
According to all informational criteria, the multivariate skew-normal distribution is the distribution that fits Cluster 1 the best. The estimated parameters for this cluster's best fitted multivariate skew-normal distribution, as determined by the maximum likelihood technique, are shown in Table 6.

According to Table 7, the best fitted distribution for Cluster 2 is also the multivariate skew-normal distribution, supported by all four information criteria. Table 8 provides estimated values for the parameters of the best-fit skew-normal distribution for Cluster 2.

To evaluate the performance of the proposed classification rule, three habitable exoplanets from the test set (Table 9) were considered. The classification rule assigns relative conditional probabilities to each



(a) Cluster 1



(b) Cluster 2

Figure 8: Correlation heat maps for the two clusters.

test observation for belonging to either of the two clusters based on the fitted multivariate probability distributions.

For KOI-3344.03, the relative conditional probability of belonging to Cluster 1 is 0.0341, while for Cluster 2, it is significantly higher at 0.9659, leading to its classification in Cluster 2. Similarly, KOI-7749.01 has relative conditional probabilities of 0.9747 for Cluster 1 and 0.0253 for Cluster 2, resulting in its

$\boldsymbol{\mu}$	(1.959695, 82.595659, 1.416086, 4580.238377)
$\boldsymbol{\Sigma}$	$\begin{bmatrix} 0.4204 & -0.1014 & -0.0091 & 0.0486 \\ -0.1014 & 107.4013 & -0.5685 & 41.8017 \\ -0.0091 & -0.5685 & 0.2880 & 0.0559 \\ 0.0486 & 41.8017 & 0.0559 & 365.1296 \end{bmatrix}$
$\boldsymbol{\delta}$	(-1.535641, 14.729441, 4.132387, 2.838772)

Table 6: Estimated parameters for multivariate skew-normal distribution fit for Cluster 1.

Method	**AIC**	**BIC**	**EDC**	**Log-Likelihood**
Normal	718.7533	736.3667	705.0306	-345.3767
Skew-Normal	710.2636	732.9093	692.6200	-337.1318
Skew-Slash	712.2557	736.1595	693.6320	-337.1278
Skew-t	712.1583	736.0621	693.5346	-337.0791
t	720.9783	739.8498	706.2754	-345.4892

Table 7: Information criterion values for Cluster 2.

classification as part of Cluster 1. Lastly, KOI-2162.02 exhibits a relative conditional probability of 0.0272 for Cluster 1 and 0.9728 for Cluster 2, which places it in Cluster 2.

4 Conclusion

As highlighted in astrophysical literature, the subpopulations identified through optimal clustering techniques accurately represent distinct exoplanet classes. This study aimed to uncover intrinsic patterns within rocky habitable exoplanets—those with Earth-like characteristics. By utilizing data with an inclusion probability of 0.5 or higher, we focused on four critical physical attributes: radius, orbital period, instellation flux, and the host star's effective temperature. The dataset was split into training and testing subsets to facilitate robust analysis.

Using hierarchical clustering with the Maximum distance metric and Ward.D2 linkage method, two optimal clusters were identified within the dataset. Cluster 1 contained 24 exoplanets, while Cluster 2 included 26, with notable distinctions between them. Exoplanets in Cluster 1 were generally smaller in size, with a mean radius ratio lower than those in Cluster 2. Orbital periods differed significantly, with 75% of Cluster 1 exoplanets having orbital periods under 207 days, whereas 75% of Cluster 2 exoplanets exhibited orbital periods exceeding 207 days. Additionally, host stars in Cluster 1 had effective temperatures ranging from 4020 K to 5100 K, while those in Cluster 2 ranged from 5200 K to 6300 K. This suggests that Cluster 2 exoplanets orbit hotter, more luminous stars, contributing to their slightly higher mean instellation flux (1.380) compared to Cluster 1 (1.113).

These findings align with the principles of habitability. Exoplanets in Cluster 2 require greater orbital distances from their host stars to maintain surface liquid water within their habitable zones, which consequently leads to extended orbital periods. The clustering results provided a robust basis for fitting multivariate distributions, with the multivariate skew-normal distribution emerging as the best fit for both clusters. Estimated parameters from this distribution were employed to develop a classification model.

The classification model demonstrated high accuracy on the test dataset, correctly classifying exoplanets based on their physical attributes. For instance, one test exoplanet was accurately assigned to Cluster 2 due to its larger radius, orbital period exceeding 207 days, host star temperature of 5495 K (within the range of Cluster 2), and instellation flux of 1.44, higher than the mean flux for Cluster 2. These results underscore the effectiveness of the clustering methodology and classification model in uncovering and validating the nature of rocky habitable exoplanets.

$\boldsymbol{\mu}$	(2.273672, 222.570347, 1.514070, 5651.258166)
$\boldsymbol{\Sigma}$	$\begin{bmatrix} 0.3115 & -0.2872 & 0.0067 & -0.1079 \\ -0.2872 & 124.5596 & -0.3063 & 54.5465 \\ 0.0067 & -0.3063 & 0.2662 & 0.1053 \\ -0.1079 & 54.5465 & 0.1053 & 297.9106 \end{bmatrix}$
$\boldsymbol{\delta}$	(-3.842935, 11.712419, 3.632052, 1.705714)

Table 8: Estimated parameters for multivariate skew-normal distribution fit for Cluster 2.

Kepler Object Identifier	Rad. ($R_p/R_\oplus$)	Per. (d)	Instel	T_{eff} (K)
KOI-3344.03	2.13	208.54	1.44	5495
KOI-7749.01	1.68	133.63	1.73	5098
KOI-2162.02	1.42	199.67	2.09	5814

Table 9: Test set of exoplanets

5 Limitations and Future Research Directions

This study provides important perspectives on the clustering and classification of rocky habitable exoplanets, it is not without certain limitations. The dataset is relatively small, with only 50 observations, which may not capture the full diversity of exoplanetary characteristics. Moreover, the analysis is limited to four physical parameters: radius, orbital period, instellation flux, and host star temperature; it does not incorporate other essential factors, notably atmospheric composition, surface pressure, and magnetic field strength, all of which could significantly impact habitability. The assumption of static habitable zones and the chosen clustering and classification techniques, including hierarchical clustering with Ward.D2 and multivariate skew-normal distributions, may not capture the full complexity of the data.
Future research in this area should focus on several key directions to enhance the understanding of rocky habitable exoplanets. First, incorporating a broader set of features, like atmospheric composition, magnetic field strength, and surface pressure, is crucial for more accurate habitability models. Expanding the dataset with larger and more diverse exoplanet samples, particularly from advanced telescopes like TESS or JWST, will enable the identification of new patterns and improve model generalizability. Dynamic modeling of habitable zones, accounting for stellar evolution, orbital dynamics, and atmospheric processes, would provide more realistic predictions of habitability over time. Exploring advanced clustering and classification techniques, such as DBSCAN or deep learning approaches, could uncover more complex relationships and non-linear patterns in the data, leading to refined models of exoplanet classification.

6 Acknowledgement

It is an expression of heartfelt gratitude and indebtedness to all those who have been associated with the development of this work. We gratefully acknowledge the data obtained from the Exoplanet Orbit Database (EOD) and the Exoplanets Data Explorer (EDE), derived from observations made by NASA's Kepler space telescope. We also extend our sincere thanks to the anonymous reviewer and the editor for their insightful comments and constructive suggestions, which have significantly improved the clarity and quality of this manuscript.

7 Conflicts of Interest

The authors declare no conflict of interest.

References

Azzalini, A. & Valle, D. (1996). The multivariate skew-normal distribution, *Biometrika*, *83*(4), 715–726. https://doi.org/10.1093/biomet/83.4.715.

Babuška, R., van der Veen, P.J. & Kaymak, U. (2002). Improved covariance estimation for Gustafson-Kessel clustering, in *Proceedings of the IEEE International Conference on Fuzzy Systems*, Honolulu, HI, USA, *2*, 1081–1085. IEEE. https://doi.org/10.1109/FUZZ.2002.1006654.

Basak, S., Mathur, A., Theophilus, A.J., Deshpande, G., & Murthy, J. (2021). Habitability classification of exoplanets: A machine learning insight, *Eur. Phys. J. Spec. Top.*, *230*, 2221–2251. https://doi.org/10.1140/epjs/s11734-021-00203-z.

Borucki, W., Koch, D., Basri, G., Batalha, N., Brown, T., et al. (2010). Kepler planet-detection mission: Introduction and first results, *Science*, *327*(5968), 977–980. https://doi.org/10.1126/science.1185402.

Bryson, S., Kunimoto, M., Kopparapu, R.K., Coughlin, J.L., Borucki, W.J., et al. (2020). The occurrence of rocky habitable-zone planets around solar-like stars from Kepler data, *Astron. J.*, *161*(1), 36–68. https://doi.org/10.3847/1538-3881/abc418.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set, *J. Stat. Softw.*, *61*(6), 1–36. https://doi.org/10.18637/jss.v061.i06.

Damiano, M., Bello-Arufe, A., Yang, J., & Hu, R. (2024). LHS 1140 b is a potentially habitable water world, *Astrophys. J. Lett.*, *968*(2), L22. https://doi.org/10.3847/2041-8213/ad5204.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybernetics*, *3*, 32–57. https://doi.org/10.1080/01969727308546046.

Fisher, T., Kim, H., Millsaps, C., Line, M., & Walker, S.I. (2022). Inferring exoplanet disequilibria with multivariate information in atmospheric reaction networks, *Astron. J.*, *164*(2), 53–93. https://doi.org/10.3847/1538-3881/ac6594.

Gustafson, D.E. & Kessel, W.C. (1979). Fuzzy clustering with a fuzzy covariance matrix, in *Proceedings of the 1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, San Diego, CA, USA, 761–766. IEEE. https://doi.org/10.1109/CDC.1978.268028.

Jana, S. & Pal, C. (2019). Clustering and classification of Kepler's confirmed exoplanets based on mixture models, *Int. J. Appl. Math. Stat.*, *58*(3), 35–51. http://www.ceser.in/ceserp/index.php/ijamas/article/view/6175.

Jiang, J.H., Rosen, P.E., Liu, C.X., Wen, Q., & Chen, Y. (2024). Analysis of habitability and stellar habitable zones from observed exoplanets, *Galaxies*, *12*(6), 86. https://doi.org/10.3390/galaxies12060086.

Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika*, *22*, 241–254. https://doi.org/10.1007/BF02289588.

Kasting, J.F., Kopparapu, R., Ramirez, R.M., & Harman, C.E. (2014). Remote life-detection criteria, habitable zone boundaries, and the frequency of Earth-like planets around M and late K stars, *Proc. Natl. Acad. Sci.*, *111*(35), 12641–12646. https://doi.org/10.1073/pnas.1309107110.

Kaufman, L. & Rousseeuw, P. J. (1990). Partitioning around medoids (Program PAM), *Wiley Series in Probability and Statistics*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 68–125. https://doi.org/10.1002/9780470316801.ch2, ISBN: 978-0-470-31680-1.

Koch, D.G., Borucki, W.J., Basri, G., Batalha, N.M., Brown, T.M., et al. (2010). Kepler mission design, realized photometric performance, and early science, *Astrophys. J. Lett.*, *713*(2), L79. https://iopscience.iop.org/article/10.1088/2041-8205/713/2/L79.

Konatham, S., Martin-Torres, J., & Zorzano, M.P. (2020). Atmospheric composition of exoplanets based on the thermal escape of gases and implications for habitability, *Proceedings of the Royal Society A*, *476*(2241), 20200148–20200169. https://doi.org/10.1098/rspa.2020.0148.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, *1*, 281–297. http://projecteuclid.org/euclid.bsmsp/1200512992.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*, Wiley InterScience, New York. https://onlinelibrary.wiley.com/doi/book/10.1002/0471721182.

Rdusseeun, L.K.P.J. & Kaufman, P. (1987). Clustering by means of medoids, *Proc. Stat. Data Anal. Based on the L1 Norm Conf.*, Neuchâtel, Switzerland, *31*, 28–40. https://www.researchgate.net/publication/243777819_Clustering_by_Means_of_Medoids.

Rogers, L.A. (2016). Current best estimates of planet populations, *Micro- and Nanotechnology Sensors, Systems, and Applications VIII*, *9836*, 983602–983615. https://doi.org/10.1117/12.2223920.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Sahu, S., Dey, D., & Branco, M. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models, *Can. J. Stat.*, *31*, 129–150. https://doi.org/10.2307/3316064.

Thompson, S.E. & Kepler Team (2018). Kepler's DR25 most Earth-like planet candidates: What to know before you go, *AAS Meeting Abstracts*, *231*, 431–05. https://ui.adsabs.harvard.edu/abs/2018AAS...23143105T/abstract.

Tian, W., Han, G., Wang, T., & Pipitpojanakarn, V. (2017). EM estimation for multivariate skew slash distribution, *Robustness in Econometrics*, *692*, 235–248. http://dx.doi.org/10.1007/978-3-319-50742-2_14.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc. Ser. B*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293.

Ward, J. (1963). Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.