Mousavi-Sadr, M.; Jassur, D. M.; Gozaliasl, G.

Revisiting mass-radius relationships for exoplanet populations : a machine learning insight

Published in:
Monthly Notices of the Royal Astronomical Society

DOI:
10.1093/mnras/stad2506

Published: 01/11/2023

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Mousavi-Sadr, M., Jassur, D. M., & Gozaliasl, G. (2023). Revisiting mass-radius relationships for exoplanet populations : a machine learning insight. *Monthly Notices of the Royal Astronomical Society*, *525*(3), 3469-3485. https://doi.org/10.1093/mnras/stad2506

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **525,** 3469–3485 (2023) https://doi.org/10.1093/mnras/stad2506
Advance Access publication 2023 August 21

Revisiting mass–radius relationships for exoplanet populations: a machine learning insight

M. Mousavi-Sadr [1]★ D. M. Jassur[1] and G. Gozaliasl [2,3]★

[1] *Faculty of Physics, Department of Theoretical Physics and Astrophysics, University of Tabriz, 5166616471, Tabriz, Iran*
[2] *Department of Computer Science, Aalto University, P.O. Box 15400, FI-00076 Espoo, Finland*
[3] *Department of Physics, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland*

Accepted 2023 August 15. Received 2023 August 13; in original form 2023 January 16

ABSTRACT

The growing number of exoplanet discoveries and advances in machine learning techniques have opened new avenues for exploring and understanding the characteristics of worlds beyond our Solar system. In this study, we employ efficient machine learning approaches to analyse a data set comprising 762 confirmed exoplanets and eight Solar system planets, aiming to characterize their fundamental quantities. By applying different unsupervised clustering algorithms, we classify the data into two main classes: 'small' and 'giant' planets, with cut-off values at $R_p = 8.13R_\oplus$ and $M_p = 52.48M_\oplus$. This classification reveals an intriguing distinction: giant planets have lower densities, suggesting higher H–He mass fractions, while small planets are denser, composed mainly of heavier elements. We apply various regression models to uncover correlations between physical parameters and their predictive power for exoplanet radius. Our analysis highlights that planetary mass, orbital period, and stellar mass play crucial roles in predicting exoplanet radius. Among the models evaluated, the Support Vector Regression consistently outperforms others, demonstrating its promise for obtaining accurate planetary radius estimates. Furthermore, we derive parametric equations using the M5P and Markov Chain Monte Carlo methods. Notably, our study reveals a noteworthy result: small planets exhibit a positive linear mass–radius relation, aligning with previous findings. Conversely, for giant planets, we observe a strong correlation between planetary radius and the mass of their host stars, which might provide intriguing insights into the relationship between giant planet formation and stellar characteristics.

Key words: software: data analysis – planets and satellites: composition – planets and satellites: dynamical evolution and stability – planets and satellites: formation – planets and satellites: fundamental parameters – planets and satellites: general.

1 INTRODUCTION

Our comprehension of new worlds beyond the Solar system, known as exoplanets, their population, and diversity come largely from the latest generation of modern satellites. The *Kepler* space mission, the *Transiting Exoplanet Survey Satellite*, the *JWST*, and many ground-based observatories make important contributions to detecting and characterizing exoplanets (Pepper et al. 2007; Borucki et al. 2010; Beichman et al. 2014). The data generated by these state-of-the-art instruments are now available to everyone. Researchers skilled in data science, data analytics, or machine learning (ML) and neural network techniques study and analyse these data to predict, identify, characterize, and classify the exoplanets (Alibert & Venturini 2019; MacDonald 2019; Barboza, Ulmer-Moll & Faria 2020; Tasker, Laneuville & Guttenberg 2020; Armstrong, Gamper & Damoulas 2021; Leleu et al. 2021a, b; Mousavi-Sadr, Gozaliasl & Jassur 2021; Schlecker et al. 2021; Van Eylen et al. 2021; Maltagliati 2023; Mishra et al. 2023b). In addition, the observational data are not only used to study exoplanets but they are also applied to peruse entire planetary science. As many planets are found around other stars, they have provided us with an opportunity to understand the main ways of planet formation and evolution and to put our Solar system in a broader context (Kipping 2018; Armitage 2020; Gilbert & Fabrycky 2020; Mishra et al. 2023a).

At this paper's writing, more than 5000 exoplanets have been discovered, and thousands of candidates are yet to be confirmed. Transit and radial velocity are two fundamental methods to discover exoplanets and determine their main parameters. The transit method regularly observes the small fraction of the star's light blocked by a transiting planet. Observing this light decrement makes it possible to calculate the planet's radius. On the other hand, using the radial velocity method, the slight movement of a star caused by an orbiting planet is measured, and the planet's mass is obtained. However, not all discovered planets have measured mass and radius as two important physical properties (Seager 2010; Deeg & Alonso 2018).

There is a clear correlation between the radius and mass of a planet that can be described with a polytropic relation (Burrows & Liebert 1993; Chabrier & Baraffe 2000). Many works have investigated the relationship between planetary parameters, particularly mass, and radius, and to deduce the composition and structure of exoplanets (Seager et al. 2007; Swift et al. 2012; Otegi, Bouchy & Helled 2020). Weiss et al. (2013) divided the planets into two groups, those with masses greater than and lower than $150M_\oplus$, and presented a power-law relation for the mass–radius distribution of each group.

★ E-mail: mahdiyar.mousavi@gmail.com (MMS); ghassem.gozaliasl@helsinki.fi (GG)

Bashi et al. (2017) revised this mass breakpoint to $124 \pm 7M_{\oplus}$ and proposed $R_p \propto M^{0.55 \pm 0.02}$ for small planets and $R_p \propto M^{0.01 \pm 0.02}$ for large planets. Assuming a power-law description of the mass–radius relation, for the first time, a probabilistic model for planets with radii lower than $8R_{\oplus}$ was presented by Wolfgang, Rogers & Ford (2016). Chen & Kipping (2017) implemented this idea to an extended data set, forecasting the mass or radius of planets. Moreover, they calculated the forecasted mass for ∼7000 Kepler Objects of Interest (Chen & Kipping 2018).

Most previous works use a power-law model to explore the mass–radius relation and have assumptions that are not pliable enough to consider principal attributes in such diagrams. Consequently, Ning, Wolfgang & Ghosh (2018) developed a non-parametric approach using a sequence of Bernstein polynomials and the sample of Wolfgang et al. (2016). The same method was used in a follow-up work to analyse the mass–radius relation of exoplanets orbiting M dwarfs (Kanodia et al. 2019).

The correlation between physical parameters in planetary systems is not limited to the planet's mass and radius. It has been demonstrated that the radius of a giant planet is related to other parameters such as the orbital semimajor axis, the planetary equilibrium temperature, the tidal heating rate, and the stellar irradiation and metallicity (Guillot et al. 2006; Fortney, Marley & Barnes 2007; Enoch, Collier Cameron & Horne 2012). Zucker & Mazeh (2002) reported a possible correlation between the mass and period of an exoplanet. They also showed that planets revolving around a binary host star might have an opposite correlation. Weiss & Marcy (2014) studied a restricted data set containing 65 exoplanets smaller than $4R_{\oplus}$ with orbital periods shorter than 100 d. They showed that planets smaller than $1.5R_{\oplus}$ are consistent with a positive linear density-radius relation, but for planets larger than $1.5R_{\oplus}$, density decreases with radius. Hatzes & Rauer (2015) presented the mass–density relationship in a logarithmic space for objects ranging from planets ($M \approx 0.01M_J$) to stars ($M > 0.08\mathrm{M}_{\odot}$). They divided the mass–density distribution into three regions based on changes in the slope of the relationship and introduced a new definition for giant planets.

Bhatti et al. (2016) used a Random Forest regression model to evaluate the influence of different physical parameters on planet radii. Applying this model to different groups of giant planets, they found that the planet's mass and equilibrium temperature has the greatest effect on determining the radius of a hot-Saturn ($0.1 < M_p < 0.5M_J$). They also showed that the equilibrium temperature is more important for more massive planets. Moreover, Ulmer-Moll et al. (2019) (hereafter, Ulmer19) introduced Random Forest as a promising algorithm for obtaining exoplanet properties. They used Random Forest to predict the exoplanet radii based on several planetary and stellar parameters. Similar to previous results, an exoplanet's mass and equilibrium temperature were the fundamental parameters.

As the number of discovered exoplanets rapidly increases, ML techniques can be used to investigate correlations between planets and their host stars. In this study, we implement various ML algorithms to find the potential relationships between physical parameters in exoplanet systems. The Markov Chain Monte Carlo (MCMC) (see Goodman & Weare 2010; Foreman-Mackey et al. 2019) is used to quantify the uncertainties of the best-fitting parameters. In addition, different ML clustering algorithms are used to group the exoplanets and study their properties. We organize this paper as follows: In Section 2, the sample data and methods of pre-processing, clustering, and modelling are introduced. Section 3 presents the results. In Section 4, we summarize the main results and conclusions.

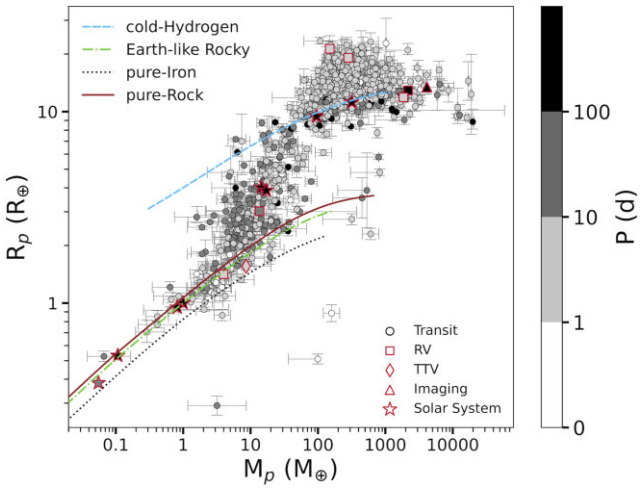

Figure 1. The mass–radius distribution of 770 planets colour coded by orbital period. This figure separates exoplanets into four groups based on detection methods: transit (black circles), radial velocity (red squares), transit timing variations (red diamond), and imaging (red triangle). The red stars show the Solar system's planets. Four sample mass–radius relations are also shown: cold-hydrogen (blue dashed line), Earth-like rocky (green dash–dotted line), pure-iron (black dotted line), and pure rocky (solid crimson line) planets (Marcus et al. 2010; Becker et al. 2014).

2 DATA AND METHODS

2.1 Data set

We use the NASA Exoplanet Archive[1] and the Extrasolar Planets Encyclopedia[2] to extract the data of exoplanets.[3] These two catalogues are comprehensive, up-to-date, and available to the public, and also provide access to relevant publications (Schneider 2011; Akeson et al. 2013). There are 762 confirmed exoplanets with reported physical parameters, including the orbital period (P) and eccentricity (e), planetary mass (M_p) and radius (R_p), and the stellar mass (M_s), radius (R_s), metallicity (Fe/H), and effective temperature (T_{eff}). Note that exoplanets with only a minimum mass are not considered. Among these 762 exoplanets, six have been discovered by the radial velocity method, and each of the imaging and transit timing variation methods has identified only one exoplanet. In general, most exoplanets have been discovered by observing the slight decrease in brightness of the host star caused by the transit of a planet in front of it. Using NASA's Planetary Fact Sheet,[4] we add the eight Solar system planets to the sample. Overall, the data set contains 770 planets. Fig. 1 shows the radius of planets plotted as a function of mass and colour coded by orbital period. It is separated into four groups based on detection methods: transit, radial velocity, transit timing variation, and imaging. Distributions of cold-hydrogen, Earth-like rocky (32.5 per cent Fe+67.5 per cent MgSiO3), pure-iron (100 per cent Fe), and pure-rock (100 per cent MgSiO3) planets are also illustrated (Marcus et al. 2010; Becker et al. 2014). We should note that, like other observational data sets, our sample is also affected by detection biases. A data set containing planets whose radius and mass have been measured suffers from the detection limits of both radial velocity and transit methods. Thus, it is impossible to draw a

[1] https://exoplanetarchive.ipac.caltech.edu/
[2] http://exoplanet.eu/
[3] The data was last extracted on March 22, 2022.
[4] https://nssdc.gsfc.nasa.gov/planetary/factsheet/

reliable conclusion about the occurrence of planets using our data set.

2.2 Data pre-processing

We aim to predict the planetary radius as the target variable, using other physical parameters as features. As the parameters have different ranges, we transfer them to a logarithmic space. In data analysis, the ultimate results might be affected by some unreliable measurements in the sample. In ML and statistics, there are diverse methods to detect these unusual observations in a data set. We choose the Local Outlier Factor (LOF) method to identify and remove observations with abnormal distances from other values. The technique is often used with multidimensional data sets, like our eight-dimensional one, which has different densities and types of outliers. The LOF uses two hyperparameters: neighbourhood size (k), which defines the neighbourhood for local density calculation, and contamination (c), which specifies the proportion of outliers in the data set (Breunig et al. 2000; Chandola, Banerjee & Kumar 2009). By choosing $k = 20$ and $c = 0.05$, the LOF method is run twice: once for all parameters including P, e, M_p, R_p, M_s, R_s, Fe/H, and $T_{\rm eff}$, then for planetary mass and radius, which are known to be highly correlated. Altogether, the LOF detects 76 data points as outliers. Appendix A describes the process of identifying outliers in detail.

Choosing appropriate features plays a vital role in building an efficient ML model. Adding extra variables or those highly correlated with each other may reduce the overall predictive ability of the model and lead to wrong results. Feature selection (hereafter FS) methods rank features based on their usefulness and effectiveness in making predictions. The FS methods can be divided into three groups: filter, wrapper, and embedded methods (Guyon et al. 2008; Chandrashekar & Sahin 2014; Jović, Brkić & Bogunović 2015; Brownlee 2016a). In filter methods, features are filtered independently of any induction algorithm and based on some performance evaluation metrics calculated directly from the data (Sánchez-Maroño, Alonso-Betanzos & Tombilla-Sanromán 2007; Cherrington et al. 2019). In contrast, the selection process in wrapper methods is based on the performance of a specific ML algorithm operating with a subset of features (Ferri et al. 1994; Kohavi & John 1997; Hall & Smith 1999). Embedded methods combine the qualities of both filter and wrapper methods. They perform the FS in the training process and are usually specific to given learning machines (Lal et al. 2006; Bolón-Canedo, Sánchez-Maroño & Alonso-Betanzos 2013). We use five FS methods to identify the most important features in our data set: Spearman's rank correlation test as a filter method, the Backward Elimination and Forward Selection as two wrapper methods, and the CART (classification and regression trees) and XGBoost (extreme gradient boosting) as two Embedded methods.

2.3 Clustering

Clustering as an unsupervised ML task involves grouping each data point with a specific type. In theory, data points belonging to a particular group should have similar properties (Xu & Wunsch 2005; Kaufman & Rousseeuw 2009; Soni Madhulatha 2012). Data clustering algorithms can be divided into hierarchical and partitional groups. Hierarchical algorithms find clusters using previously established clusters, while partitional algorithms find all clusters at once (Kononenko & Kukar 2007; Soni Madhulatha 2012). We aim to group planets into distinct, non-overlapping clusters, similar to exclusive clustering (Jain & Dubes 1988). We use 10 ML clustering algorithms to include a wide range of clustering methods and examine their performance in exoplanet data. Since many diverse exoplanets have been discovered, implementing clustering algorithms can find potential exoplanet groups to investigate their characteristics. The algorithms are available in the SCIKIT-LEARN software ML library (Pedregosa et al. 2011) and are as follows: Affinity Propagation, BIRCH (balanced iterative reducing and clustering using hierarchies), DBSCAN (density-based spatial clustering of applications with noise), Gaussian Mixture Model, Hierarchical Clustering, K-Means, Mean Shift, Mini-Batch K-Means, OPTICS (ordering points to identify the clustering structure), and Spectral Clustering (Davies & Bouldin 1979; Ester et al. 1996; Zhang, Ramakrishnan & Livny 1996; Ankerst et al. 1999; Halkidi, Batistakis & Vazirgiannis 2001; Comaniciu & Meer 2002; Frey & Dueck 2007; Von Luxburg 2007; Schubert et al. 2017).

BIRCH, Gaussian Mixture Model, Hierarchical Clustering, K-Means, Mini-Batch K-Means, and Spectral Clustering are algorithms that do not learn the number of clusters (K) from data. Therefore, we first perform the Elbow and Silhouette methods to find the optimal number of clusters. The Elbow method runs the K-Means clustering algorithm for K values. Then, for each K, it computes the sum of squared distances (SSD) between data points and their assigned cluster centroids and uses them to propose an optimal number of clusters. The Silhouette method determines the degree of separation between clusters by choosing a range of K values and calculating a coefficient for each K. The silhouette coefficient for a particular data point is calculated by $(b^i - a^i)/max(a^i, b^i)$. Here, a^i represents the average distance from all data points in the same cluster, whereas b^i is the average distance from data points that belong to the closest cluster. Provided that the sample is on or near the decision boundary between two neighbouring clusters, the silhouette coefficient becomes 0. A coefficient close to $+1$ indicates that the sample is far from neighbouring clusters. A negative coefficient value indicates that samples may have been assigned to the wrong cluster (Rousseeuw 1987; Soni Madhulatha 2012).

2.4 Modelling

In ML, different algorithms allow machines to learn information from a given data set, uncover relationships, and make predictions (Brownlee 2016a, b). In our case, we apply ML models to predict planet radii when other parameters are given. It is also possible to see how efficiently each parameter uses these models. The algorithms used to perform regression tasks are as follows: Decision Tree, K-Nearest Neighbors, Linear Regression, Multilayer Perceptron, M5P, and Support Vector Regression (SVR) (Hinton 1990; Quinlan 1992, 1993; Hastie et al. 2009; Chang & Lin 2011). Bootstrap Aggregation and Random Forest are also used as ensemble algorithms that combine the predictions from multiple models (Breiman 1996, 2001). These algorithms are available in the Weka tool environment (Witten et al. 2005; Hall et al. 2009).

Linear Regression and M5P are two algorithms that can extract parametric equations. Linear Regression algorithm fits a linear model to the entire data. M5P performs a multiple linear regression model. This tree-based algorithm allocates linear regressions at the terminal nodes. It divides the entire data set into several smaller subsets and fits a linear model to each subset (Quinlan 1992). They both, consequently, result in a basic linear equation like equation (1),

$$Y = C + \sum_{i=1}^{N} A_i X_i \quad (1)$$

where A and C are fitting parameters, Y is the dependent variable, X is the independent variable, and N represents the total number of independent variables. In our case, Y is the planet's radius, and X represents other physical parameters. For these two algorithms, we use the MCMC to quantify the uncertainties of the best-fitting parameters (Goodman & Weare 2010; Foreman-Mackey et al. 2019).

To evaluate the quality of the predicted radius (R_{pre}) compared to the observed radius (R_{obs}) and to compare the efficiency of the models, root means square error (RMSE), mean absolute error (MAE), and coefficient of determination (ρ^2) are calculated. Equations (2)–(4) define RMSE, MAE, and ρ^2, respectively,

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(R_{obs} - R_{pre})^2}{n}}, \tag{2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \mid R_{obs} - R_{pre} \mid, \tag{3}$$

$$\rho^2 = 1 - \frac{\sum_{i=1}^{n} (R_{obs} - R_{pre})^2}{\sum_{i=1}^{n} (R_{obs} - R_{mean})^2}, \tag{4}$$

where R_{mean} is the mean of the R_{obs} values and n represents the total number of samples. Lower values of RMSE and MAE and higher values of ρ^2 indicate better accuracy of the models. It should be noted that hyperparameters specific to each model are tuned to have the best performance. Also, a 10-fold cross-validation procedure, as a data resampling method, is used to evaluate the performance of models. Furthermore, each algorithm is executed for original and logarithmic data sets to understand the effect of data re-scaling.

3 RESULTS

Before applying the clustering and predictive methods, we use the LOF algorithm to identify outliers in the data set containing 770 data points. The LOF algorithm marks 76 data points as outliers, resulting in a study data set of 694 planets. Appendix A describes finding outliers and their effect on prediction accuracy. In general, we find that all regression models perform poorly considering the outliers in the data set. Additionally, the effect of data re-scaling on planetary radius predictions is discussed in Appendix B. As a result, regression algorithms provide better results on a logarithmic scale.

3.1 Feature importance

It has been known that having inefficient and unnecessary features can cause declination in the performance of an ML model and lead to inaccurate results (Chandrashekar & Sahin 2014). We apply five FS methods to find and eliminate the least important planetary, stellar, and orbital parameters in predicting planet radii. They are the Spearman's rank correlation, Backward Elimination, Forward Selection, CART, and XGBoost. Features are orbital period (P), eccentricity (e), planetary mass (M_p), stellar mass (M_s), stellar radius (R_s), metallicity (Fe/H), and effective temperature ($T_{\rm eff}$), and the target variable is planetary radius (R_p).

Spearman's rank correlation (r_s) is a number between –1 and 1 that measures the monotonic correlation between two variables. This filter method reveals parameters that are strongly correlated with planetary radius: M_p with a coefficient of 0.780 has the highest correlation, followed by M_s with $r_s = 0.590$. Furthermore, $T_{\rm eff}$ with a coefficient of 0.568 and R_s with a coefficient of 0.552 are in third and fourth place, respectively. Orbital period with $r_s = -0.389$ and eccentricity with $r_s = -0.151$ are two features that have a negative correlation with R_p. In addition, highly correlated stellar parameters (R_s, M_s, and $T_{\rm eff}$) are also indicated by coefficients greater than 0.820. We should note that the estimated p-values are less than 0.001, which indicates strong certainty in the results. As an exception, the p-value corresponding to the coefficient between Fe/H and R_p ($r_s = -0.037$) is greater than 0.1, which is statistically uncertain. To calculate uncertainties in the value of correlation coefficients, we apply the Monte Carlo error analysis, considering errors in each measurement (Curran 2015). Fig. 2 illustrates the distribution of r_s for each feature except Fe/H, which has a p-value greater than 0.1. By taking the standard deviation of distributions, we report the mean and uncertainty values in the upper right corner of each panel. As it is seen, the distributions of Spearman correlation coefficients for M_p, M_s, R_s, and $T_{\rm eff}$ are constrained to positive values. In contrast, the distribution for P is limited to negative values. In the case of e, even though the mean of the distribution is slightly negative, it is exceptionally wide, taking both negative and positive values. This more extensive distribution results from higher amounts of error in eccentricity measurement.

Forward Selection and Backward Elimination are two wrapper FS methods. The procedure for Forward Selection starts with an empty set of features. Then, the best feature is determined and added to the set by applying a Random Forest regressor. In each subsequent iteration, the best remaining feature is determined and added until a complete set of features is reached. In contrast, Backward Elimination starts with a complete set of features and, at each step, eliminates the worst feature remaining in the set. Using a 10-fold cross-validation method, ρ^2 values, and corresponding standard errors are calculated for each step and shown in Fig. 3, where the left-hand panel is Forward Selection, and the right-hand panel is Backward Elimination. Both methods highlight M_p, P, and R_s as three important parameters.

Applied embedded methods include CART, which uses a decision tree regressor, and XGBoost, which implements a gradient boosting trees algorithm. These techniques score features based on their importance in computing the target variable. The ranking (and scores) obtained by the CART method are as follows: M_p (0.905), P (0.031), R_s (0.026), M_s (0.012), $T_{\rm eff}$ (0.011), e (0.008), and Fe/H (0.006). The XGBoost method ranks (and scores) features as follows: M_p (0.851), R_s (0.038), P (0.029), M_s (0.025), $T_{\rm eff}$ (0.024), Fe/H (0.021), and e (0.012). Similar importance is assigned to all features except planetary mass by CART and XGBoost. Finally, we conclude that a set of features including M_p, P, and one of the stellar parameters (M_s, R_s, or $T_{\rm eff}$) works well. Therefore, we select planetary mass, orbital period, and stellar mass as the main features.

3.2 Clusters

FS methods highlight planetary mass, stellar mass, and orbital period as vital features to estimate planet radii. We use ML clustering algorithms to investigate potential groups of exoplanets in a four-dimensional logarithmic space consisting of planetary mass and radius, stellar mass, and orbital period. The algorithms are Affinity Propagation, BIRCH, DBSCAN, Gaussian Mixture Model, Hierarchical Clustering, K-Means, Mean Shift, Mini-Batch K-Means, OPTICS, and Spectral Clustering.

3.2.1 Number of clusters

When the number of clusters (K) is not known in advance, as in our case, hierarchical clustering is an appropriate technique to adopt

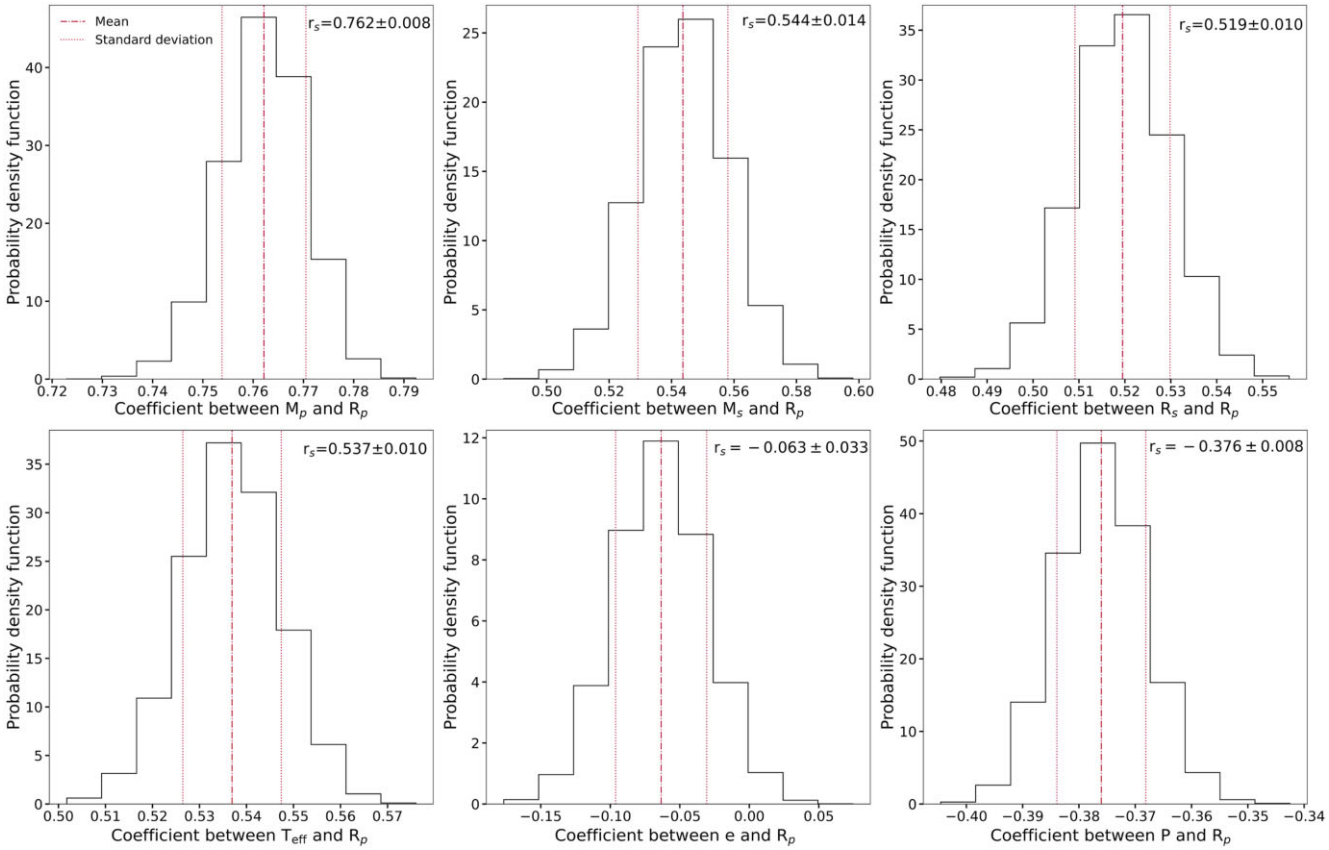

Figure 2. Distribution of correlation coefficient (r_s) between planetary radius (R_p) and other physical parameters obtained by the Monte Carlo analysis. Parameters are planetary mass (M_p), stellar mass (M_s), radius (R_s), and effective temperature ($T_{\rm eff}$), and orbital eccentricity (e) and period (P). Stellar metallicity (Fe/H), which displays a p-value greater than 0.1, has been excluded. The red dash–dotted line is the mean, and two red dotted lines represent uncertainties around it. The mean and uncertainty values are presented in the upper right corner of each panel. The absolute value of the coefficients determines the strength of the relationship. The larger the number, the stronger the relationship. It marks M_p as the most and e as the least relevant parameters to the planetary radius.

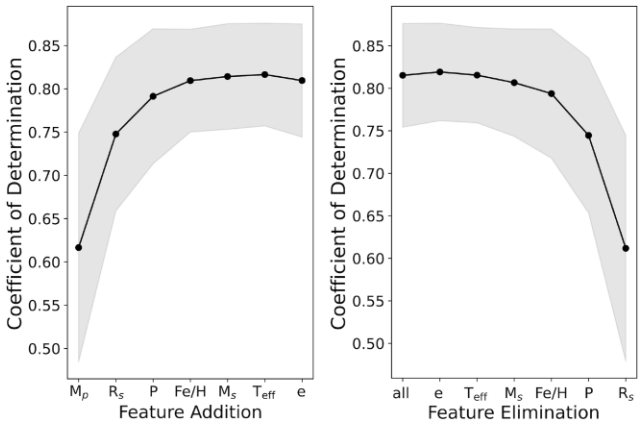

Figure 3. Wrapper FS methods. Left: coefficients of determination (ρ^2) against feature sets for the Forward Selection technique. In the first step, it determines M_p as the best feature, and in each following iteration, the best remaining feature is added to the set. Right: ρ^2 values against feature sets for the Backward Elimination technique. Unlike Forward Selection, it starts with all features and removes the worst one at each step. Features are planetary mass (M_p) and radius (R_p), orbital period (P) and eccentricity (e), and the stellar mass (M_s), radius (R_s), metallicity (Fe/H), and effective temperature ($T_{\rm eff}$). The grey areas demonstrate standard errors. Both methods highlight M_p, P, and R_s as three important parameters in predicting the planetary radius.

(Landau et al. 2011). Using the Hierarchical Clustering algorithm as a distance-based method, one can have an assumption of K. Its agglomerative algorithm assigns each data point to an individual partition; then, at each iteration, the closest pair of partitions are merged until all data belong to a single partition. Fig. 4 shows the Hierarchical Clustering dendrogram, which records the sequences of merges. The greater the height of the vertical lines in the dendrogram, the greater the distance between the clusters. Clusters can be defined by trimming a dendrogram with a distance threshold. However, there is no universal method for setting thresholds. In general, a distance threshold on the dendrogram is set to intersect the longest vertical line. The number of vertical lines intersecting the threshold line indicates the K value. Two distance thresholds are set to the dendrogram of our data set (red dashed lines in Fig. 4). The larger threshold results in two clusters with a Euclidean distance of 30.6, while the smaller threshold splits the larger cluster into two parts with a distance of 11.9 between them.

For algorithms that do not learn the K from data, we use Elbow and Silhouette methods to find the optimal value of K and use it as an input parameter in clustering algorithms. The Elbow method performs the K-Means clustering algorithm for different values of K. Then, it calculates the SSD between data points and their cluster centroids each time. Fig. 5 demonstrates SSD values as a function of K. As shown, the curve starts to flatten out and form an elbow shape in K=2, chosen as the optimal number of clusters.

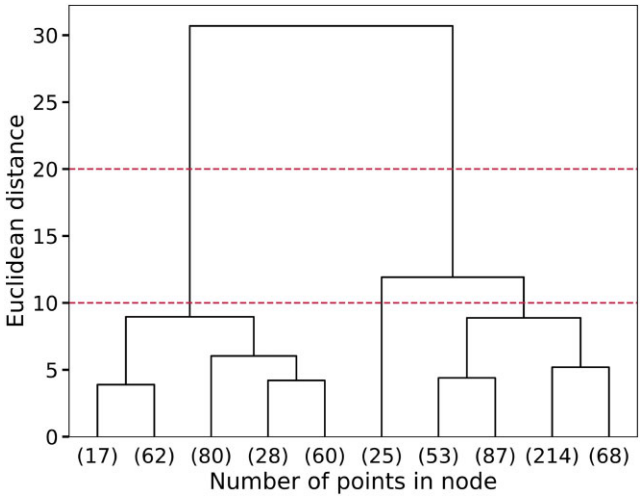

Figure 4. Hierarchical Clustering dendrogram. The greater the height of the vertical lines, the greater the distance between the clusters. Two red dashed lines are distance thresholds. The number of vertical lines intersecting the threshold line indicates the number of clusters. The larger threshold results in two clusters, while the smaller threshold results in three clusters. Due to illustration purposes, lower sequences have not been shown.

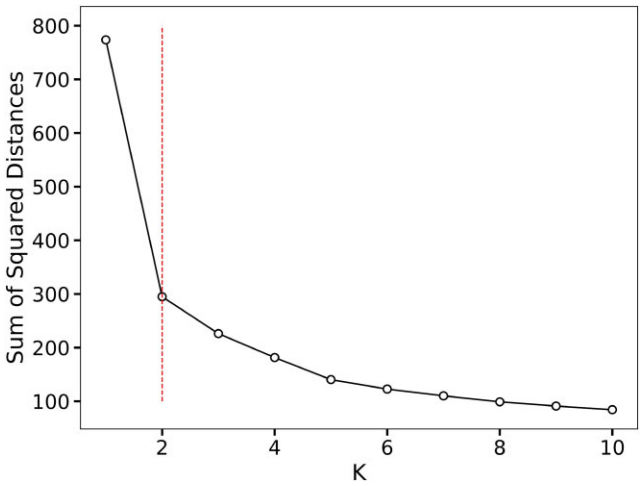

Figure 5. SSD between data points and their assigned cluster centroids against the number of clusters (K), calculated by the Elbow method. The red vertical dashed line corresponds to K=2, where the curve starts to flatten out. This point is chosen as the optimal number of clusters.

Furthermore, the average Silhouette width can evaluate clustering reliability and may be used to estimate K value (Rousseeuw 1987). The Silhouette method computes a coefficient for different values of K and uses it to determine the degree of separation between clusters. The coefficient becomes negative if the sample is assigned to the wrong cluster. Provided that the sample is far from neighbouring clusters, the coefficient will be close to +1. If the sample is on or near the decision boundary between two neighbouring clusters, the coefficient becomes 0. Fig. 6 presents the Silhouette plots for K=2, 3, 4, and 5. In this figure, the thickness of the plots represents the cluster size, and each red vertical dashed line belongs to an average Silhouette score. K=3, 4, and 5 are not appropriate due to clusters with lower-than-average Silhouette scores and wide fluctuations in the size of plots. Like the Elbow method, this method proposes K=2 as the proper number of clusters. Fig. 7 illustrates a matrix of

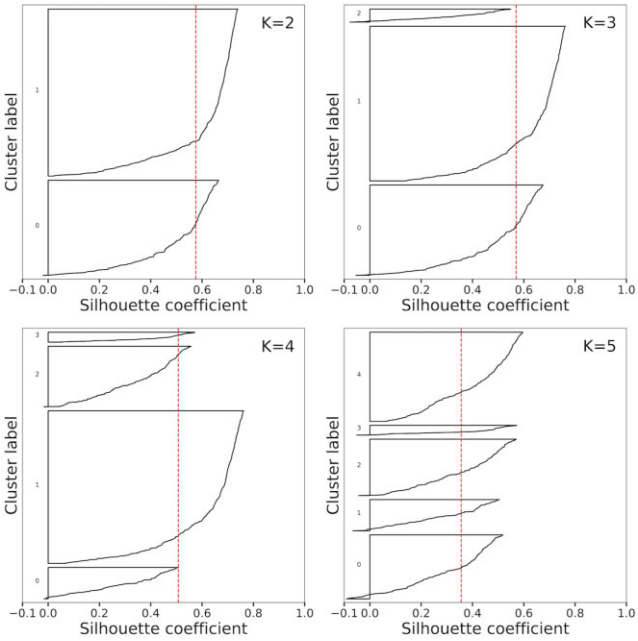

Figure 6. Silhouette plots for different numbers of clusters (K). The thickness of the plots indicates the cluster size and red vertical dashed lines represent the corresponding average Silhouette coefficients. If the sample is far from neighbouring clusters, the coefficient becomes close to +1. The coefficient can be negative if the sample is assigned to the wrong cluster. Providing that the sample is on or near the decision boundary between two neighbouring clusters, the coefficient becomes 0. Due to clusters with lower-than-average Silhouette scores and wide fluctuations in the size of plots, K=3, 4, and 5 are not appropriate. Scatter plots for each Silhouette plot are shown in Fig. 7.

scatter plots for Silhouette plots, where each row has been assigned to a specific K, and columns 1, 2, and 3 show the distribution of planetary radius versus planet's mass, orbital period, and star's mass, respectively. Like exclusive clustering algorithms where each data point belongs exclusively to one cluster (Jain & Dubes 1988), we aim to group planets into distinct non-overlapping clusters. On the one hand, if $K = 2$ (see the first row in Fig. 7), two clusters are almost well separated in all three spaces, with only a few data points overlapping. On the other hand, when planets are divided into more than two clusters (see the second, third, and fourth rows in Fig. 7), the members of the clusters become less distinguishable from each other.

For $K = 3$, planets with a longer orbital period are defined as a new cluster (black-filled squares). Although three clusters are separated in the R_p-P space, in the R_p-M_p and R_p-M_s spaces, most members of cluster 2 are distributed over the other two clusters, especially cluster 1. Likewise, in the case of $K = 4$, the cluster separation is better in the R_p-P space than in the other two spaces, where clusters overlap.

For $K = 5$, although clusters 0 and 4 are well distinguished in the planet's mass–radius distribution, they overlap a lot in the R_p-P and R_p-M_s spaces. In the R_p-P space, members of clusters 0, 1, 2, and 3 are almost separated; nevertheless, members of cluster 4 extremely overlap with those of cluster 0. In addition, cluster 3 members overlap with members of clusters 0 and 4 in the R_p-M_p and R_p-M_s spaces.

Consequently, we choose $K = 2$ based on the results from the Hierarchical, Elbow, and Silhouette methods and the idea that planetary clusters are separate groups that do not overlap. Similarly, the Affinity Propagation and Mean Shift algorithms, which do not need to specify the number of clusters, give two clusters. Another

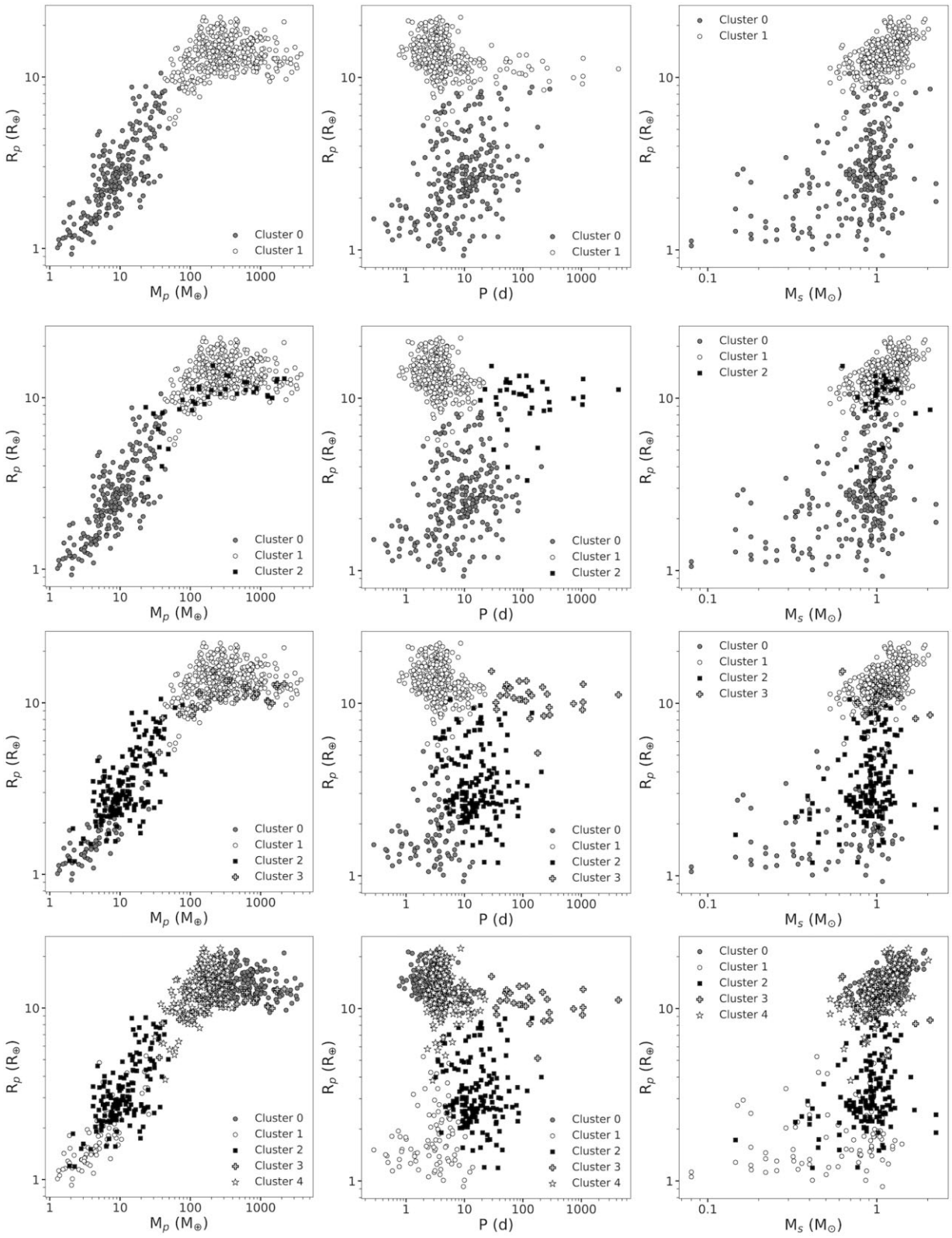

Figure 7. Matrix of scatter plots corresponding to Fig. 6. Rows 1, 2, 3, and 4 represent the number of clusters (K) equal to two, three, four, and five, respectively. Columns 1, 2, and 3 illustrate the distribution of planetary radius (R_p) versus the planet's mass (M_p), orbital period (P), and star's mass (M_s), respectively. The grey-filled circles, white circles, black-filled squares, light grey-filled pluses, and white stars represent members of clusters 0, 1, 2, 3, and 4, respectively.

Table 1. Clustering algorithms, breakpoints of radius ($B_{\rm Radius}$) and mass ($B_{\rm Mass}$) in logarithmic space, number of planets in the first (N_1) and second (N_2) clusters along with adjusted hyperparameters introduced in the Scikit-learn library (Pedregosa et al. 2011). The one-dimensional Gaussian distribution is used to find the intersection point and introduce the relevant breakpoints. Default values are set for hyperparameters that are not listed. The Elbow method and Silhouette score have been used to find the optimal number of clusters equal to 2. DBSCAN and OPTICS algorithms fail to provide appropriate clusters.

Algorithm	$B_{\rm Radius}$	$B_{\rm Mass}$	N_1	N_2	Adjusted parameter
Affinity propagation	0.93	1.73	254	440	damping=0.9, preference=−60
BIRCH	0.90	1.72	247	447	n_clusters=2, threshold=0.01
DBSCAN	–	–	–	–	eps=0.2, min_samples=25
Gaussian Mixture Model	0.95	1.80	271	423	n_components=2
Hierarchical Clustering	0.90	1.72	247	447	n_clusters=2
K-Means	0.92	1.72	252	442	n_clusters=2
Mean shift	0.93	1.73	257	437	bandwidth=0.9
Mini-Batch K-Means	0.92	1.72	252	442	n_clusters=2
OPTICS	–	–	–	–	min_samples=40
Spectral clustering	0.89	1.64	239	455	n_clusters=2

point that should be taken into account is that this number of clusters chosen is consistent with published works (e.g. Weiss et al. 2013 and Bashi et al. 2017), where two regimes were introduced in the planetary mass–radius relation by different techniques.

As for DBSCAN and OPTICS algorithms, they give K=2 but cannot separate clusters well; thus, we exclude them from the analysis. The effectiveness of clustering methods depends on several factors, including the data set's characteristics and underlying distribution. Different clustering algorithms make certain assumptions about the data structure and employ distinct approaches to identify clusters. Accordingly, their performance can vary based on how well these assumptions align with the data set's properties. In cases where clustering algorithms fail to find appropriate clusters, there may be several potential explanations. One important factor is the distribution of the exoplanet data, which might not conform to the assumptions made by certain clustering algorithms. DBSCAN and OPTICS are two density-based methods that execute clustering by finding areas where data points are concentrated. They can discover arbitrarily shaped clusters, including non-spherical ones; they, however, might fail when the data set is too sparse, and the density varies across the data, like in the case of exoplanet data (Moreira, Santos & Carneiro 2005; Ahmad & Dang 2015).

3.2.2 *Planet classes*

The planets are divided into two groups by choosing $K = 2$ for BIRCH, Gaussian Mixture Model, Hierarchical Clustering, K-Means, Mini-Batch K-Means, and Spectral Clustering. Moreover, the Affinity Propagation and Mean Shift algorithms, which learn the number of clusters from data, result in two clusters. To introduce a boundary between two groups in the planet's mass–radius space, we construct a Gaussian kernel density estimation for each cluster and find the intersection point. Table 1 lists the results of the clustering algorithms, which are almost similar (except for DBSCAN and OPTICS, as discussed in Section 3.2.1). Ultimately, we separate data into two classes using an average value of $\log R_p = 0.91$ ($R_p = 8.13R_\oplus$) for radius breakpoint (B_{Radius}) and $\log M_p = 1.72$ ($M_p = 52.48M_\oplus$) for mass breakpoint (B_{Mass}). Exoplanets with $R_p \leq 8.13R_\oplus$ and $M_p \leq 52.48M_\oplus$ are defined as small planets, and those with $R_p > 8.13R_\oplus$ and $M_p > 52.48M_\oplus$ as giant planets. Fig. 8 shows the mass–radius distribution of clustered and outlier data. For several planets that lie outside the boundaries (B_{Radius} and B_{Mass}), we use the criterion of their closeness to the boundaries to assign them to either of the classes.

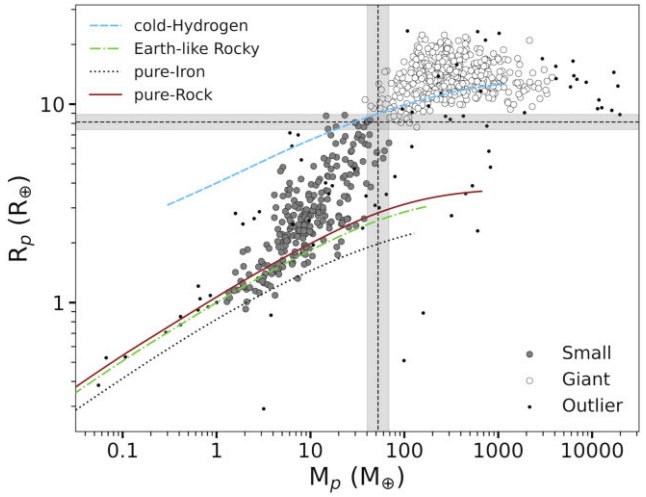

Figure 8. The mass–radius distribution of clustered data. Data are separated into two classes using $R_p = 8.13R_\oplus$ (horizontal dashed line) and $M_p = 52.48M_\oplus$ (vertical dashed line). The grey areas demonstrate mean errors of mass and radius. Exoplanets with $R_p \leq 8.13R_\oplus$ and $M_p \leq 52.48M_\oplus$ are defined as small planets (grey circles), and those with $R_p > 8.13R_\oplus$ and $M_p > 52.48M_\oplus$ as giant planets (white circles). There are 254 small planets and 440 giant planets. The black dots are outlier planets found by the LOF method (see Appendix A). Four iso-density curves are also drawn: cold-hydrogen (blue dashed line), Earth-like rocky (green dash–dotted line), pure-iron (black dotted line), and pure rocky (solid crimson line) planets (Marcus et al. 2010; Becker et al. 2014).

According to a traditional definition, small exoplanets are planets with radii smaller than $4R_\oplus$ and masses lower than $\sim 30M_\oplus$ (Howard et al. 2010; Marcy et al. 2014; Weiss & Marcy 2014); however, this customary definition of small and large planets does not exactly match previous studies that have investigated a transition point in the mass–radius distribution of exoplanets. Table 2 compares our breakpoints with those found by others in the literature and shows a considerable difference between mass breakpoints. Besides an increasing number of exoplanets and the evolution of their mass–radius distribution, this difference could result from applying different methods to find a cut-off point in planetary masses. Performed methods vary from a simple visual investigation of mass–radius and

Table 2. Breakpoints of mass ($B_{\rm Mass}$) and radius ($B_{\rm Radius}$) derived by previous studies and in this work.

Study	$B_{\rm Mass}$ ($M_\oplus$)	$B_{\rm Radius}$ ($R_\oplus$)
Weiss et al. (2013)	150	–
Hatzes & Rauer (2015)	95	–
Chen & Kipping (2017)	130 ± 22	–
Bashi et al. (2017)	124 ± 7	12.1 ± 0.5
This work	52.48	8.13

mass–density distributions (Weiss et al. 2013) to using different slope criteria in the planetary parametric relations (Hatzes & Rauer 2015; Bashi et al. 2017; Chen & Kipping 2017). As a result, the mass and radius breakpoints identified in this work ($52.48M_\oplus$ and $8.13R_\oplus$, respectively) are closer to traditional breakpoints than those found in previous studies.

There are 254 small planets and 440 giant planets. The distributions of the orbital period, stellar mass, average density, and calculated equilibrium temperature for small and giant planets are demonstrated in Fig. 9. The upper left-hand panel demonstrates that most giant planets are closer to their host star than small planets are. They, none the less, have a P varying from 0.77 to 4331.01 d, while for small planets, it is between 0.28 and 207.62 d. The upper right-hand panel shows that the host star mass is frequently in a range around the Solar mass for most planets. It results from a selection effect: exoplanet-search programs often concentrate on Sun-like stars. In addition to this, lower mass stars are not as likely to host exoplanets with sufficient mass to be identified by the radial-velocity technique (Bonfils et al. 2005; Cumming et al. 2008). Small planets revolve around stars with M_s between 0.08 and 2.24 $M_\odot$, while, for host stars of giant planets, they vary from 0.53 to 2.07 $M_\odot$.

The lower left-hand panel compares small and giant planets' average density distribution. As expected, giant planets are generally less dense than small planets. The density distribution of giant planets peaks at about Saturn's density. Hence, a significant fraction of giant planets has an average density similar to Saturn, the least dense

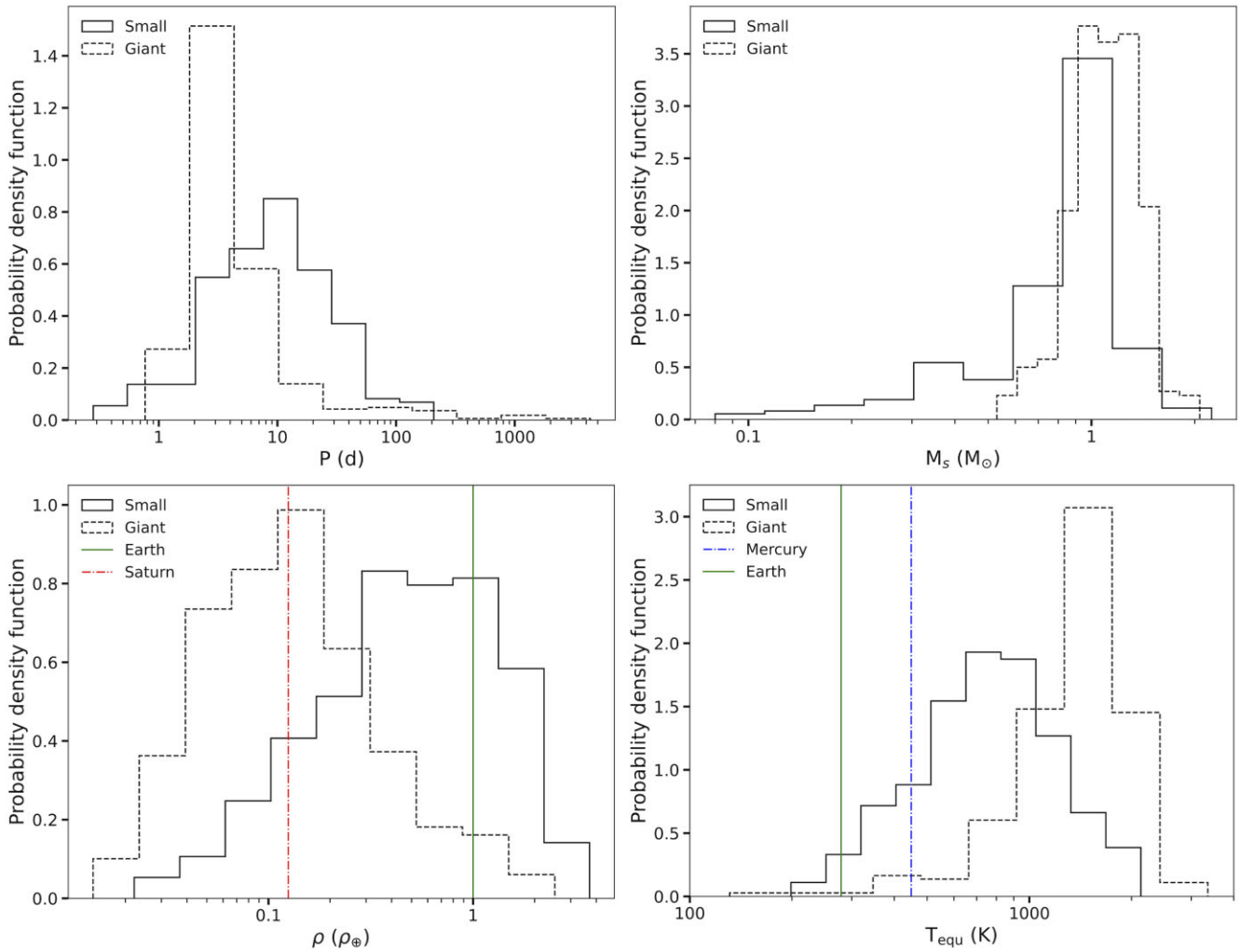

Figure 9. Histograms of the orbital period (upper left-hand panel), stellar mass (upper right-hand panel), average density (lower left-hand panel), and equilibrium temperature (lower right-hand panel) for small (solid-border bars) and giant (dashed-border bars) planets. The red dash–dotted line in the lower left-hand panel represents Saturn as the least dense planet, and the green solid line is Earth as the densest planet in the Solar system. In the lower right-hand panel, the green solid line represents Earth's equilibrium temperature, and the blue dash–dotted line is Mercury, which has the highest equilibrium temperature in the Solar system. The orbital period distribution shows that giant planets are closer to their host star than small planets. The stellar mass distribution demonstrates that, for most planets, the host star's mass is around the Sun's mass. Comparing planets' average density and calculated equilibrium temperature demonstrate that giant planets are hotter and less dense than small planets.

planet in the Solar system. On the contrary, small planets cover a higher density range that includes Earth, the densest planet in the Solar system. This, in turn, suggests that giant planets are composed mainly of hydrogen and helium envelopes, whereas heavier elements dominate small planets.

Comparing the calculated equilibrium temperature of planets (lower right-hand panel) demonstrates that giant planets are hotter than small planets. This temperature difference between the two planet classes is expected because giant planets are closer to their host star (as shown in the upper left-hand panel); in addition to this, the host stars are almost of the same spectral type (equivalently, stellar mass, as shown in the upper right-hand panel).

It is important to note that our sample of giant planets is dominated by gas giant exoplanets with orbital periods of less than 10 d, commonly referred to as hot Jupiters. It is believed that these planets are more likely to be found around metal-rich stars than around stars with low stellar metallicity (Maldonado, Villaver & Eiroa 2018; Osborn & Bayliss 2020; Yee & Winn 2023).

3.3 Prediction of the planetary radius

To find the radius of a planet based on physical parameters, ML predictive algorithms are applied to our data set of 694 planets. The physical parameters are the planet's mass (M_p), orbital period (P), and host star's mass (M_s), selected by FS methods. Bootstrap Aggregation, Decision Tree, K-Nearest Neighbors, Linear Regression, Multilayer Perceptron, M5P, Random Forest, and SVR are implemented algorithms. These algorithms are applied separately to entire, small, and giant planets. To have the best performance of algorithms, the hyperparameters are tuned. Furthermore, a 10-fold cross-validation procedure is used to assess the performance of models. The RMSE, mean absolute error (MAE), and coefficient of determination (ρ^2) are calculated as validation metrics (see equations 2, 3, and 4).

RMSE, MAE, and ρ^2 of entire, small, and giant planets, along with the tuned hyperparameters, are listed in Table 3. Fig. 10 shows the box plots of accuracy in the 10-fold cross-validation for models. SVR with an RMSE of 0.093 is the best-performing model for the entire data set, followed by Bootstrap Aggregation, Random Forest, and M5P with RMSEs of 0.096, 0.097, and 0.098, respectively. They also have lower MAE and higher ρ^2 values than other models. The SVR performs better than other algorithms for both subsets of small and giant planets.

Fig. 11 compares the observed ($R_{\rm obs}$) and predicted ($R_{\rm pre}$) radius (upper panel) along with residual values (lower panel) obtained by SVR as the best-performing model. In this figure, the model has been applied separately to small and giant planets, resulting in a gap and a relatively higher dispersion around $\sim 8R_{\oplus}$. ρ^2 value is 0.710 for small planets and 0.510 for giant planets. The normalized median absolute deviation (NMAD) has been calculated for small and giant planets predictions. NMAD is $R_{pre} = R_{obs} \pm \sigma(1 + R_{obs})$, where $\sigma = 1.48 \times median[|R_{pre} - R_{obs}|/(1 + R_{obs})]$ (Hoaglin, Mosteller & Tukey 1983). As observed in the lower panel of Fig. 11, a distinct linear trend with a slope of 0.459 is noticeable in the residuals between predicted and observed radii of giant planets. This trend could be attributed to factors such as systematic observation errors, the determination of physical parameters, and calculation issues including hyperparameter tuning, algorithm characteristics and limitations, and selected features. We adjust the learning algorithms to address this challenge by re-configuring their hyperparameters. Interestingly, a similar pattern emerges in the residual values across all eight predictive algorithms. This suggests that the constraints

Table 3. Results obtained by different regression algorithms. These algorithms have been separately applied to entire, small, and giant planets. Column 1 presents the name of the algorithm. Columns 2, 3, and 4 present the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (ρ^2) for the whole data set. Columns 5 to 10 present RMSE, MAE, and ρ^2 for small and giant planets, respectively. The last column lists adjusted hyperparameters introduced in the Weka environment (Witten et al. 2005; Hall et al. 2009).

Algorithm	RMSE	MAE	ρ^2	RMSE_1	MAE_1	ρ_1^2	RMSE_2	MAE_2	ρ_2^2	Adjusted parameter
Bootstrap Aggregation	0.096	0.068	0.933	0.126	0.096	0.692	0.064	0.047	0.489	numIterations=100 classifier=REPTree
Decision Tree	0.108	0.077	0.916	0.142	0.109	0.616	0.071	0.051	0.392	maxDepth=−1 minNum=2
K-Nearest Neighbors	0.104	0.075	0.921	0.140	0.106	0.627	0.072	0.053	0.390	KNN=3 distanceFunction=Euclidean
Linear Regression	0.155	0.127	0.822	0.128	0.099	0.683	0.069	0.051	0.402	eliminateColinearAttributes=True attributeSelectionMethod=M5
Multilayer Perceptron	0.112	0.084	0.909	0.131	0.103	0.670	0.064	0.048	0.491	learningRate=0.1 decay=False momentum=0.1 hiddenLayers=a
M5P	0.098	0.071	0.930	0.130	0.101	0.688	0.073	0.053	0.391	unpruned=False
Random Forest	0.097	0.068	0.932	0.128	0.096	0.682	0.064	0.047	0.485	numIterations=100 maxDepth=0 numFeatures=2
SVR	0.093	0.065	0.937	0.123	0.088	0.710	0.063	0.046	0.510	c=1 kernel=Puk

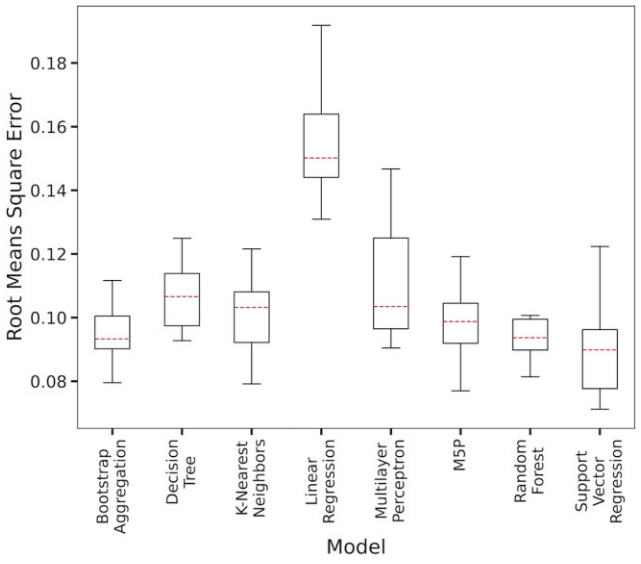

Figure 10. Box plots showing the spread of accuracy in the 10-fold cross-validation for predictive algorithms. On each box, the red dashed mark is the median, and the edges of the box are the 25th and 75th percentiles. SVR with an RMSE of 0.093 is the best-performing model for the entire data set, followed by Bootstrap Aggregation, Random Forest, and M5P with RMSEs of 0.096, 0.097, and 0.098, respectively.

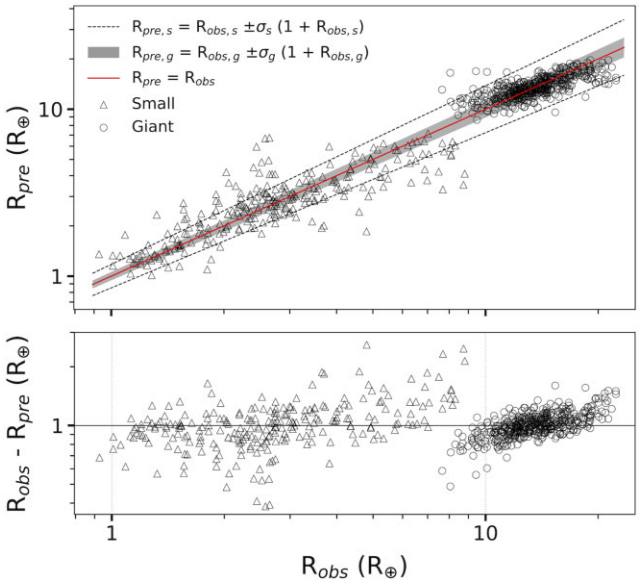

Figure 11. Comparison between observed ($R_{\rm obs}$) and predicted ($R_{\rm pre}$) planetary radius (upper panel) along with residual values (lower panel) obtained by the SVR model, which has been applied separately to small (triangles) and giant (circles) planets. The red line indicates $R_{pre} = R_{\rm obs}$ along with two dashed lines and a grey area that illustrate the NMAD for small and giant planets, respectively. Considering Hoaglin et al. (1983), NMAD is calculated using $R_{\rm pre} = R_{\rm obs} \pm \sigma(1 + R_{\rm obs})$, where $\sigma = 1.48 \times median[|R_{\rm pre} - R_{\rm obs}|/(1 + R_{\rm obs})]$.

of SVR alone cannot explain this pattern. Moreover, we discover that excluding or including the parameter M_p significantly affects the slope of the linear trend (± 0.08) compared to other parameters. However, modifying the subset of features does not eliminate this trend. The most gradual trend appears when utilizing the feature subset of M_p, P, and M_s.

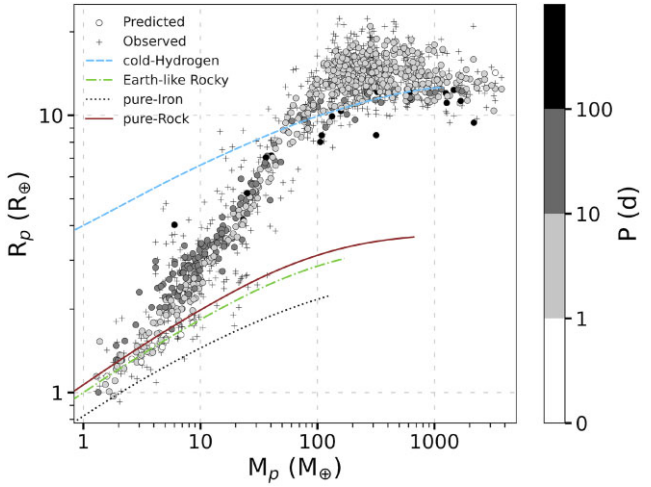

Figure 12. Predicted (circles) and observed (pluses) radius as a function of a mass and orbital period obtained by the SVR model for all planets in the sample. The distribution of cold-hydrogen (blue dashed line), Earth-like rocky (green dash–dotted line), pure-iron (black dotted line), and pure rocky (solid crimson line) planets are also illustrated (Marcus et al. 2010; Becker et al. 2014).

It is important to note that this trend minimally impacts radius prediction. Additionally, the residuals between predicted and observed radii of giant planets fall within the range seen for small planets. It is plausible that this trend is linked to systematic issues in exoplanet observations or the determination of observed parameters (e.g. mass, orbital period, and radius). As the current sample of exoplanets detected through the transit method is substantial, further investigation of this effect could be undertaken when a statistically significant sample of exoplanets detected through other detection methods becomes available.

Fig. 12 shows the predicted and observed radii as a function of a mass and orbital period obtained by the SVR model. The model has been applied to the entire sample in this figure, resulting in an ρ^2 of 0.937. SVR can efficiently reproduce the spread in radius.

Linear Regression and M5P can derive parametric equations between physical parameters by fitting linear models to the exoplanet data. A linear model is fitted to the real data in the Linear Regression algorithm. At the same time, M5P splits the entire data set into several subsets and fits a multivariate linear function to each subset. Equation (5) presents a linear fit between the planetary radius, planetary mass, orbital period, and stellar mass derived by Linear Regression and M5P, where A_{M_p}, A_P, A_{M_s}, and C are fitting parameters.

$$\log\left(\frac{R_p}{R_\oplus}\right) = A_{M_p}\log\left(\frac{M_p}{M_\oplus}\right) + A_P\log\left(\frac{P}{d}\right) + A_{M_s}\log\left(\frac{M_s}{M_\odot}\right) + C. \quad (5)$$

Best-fitting parameters obtained by Linear Regression and M5P are listed in Table 4. Row 1 presents the linear fit of all planets provided by the Linear Regression algorithm. Running this algorithm independently for clusters produces individual linear fit for each cluster (rows 2 and 3). For small planets $A_{M_s} = 0$, and for giant planets $A_{M_p} = 0$, implying no dependence between the planetary radius and stellar mass of small planets, as well as between the radius and mass of giant planets.

M5P divides the planets into two groups using a mass breakpoint of $\log M_p = 1.717$ ($M_p = 52.12 M_\oplus$). Interestingly, clustering algorithms

Table 4. Parametric equations obtained by different regression algorithms. Rows 1 to 7 present a linear fit, which equates the logarithm of planetary radius (R_p) to logarithms of planetary mass (M_p), orbital period (P), and stellar mass (M_s) plus a constant term (C) (see equation (5)). Row 8 presents a linear fit between planetary mass and radius as $\log(R_p/R_\oplus) = A_{M_p} \log(M_p/M_\oplus) + C$. The linear fit between the planetary radius and stellar mass is also presented in row 9 as $\log(R_p/R_\oplus) = A_{M_s} \log(M_s/\mathrm{M}_\odot) + C$. The first column shows the row number. Column 2 presents the used data set: the entire data set, small planets, or giant planets. Best-fitting parameters are listed in columns 3 to 6. The last column presents the applied algorithm: Linear Regression, M5P, or MCMC.

#	Data set	A_{M_p}	A_P	A_{M_s}	C	Algorithm
1	Entire	0.367	−0.030	0.280	0.191	Linear Regression
2	Small	0.467	0.090	0	−0.103	Linear Regression
3	Giant	0	−0.069	0.480	1.157	Linear Regression
4	Small	0.481	0.076	0.016	−0.095	M5P
5	Giant	0.012	−0.067	0.489	1.123	M5P
6	Small	$0.482^{+0.025}_{-0.024}$	$0.078^{+0.017}_{-0.016}$	$0.031^{+0.041}_{-0.042}$	$-0.099^{+0.030}_{-0.030}$	M5P and MCMC
7	Giant	$0.013^{+0.010}_{-0.009}$	$-0.070^{+0.007}_{-0.007}$	$0.492^{+0.036}_{-0.036}$	$1.121^{+0.024}_{-0.024}$	M5P and MCMC
8	Small	$0.497^{+0.023}_{-0.022}$	–	–	$-0.050^{+0.024}_{-0.024}$	MCMC
9	Giant	–	–	$0.480^{+0.036}_{-0.037}$	$1.109^{+0.004}_{-0.004}$	MCMC

also find this breakpoint (see Table 1). Rows 4 and 5 present multivariate linear fits of small and giant planets produced by the M5P algorithm. Splitting the data provides M5P with much better results than Linear Regression (see Table 3). To estimate the uncertainty values of the M5P's best-fitting parameters, we use the MCMC method. The likelihood function and initial values implemented in the MCMC analysis are the same as those acquired by the M5P. In Table 4, rows 6 and 7 present the linear fits of small and giant planets, respectively, together with the uncertainty values obtained by the MCMC method. Fig. 13 presents the distributions of the predicted and observed radii as a function of mass and orbital period, obtained by the M5P model for small (lower panel) and giant (upper panel) planets. The value of ρ^2 for the whole sample is 0.930. The predicted radii reproduce the spread in radius, especially for giant planets.

3.4 Dependence of planetary radius on host star's mass

There are inconsistent assertions in the literature about the dependence of planetary parameters on the host star's mass. Pascucci et al. (2018) claimed that the mass of the most common exoplanets depends on their host star mass. They investigated G, K, and M stars and suggested that planets around relatively low-mass stars (with a mass lower than $1\mathrm{M}_\odot$) are lower in mass and smaller in radius. In contrast, Neil & Rogers (2018) showed that the mass–radius relation of small planets has no strong dependence on stellar mass. Wu (2019) discussed a linear relationship between exoplanet mass and host star mass and the lack of correlation between the planetary radius and stellar metallicity. In addition, Lozovsky et al. (2021) studied exoplanets with radii up to $8R_\oplus$ and masses up to $20M_\oplus$ surrounding G and K stars. They confirmed that exoplanets revolving around more massive stars tend to be larger and more massive.

As can be seen in Table 4, the radius of a small planet shows a strong dependency on its mass. Furthermore, there is no strong correlation between stellar mass and planetary radius for small planets. In comparison, the radius of a giant planet depends weakly on its mass because above $\sim 8R_\oplus$ the electron degeneracy pressure dominates (Zapolsky & Salpeter 1969; Seager et al. 2007; Swift et al. 2012). In addition to this, the planetary radius and stellar mass of giant planets have a strong linear correlation. We apply the MCMC as a supportive method to find the best scaling relations between the radius and mass of small planets and between the radius and

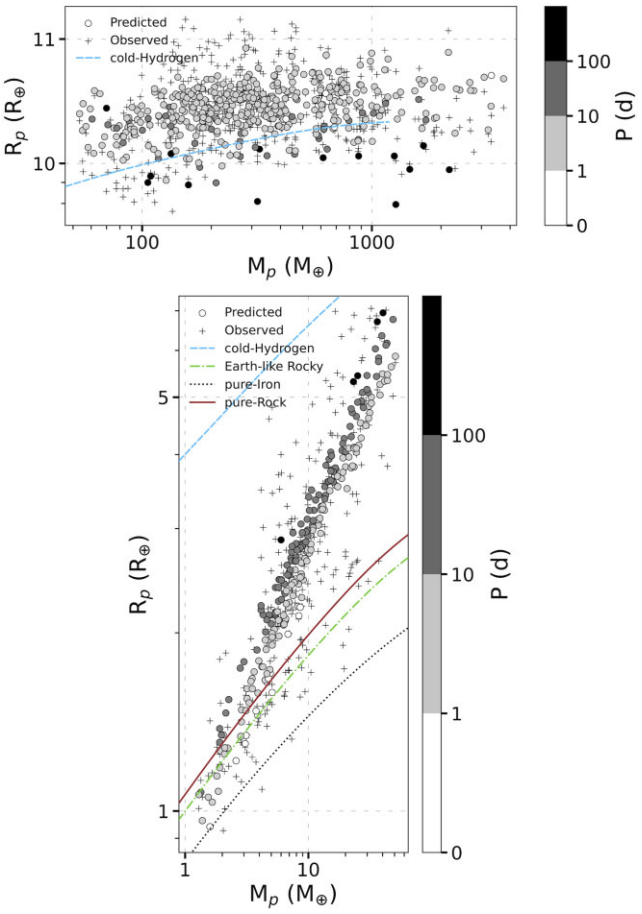

Figure 13. Predicted (circles) and observed (pluses) radius as a function of a mass and orbital period obtained by the M5P model for small (lower panel) and giant (upper panel) planets. Four iso-density curves are also drawn: cold-hydrogen (blue dashed line), Earth-like rocky (green dash–dotted line), pure-iron (black dotted line), and pure rocky (solid crimson line) planets (Marcus et al. 2010; Becker et al. 2014).

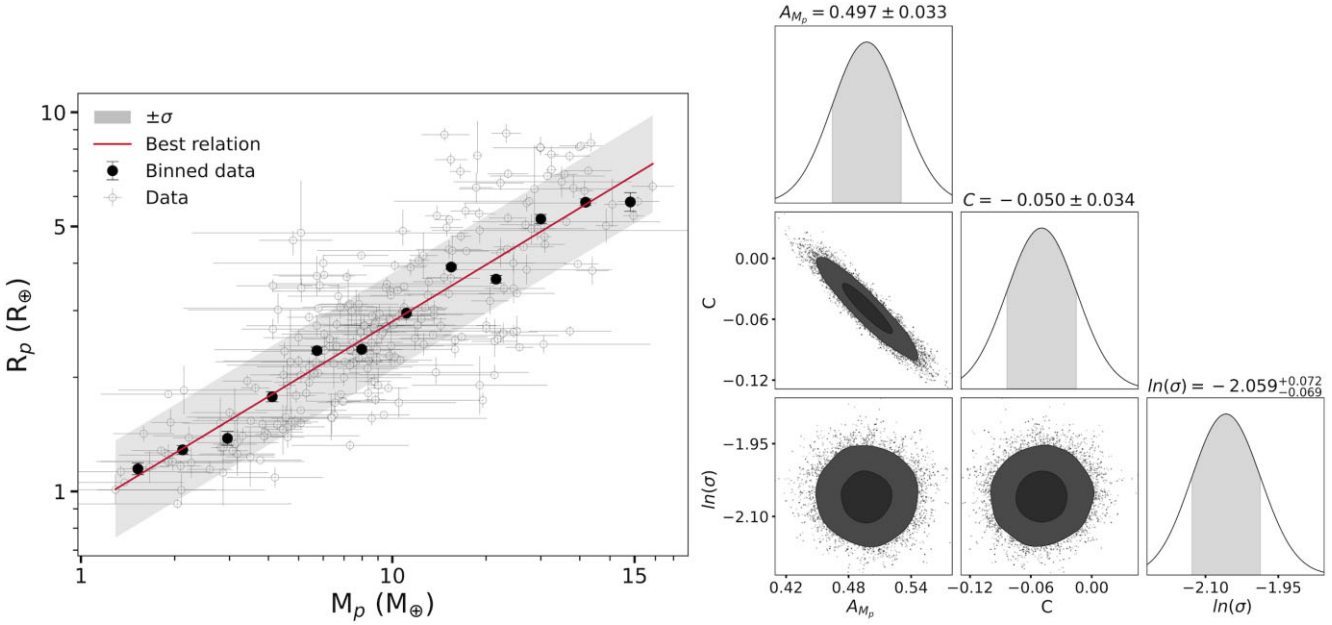

Figure 14. Left-hand panel: the relation between mass and radius of small planets obtained by the MCMC method. The red line is the best scaling relation plotted using $\log(R_p/R_\oplus) = 0.497\log(M_p/M_\oplus) - 0.050$ (see Table 4, row 8). The grey area is $\pm 1\sigma$ uncertainties around the best scaling relation. For better illustration, the data points have been binned with a width of 0.05 $\log M_p$. The binned and actual data are shown in black and grey circles. Right-hand panel: the one- and two-dimensional marginalized posterior distributions of the scaling relation parameters obtained by the MCMC method. A_{M_p} and C are the slope and intercept, respectively. The uncertainty around the best scaling relation is shown by σ.

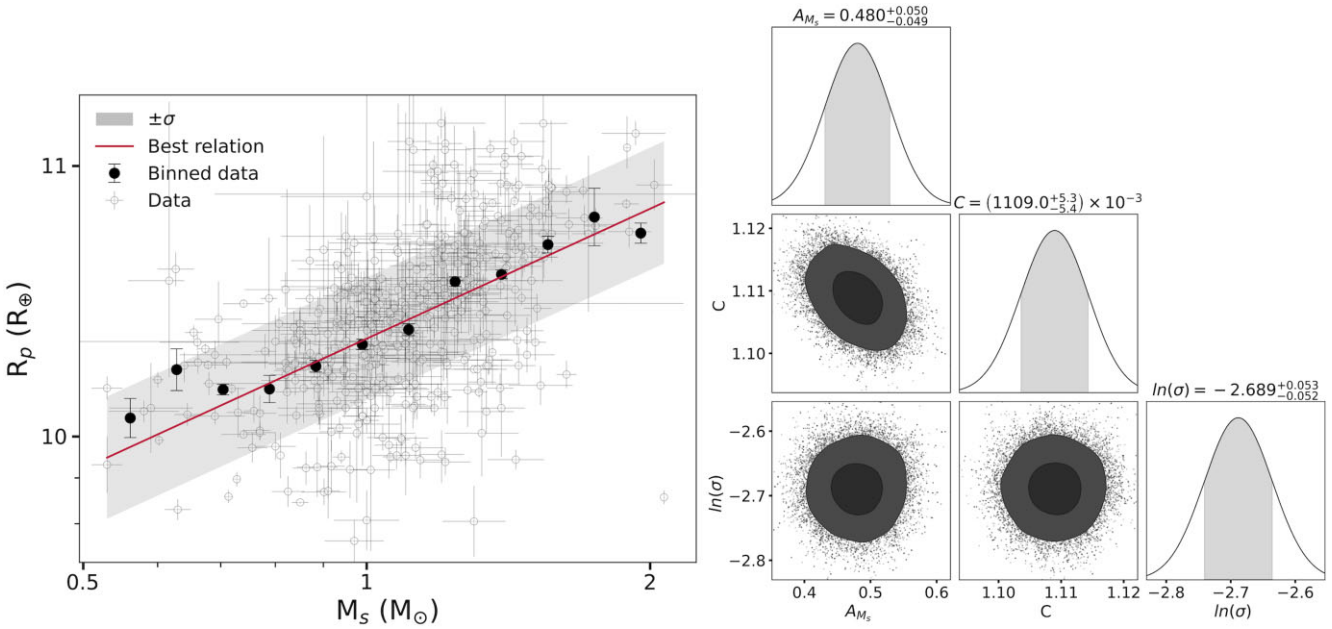

Figure 15. Left-hand panel: the relation between the radius of giant planets and the host star's mass obtained by the MCMC method. The red line is the best scaling relation plotted using $\log(R_p/R_\oplus) = 0.480\log(M_s/M_\odot) + 1.109$ (see Table 4, row 9). The grey area is $\pm 1\sigma$ uncertainties around the best scaling relation. For better illustration, the data points have been binned with a width of 0.05 $\log M_s$. The binned and actual data are shown in black and grey circles. Right-hand panel: the one- and two-dimensional marginalized posterior distributions of the scaling relation parameters obtained by the MCMC method. A_{M_s} and C are the slope and intercept, respectively. The uncertainty around the best scaling relation is shown by σ.

stellar mass of giant planets while considering the reported errors of physical values. Row 8 of Table 4 presents the linear fit between the radius and mass of small planets. Additionally, the linear fit between the radius of giant planets and the mass of their host stars is presented in row 9. The related diagrams are depicted in Figs 14 and 15.

Giant planets are less dense than small planets (see the lower left-hand panel of Fig. 9) and mostly composed of volatile elements (hydrogen and helium envelopes). On the other hand, giant planets orbit stars more massive than $\sim 1 M_\odot$, whereas the hosts of small planets include low-mass stars, that is, they have a mass greater

than 0.08$M_\odot$ (see the upper right-hand panel of Fig. 9). Concentrating on a limited sample of exoplanets, Lozovsky et al. (2021) concluded that planets forming around massive stars accrete more H–He atmospheres than those that form around low-mass stars. Our extensive sample suggests a similar scenario for giant planets. Hence, the dependence of the radius of giant planets on the host star's mass may result from their different planetary composition. It should be noted that, in addition to naturally correlated parameters, this trend with stellar mass might be a consequence of observational biases. The larger transiting exoplanets are more detectable around luminous stars with larger masses (Bhatti et al. 2016).

3.5 Effect of equilibrium temperature, semimajor axis, and luminosity

Bhatti et al. (2016) used a Random Forest model to assess the effect of different physical parameters on predicting a planet's radius. They concluded that the planet's mass and equilibrium temperature have the greatest effect. In a similar work, Ulmer19 presented planetary mass, equilibrium temperature, semimajor axis, stellar radius, mass, luminosity, and effective temperature as important parameters, and stellar metallicity, orbital period, and eccentricity as the three least important parameters. We add orbital periods to the data set collected by Ulmer19 and transfer it to a logarithmic space. Hence, a new data set consisting of 506 planets is provided. The features of this data set are as follows: orbital period (P), planetary mass (M_p), semimajor axis (a), equilibrium temperature ($T_{\rm equ}$), luminosity (L), stellar mass (M_s), stellar radius (R_s), and effective temperature ($T_{\rm eff}$). To evaluate the effect of equilibrium temperature, semimajor axis, and luminosity in predicting planetary radius and to compare the performance of Random Forest and SVR models in Ulmer19's data set, we implement models on the nine feature combinations. As the most important parameter, planetary mass is added to all combinations. Fig. 16 presents the RMSE values obtained by Random Forest and SVR versus different feature combinations. The SVR model performs better than Random Forest for all combinations.

According to Kepler's third law, the orbital period and semimajor axis are correlated to each other (Cox 2015). So, as expected, considering the SVR model, the RMSE values of the second (0.104) and third (0.103) combinations are not significantly different. The second combination consists of M_p and a, while the third set includes M_p and P. Moreover, the equilibrium temperature of a planet can be calculated using equation (6) without considering the effect of albedo and eccentricity (Laughlin & Lissauer 2015).

$$T_{\rm equ} = \sqrt{\frac{R_s}{2a}} \times T_{\rm eff}. \tag{6}$$

The fourth combination includes planetary mass and equilibrium temperature, and the fifth combination includes planetary mass and constituent parameters of equilibrium temperature (a, R_s, and $T_{\rm eff}$). The SVR model's RMSE values corresponding to the fourth and fifth sets are almost identical (0.096).

The luminosity of a star is correlated to its radius and effective temperature ($L \propto R_s^2 \times T_{\rm eff}^4$). The sixth combination is planetary mass and luminosity, whose RMSE value (0.100) is almost the same as the seventh combination, which includes planetary mass and constituent parameters of luminosity ($T_{\rm eff}$ and R_s). Additionally, the eighth set corresponds to features selected by Ulmer19, including M_p, $T_{\rm equ}$, a, R_s, M_s, L, and $T_{\rm eff}$, and the last set consists of M_p, P, and M_s, which we have selected. Using the SVR model, it is clear that there is no remarkable difference between the results obtained by these two feature sets. RMSE of our feature set equals 0.095 while 0.096 for that used by Ulmer19.

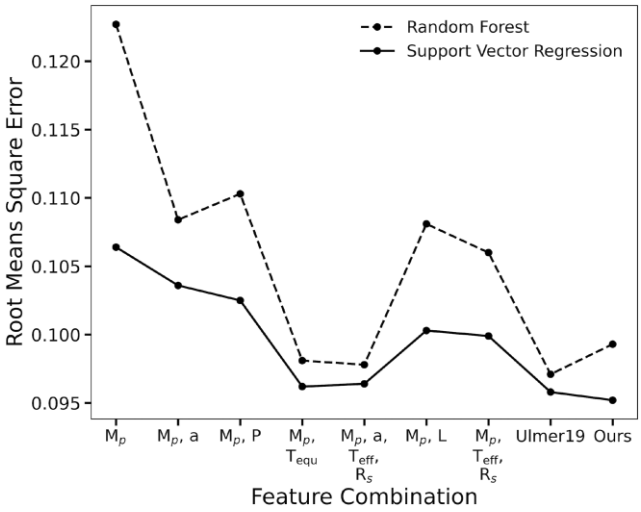

Figure 16. Comparison of the performance of Random Forest and SVR models for different feature combinations. The data set consists of 506 planets collected by Ulmer19 and transferred to a logarithmic space. The dashed and solid lines represent RMSEs obtained by Random Forest and SVR, respectively. Features are as follows: orbital period (P), planetary mass (M_p), semimajor axis (a), equilibrium temperature ($T_{\rm equ}$), luminosity (L), stellar mass (M_s), stellar radius (R_s), and effective temperature ($T_{\rm eff}$). The eighth set consists of M_p, $T_{\rm equ}$, a, R_s, M_s, L, and $T_{\rm eff}$, selected by Ulmer19. The last set is our feature combination, which contains M_p, P, and M_s. The SVR model performs better than the Random Forest model for all combinations.

Although Ulmer19 considered the orbital period as an inconsequential parameter, here we show that the orbital period (or semimajor axis), along with the planet's mass and one of the stellar parameters, have remarkable effects on predictions. Moreover, contrary to the results acquired by Bhatti et al. (2016) and Ulmer19, it seems that considering stellar luminosity and planetary equilibrium temperature as features do not improve the accuracy of planetary radius predictions. Luminosity and equilibrium temperature are two physically dependent parameters. Thus, similar results can be achieved only by considering their constituent parameters.

4 SUMMARY AND CONCLUSIONS

In this study, we conduct a comprehensive analysis of a sample comprising 762 exoplanets and eight Solar system planets. Our main objective is to investigate the characteristics of these exoplanets and explore the correlations between various features. The data set includes essential parameters such as orbital period (P) and eccentricity (e), planetary mass (M_p), and radius (R_p), and the stellar mass (M_s), radius (R_s), metallicity (Fe/H), and effective temperature ($T_{\rm eff}$).

To ensure the reliability of our analysis, we employ the LOF algorithm, which allows us to identify and filter out data points that deviate significantly from the overall data set. This process leads us to a refined data set consisting of 76 anomalous objects, which can be considered as robust and reliable measurements for our subsequent analysis.

By utilizing FS methods, we determine the most influential factors in predicting the radius of exoplanets. Our findings highlight that planetary mass (M_p) plays a pivotal role in this regard, whereas

eccentricity (e) and metallicity (Fe/H) demonstrate relatively lesser significance in the prediction process.

To further understand the underlying structure of the data set, we employ various clustering algorithms and evaluation techniques such as the Elbow, Silhouette, and Hierarchical methods. Based on the outcomes of these analyses and in alignment with the conclusions drawn by Bashi et al. (2017), we opt to divide the data set into two distinct clusters: small and giant planets. Notably, we observe distinct breakpoints in the mass–radius space at $M_p = 52.48M_\oplus$ and $R_p = 8.13R_\oplus$ for these clusters.

Our analysis uncovers significant disparities between small and giant planets. Giant planets tend to exhibit higher masses, larger radii, and lower densities, suggesting a prevalence of volatile-rich exoplanets in this category. Additionally, these giant planets tend to orbit their host stars at closer distances and possess higher equilibrium temperatures. On the other hand, small planets predominantly consist of elements heavier than hydrogen and helium, exhibiting lower equilibrium temperatures.

To predict the planetary radius, we employ various ML regression models. Among these models, the SVR demonstrates superior performance, yielding an RMSE of 0.093. A discernible linear trend appears in the residuals between predicted and observed radii of giant planets, which is not attributed to the restrictions of the predictive models or calculation issues. This evident trend has no significant impact on the radius predictions and is possibly related to the systematic issues in the exoplanet observations or the determination of physical parameters.

Additionally, utilizing Linear Regression, M5P, and MCMC methods, we establish a positive linear mass–radius relationship for small planets. In contrast, the radius of giant planets exhibits a positive correlation with the mass of their host stars, consistent with the findings presented by Lozovsky et al. (2021), which suggest a connection between volatile-rich planets and more massive host stars. None the less, as most of our sample consists of transiting exoplanets, besides naturally correlated parameters, the observational bias in the detection method can explain this result.

Furthermore, our analysis reveals that a carefully selected subset of features, encompassing planetary mass, orbital period, and one of the stellar parameters (stellar mass, radius, or effective temperature), is sufficient for accurate radius prediction. The inclusion of additional features such as semimajor axis, equilibrium temperature, and luminosity does not yield substantial improvements in the predictive capability.

Looking ahead, a comprehensive understanding of exoplanet composition and structure, as well as the testing of theories related to planetary formation and evolution, necessitates further follow-up observations. The *JWST* and future missions such as the Extremely Large Telescope (ELT), the *Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL)*, and the *Planetary Transits and Oscillations of Stars (PLATO)* mission will undoubtedly contribute invaluable insights into the atmospheric characteristics of exoplanets and the determination of stellar ages, thereby facilitating a more detailed exploration of exoplanetary systems.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers who helped with their valuable and priceless comments towards improving the manuscript. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. It has also made use of data obtained from or tools provided by the portal exoplanet.eu of The Extrasolar Planets Encyclopedia. Furthermore, this research has used the NASA's Planetary Fact Sheet. The authors acknowledge the usage of the `Scikit-learn` library (Pedregosa et al. 2011) and the `WEKA` software (Witten et al. 2005; Hall et al. 2009). They also acknowledge the usage of the following python packages, in alphabetical order: `ASTROPY` (Astropy Collaboration 2013, 2018), `CHAINCONSUMER` (Hinton 2019), `EMCEE` (Foreman-Mackey et al. 2019), `MATPLOTLIB` (Hunter 2007), `NUMPY` (van der Walt, Colbert & Varoquaux 2011), and `SCIPY` (Virtanen et al. 2020).

DATA AVAILABILITY

The data underlying this article were derived from sources in the public domain: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu/), Extrasolar Planets Encyclopedia (http://exoplanet.eu/), and NASA's Planetary Fact Sheet (https://nssdc.gsfc.nasa.gov/planetary/factsheet/).

REFERENCES

Ahmad H. P., Dang S., 2015, Int. J. Adv. Res. Comput. Sci. Manage. Stud., 8
Akeson R. L. et al., 2013, PASP, 125, 989
Alibert Y., Venturini J., 2019, A&A, 626, A21
Ankerst M., Breunig M. M., Kriegel H.-P., Sander J., 1999, ACM Sigmod Record, 28, 49
Armitage P. J., 2020, Astrophysics of Planet Formation, 2nd edn. Cambridge Univ. Press, Cambridge
Armstrong D. J., Gamper J., Damoulas T., 2021, MNRAS, 504, 5327
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Barboza A., Ulmer-Moll S., Faria J., 2020, Europlanet Science Congress 2020. p. EPSC2020–833
Bashi D., Helled R., Zucker S., Mordasini C., 2017, A&A, 604, A83
Becker A., Lorenzen W., Fortney J. J., Nettelmann N., Schöttler M., Redmer R., 2014, ApJS, 215, 21
Beichman C. et al., 2014, PASP, 126, 1134
Bhatti W. et al., 2016, preprint (arXiv:1607.00322)
Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., 2013, Knowl. Inf. Syst., 34, 483
Bonfils X. et al., 2005, A&A, 443, L15
Borucki W. J. et al., 2010, Science, 327, 977
Breiman L., 1996, Mach. Learn., 24, 123
Breiman L., 2001, Mach. Learn., 45, 5
Breunig M. M., Kriegel H.-P., Ng R. T., Sander J., 2000, SIGMOD '00: Proc. 2000 ACM SIGMOD International Conference on Management of Data. ACM, New York, p. 93
Brownlee J., 2016a, Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End. Machine Learning Mastery
Brownlee J., 2016b, Machine Learning Mastery with Weka: Analyze Data, Develop Models, and Work Through Projects. Machine Learning Mastery
Burrows A., Liebert J., 1993, Rev. Mod. Phys., 65, 301
Chabrier G., Baraffe I., 2000, ARA&A, 38, 337
Chandola V., Banerjee A., Kumar V., 2009, ACM Comput. Surv., 41, 1
Chandrashekar G., Sahin F., 2014, Comput. Electr. Eng., 40, 16
Chang C.-C., Lin C.-J., 2011, ACM Trans. Intell. Syst. Technol., 2, 1
Chen J., Kipping D., 2017, ApJ, 834, 17
Chen J., Kipping D. M., 2018, MNRAS, 473, 2753
Cherrington M., Thabtah F., Lu J., Xu Q., 2019, in 2019 International Conference on Computer and Information Sciences (ICCIS). IEEE, USA, p. 1
Comaniciu D., Meer P., 2002, IEEE Trans. Pattern Anal. Mach. Intell., 24, 603
Cox A. N., 2015, Allen's Astrophysical Quantities. Springer, New York

Cumming A., Butler R. P., Marcy G. W., Vogt S. S., Wright J. T., Fischer D. A., 2008, PASP, 120, 531
Curran P. A., 2015, Astrophysics Source Code Library, record ascl:1504.008
Davies D. L., Bouldin D. W., 1979, IEEE Trans. Pattern Anal. Mach. Intell., PAMI–1, 224
Deeg H. J., Alonso R., 2018, in Deeg H. J., Belmonte J. A., eds, Handbook of Exoplanets. Springer, Cham, p. 117,
Enoch B., Collier Cameron A., Horne K., 2012, A&A, 540, A99
Ester M., Kriegel H.-P., Sander J., Xu X. 1996, in KDD'96: Proc. Second Int. Conf. Knowl. Discov. Data Mining. ACM, p. 226
Ferri F. J., Pudil P., Hatef M., Kittler J., 1994, in Gelsema E. S., Kanal L. S., eds, Machine Intelligence and Pattern Recognition, Vol. 16. Elsevier, Amsterdam, The Netherlands, p.403
Foreman-Mackey D. et al., 2019, J. Open Source Softw., 4, 1864
Fortney J. J., Marley M. S., Barnes J. W., 2007, ApJ, 659, 1661
Frey B. J., Dueck D., 2007, Science, 315, 972
Gilbert G. J., Fabrycky D. C., 2020, AJ, 159, 281
Goodman J., Weare J., 2010, Commun. Appl. Math. Comp. Sci., 5, 65
Guillot T., Santos N. C., Pont F., Iro N., Melo C., Ribas I., 2006, A&A, 453, L21
Guyon I., Gunn S., Nikravesh M., Zadeh L. A., 2008, Feature Extraction: Foundations and Applications. Vol. 207, Springer, Berlin, Heidelberg
Halkidi M., Batistakis Y., Vazirgiannis M., 2001, J. Intell. Inf. Syst., 17, 107
Hall M. A., Smith L. A., 1999, in Kumar A. N., Russell I., FLAIRS Conference: Proc. 20th International Florida Artificial Intelligence Research Society Conference. AAAI Press, Orlando, Florida, p. 235
Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., 2009, ACM SIGKDD Explor. Newsl., 11, 10
Hastie T., Tibshirani R., Friedman J. H., Friedman J. H., 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2, Springer, New York
Hatzes A. P., Rauer H., 2015, ApJ, 810, L25
Hinton G. E., 1990, in, Machine Learning. Elsevier, p. 555
Hinton S. R., 2019, Astrophysics Source Code Library, record ascl:1910.017
Hoaglin D. C., Mosteller F., Tukey J. W., 1983, Understanding Robust and Exploratory Data Anlysis. John Wiley & Sons, New York
Howard A. W. et al., 2010, Science, 330, 653
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jain A. K., Dubes R. C., 1988, Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey
Jović A., Brkić K., Bogunović N., 2015, in Sokolic P., ed., 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, Opatija, Croatia, p. 1200
Kanodia S., Wolfgang A., Stefansson G. K., Ning B., Mahadevan S., 2019, ApJ, 882, 38
Kaufman L., Rousseeuw P. J., 2009, Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, Hoboken, NJ
Kipping D., 2018, MNRAS, 473, 784
Kohavi R., John G. H., 1997, Artif. Intell., 97, 273
Kononenko I., Kukar M., 2007, Chapter 12-Cluster Analysis
Lal T. N., Chapelle O., Weston J., Elisseeff A., 2006, in Feature Extraction. Springer, Berlin, p. 137
Landau S., Leese M., Stahl D., Everitt B. S., 2011, Cluster Analysis. John Wiley and Sons, Chichester
Laughlin G., Lissauer J. J., 2015, in Schubert G., ed., Treatise on Geophysics. Elsevier, Oxford, p. 673
Leleu A. et al., 2021a, A&A, 649, A26
Leleu A., Chatel G., Udry S., Alibert Y., Delisle J. B., Mardling R., 2021b, A&A, 655, A66
Lozovsky M., Helled R., Pascucci I., Dorn C., Venturini J., Feldmann R., 2021, A&A, 652, A110
MacDonald M. G., 2019, MNRAS, 487, 5062
Maldonado J., Villaver E., Eiroa C., 2018, A&A, 612, A93
Maltagliati L., 2023, Nat. Astron., 7, 8
Marcus R. A., Sasselov D., Hernquist L., Stewart S. T., 2010, ApJ, 712, L73
Marcy G. W., Weiss L. M., Petigura E. A., Isaacson H., Howard A. W., Buchhave L. A., 2014, Proc. Natl. Acad. Sci., 111, 12655
Mishra L., Alibert Y., Udry S., Mordasini C., 2023a, A&A, 670, A69
Mishra L., Alibert Y., Udry S., Mordasini C., 2023b, A&A, 670, A68
Moreira A., Santos M. Y., Carneiro S., 2005, University of Minho-Portugal, Braga, 1, 18
Mousavi-Sadr M., Gozaliasl G., Jassur D. M., 2021, Publ. Astron. Soc. Aust., 38, e015
Neil A. R., Rogers L. A., 2018, ApJ, 858, 58
Ning B., Wolfgang A., Ghosh S., 2018, ApJ, 869, 5
Osborn A., Bayliss D., 2020, MNRAS, 491, 4481
Otegi J. F., Bouchy F., Helled R., 2020, A&A, 634, A43
Pascucci I., Mulders G. D., Gould A., Fernandes R., 2018, ApJ, 856, L28
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pepper J. et al., 2007, PASP, 119, 923
Quinlan J., 1992, in Adams A., Sterling L., eds, AI'92: Proc. 5th Australian Joint Conference on Artificial Intelligence, Learning with Continuous Classes. World Scientific, Hobart
Quinlan J. R., 1993, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Burlington, Massachusetts
Rousseeuw P. J., 1987, J. Comput. Appl. Math., 20, 53
Sánchez-Maroño N., Alonso-Betanzos A., Tombilla-Sanromán M., 2007, in Yin H., Tino P., Corchado E., Byrne W., Yao X., eds, Lecture Notes in Computer Science, Vol. 4881, Intelligent Data Engineering and Automated Learning - IDEAL 2007. Springer, Berlin, Heidelberg, p. 178
Schlecker M. et al., 2021, A&A, 656, A73
Schneider J., Dedieu C., Le Sidaner P., Savalle R., Zolotukhin I., 2011, A&A, 532, A79
Schubert E., Sander J., Ester M., Kriegel H. P., Xu X., 2017, ACM Trans. Database Syst., 42, 1
Seager S., 2010, Exoplanets. Univ. Arizona Press, Tucson
Seager S., Kuchner M., Hier-Majumder C. A., Militzer B., 2007, ApJ, 669, 1279
Soni Madhulatha T., 2012, preprint (arXiv:1205.1117)
Swift D. C. et al., 2012, ApJ, 744, 59
Tasker E. J., Laneuville M., Guttenberg N., 2020, AJ, 159, 41
Ulmer-Moll S., Santos N. C., Figueira P., Brinchmann J., Faria J. P., 2019, A&A, 630, A135
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Van Eylen V. et al., 2021, MNRAS, 507, 2154
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Von Luxburg U., 2007, Stat. Comput., 17, 395
Weiss L. M., Marcy G. W., 2014, ApJ, 783, L6
Weiss L. M. et al., 2013, ApJ, 768, 14
Witten I. H., Frank E., Hall M. A., Pal C. J., DATA M., 2005, in Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam, The Netherlands
Wolfgang A., Rogers L. A., Ford E. B., 2016, ApJ, 825, 19
Wu Y., 2019, ApJ, 874, 91
Xu R., Wunsch D., 2005, IEEE Trans. Neural Netw., 16, 645
Yee S. W., Winn J. N., 2023, ApJ, 949, L21
Zapolsky H. S., Salpeter E. E., 1969, ApJ, 158, 809
Zhang T., Ramakrishnan R., Livny M., 1996, ACM Sigmod Record, 25, 103
Zucker S., Mazeh T., 2002, ApJ, 568, L113

APPENDIX A: DATA-CLEANING AND ITS IMPACT ON PREDICTION ACCURACY

Our data set contains 770 data points. The LOF method is chosen to identify outlier observations. It is first applied to all parameters including P, e, M_p, R_p, M_s, R_s, Fe/H, and $T_{\rm eff}$, and then to R_p and M_p. The first step determines 39 outliers with an average score of 1.951, where the higher the LOF score, the more abnormal the data point. In comparison, the average score of inliers is 1.113. The second step determines 37 outliers and assigns average scores of 2.113 and 1.061 to the outlier and inlier data points, respectively.

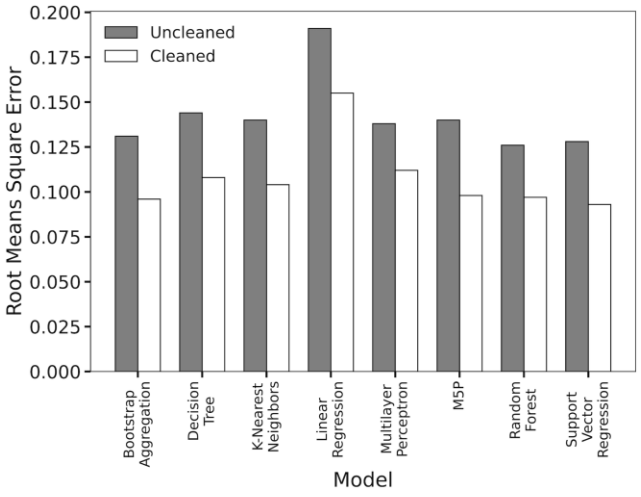

Figure A1. RMSE values of different ML regression models implemented in uncleaned (grey bars) and cleaned (white bars) data sets. All models have higher RMSE values when outliers are included in the data set.

In total, the LOF marks 76 data points as outliers. The outliers have an average Mahalanobis distance of 18.050, compared to 6.888 for the inliers. This indicates that the 76 outlier data points are farther away from the data set's central point than the inliers. It is interesting to note that identified outliers have higher uncertainties than inliers. In a logarithmic scale, outlier data points have an average uncertainty of 0.19 for planetary mass and 0.05 for planetary radius. In comparison, these values for inlier data points are 0.11 and 0.04, respectively.

To quantify the impact of outliers on prediction precisions, we run ML regression models on the data set containing all 770 data points. Fig. A1 compares the prediction accuracy obtained from the uncleaned (with outliers) and cleaned (without outliers) data sets. As can be seen, all models perform remarkably better when outliers are removed from the data set, demonstrating the significance of the data-cleaning step in predicting the planetary radius.

APPENDIX B: IMPACT OF DATA RE-SCALING ON PREDICTION ACCURACY

We process both logarithmic and non-logarithmic data sets to understand the effect of data re-scaling on predicting planetary radius.

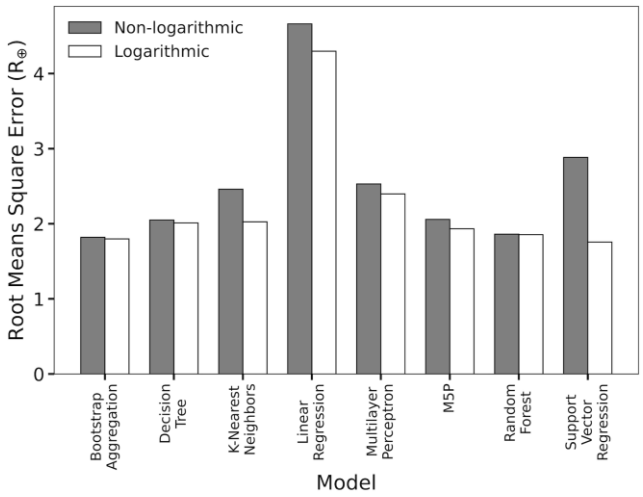

Figure B1. RMSE values of different ML regression models implemented in logarithmic (white bars) and non-logarithmic (grey bars) data sets. Bootstrap Aggregation and Random Forest models do not show a remarkable difference between logarithmic and non-logarithmic scaling. In contrast, other algorithms, particularly the SVR, provide better results on a logarithmic scale.

Fig. B1 compares the RMSE values corresponding to different models applied in non-logarithmic and logarithmic data sets. Bootstrap Aggregation and Random Forest slightly differ between logarithmic and non-logarithmic scaling. In contrast, other algorithms, particularly SVR, provide better results on a logarithmic scale. Transforming the exoplanet data into a logarithmic space helps handle the wide range of values by compressing them, allowing the ML model better to capture the underlying patterns and relationships within the data. Moreover, logarithm transformation efficiently addresses data skewness and outliers. When data do not follow a normal distribution and contain extreme values, the logarithmic space helps mitigate the influence of outliers by compressing their impact and making the data more symmetrical.

This paper has been typeset from a TEX/LATEX file prepared by the author.