Planetary Science and Exoplanets in the Era of JWST
Proceedings IAU Symposium No. 393, 2026
Z. Benkhaldoun, Y. Moulane & A. Szentgyorgyi, eds.
doi:10.1017/S1743921324003168

Refining the exoplanet mass-radius relation using machine learning

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

[1]School of Astronomy, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
email: mahdiyar@ipm.ir

[2]Department of Theoretical Physics and Astrophysics, Faculty of Physics, University of Tabriz, Tabriz, Iran

[3]Department of Computer Science, Aalto University, P. O. Box 15400, Espoo, FI-00 076, Finland

Abstract. The rising number of exoplanet discoveries and advances in machine learning (ML) techniques present new possibilities for exploring and understanding the characteristics of worlds beyond our solar system. This research examines the exoplanet dataset by applying ML techniques to categorize these systems, uncover relationships among their physical features, and predict the exoplanet radius. We group the data into two primary categories: 'small' and 'giant' planets, with thresholds at $R_p = 8.13R_\oplus$ and $M_p = 52.48M_\oplus$. Our study indicates that the planetary mass, orbital period, and stellar mass play critical roles in predicting the exoplanet radius. A notable finding of our research is that small planets exhibit a positive linear mass-radius relationship, consistent with other studies. Conversely, for giant planets, we observe a strong correlation between planetary radius and the mass of their host stars, potentially providing significant insights into the relationship between giant planet formation and stellar properties.

Keywords. planets and satellites: fundamental parameters, planets and satellites: composition, planets and satellites: formation, planets and satellites: dynamical evolution and stability, planets and satellites: general, software: data analysis

1. Introduction

The exoplanet data generated by the state-of-the-art telescopes is accessible to all. Researchers skilled in data analysis and data science analyze these data to anticipate, discover, and characterize exoplanets, and to improve theories of planet formation and evolution (Armstrong, Gamper, & Damoulas 2021; Mousavi-Sadr, Gozaliasl, & Jassur 2021; Mousavi-Sadr 2024). A polytropic relationship can be used to describe the relationship that exists between the mass and radius of a planet (Chabrier & Baraffe 2000). Many studies have examined the correlation between planetary parameters, specifically mass and radius, in order to infer the composition and structure of exoplanets (Otegi, Bouchy, & Helled 2020). Weiss et al. (2013) separated the planets into two categories based on mass, those with masses higher than and below $150M_\oplus$, then provided a power law relation for the mass-radius distribution of each group. The radius of a giant planet is related to other parameters such as the planetary equilibrium temperature ($T_{\rm equ}$), the semi-major orbital axis, the tidal heating rate, and stellar metallicity and irradiation (Enoch, Collier Cameron, & Horne 2012). Bhatti et al. (2016) assessed the impact of several physical parameters on the radii of the planet using a random forest regression model. They found that the parameters that most influence the radius of

© The Author(s), 2026. Published by Cambridge University Press on behalf of International Astronomical Union.

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press

a hot Saturn ($0.1 < M_p < 0.5 M_J$) are the planet's mass and equilibrium temperature. Moreover, Ulmer-Moll et al. (2019) used the Random Forest model to predict the exoplanet radii based on several stellar and planetary parameters. The important parameters were, as in earlier findings, the mass and equilibrium temperature of an exoplanet. In this work, we apply different machine learning methods to identify plausible correlations between the physical properties of exoplanet systems. Furthermore, the exoplanets are grouped and their features are studied using various machine learning clustering algorithms.

2. Data and method

We use the NASA Exoplanet Archive and the Extrasolar Planets Encyclopedia to extract the data of exoplanets with reported physical parameters including the orbital period (P) and eccentricity (e), the planetary mass (M_p), and the radius (R_p), and the stellar mass (M_s), the radius (R_s), the metallicity (Fe/H), and the effective temperature ($T_{\rm eff}$). We have eight solar system planets and 762 confirmed exoplanets in our data set. The target variable is the planetary radius (R_p), and the features are the orbital period (P), eccentricity (e), planetary mass (M_p), stellar mass (M_s), stellar radius (R_s), metallicity (Fe/H), and effective temperature ($T_{\rm eff}$). It is generally accepted that unnecessary and ineffective features can negatively impact the performance of an ML model and produce unreliable results (Amr Abdallah & de La Iglesia 2015). Therefore, we use five Feature Selection (FS) approaches to identify and exclude the least significant planetary, stellar, and orbital features: Spearman's rank correlation test is a filter method, where features are selected independently of any induction algorithm, relying on specific performance evaluation metrics that are computed directly from the data, the Forward Selection and Backward Elimination as two wrapper methods, where the selection process relies on the effectiveness of a particular ML algorithm functioning with a specific set of features, and the CART and XGBoost as two Embedded methods. Embedded methods integrate the advantages of both filter and wrapper techniques. They conduct the FS during the training process and are typically specific to given learning algorithms (Belanche & González 2011). Similar to exclusive clustering, where each data point belongs exclusively to one cluster (Jain & Dubes 1988), we want to group exoplanets into distinct, non-overlapping clusters. Data clustering algorithms can be categorized into two main types: hierarchical and partitional. Hierarchical algorithms identify clusters by utilizing previously formed clusters, whereas partitional algorithms determine all clusters simultaneously. We apply the following ML clustering algorithms to identify potential exoplanet groups and explore their properties: Affinity Propagation, BIRCH, DBSCAN, Gaussian Mixture Model, Hierarchical Clustering, K-Means, Mean Shift, Mini Batch K-Means, OPTICS, and Spectral Clustering (Kaufman & Rousseeuw 1990). We also apply ML models to predict planet radii when other parameters are given. The algorithms used to perform regression tasks are Decision Tree, K-Nearest Neighbors, Linear Regression, Multilayer Perceptron, M5P, and Support Vector Regression (SVR). Bootstrap Aggregation and Random Forest serve as ensemble algorithms that integrate the predictions from various models (Breiman 2001). The hyperparameters are tuned to achieve optimal algorithm performance. A 10-fold cross-validation procedure is also used to evaluate the performance of the models. The mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (ρ^2) are calculated as validation metrics. Linear regression and M5P are two algorithms that can extract parametric equations. To quantify the uncertainties of the best-fitting parameters of linear regression and M5P, we use the Markov Chain Monte Carlo (MCMC) (Goodman & Weare 2010; Foreman-Mackey et al. 2019).

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press

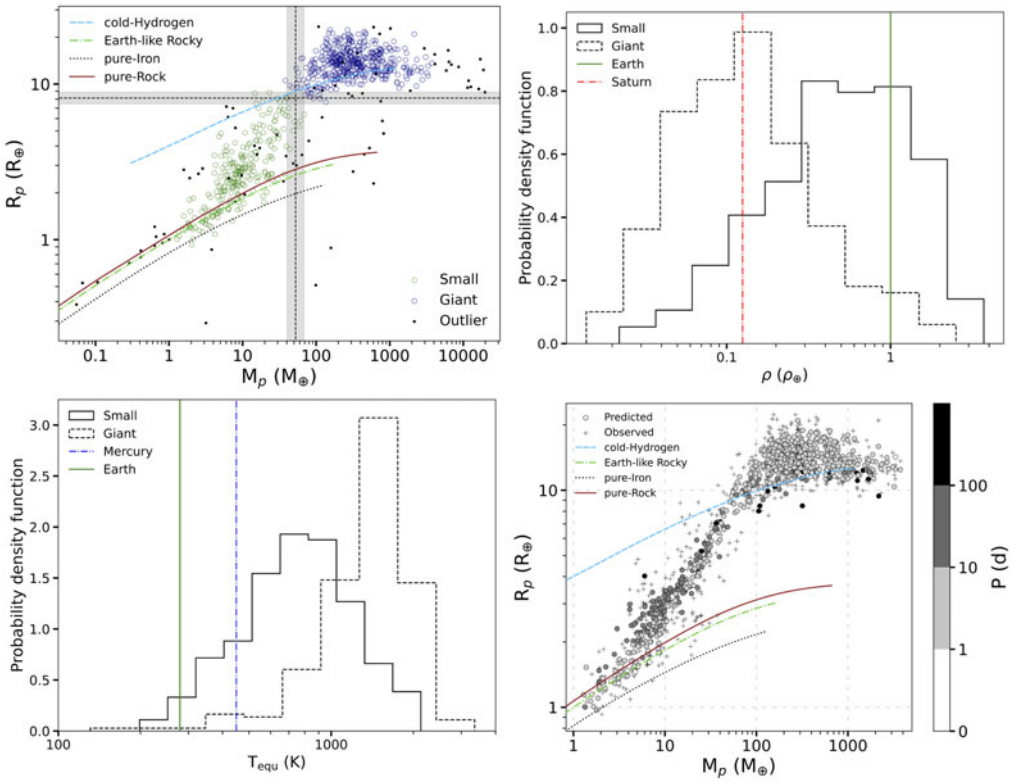


Figure 1. The mass-radius distribution for clustered data is shown in the upper left panel. The dataset is divided into two categories by $R_p = 8.13R_\oplus$ (horizontal dashed line) and $M_p = 52.48M_\oplus$ (vertical dashed line). Mean errors in mass and radius are represented by gray regions. The upper right and lower left panels show histograms of the average density (ρ) and equilibrium temperature ($T_{\rm equ}$), respectively, for small (solid-border bars) and giant (dashed-border bars) planets. A comparison of the average densities and equilibrium temperatures of these planets reveals that giant planets tend to be hotter and less dense than their smaller counterparts. The lower right panel represents the predicted and observed radius as a function of mass and orbital period obtained by the SVR model for all planets in the sample. Iso-density curves are plotted based on Marcus et al. (2010) and Becker et al. (2014).

3. Results

Before applying the clustering and regression models, we use the Local Outlier Factor (LOF) method to identify outliers, which marks 76 data points as outliers in our sample. The LOF employs two hyperparameters: the neighborhood size, which defines the neighborhood for local density calculation, and contamination, which indicates the percentage of outliers in the dataset. Then, it identifies and filters out data points that deviate significantly from the overall dataset. To enhance the efficacy of predictive models, we remove these outliers and transform the data into a logarithmic space. The impact of data-cleaning and data re-scaling on prediction accuracy are discussed in Appendixes A and B of Mousavi-Sadr, Jassur, & Gozaliasl (2023) (hereafter, Sadr2023). FS methods highlight that planetary mass plays an important role in predicting the radius of exoplanets, whereas metallicity and eccentricity demonstrate relatively less significance. So, we exclude metallicity and eccentricity from the feature list. Moreover, all stellar parameters (M_s, R_s, and $T_{\rm eff}$) are of similar importance. Eventually, we choose a set of parameters

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press

Table 1. Results obtained by different regression algorithms. These algorithms have been separately applied to the whole dataset (columns 2 to 4), small planets (columns 5 to 7), and giant planets (columns 8 to 10) planets.

Algorithm	**RMSE**	**MAE**	ρ^2	$\mathbf{RMSE}_1$	$\mathbf{MAE}_1$	ρ_1^2	$\mathbf{RMSE}_2$	$\mathbf{MAE}_2$	ρ_2^2
Bootstrap Aggregation	0.096	0.068	0.933	0.126	0.096	0.692	0.064	0.047	0.489
Decision Tree	0.108	0.077	0.916	0.142	0.109	0.616	0.071	0.051	0.392
K-Nearest Neighbors	0.104	0.075	0.921	0.140	0.106	0.627	0.072	0.053	0.390
Linear Regression	0.155	0.127	0.822	0.128	0.099	0.683	0.069	0.051	0.402
Multilayer Perceptron	0.112	0.084	0.909	0.131	0.103	0.670	0.064	0.048	0.491
M5P	0.098	0.071	0.930	0.130	0.101	0.688	0.073	0.053	0.391
Random Forest	0.097	0.068	0.932	0.128	0.096	0.682	0.064	0.047	0.485
SVR	0.093	0.065	0.937	0.123	0.088	0.710	0.063	0.046	0.510

that include planetary mass, orbital period, and stellar mass as the main features (for more details, refer to Section 3.1 in Sadr2023).

We select two clusters (K) for the Gaussian Mixture Model, BIRCH, K-Means, Mini-Batch K-Means, Spectral clustering, and Hierarchical clustering. The Affinity Propagation and Mean Shift algorithms, which determine K from the data, also result in two clusters. Although DBSCAN and OPTICS provide $K = 2$, their inability to distinctly separate clusters leads us to exclude them from the analysis. To distinguish between the two groups in the mass-radius space of the planet, we construct a Gaussian kernel density estimate for each cluster and identify the intersection point. The clustering results are largely consistent across the algorithms except for DBSCAN and OPTICS (refer to Table 1 in Sadr2023). We used an average value of $\log M_p = 1.72$ ($M_p = 52.48 M_\oplus$) for the mass break point and $\log R_p = 0.91$ ($R_p = 8.13 R_\oplus$) for the radius breakpoint to categorize the planets into two groups. Planets with $R_p > 8.13 R_\oplus$ and $M_p > 52.48 M_\oplus$ are classified as giant planets, while those with $R_p \leq 8.13 R_\oplus$ and $M_p \leq 52.48 M_\oplus$ are termed small planets. There are 254 small planets and 440 giant planets. The mass-radius distribution of small and giant planets is illustrated in the upper left panel of Figure 1. Our findings indicate significant differences between small and giant planets. Giant planets exhibit larger masses, larger radii, and lower densities (see the upper right panel of Figure 1), suggesting a higher frequency of volatile-rich exoplanets in this category. Additionally, giant planets typically orbit closer to their parent stars and possess higher equilibrium temperatures (see the lower left panel of Figure 1). In contrast, small planets consist mainly of elements heavier than hydrogen and helium and have lower equilibrium temperatures.

We apply predictive models separately to entire, small, and giant planets. The SVR model delivers the best performance among the ML regression models, achieving an RMSE of 0.093 (see Table 1). Using Linear Regression and M5P, which can derive parametric equations, we find a positive linear mass-radius relationship for small planets and a positive linear correlation between the radius of giant planets and the mass of their host stars (see Table 4 in Sadr2023 for more details). The left panel of Figure 2 displays the correlation between the radius and mass of small planets derived from the MCMC. The best scaling relation, represented by the red line, is illustrated using Equation 3.1. The correlation between the radius of giant planets and the stellar mass is displayed in the right panel of Figure 2, where the best scaling relation is illustrated using Equation 3.2.

$$\log \frac{R_p}{R_\oplus} = 0.497^{+0.023}_{-0.022} \times \log \frac{M_p}{M_\oplus} - 0.050^{+0.024}_{-0.024}. \tag{3.1}$$

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press

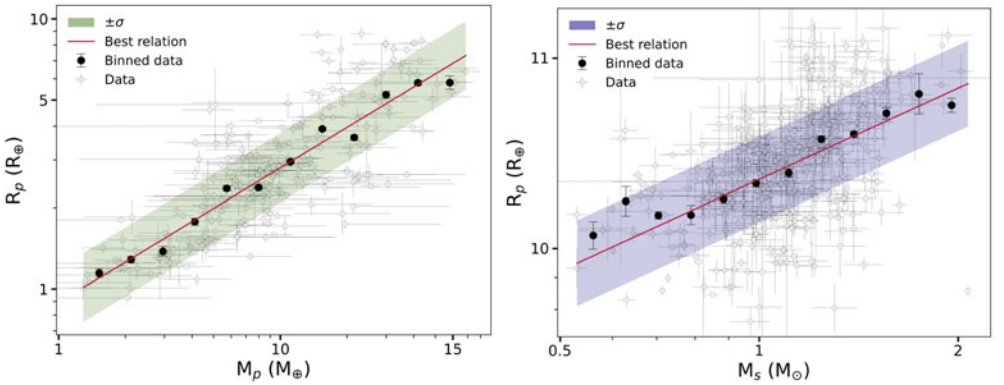


Figure 2. The mass-radius relation of small planets (left panel) and the relation between the radius of giant planets and the host star's mass (right panel) obtained by the MCMC method. Equation 3.1 and Equation 3.2 represent the best scaling relations (red lines). The highlighted area is $\pm 1\sigma$ uncertainties around the best relation. For better illustration, the data points have been binned.

$$\log \frac{R_p}{R_\oplus} = 0.480^{+0.036}_{-0.037} \times \log \frac{M_s}{M_\odot} + 1.109^{+0.004}_{-0.004}. \tag{3.2}$$

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1743921324003168

References

Amr Abdallah, T., de La Iglesia, B., 2015, arXiv, arXiv:1510.02892. doi:10.48550/arXiv.1510.02892

Armstrong, D. J., Gamper, J., Damoulas, T., 2021, MNRAS, 504, 5327. doi:10.1093/mnras/staa2498

Becker, A., Lorenzen, W., Fortney, J. J., Nettelmann, N., Schöttler, M., Redmer, R., 2014, ApJS, 215, 21. doi:10.1088/0067-0049/215/2/21

Belanche, L. A., González, F. F., 2011, arXiv, arXiv:1101.2320. doi:10.48550/arXiv.1101.2320

Bhatti, W., Bakos, G. Á., Hartman, J. D., Zhou, G., Penev, K., Bayliss, D., Jordán, A., et al., 2016, arXiv, arXiv:1607.00322. doi:10.48550/arXiv.1607.00322

Breiman, L., 2001, MachL, 45, 5. doi:10.1023/A:1010933404324

Chabrier, G., Baraffe, I., 2000, ARA&A, 38, 337. doi:10.1146/annurev.astro.38.1.337

Enoch, B., Collier Cameron, A., Horne, K., 2012, A&A, 540, A99. doi:10.1051/0004-6361/201117317

Foreman-Mackey, D., Farr, W., Sinha, M., Archibald, A., Hogg, D., Sanders, J., Zuntz, J., et al., 2019, JOSS, 4, 1864. doi:10.21105/joss.01864

Goodman, J., Weare, J., 2010, CAMCS, 5, 65. doi:10.2140/camcos.2010.5.65

Jain, A. K., Dubes, R. C., 1988, acd..book

Kaufman, L., Rousseeuw, P. J., 1990, fgda.book

Lozovsky, M., Helled, R., Pascucci, I., Dorn, C., Venturini, J., Feldmann, R., 2021, A&A, 652, A110. doi:10.1051/0004-6361/202140563

Marcus, R. A., Sasselov, D., Hernquist, L., Stewart, S. T., 2010, ApJL, 712, L73. doi:10.1088/2041-8205/712/1/L73

Mousavi-Sadr, M., Gozaliasl, G., Jassur, D. M., 2021, PASA, 38, e015. doi:10.1017/pasa.2021.9

Mousavi-Sadr, M., Jassur, D. M., Gozaliasl, G., 2023, MNRAS, 525, 3469. doi:10.1093/mnras/stad2506

Mousavi-Sadr, M., 2024, PhDT. doi:10.48550/arXiv.2402.17898

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press

Otegi, J. F., Bouchy, F., Helled, R., 2020, A&A, 634, A43. doi:10.1051/0004-6361/201936482
Ulmer-Moll, S., Santos, N. C., Figueira, P., Brinchmann, J., Faria, J. P., 2019, A&A, 630, A135. doi:10.1051/0004-6361/201936049
Weiss, L. M., Marcy, G. W., Rowe, J. F., Howard, A. W., Isaacson, H., Fortney, J. J., Miller, N., et al., 2013, ApJ, 768, 14. doi:10.1088/0004-637X/768/1/14

Discussion

In our study of a sample comprising 770 planets, we find that the size of giant planets is positively linked to the mass of the stars they orbit. This aligns with the results reported in Lozovsky et al. (2021), indicating an association between planets rich in volatile substances and stars of larger mass. However, since the majority of our sample comprises transiting exoplanets, this result can be explained by the observational bias in the detection method, in addition to naturally correlated parameters. To understand exoplanet composition and structure and test theories about how planets form and evolve, further follow-up programs like the Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL), the Planetary Transits and Oscillations of Stars (PLATO) mission, and the Extremely Large Telescope (ELT) are needed.

https://doi.org/10.1017/S1743921324003168 Published online by Cambridge University Press