Do Stellar-Host parameters have any significance in devising Exoplanet characteristics?

Final Report

1st Rohan Sai Nalla
Department of Data Science
University of Colorado Boulder
Boulder, CO, USA
rona8789@colorado.edu

2nd Kaushik Narasimha Bukkapatnam
Department of Data Science
University of Colorado Boulder
Boulder, CO, USA
kabu9868@colorado.edu

***Abstract*—The genesis and discovery of exoplanets are intricately linked to the characteristics of their respective host stars. The present study endeavors to examine the correlation between diverse parameters of stellar hosts and the characteristics of exoplanets through the application of regression analysis. We plan to collect data on exoplanets and their host stars from online databases and use statistical techniques such as linear regression, multiple regression, and model selection to model the relationship between the host parameters (such as the star's mass, radius, metallicity, age, and activity level) and the exoplanet characteristics (such as mass, radius, and composition). Our study investigated the host parameters that exert the most significant influence on the characteristics of exoplanets. Additionally, we analyzed how these associations differ across various categories of exoplanets. The outcomes of this endeavor will enhance our comprehension of the genesis of exoplanets and the impact of stellar host parameters on the features of exoplanets.**

***Index Terms*—Regression, statistical-significance, Random Forest, Generalised Additive Models, Semiparametric Models, Exoplanet Radius, Multiple T-test, ANOVA, Correlation, Hypothesis.**

I. Introduction

Exoplanets are celestial bodies that orbit stars located outside of our solar system. Astronomers have discovered numerous exoplanets through diverse observational techniques subsequent to the initial discovery of an exoplanet in 1995. The aforementioned discoveries have significantly transformed our comprehension of planetary systems, and have also raised doubts regarding our preconceived notions about the essential conditions for planetary genesis and sustainability.

There exists a strong correlation between the characteristics of exoplanets and those of their respective host stars. The atmospheric and surface conditions of a planet may be influenced by the radiation and magnetic fields emanating from the host star, in addition to the gravitational force that maintains the planet's orbital trajectory. Consequently, possessing knowledge of the attributes of the host star is imperative in comprehending the attributes of any potential planets that may be present within its orbital path.

The crucial properties of a stellar host that have an impact on exoplanet characteristics include its mass, radius, temperature, metallicity, age, and activity level. These characteristics may affect the exoplanet's size, mass, composition, and habitability. As an illustration, a star of greater mass could conceivably possess a larger habitable zone; however, it could also generate a greater amount of radiation, which may result in the depletion of an atmosphere from a planet. Analogous to the aforementioned, stars exhibiting higher metallicities may offer greater accessibility to heavier elements for the purpose of planetary formation, thereby resulting in the development of planets with relatively larger dimensions.

Astronomical researchers utilize various observational techniques such as radial velocity measurements, transit photometry, and direct imaging to explore the correlation between the properties of star hosts and the characteristics of exoplanets. They endeavor to enhance their comprehension of the origin and progression of planetary systems, along with the prerequisites for sustaining life on planets, by scrutinizing the attributes of exoplanets and their corresponding host stars. To conclude, the investigation of exoplanets and their corresponding stars is a rapidly growing field in astronomy that possesses the capability to fundamentally transform our understanding of the universe and our place within it.

II. Related Work

Valuable insights and contextual information for our investigation can be obtained from prior research on the characterization of exoplanets and the parameters of their host stars. The subsequent enumeration comprises relevant fields of study. Catalogs of Exoplanets: Numerous substantial exoplanet catalogs, such as the NASA Exoplanet Archive and the exoplanet.eu database, furnish an abundance of information and particulars regarding identified exoplanets. These catalogues can be utilized to expand our collection, provide additional information, and facilitate comparisons.

Empirical investigations have been conducted to explore the correlation between the size of a planet and various factors such as the effective temperature, metallicity, and mass of the parent star. The comprehension of correlations between host-star parameters and exoplanet properties may facilitate the identification of the latter's crucial features.

Various statistical methodologies, such as machine learning, Bayesian analysis, and linear regression, have been utilized to scrutinize exoplanetary data. The identification of optimal study techniques for personal use can be facilitated through a comprehensive review of prior research endeavors that have employed analogous methodologies.

Based on our research, we have identified several studies that align with the analysis we intend to conduct in this paper.

The study conducted by Batalha et al. in 2013 [1]. The present manuscript presents a comprehensive account of the Kepler mission and its principal objective of identifying and characterizing exoplanets. The discourse also expounds upon certain statistical methodologies employed for the analysis of Kepler data, such as regression analysis and cluster analysis.

According to Coughlin and colleagues (2016), as cited in reference [2],... The present study centers on the quantification of host-star parameters pertaining to exoplanetary systems identified by the Kepler mission. The authors employed a machine learning algorithm to categorize Kepler stars according to their spectral attributes. Additionally, they expounded on the statistical techniques employed to authenticate their findings.

According to Buchhave and colleagues (2012), as cited in reference [3],... The present study investigates the correlation between the frequency of exoplanet occurrence and the metallicity of the host star, which is indicative of the star's concentration of heavy elements. The researchers employ statistical techniques to examine data obtained from the Kepler mission and detect indications of a favorable association between the incidence of exoplanets and the metallicity of their host stars.

III. The Data

The primary source of data is an extensive digital repository of observations and revelations pertaining to exoplanets. NASA's Exoplanet Archive offers a wealth of information to both scholars and the wider community regarding celestial bodies beyond our own solar system. The Planetary Systems table, which is a crucial database that can be accessed through the Exoplanet Archive, furnishes comprehensive information regarding the characteristics of identified exoplanetary systems, encompassing the quantity of planets, their respective masses, radii, and orbital parameters. The present dataset, subject to periodic updates as new discoveries are made, was compiled from diverse sources of observations, encompassing both terrestrial surveys and spaceborne expeditions such as Kepler and TESS. The Planetary Systems table serves as a valuable point of reference for astronomers engaged in research pertaining to the genesis and progression of planetary systems. Additionally, it is a resource of interest to individuals seeking to expand their knowledge of the diverse array of exoplanets that have been discovered to date.

The Exoplanet Archive is a repository that was established and is currently maintained by the NASA Exoplanet Science Institute. The resource mentioned above functions as a centralized repository of information regarding all identified exoplanets and their corresponding host stars. This encompasses exhaustive information regarding their characteristics, modalities of identification, and other relevant particulars.

The Exoplanet Archive comprises observational data obtained from measurements and observations of exoplanets and their host stars, conducted by diverse telescopes and observatories. The information contained within the archive has been acquired through a range of observational techniques, including but not limited to the transit method, radial velocity method, and direct imaging. The attributes delineated in the data are inferred from empirical observations of the exoplanets and their respective host stars, rather than being derived from controlled laboratory experiments. As the data in question is of an observational nature, it is susceptible to various sources of error, including measurement error, sampling error, and natural variability. It is essential to take these factors into account when conducting any analysis.

The fundamental aim underlying the establishment of the Exoplanet Archive was to provide an all-encompassing and up-to-date depository of exoplanetary information to both the scientific fraternity and the general public. The data provided is intended to facilitate a wide range of inquiries and assessments pertaining to exoplanetary populations, attributes, and origins.

The Archive plays a crucial role in identifying new exoplanetary candidates and prioritizing targets for subsequent observations and research by upcoming telescopes and missions. As new discoveries of exoplanets are revealed, the data contained within the database is continually updated and expanded.

Among the variables present in the dataset, it is possible to explore several connections. The aforementioned are notable associations.

The correlation between an exoplanet's radius and the properties of its host star, such as its effective temperature, mass, and radius, is a subject of anticipation. The genesis and development of a planet are significantly influenced by the characteristics of its host star. As an illustration, a star with a greater mass may possess a larger protoplanetary disk, thereby leading to the formation of larger planets. Analogous to the aforementioned scenario, a host star with higher temperature may induce the process of atmospheric evaporation in the planet, leading to a decrease in its radius.

The correlation between a planet's radius and the amount of insolation flux received from its host star is a subject of prediction. The aforementioned phenomenon can be attributed to the influence of insolation flux on the atmospheric temperature of the planet, which could potentially affect its radius. A celestial body possessing a dense gaseous envelope may undergo an expansion in response to the absorption of additional thermal energy from its parent star, thereby augmenting its overall radius.

It is hypothesized that there exists a correlation between a planet's radius and the eccentricity of its orbit, which refers to the degree of ellipticity. This phenomenon can be attributed to alterations in a planet's radius that arise due to significant fluctuations in temperature resulting from an eccentric orbit. As an illustration, in the event that the orbit of a planet brings it in close proximity to its host star, the substantial heat generated

may result in the expansion of its atmosphere, thereby causing an increase in the planet's radius.

It is anticipated that the variables within the dataset will exhibit some form of interdependence due to a multitude of theoretical rationales. However, further investigation is required to comprehensively comprehend the exact characteristics of these interactions, which could potentially depend on the distinctive features of individual exoplanetary systems.

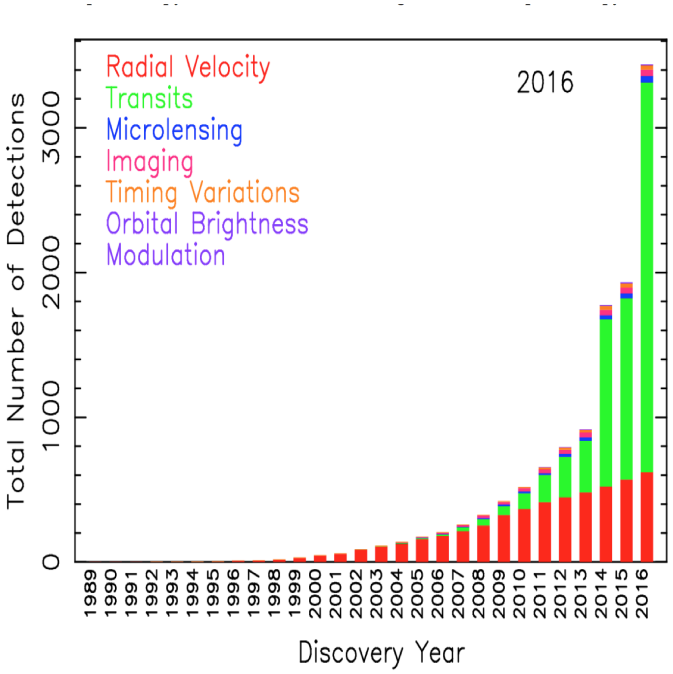

Fig. 1. Discovery Methods and the Temporal chart of Exoplanet Discoveries

IV. Curiosity around the Topic

The examination of exoplanets and their corresponding host stars is a vital area of investigation in the field of exoplanetary research. Understanding the association between the attributes of exoplanets and the features of their host stars is crucial in acquiring significant knowledge regarding the genesis and progression of planetary systems. By means of scrutinizing data obtained from sources such as the NASA Exoplanet collection, one can extract noteworthy metrics such as the radius of exoplanets, in addition to the mass, radius, temperature, and metallicity of the host star. The metrics mentioned above possess the capability to augment our understanding of the necessary conditions for the formation and sustainability of exoplanets. Consequently, this can broaden our understanding of the universe in which we reside.

Through the analysis of this data, our aim is to address pertinent research inquiries.

- What is the relationship between the exoplanet radius and the host star's properties such as mass, radius, temperature, and metallicity?
- Are there any significant correlations between the exoplanet radius and other host parameters such as eccentricity, insolation flux, and equilibrium temperature?
- Can we predict the exoplanet radius based on the properties of the host star?
- Are there any interesting outliers or anomalies in the data that could provide clues to unusual exoplanet formation or evolution scenarios?

Answering these might establish as concrete evidence that there's indeed a relation between host star and the characteristics of the exoplanet.

V. Analysis

The primary aim of this analysis is to investigate whether the characteristics of a planet are influenced by its stellar host, which refers to the star of the exoplanetary system. The inquiry at hand presents a dichotomous nature, comprising two composite questions. The first pertains to the identification of parameters that exhibit the greatest degree of significance in relation to the characteristics under consideration. The second question concerns the existence of a noteworthy predictive capacity within specific parameters of the data, with respect to the estimation of the characteristics of the exoplanet in question. The present analysis is primarily concerned with the estimation of the radius of exoplanets in Earth units, as well as the derivation of significant parameters that exert an influence on this parameter.

A. Data Exploration

Given the vast and diverse range of datasets available in the exoplanet archive, we have opted to utilize the Planetary Systems Composite Parameters Planet Data table (PSCompPars). The document comprises a comprehensive inventory of verified exoplanetary systems, including pertinent details regarding the respective celestial bodies, such as the characteristics of the star and its orbiting planets. The objective of this tabular representation is to provide a more scholarly perspective on the documented number of exoplanets and their respective environmental conditions. The present version of the dataset initially contained a total of 34 variables. For the purpose of this analysis, we have selectively reduced the number of variables to 13. These variables are listed as follows: The parameters of interest in this study include the orbital period in days, planet radius in Earth radius units, planet mass or mass*sin(i) in Earth mass units, eccentricity, planet mass or mass*sin(i) in Jupiter mass units, insolation flux in Earth flux units, equilibrium temperature in Kelvin, stellar effective temperature in Kelvin, stellar radius in Solar radius units, stellar mass in Solar mass units, stellar metallicity in dex units, and stellar surface gravity in log10(cm/s**2) units. The variables of interest in this study are the distance in parsecs, the V magnitude in the Johnson photometric system, the Ks magnitude in the 2MASS photometric system, and the Gaia magnitude.

This dataset encompasses several methods of exoplanet discovery, including but not limited to: The preponderance of these methods is transit-based, as illustrated in Figure 2. This approach is widely employed and represents one of the

most commonly utilized means of detecting and categorizing exoplanets and their associated planetary systems.

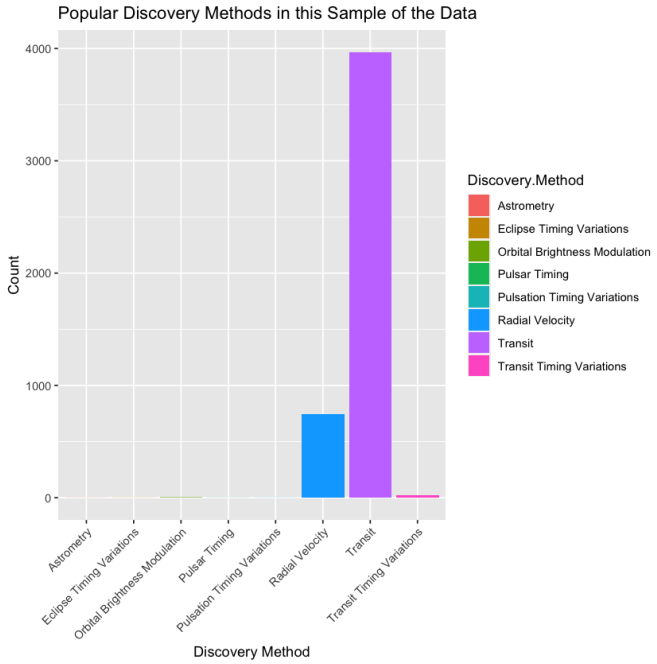

Fig. 2. Discovery Methods

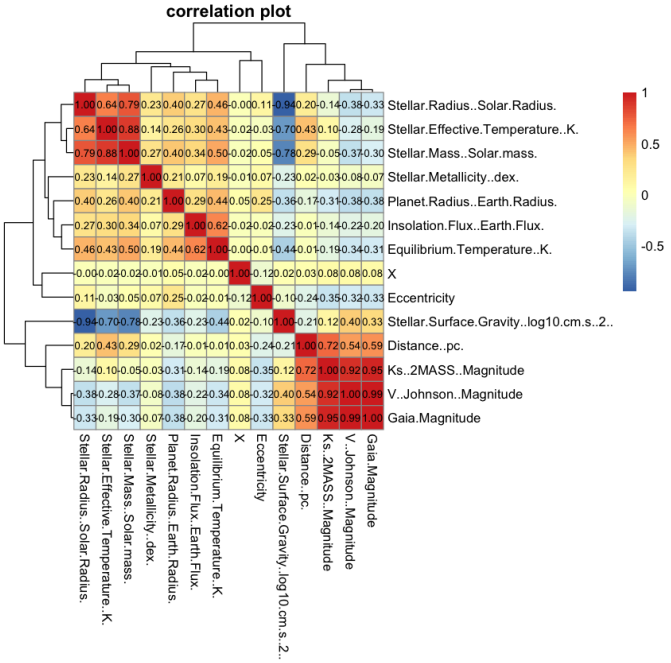

Fig. 3. Correlation Matrix

Subsequently, we undertook data preprocessing, encompassing the handling of outliers and anomalies within the dataset. It is noteworthy that certain data points, although not deemed as outliers in actuality, are statistically classified as such due to their impact on the analysis. Subsequently, we initiated an examination of the interrelationships among the attributes with the aim of identifying the characteristics to be incorporated in our analysis via a correlation matrix.

First, there is a moderately positive correlation of 0.622 between the Equilibrium Temperature (K) and the Insolation Flux (Earth Flux). This means that the amount of energy a planet gets from its mother star affects the temperature at which it is in equilibrium. Higher insolation flux leads to higher average temperatures, which shows how important star energy is for figuring out how hot or cold a world is.

With a strong positive connection of 0.786, Stellar Radius (Solar Radius) and Stellar Mass (Solar Mass) are also linked in a way that is important. This relationship shows how close the size and mass of stars are related. In general, stars that are bigger tend to have more mass, which shows that there is a strong link between these two qualities.

The strong negative link of -0.942 between Stellar Surface Gravity (log10 cm.s-2) and Stellar Radius (Solar Radius) is something that stands out. This shows that the mass and size of a star are closely linked to its surface gravity. As a star gets bigger, its surface gravity goes down. This shows how the size of a star affects its gravitational force.

Also, there is a strong positive correlation of 0.916 between V Johnson Magnitude and Ks 2MASS Magnitude. Both measures show how bright a star is, but they do so in different ways. The strong positive association means that stars that look brighter in one system tend to look brighter in the other system, too. This shows that different ways of measuring the brightness of stars are similar.

Lastly, the relationship between Stellar Effective Temperature (K) and Stellar Surface Gravity (log10 cm.s-2) is important, with a correlation of -0.699 that is a moderately negative. This shows that there is a negative link between these two things. Higher effective temperatures are linked to lower surface gravities, which shows that a star's energy output and mass are linked.

These key patterns tell us a lot about how things work together and depend on each other in exoplanetary systems. By knowing these connections, scientists can learn more about the things that affect the traits of planets and stars. This adds to what we know about planetary systems and their features as a whole.

After understanding the relations between the variables, we proceeded to check the actual distributions for the important variables. The following grid of histograms are the distributions of Stellar-Host parameters and the Exoplanet parameters.

Upon examination, it was determined that certain parameters do not possess the requisite distributions for modeling purposes. To address this matter, we conducted two distinct transformations, namely a logarithmic transformation and a square root transformation.

The square root transformation is a commonly employed mathematical technique in statistical analysis and data analytics. The procedure involves the calculation of the square root of each individual data point, resulting in the derivation of a

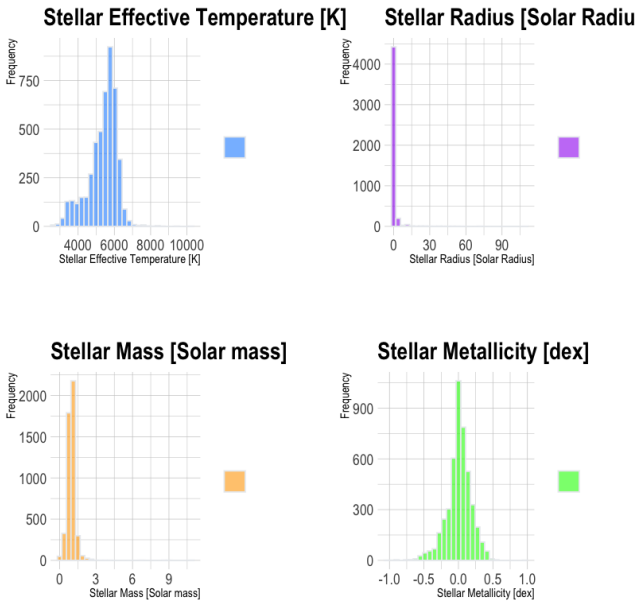

Fig. 4. Distribution Plots for the Stellar-Host Parameters

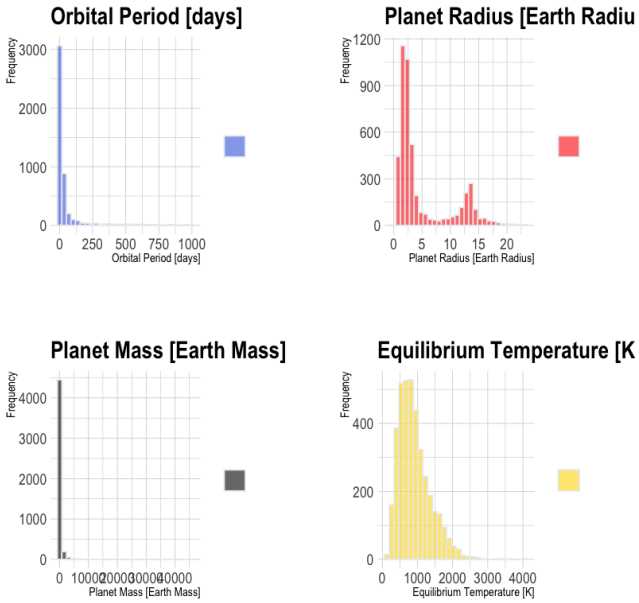

Fig. 5. Distribution Plots for the Exoplanet Parameters

new transformed variable. The technique mentioned above is commonly utilized to address data sets that display a skewed or non-normal distribution. The reason for this is that it possesses the capability to make the data more symmetrical and improve its adherence to the fundamental assumptions of linear regression models. The implementation of the square root transformation possesses the capability to alleviate the impact of outliers and amplify fluctuations in lower values, thereby rendering the data more suitable for analytical and modeling purposes. It is advisable to employ a straightforward and efficient approach to achieve data normalization and address issues related to skewness and heteroscedasticity.

The utilization of logarithmic transformation is a commonly utilized methodology in the analysis of linear regression for diverse objectives. One of the main justifications is to challenge the assumption of linearity. The linear regression model posits that the relationship between the dependent variable and the independent variables is linear in nature. However, in pragmatic scenarios, this presumption may not be feasible. The utilization of a logarithmic transformation on the variables often leads to a heightened linearity in the relationship between the transformed variables and the response variable. This approach has the potential to improve the accuracy and reliability of the regression model.

In this particular context, log transformations were executed on the necessary variables, as illustrated in Figures 5, 6, 7, and 8.

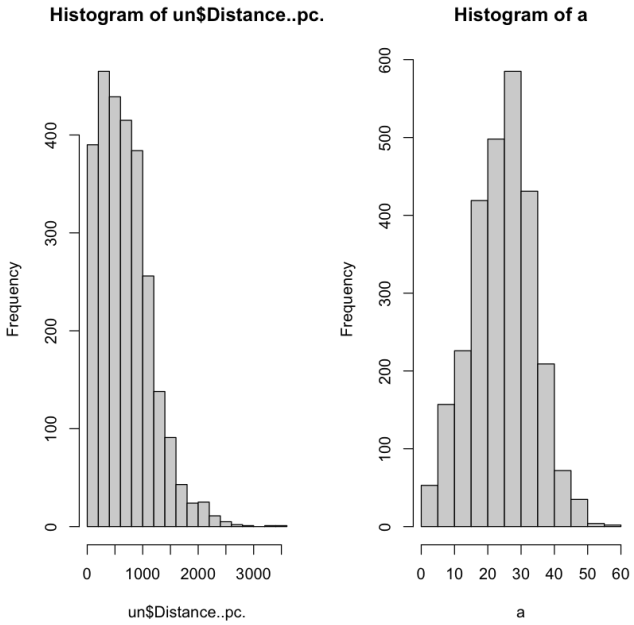

Fig. 6. Square Root Transformation for the Distance Variable

In terms of modeling for the analysis, across all the characteristics of the Exoplanet available in the dataset, we intended to estimate for the Exoplanet Radius [Earth Radius].

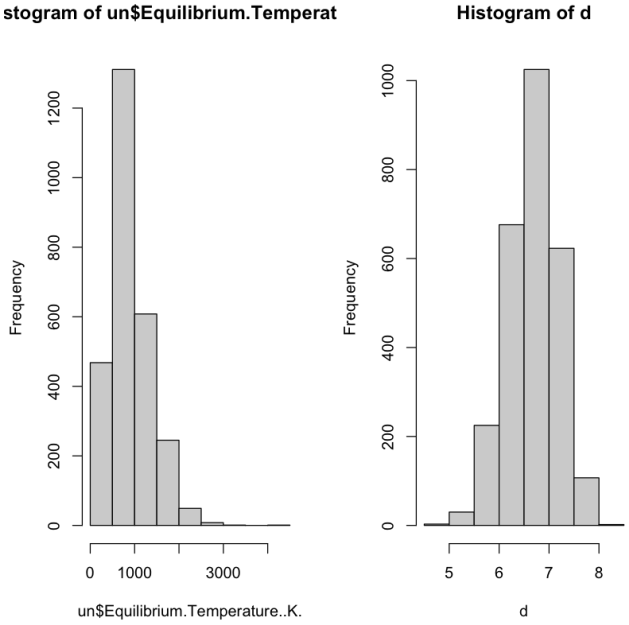

Fig. 7. Log Transformation for the Equilibrium Temperature Variable

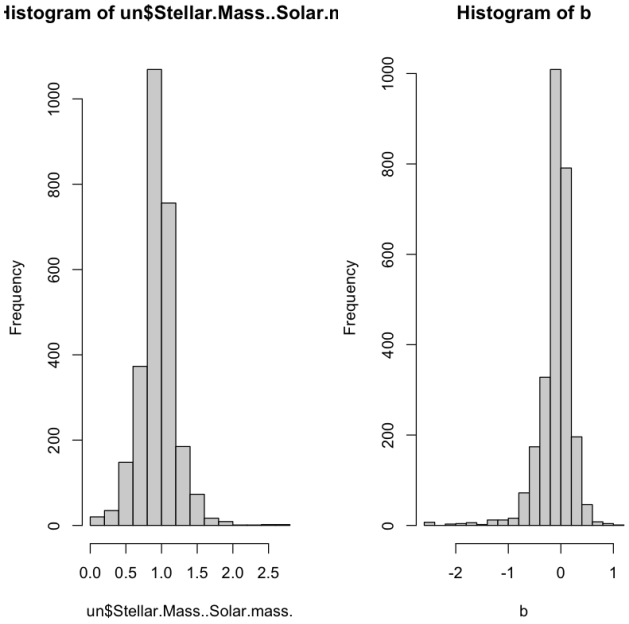

Fig. 8. Log Transformation for the Solar Mass Variable

B. Causality

Causality pertains to the correlation between cause and effect, wherein a modification in one variable has a direct impact on the alteration of another variable. The establishment of causality necessitates the consideration of temporal sequence, correlation, and the elimination of alternative explanations or confounding variables. Although correlation may suggest an association between variables, it does not establish a causal relationship. In order to bolster causal assertions, scholars utilize experimental methodologies and empirical data to isolate and exhibit the direct influence of one variable upon another. The concept of causality entails comprehending the causal mechanisms and furnishing substantial evidence to substantiate the assertion that alterations in one variable are the direct cause of modifications in another.

It is imperative to acknowledge that observational studies alone cannot establish causality within the exoplanet dataset. The reason for this is that the variables of interest, such as the distance between the exoplanet and its host star, cannot be manipulated to establish a causal relationship. However, it is only possible for us to make observations of correlations that exist between variables.

Statistical methods, such as regression analysis, can be employed to comprehend the interrelationships among the various variables in the exoplanet dataset. Through the implementation of control measures for extraneous variables, it is possible to ascertain the presence of a statistically significant correlation between the characteristics of an exoplanet, including its mass, radius, and temperature, and the attributes of its parent star. Nevertheless, it is important to note that causality cannot be established solely on the basis of these analyses.

In order to establish causality within the exoplanet dataset, it is imperative to conduct meticulously controlled experiments that manipulate a single variable while keeping all other variables constant. At present, conducting such experiments in the field of exoplanetary research is not feasible due to the intricate nature of the physical systems involved, which pose significant challenges in terms of their controllability within a laboratory environment. Hence, it can be inferred that observational studies, which scrutinize the associations between variables, will persist as a crucial mechanism for comprehending exoplanetary characteristics in the foreseeable future.

C. Methods and Modelling

1) Multiple Linear Regression: A statistical approach used to represent the connection between a dependent variable and numerous independent variables is multiple linear regression. It expands the notion of basic linear regression by taking into account the effects of many predictors on the response variable at the same time. The purpose of multiple linear regression is to identify the best-fitting linear equation that predicts the value of the dependent variable based on the independent variables' values. This approach is frequently used in a variety of domains, including as economics, social sciences, and data analysis, to comprehend and quantify the

interactions between several variables and their influence on a certain result. Multiple linear regression gives useful insights into the relative importance and direction of their impacts by assessing the coefficients and significance levels of the independent variables, enabling researchers to make educated predictions and draw meaningful conclusions. Below, is the model we initially tried.

$$\begin{aligned}\phi I_1 =&\beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D \\ &+ \beta_5 E + \beta_6 F + \beta_7 G + \beta_8 H + \beta_9 I \\ &+ \beta_{10} J + \beta_{11} K + \beta_{12} L + \epsilon\end{aligned}$$

Where, A=Eccentricity, B = Insolation.Flux..Earth.Flux., C = Equilibrium.Temperature..K., D = Stellar.Effective.Temperature..K., E = Stellar.Radius..Solar.Radius. , F = Stellar.Mass..Solar.mass. , and G = Stellar.Metallicity..dex., H = Stellar.Surface.Gravity..log10.cm.s..2.., I = Distance..pc., J = V..Johnson..Magnitude, K = Ks..2MASS..Magnitude, and L = Gaia.Magnitude.

Looking at the summary of the model, we could see that the R-squared value of the model is 0.3512, which means that 35.12% of the variation in planet radius is explained by the independent variables. The adjusted R-squared value is 0.3483, indicating that the model is not over-fitting the data. In order to check if there's a presence of Multi-collinearity within the model we leverage the VIF values. The degree of multicollinearity in a regression model may be determined with the use of the Variance Inflation Factor (VIF), which is a measurement. When two or more predictors in a regression model have a strong correlation with one another, this is known as multicollinearity, and it results in unstable and incorrect estimations of the regression coefficients.

A VIF score of 1 typically denotes a predictor-to-predictor relationship that is uncorrelated. A VIF rating between 1 and 5 is regarded as low and indicates little to no multicollinearity. The occurrence of moderate multicollinearity is suggested by VIF values between 5 and 10, which are regarded as moderate. Severe multicollinearity is seen when the VIF value exceeds 10.

It might be challenging to determine the importance of the model's predictors when multicollinearity is present because big regression coefficient standard errors result in low t-statistics and high p-values. Multicollinearity may, under extreme circumstances, result in unstable and incorrect estimations of the regression coefficients, which can negatively impact the model's accuracy and dependability. Regression models must thus be aware of and corrected for multicollinearity.

The VIF values for, V..Johnson..Magnitude, Ks..2MASS..Magnitude were greater than 100 indicating a strong possibility of the presence of multi-collinearity. Leveraging the correlation plot from above, we decided to drop the variables Ks..2MASS..Magnitude and then re-fit the model upon which the VIF values for the V..Johnson..Magnitude reduced below 5 which indicates the elimination of multi-collinearity. The performance of this model however wasn't much different, as indicated by the adjsuted R squared and MSPE values. In order to determine which predictors are significant for predicting the target variable we leveraged on the Multiple T-test at 95% confidence interval wherein the null hypothesis being that the predictor is significant and the alternate hypothesis is otherwise. As the P values for "Isolation Flux", "Stellar Radius", "Stellar Surface Gravity", and "Gaia Magnitude" are fairly greater than the 0.05 we failed to accept our null hypothesis and hence we decided to remove these variables. We then proceeded to model the below Reduced MLR.

$$\begin{aligned}\phi I_2 =&\beta_0 + \beta_1 A + \beta_2 C + \beta_3 D \\ &\beta_4 F + \beta_5 G + \beta_6 I \\ &+ \beta_7 J + \epsilon\end{aligned}$$

Looking at the summary of this model, we could see that the coefficient of determination (R-squared) is 0.3334, indicating that 33.34% of the variability in planet radius can be explained by the independent variables. This suggests that the model exhibits a moderate level of efficacy in forecasting the radius of a planet, taking into account the provided independent variables.

Furthermore, the adjusted R-squared coefficient is 0.3319, which accounts for the number of independent variables incorporated in the model. This numerical value indicates that the model is not exhibiting over-fitting tendencies towards the data. In other words, the model has not incorporated an excessive number of independent variables and is still proficient in its ability to accurately forecast the radius of planets. In general, the model exhibits a plausible correspondence with the data, albeit with potential for further enhancement.

To determine which of the two models is better one we leveraged the F-Test and the Root Mean Squared Error (RMSE) values. For the F-test at 95% significance level our Null Hypothesis was that the reduced model is sufficient to the variation in the data. The P value from the F-test resulted in a value less than 0.05. So, we do not have evidence to reject the Null Hypothesis. Hence, we considered our reduced model to be better to predict the Planet's Radius.

From the Residuals Vs Fitted values plot in Figure 9, we could see that the model violates the assumption of Linearity as the plot clearly shows the curvature of the line. Furthermore, it can be inferred from the plot that the model is in violation of the assumption of Homoscedasticity. This is due to the fact that the points initially exhibit a lower degree of dispersion, whereas towards the end of the plot, the points are more widely dispersed, indicating that the variance is not constant. From the Normal Q-Q plot, we can see that the model also violates the assumption of normality as, most of the data points towards the lower end and tail-end deviates from the normal line.

As the Reduced MLR model violates the assumptions of a linear regression model and moreover the overall explanatory

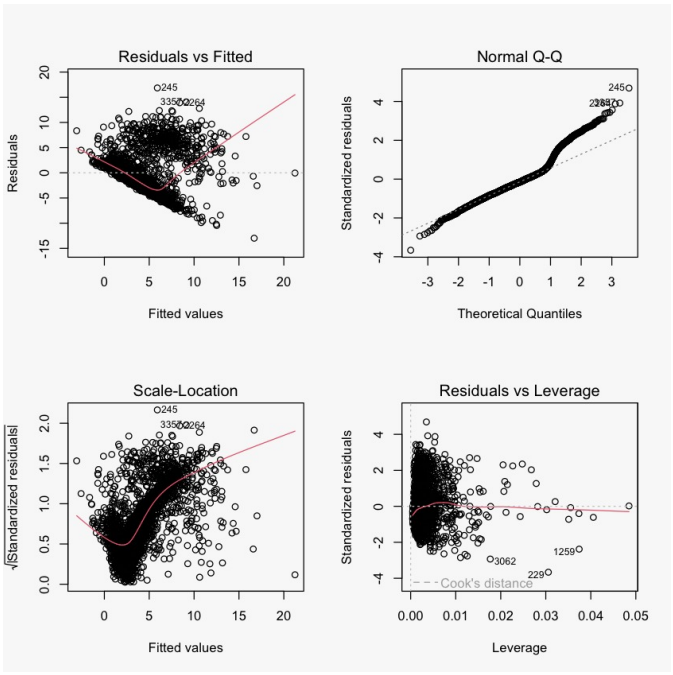

Fig. 9. Standard plots for Validating the Assumptions of the Reduced MLR

power of the parameters involved in this model is not satisfactory. This prompted to us to test our data other regression models to get better understanding for the fit of the data. Below are the additional models we tried for the analysis.

2) Generalised Additive Models: The Generalized Additive Model (GAM) is a statistical modeling technique that enhances the capabilities of Generalized Linear Models (GLMs) by enabling the incorporation of non-linear associations between the response variable and predictors. Generalized Additive Models (GAMs) employ smooth functions to effectively capture intricate patterns and non-linearities present in the data, thereby facilitating more precise modeling of diverse phenomena. Generalized Additive Models (GAMs) are capable of accommodating non-linear, non-monotonic, and interaction effects by means of incorporating smooth functions. As a result, they have gained significant importance in various fields, including ecology, epidemiology, finance, and social sciences. The process of estimating a Generalized Additive Model (GAM) entails the identification of the most suitable smoothing parameters for each predictor variable. This is commonly accomplished through the use of iterative optimization methods. Generalized Additive Models (GAMs) provide a versatile analytical tool for comprehending and interpreting intricate relationships, thereby enabling more accurate predictions and deeper insights into the underlying processes. Nevertheless, meticulous validation and interpretation of the model are imperative to prevent over-fitting and guarantee dependable outcomes.

Generalized Additive Models (GAMs) lay back the assumption that the response variable can be expressed as a linear combination of predictors by allowing for the use of a non-linear combination of predictors, represented by the symbol 's' denoting a 'smooth function'.

The equation for a Generalised Additive Model (GAM) is a formal mathematical expression that describes the relationship between a response variable and one or more predictor variables.

$$I = s_0x_0 + s_1x_1 + ... + s_nx_n$$

We developed a GAM model with the Gaussian Family and the identity link function fitting the all parameters. The statistical model incorporates smooth terms for every predictor variable, signifying a non-linear association between the predictor and the outcome. The equation is for this model is as represented as follows

$$\begin{aligned}\phi G_1 =& \beta_0 + S_1A + S_2B + S_3C + S_4D \\ &+ S_5E + S_6F + S_7G + S_8H + S_9I \\ &+ S_{10}J + S_{11}K + S_{12}L + \epsilon\end{aligned}$$

The results of the GAM model suggest that the predictor variables Eccentricity, Isolation Flux, Equilibrium Temperature, Stellar Mass, Stellar Metallicity, Distance, Ks (2MASS Magnitude), and Gaia Magnitude have a statistically significant association with the outcome variable Planet Radius.

The adjusted R-squared value is 0.52, which indicates that the model explains 52% of the variation in Planet Radius after adjusting for the number of predictors. The GCV (generalized cross-validation) is 9.5465, and the scale estimate is 9.3055, which provide information on the accuracy of the model.

The Generalized Cross-Validation (GCV) method is a statistical technique utilized to determine the optimal smoothing parameter in a smoothing spline regression model. This method estimates the mean squared prediction error of the model, thereby aiding in the selection of the most appropriate smoothing parameter. In contrast, the scale estimate offers an approximation of the scale parameter pertaining to the distribution of errors. Collectively, these metrics can provide an indication of the degree to which the model conforms to the data, with diminished values signifying superior model fit.

In this case, the GCV is 9.5465 and the scale estimate is 9.3055, which suggest that the model has moderate predictive accuracy.

Similar to the MLR, we leveraged on the Multiple T-test to develop a reduced GAM model with only significant features. The equation of this reduced model is as follows

$$\begin{aligned}\phi G_2 =& \beta_0 + S_1A + S_2B \\ &+ S_3E + S_4F + S_5G + S_6I \\ &+ S_7J + S_8K + S_9L + \epsilon\end{aligned}$$

The models exhibit comparable R-squared values, suggesting that they account for a similar proportion of the variability in the dependent variable. Nevertheless, the second

model presents a slightly lower degree of deviance explained and generalized cross-validation (GCV) values, indicating a potential decrease in its predictive capacity compared to the first model. We also compared the RMSE values for the two models to which we could see that both the models have very similar RMSE scores.

In order to choose between the two models, we used a model selection technique namely AIC to determine which model provides the best balance between complexity and predictive power. The acronym AIC denotes the Akaike Information Criterion. The statistical measure employed for model selection involves comparing the goodness of fit of various models that have been estimated for a specific dataset. The Akaike Information Criterion (AIC) is founded on the principle that an effective model ought to exhibit a high degree of congruence with the data, while simultaneously being as parsimonious as possible, that is, avoiding undue complexity.

The Akaike Information Criterion (AIC) quantifies the relative merit of a statistical model, enabling comparisons between models of varying types or with differing parameter counts.

On comparing the AIC values for our two GAM models, we could see that, the AIC value for the second model is slightly lower than the first model. So we determined that our best GAM model is the reduced model.

Low p-values for the smooth terms suggest that these predictors have substantial statistical relevance for Planet Radius. Although the model appears to fit the data well, further research. looking at the plots in the Figure 10, we could see that the Smooth term for the parameter V Johnson Magnitude shows an estimated linear relationship between the predictor and the response variable. So, we decided to model a semiparametric penalised generalized additive model by removing the smoothening term for "V Johnson Magnitude".

D. Semiparametric Penalised Generalized Additive Model

The Semiparametric Penalized Generalized Additive Model (SP-GAM) is a statistical modeling technique that offers a high degree of flexibility and robustness by integrating the strengths of both parametric and nonparametric methods. The present study expands upon the conventional generalized additive model (GAM) by integrating penalties on the smooth functions, thereby achieving a balance between flexibility and regularization. The SP-GAM model is capable of estimating linear and nonlinear associations between the predictors and the response variable, rendering it appropriate for capturing intricate and nonlinear patterns in the data. The inclusion of a penalization term serves the purpose of regulating the complexity of the model and mitigating the risk of overfitting by inducing a reduction in the magnitudes of the coefficients associated with the smooth functions, ultimately driving them towards zero. The SP-GAM model is especially advantageous in scenarios where the association between the predictors and the response variable is anticipated to be non-linear, and where there is a preference to circumvent an overabundance of intricacy in the model. The utilization of this methodology is prevalent across diverse domains, including but not limited to

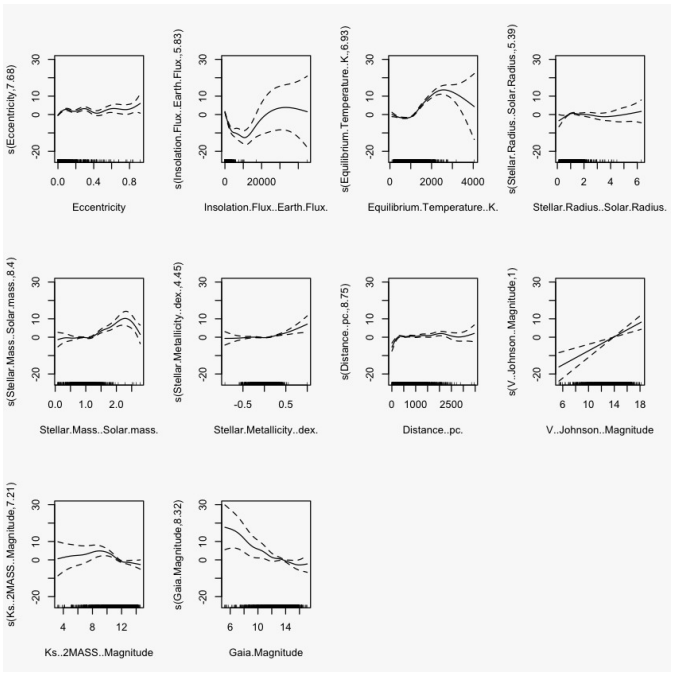

Fig. 10. Visualised Smooth Terms in the Model

environmental science, economics, and epidemiology, with the purpose of scrutinizing intricate data structures and generating precise prognostications.

The Semi-parametric Model that we developed is represented as follows,

$$\begin{aligned}\phi SG_1 =&\beta_0 + S_1A + S_2B + S_3C + S_4D \\ &+ S_5E + S_6F + S_7G + S_8H + S_9I \\ &+ \beta_1 J + S_{11}K + S_{12}L + \epsilon\end{aligned}$$

This model is characterized by a single parametric predictor, namely V. Johnson.The magnitude of the variable in question exhibits a statistically significant positive coefficient at the 0.001 level. This implies that a rise in V.Johnson.Magnitude is correlated with an augmentation in the radius of the planet.

The aforementioned model comprises of several smooth predictors, each of which has been assigned an estimated degrees of freedom (edf) and p-value. Insolation is the smooth term that exhibits the highest EDFs.Flux..Earth.The predictors of Flux, Distance, and Stellar Mass in terms of Solar Mass indicate a more intricate association with the response variable.

The adjusted R-squared value, which stands at 0.52, denotes that the model accounts for 52% of the variability in the response variable, while taking into account the number of predictors. The model's deviance explained, which amounts to 53.2%, indicates that it offers a satisfactory level of fit to the data. The obtained result of the generalized cross-validation (GCV) value, which is 9.5347, implies that the model possesses a moderate level of complexity. Additionally, the scale estimate of 9.3045 denotes the estimated standard deviation of the model errors.

However, the MSPE and RMSE values slightly less but quite comparable to the GAM reduced model. So, out attempt at improving the GAM model was not sucessful. After establishing this we moved on to try other regressor models to try and improve the fit of the data.

E. Support Vector Regressor

The support vector regressor (SVR) is a robust machine learning algorithm that is frequently employed for regression tasks. The algorithm in question is a variant of the support vector machine (SVM) and is known for its efficacy in addressing non-linear and intricate regression problems. Support Vector Regression (SVR) operates by identifying an optimal hyperplane that maximizes the margin between the predicted values and the actual target values. The employed methodology involves the utilization of a kernel function to facilitate the mapping of the input data into a space of higher dimensionality, thereby enabling the capture of non-linear relationships. The objective of Support Vector Regression (SVR) is to minimize the discrepancy between the predicted and observed values, while also accommodating a specific level of permissible error, as determined by the parameter epsilon.

The general form of a Support Vector Regressor is represented as follows,

$$\text{Minimize:} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{Subject to:} \quad \begin{cases} y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i \\ \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

We modelled our entire data using on a support vector regressor, with a radial kernel function and an epsilon value of 0.1. The cost parameter is set to 1 and the gamma parameter is set to 0.2.

The model is comprised of 1800 support vectors, which represent the training samples that establish the decision boundary of the model.

This model has a mean squared prediction error (MSPE) of 13.23263 adjusted R squared of 0.562. The MSPE is than that observed in the case of our previous models which are Reduced GAM and Semiparametric GAM. Henceforth, we cannot consider this as a good model for our objective.

F. Decision Tree Regressor

The Decision Tree Regressor is a widely employed algorithm utilized for regression tasks, which can proficiently model intricate associations between predictors and continuous target variables. The modus operandi involves the iterative division of the feature space into subgroups, guided by a set of decision rules. At every internal node of the tree, a splitting criterion is employed to identify the most optimal split that maximizes the reduction in variance or any other appropriate metric. The terminal nodes, commonly referred to as leaves, correspond to the ultimate predicted outcomes. Decision trees possess an innate interpretability and are capable of accommodating predictors of both numerical and categorical nature. The algorithm exhibits the capability to effectively manage missing values and outliers, while also demonstrating robustness in the presence of nonlinear relationships. Nevertheless, decision trees have a tendency to overfit the training data. Consequently, ensemble techniques such as Random Forests or Gradient Boosting are frequently utilized to enhance predictive accuracy.

General equation for the Decision Tree Regressor is as follow,

$$\hat{y}(x) = \sum_{j=1}^{J} c_j \cdot I(x \in R_j)$$

This model comprises a primary node and an additional 13 nodes. The variables that exhibit the highest degree of significance in the model are those pertaining to Equilibrium. The variables of interest are temperature and insolation.The topic of discussion pertains to Flux and Ks.2MASS.The topics of interest are the concepts of magnitude, Gaia.Magnitude, and V.Johnson.Magnitude.

The root node exhibits a mean squared error (MSE) of 19.38891 and a mean response variable value of 4.075931. The employed methodology involves recursive partitioning to generate reduced subsets of data, subsequently utilizing the average of each subset to predict the response variable. The model partitions the dataset at every node by selecting the variable that provides the greatest enhancement in forecasting the response variable. The model's complexity parameter has been determined to be 0.25, with an optimal number of 12 splits. This model has an adjusted R squared of 0.495.

The model exhibits an RMSE value of 10.4616, indicating that the average deviation between the model's predictions and the actual values is approximately 10.5 Planet radii. The Root Mean Square Error (RMSE) value in question is comparatively lower than the values obtained from the other models that were subjected to testing. Nevertheless, it is crucial to acknowledge that the Root Mean Square Error (RMSE) ought to be construed within the framework of the specific problem being addressed. In our study, where we are making an estimation of the radius of the exoplanet in Earth units, a deviation of 10 is considered significant. Therefore, we cannot regard this prediction as ideal.

Regardless, given the better results of the aforementioned model with respect to the evaluated metrics, we proceeded to assess the ensemble variant of the Decision Tree Regressor, namely the Random Forests.

G. Random Forest Regressor

The Random Forest Regressor is an ensemble learning technique that utilizes decision trees as its foundation. It amalgamates several decision trees to generate predictions. The operational mechanism involves the creation of numerous decision trees in the course of training, with each tree utilizing

a random subset of features to forecast the output value. The ultimate forecast is derived through the process of averaging the prognostications of each of the individual trees.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} f_t(x_i) \right)$$

Similarly to the aforementioned models, using all the variables in the data we modeled a random forest regressor. The results suggest that, the model has the RMSE of 2.6218 and an Adjusted R Squared of 0.6708.

Based on the observed RMSE value, it is determined that the model is performing significantly better than the other models, in terms of estimating the exoplanet radius. A value of 2.6 suggests the maximum deviation of the model while predicting this variable. Adjusted R squared value indicates that the model explains 67 percent of the variation in the data.

Compiling all the performance metrics across all the observed models, it is evident that the Random Forest regressor model has the optimal performance out of all.

TABLE I
PERFORMANCE METRICS TABLE

Model	RMSE	Adjusted R-Squared
Support Vector Regressor	3.63	0.562
Decision Tree Regressor	3.23	0.494
Random Forest Regressor	2.62	0.670
Generalized Additive Model	3.26	0.52
Semiparametric GAM	3.26	0.52
Reduced MLR	3.67	0.331

VI. CONCLUSION

In conclusion, this analysis on devising the exoplanet characteristics (Planet Radius in this case) through the Stellar-Host parameters is only partly explainable and is not consistent considering the fact that the regression analysis revealed that its difficult to find the relationship of stellar parameters having any significant affect on the Exoplanet characteristics. This validates the reality assumption that this methodology is not the ideal way to quantify and catalogue the discovered exoplanet features by the Science administrations worldwide.

Through the analysis we discovered that Random Forest Regressor has exhibited superior performance in estimating the planet radius, as evidenced by the performance metrics table for exoplanet data analysis. This is demonstrated by a mean squared error (MSE) of 2.62 and a coefficient of determination (R-squared) of 0.670. The Support Vector Regressor exhibits notable performance, attaining a Mean Squared Error (MSE) of 3.63 and an R-squared value of 0.562. The Decision Tree Regressor model exhibits a Mean Squared Error (MSE) of 3.23 and an R-squared value of 0.494. Both the Generalized Additive Model and Semiparametric GAM exhibit comparable performance, attaining a Mean Squared Error (MSE) of 3.26 and a coefficient of determination (R-squared) of 0.52. Ultimately, the Reduced Multiple Linear Regression (MLR) model demonstrates the most elevated Mean Squared Error (MSE) of 3.67, accompanied by an R-squared value of 0.331, thereby signifying the least effective performance when compared to the other evaluated models. Having said that, interpreting the Random Forest Regressor Model is difficult and complicated. When it comes to choosing the best model which holds the balance for explanability and predictability, it is crucial to take into account additional aspects such as interpretability and computational complexity while determining the optimal model for a given application.

Through our analysis we realised that the Exoplanet dataset is highly complex and enormous and the natural variations in this data cannot be quantified and explained to the core. However this study establishes that there is some level of relationship between the host and and planet's features concreting the reality of how things work in a proto-planetary system. In conclusion, we are unable to determine a consistent aspect through which we can establish that there is indeed a cause and affect in devising the exoplanet features through the stellar-host parameters. The future scope of this project relies totally on better understanding the data. Domain knowledge in this topic is essential to perform any exploratory analysis. Further, it required to consider sophisticated preprocessing steps to address the high variability.

REFERENCES

[1] Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. (2013). Exploring exoplanet populations with NASA's Kepler Mission. Proceedings of the National Academy of Sciences, 110(48), 19273-19280.

[2] Coughlin, J. L., Mullally, F., Thompson, S. E., et al. (2016). Stellar Parameters of Kepler's Stars with the Robovetter. The Astrophysical Journal, 750(1), 44.

[3] Buchhave, L. A., Latham, D. W., Johansen, A., et al. (2012). An abundance of small exoplanets around stars with a wide range of metallicities. Nature, 486(7403), 375-377.

[4] "Data Resources in the Exoplanet Archive." Exoplanetarchive.ipac.caltech.edu, exoplanetarchive.ipac.caltech.edu/docs/data.html.