

16º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2025

INVESTIGAÇÃO DA HABITABILIDADE PLANETÁRIA DE EXOPLANETAS DO TIPO MINI NETUNOS POR MEIO DE TÉCNICAS DE DATA MINING E MACHINE LEARNING

PEDRO DE SOUZA LA GAMBA¹, LUANDER BERNARDES²

¹ Graduando do Curso de Engenharia Elétrica, IFSP, Campus Sertãozinho, la.gamba@aluno.ifsp.edu.br.

² Professor Doutor do Curso Engenharia Elétrica, IFSP, Sertãozinho, bernardes.luander@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.04.02.00-4 Astrofísica Estelar

RESUMO: Nas últimas décadas, vários exoplanetas do tipo Mini Netunos foram descobertos orbitando estrelas dos tipos M, K e G. Entretanto, pouco se conhece sobre a constituição desses objetos, tampouco se eles são aptos a abrigar algum tipo de vida. Sendo assim, essa pesquisa teve como objetivo analisar as condições de habitabilidade planetária de 795 planetas, por meio de técnicas de Data Mining e Machine Learning. Através da Análise por Componentes Principais (PCA), as informações de uma população seleta de planetas, contidas no repositório público da NASA, sofreram um processo de redução de dimensionalidade para que, posteriormente, fosse empregado o método de clusterização não hierárquico *K-Means*, que teve como propósito agrupar corpos semelhantes e escolher entre eles aqueles mais propensos a fornecerem condições otimizadas para a vida se desenvolver. A análise desses agrupamentos revelou que um conjunto específico de exoplanetas apresentava características mais adaptadas à existência de uma biosfera, capaz de possibilitar a identificação de bioassinaturas ligadas à vida, já que possuíam valores de Índice de Similaridade com a Terra (ESI) elevados. O estudo desse aglomerado de planetas extrassolares permitiu a identificação de nove alvos prioritários para futuras missões de busca por assinaturas biológicas, demonstrando a eficácia da combinação de técnicas de análise de dados no auxílio de missões astronômicas, que visam encontrar sinais de vida extraterrestre na atmosfera de exoplanetas.

PALAVRAS-CHAVE: exoplanetas; habitabilidade; machine learning; data mining; Mini Netunos.

INVESTIGATION OF THE PLANETARY HABITABILITY OF MINI-NEPTUNE EXOPLANETS USING DATA MINING AND MACHINE LEARNING TECHNIQUES

ABSTRACT: In recent decades, several Mini Neptune-type exoplanets have been discovered orbiting M, K, and G-type stars. However, little is known about the composition of these objects, nor whether they are capable of supporting any form of life. Therefore, this research aimed to analyze the planetary habitability conditions of 795 planets using data mining and machine learning techniques. Through Principal Component Analysis (PCA), the information on a select population of planets contained in NASA's public repository underwent a dimensionality reduction process so that the non-hierarchical K-Means clustering method could then be employed. The purpose of this method was to group similar bodies and choose among them those most likely to provide optimal conditions for life to develop. Analysis of these clusters revealed that a specific set of exoplanets had characteristics more suited to the existence of a biosphere, capable of enabling the identification of biosignatures linked to life, as they had high Earth Similarity Index (ESI) values. The study of this cluster of extrasolar planets allowed the identification of nine priority targets for future missions to search for biological signatures, demonstrating the effectiveness of combining data analysis techniques to aid astronomical missions that aim to find signs of extraterrestrial life in the atmosphere of exoplanets.

KEYWORDS: Exoplanets; habitability; machine learning; data mining; mini-Neptunes.

INTRODUÇÃO

A classificação de um planeta como habitável baseia-se na observação das condições de habitabilidade existentes na Terra. Nessa definição, é fundamental a presença de água líquida, uma atmosfera densa e um campo magnético capaz de protegê-lo das partículas e radiações oriundas do espaço (vento estelar, radiação cósmica etc.). De forma simplificada, a “Zona Habitável” é a região ao redor de uma estrela onde pode existir água líquida na superfície de um planeta ou lua. Contudo, a descoberta de micro-organismos extremófilos ampliou esse conceito, permitindo a definição da “Zona Extremófila”, que inclui, por exemplo, ambientes que apresentam características mais específicas: variações de temperatura mais amplas, atmosferas mais tênues etc., isto é, condições mais hostis à vida. O estudo de exoplanetas nessa região é importante, pois eles podem possuir envoltórios gasosos contaminados com marcadores que denunciem a presença da vida.

Atualmente, o Programa de Exploração de Exoplanetas da NASA já confirmou a existência de mais de 6000 exoplanetas, demonstrando a vasta diversidade de sistemas planetários. Em breve, os telescópios espaciais estarão aptos a investigar de maneira robusta a atmosfera desses planetas. Nesse cenário, o estudo dos Mini Netunos, exoplanetas com raios entre 1,6 e 2,5 raios terrestres, torna-se relevante devido ao seu potencial para possuir atmosferas de hidrogênio ou vastos oceanos cobertos por água, condições essenciais para o possível desenvolvimento da vida.

Alguns trabalhos já propuseram a utilização de técnicas de *Machine Learning* para investigar sistemas planetários (Manfredi et al. 2020; Bernardes, 2023), mas este estudo dedica-se exclusivamente a análise das condições de habitabilidade planetária de Mini Netunos, buscando apontar os melhores alvos para a busca de biomarcadores que denunciem a presença de atividade biológica. Assim, foram aplicadas algumas técnicas de aprendizado de máquina na tentativa de se identificar agrupamentos de exoplanetas com características mais adaptadas ao surgimento da vida, considerando a Terra como modelo, já que é o único planeta onde ela existe.

MATERIAL E MÉTODOS

Os dados utilizados na pesquisa se referem a um conjunto de 795 exoplanetas do tipo Mini Netunos. Eles foram obtidos a partir da exploração do banco de dados disponibilizado pela NASA, que apresenta várias características de exoplanetas e de suas estrelas hospedeiras como, por exemplo, o raio e a massa, a temperatura superficial estimada, o índice de similaridade com a Terra (ESI), a massa estelar, a metalicidade estelar, a luminosidade estelar etc. Como o foco do trabalho é o estudo da habitabilidade planetária, esta pesquisa analisou planetas orbitando estrelas do tipo espectral M (2400 – 3700 K), K (3700 – 5200 K) e G (5200 – 6050 K). A tabela 1 mostra os parâmetros estelares utilizados para se investigar certos padrões dos exoplanetas do tipo Mini Netunos.

TABELA 1. Variáveis estelares selecionadas. Fonte: Dados originais de pesquisa oriundos do banco de dados da NASA (NASA, 2024). A massa e o raios são medidos em relação aos valores solares, já temperatura é medida em Kelvins, a idade em Ganos (10^9 anos) e a luminosidade em função do logaritmo da luminosidade solar.

Nome do Planeta	M*	R*	[Fe/H]*	L*	T*	Idade*
Kepler 829 b	0,98	0,97	0,03	0,04	5698,00	4,68
Kepler 477 b	0,87	0,79	- 0,04	- 0,37	5240,00	2,09
HD 108236 c	0,87	0,88	- 0,28	- 0,15	5660,00	6,70
HD 108236 f	0,87	0,88	- 0,28	- 0,15	5660,00	6,70
Pi Men c	1,10	1,19	0,08	1,24	6037,00	3,92
⋮	⋮	⋮	⋮	⋮	⋮	⋮
TOI 712 b	0,73	0,67	- 0,02	- 0,73	4622,00	0,83

Os dados contidos na tabela 1 foram submetidos a uma redução de dimensionalidade a partir da utilização da técnica Análise por Componentes Principais (PCA). O objetivo da PCA é reduzir o número de variáveis e trabalhar com duas dimensões, PC1 e PC2, que carregam consigo grande parte das informações contidas nas variáveis originais (Fávero e Belfiore, 2017).

Os fatores referentes às características das estrelas que hospedam os 795 exoplanetas foram agrupados por semelhança utilizando o método *K-Means*, que consiste no agrupamento de elementos amostrais em determinado *cluster* cujo centróide (vetor de média amostral) é o mais próximo do vetor

de valores observados para o respectivo elemento (Hartigan and Wong, 1979). O objetivo da utilização da técnica está relacionado à investigação das características dos planetas que constituem as aglomerações e que podem estar associadas a habitabilidade. Como a técnica exige que se escolha previamente a quantidade de *clusters*, foram utilizados o método Hierárquico Aglomerativo e o método de *Elbow* para definir esse valor (Bernardes, 2023). Assim, os exoplanetas foram agrupados em *clusters*, o que facilitou a compreensão das relações existentes entre eles. A análise busca entender como as características estelares influenciam a habitabilidade dos sistemas planetários estudados.

RESULTADOS E DISCUSSÃO

As seis variáveis descritas na tabela 1 foram reduzidas a duas por meio da utilização da técnica de Análise de Componentes Principais (PCA), isto é, passou-se de cinco para duas dimensões (Componente 1 e Componente 2). A primeira componente representa 63,13 % das informações contidas nas variáveis originais, enquanto a segunda, 17,09 %, totalizando aproximadamente 80,21 %. A garantia de que os dados se adequavam à análise fatorial ocorreu por meio do teste de esfericidade de *Bartlett* (vide tabela 2), onde fica evidenciado que para um nível de significância de 5%, pode se rejeitar a hipótese nula de que a matriz de correlações de *Pearson* para as variáveis de estudo seja igual a uma matriz identidade de mesma ordem, ou seja, de que todas as correlações entre as variáveis sejam estatisticamente iguais a zero (Fávero e Belfiore, 2017).

TABELA 2. Teste de esfericidade de *Bartlett*. Fonte: Dados originais de pesquisa.

	χ^2	P-valor
Esfericidade de <i>Bartlett</i>	4885,43	0

A análise dos fatores obtidos e a variância compartilhada capturada pelas duas primeiras componentes é mostrada na Tabela 3.

TABELA 3. Autovalores e variância compartilhada. Fonte: Dados originais de pesquisa.

Fatores	Autovalor	Variância	Variância Acumulada
Fator 1	3,787587	0,631264	0,631264
Fator 2	1,025130	0,170855	0,802119

A redução de dimensionalidade é justificada pelo critério de *Kaiser*, que adota fatores com autovalor superior a um (ver Tabela 3).

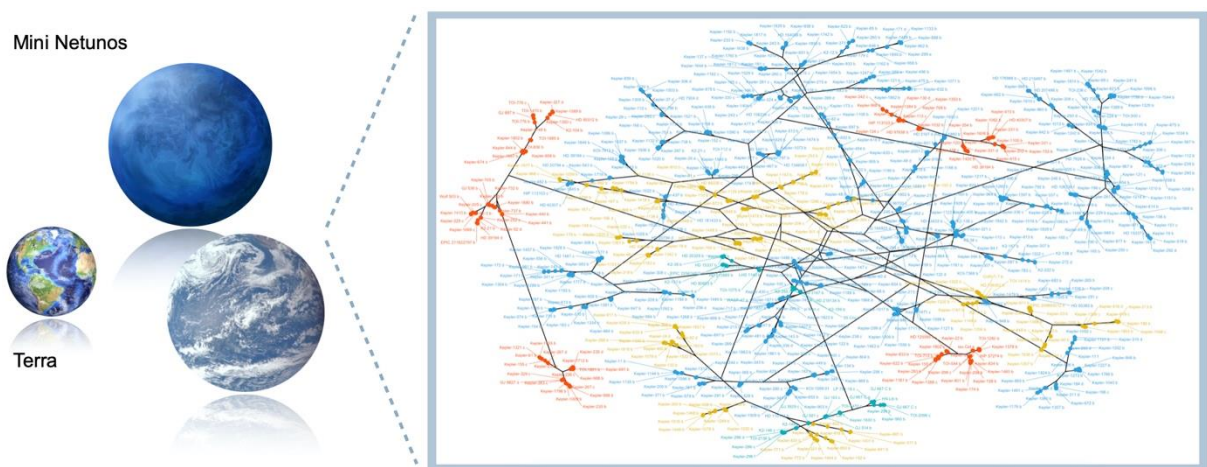
O método de aglomeração não hierárquico *K-Means* foi utilizado para a formação dos *clusters* de exoplanetas do tipo Mini Netunos. O procedimento requer que a quantidade de *clusters* seja definida a priori e, nesta pesquisa, optou-se por utilizar o método Hierárquico Aglomerativo (distância euclidiana / método *Complete Linkage*), que indicou a formação de quatro *clusters*, em conjunto com uso do “Método de *Elbow*” (ou método do Cotovelo), que também sugeriu a formação de quatro *clusters*. As variáveis escolhidas para clusterização foram os fatores F1 e F2 obtidos por meio da utilização da técnica de PCA. A figura 1 mostra uma concepção artística dos exoplanetas estudados nessa pesquisa e o resultado do Método Hierárquico apontando a formação de quatro *clusters* (em formato de árvore para uma melhor visualização).

O teste de significância estatística relativos aos fatores F1 e F2 são mostrados na tabela 4. Os *clusters* formados a partir da utilização do método *K-Means* e dos fatores um e dois oriundos da PCA são mostrados na tabela 5.

TABELA 4. Teste de significância estatística. Fonte: Dados originais de pesquisa.

Dados Cluster	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fator 1	1	6,7	6,679	6,719	0,00972
Fator 2	1	538,2	538,2	1662	2e-16

FIGURA 1. Concepção artística dos Mini Netunos e a indicação da quantidade de *clusters* por meio do método hierárquico. Fonte: dados originais de pesquisa.



A partir do processo de clusterização foi possível organizar e analisar as características dos exoplanetas que constituem os quatro aglomerados. Elas podem ser observadas na tabela 5. Ressalta-se a importância da análise desses parâmetros, já que alguns deles estão intimamente ligados ao conceito de habitabilidade planetária como, por exemplo, o índice de similaridade com a Terra (ESI), a temperatura superficial, a densidade do planeta etc.

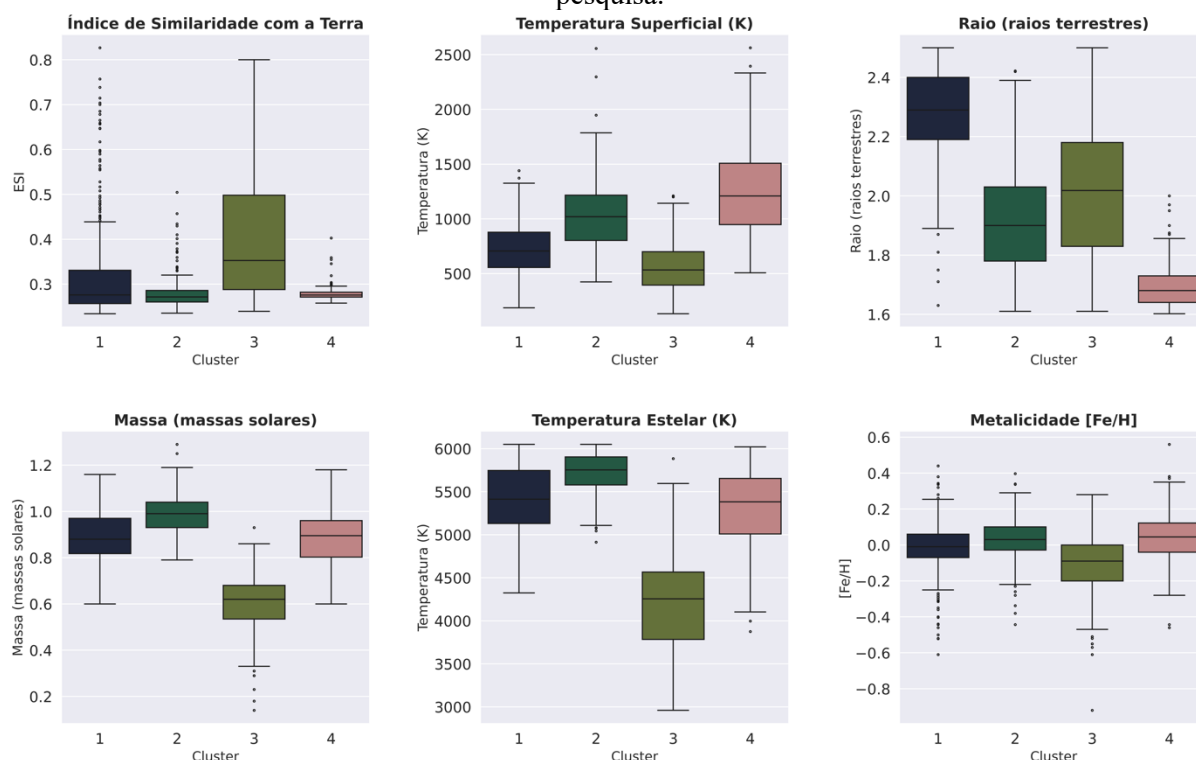
TABELA 5. Características dos *clusters* planetários formados pelo algoritmo *K-means*. Fonte: dados originais de pesquisa. Nota: D_p é a densidade do exoplaneta (g/cm^3) e T_p a estimativa de sua temperatura superficial.

Dados estatísticos	Cluster 1					Cluster 2				
	ESI	R_p	T_p	D_p	[Fe/H]*	ESI	R_p	T_p	D_p	[Fe/H]*
Média	0,32	2,28	716,92	0,49	- 0,02	0,28	1,92	1031,71	0,63	0,03
Desv. Padrão	0,11	0,15	232,70	0,06	0,15	0,04	0,17	326,04	0,08	0,12
Valor mínimo	0,23	1,63	187,27	0,13	- 0,61	0,24	1,61	423,07	0,33	- 0,44
25 %	0,26	2,19	555,45	0,46	- 0,07	0,26	1,78	802,36	0,58	- 0,03
50 %	0,28	2,29	706,27	0,49	- 0,01	0,27	1,90	1018,99	0,62	0,03
75 %	0,33	2,40	777,47	0,52	0,06	0,29	2,03	1216,23	0,68	0,10
Valor máximo	0,82	2,50	1274,78	0,76	0,44	0,50	2,42	2556,37	0,93	0,39
Nº de planetas	308					238				
Dados estatísticos	Cluster 3					Cluster 4				
	ESI	R_p	T_p	D_p	[Fe/H]*	ESI	R_p	T_p	D_p	[Fe/H]*
Média	0,41	2,01	568,24	0,61	-0,13	0,28	1,69	1279,07	0,87	0,05
Desv. Padrão	0,16	0,23	223,58	0,11	0,19	0,01	0,08	444,46	0,29	0,15
Valor mínimo	0,24	1,61	132,81	0,35	- 0,92	0,25	1,60	506,78	0,63	- 0,46
25 %	0,29	1,83	394,49	0,53	- 0,20	0,27	1,64	947,12	0,72	- 0,04
50 %	0,35	2,02	471,77	0,59	- 0,09	0,28	1,68	1208,84	0,75	0,05
75 %	0,49	2,28	699,76	0,67	0,00	0,28	1,73	1507,43	0,77	0,12
Valor máximo	0,79	2,50	1210,67	1,08	0,28	0,40	2,00	2562,95	1,90	0,56
Nº de planetas	139					110				

A partir das observações contidas na tabela 5 foi possível construir um gráfico do tipo “boxplot” com algumas variáveis importantes (vide figura 2). É possível perceber que o *cluster 3* possui, em média, o maior valor para a variável ESI, o que indica que os planetas pertencentes a esse grupo possuem características mais parecidas com aquelas encontradas na Terra. Quando se pensa em temperatura, o

cluster 3 é aquele que apresenta os planetas com as temperaturas superficiais mais baixas (em média), o que aumenta a possibilidade da existência de vida. Fica evidente também, que os demais *clusters* possuem a maioria de seus planetas possuindo temperaturas superficiais muito elevadas, impossibilitando ou dificultando a presença de vida. A análise da metalicidade estelar para o *cluster 3* deixa evidente que os sistemas planetários que o compõe podem ter uma probabilidade baixa de abrigarem planetas rochosos, já que essa medida está correlacionada com a disponibilidade de material diferente do hidrogênio e do hélio disponível.

FIGURA 2. *Boxplot* contendo algumas variáveis dos *clusters* analisados. Fonte: dados originais de pesquisa.



Ainda considerando o *cluster 3*, nota-se que a média da massa gira em torno de 4,94 massas terrestres com um valor mínimo de 2,89 e um valor máximo de 9,70 massas terrestres; já a mediana encontra-se em torno 4,74 massas terrestres. A densidade média desses planetas é de 0,61 g/cm³, com valores mínimo e máximo variando na faixa entre 0,35 e 1,08 g/cm³; a mediana possui valor igual a 0,59 g/cm³. Densidades baixas é um indicativo de que eles possuem atmosferas espessas de hidrogênio e hélio ou sejam mundos oceânicos. Esse *cluster* agrega exoplanetas com períodos orbitais de poucos dias (estão muito próximos de sua estrela hospedeira). Como essas estrelas são majoritariamente anãs vermelhas (frias e de baixa massa), isso dificulta a manutenção de condições favoráveis à vida, pois elas podem emitir fluxos radioativos tão intensos, que seriam capazes de até mesmo erodir as atmosferas de planetas mais próximos.

O banco de dados reduzido contendo somente as informações do *cluster 3* foi manipulado para que um filtro fosse aplicado. O filtro utilizou a variável ESI (Schulze-Makuch et al., 2011) para selecionar planetas com mais alto grau de semelhança com a Terra, e consequentemente, alvos prioritários na busca de sinais de vida extraterrestre. O corte adotado é para $ESI \geq 0,7$, considerando que valores próximos a 1,0 indicam maior similaridade; o corte nesse valor foi adotado já que planetas com valores de ESI maiores do que 0,5, geralmente, são considerados potencialmente habitáveis, mesmo que seja consenso comum entre os cientistas de que critérios de habitabilidade são independentes desse índice. A tabela 6 mostra os resultados da aplicação dessa filtragem. Vale lembrar que esse índice varia na faixa de zero a um e quanto mais próximo de um, mais parecido com a Terra o exoplaneta é.

TABELA 6. Planetas alvo de busca por sinais de vida. Fonte: dados originais de pesquisa. Nota: P_p é o período orbital do exoplaneta em dias.

Nome do Planeta	ESI	R_p	M_p	P_p	D_p	T_p	[Fe/H]*	Idade*
Tau Cet e	0,72	1,81	3,93	162,87	0,66	328,39	- 0,52	12,12
Kepler-705 b	0,73	2,11	5,10	56,05	0,54	269,11	0,07	3,89
Kepler-1544 b	0,79	1,78	3,82	168,81	0,67	275,66	0,02	2,34
Kepler-1410 b	0,79	1,78	3,82	60,86	0,67	292,92	0,01	4,07
Kepler-283 c	0,79	1,82	3,97	92,74	0,65	279,76	- 0,17	3,09
GJ 667 C c	0,79	1,77	3,80	28,14	0,68	278,42	- 0,55	2,00
Kepler-22 b	0,72	2,10	9,10	289,86	0,98	275,41	- 0,25	7,00
Kepler-440 b	0,75	1,86	4,12	101,11	0,64	307,80	- 0,30	1,30
Kepler-155 c	0,72	2,24	5,65	52,66	0,50	291,41	- 0,14	2,90

Vale a pena salientar que a idade estelar é um parâmetro importante quando se estuda a habitabilidade planetária. Na Terra existem indícios fósseis de organismos datados de aproximadamente 3,5 bilhões de anos. Logo, é interessante estudar sistemas planetários cuja suas estrelas tenham idade superior a essa, pois se a origem da vida nesses ambientes extraterrestres seguir os mesmos caminhos terrestres, então, há uma maior chance de encontrá-la em sistemas mais antigos. Entre os planetas selecionados há alguns que se enquadram nesse limite. Os exoplanetas listados na tabela 6 seriam os alvos prioritários de missões astronômicas na busca por bioassinaturas que denunciem a presença de vida.

CONCLUSÕES

O uso de *Data Mining* e *Machine Learning* mostrou-se eficaz no estudo da habitabilidade planetária e na escolha de alvos prioritários na busca de vida extraterrestre. Os processos de redução da dimensionalidade do banco de dados e da clusterização permitiram que exoplanetas do tipo Mini Netunos fossem agrupados a partir da consideração de que era necessário encontrar grande variabilidade entre os agrupamentos formados, porém, pequena variabilidade dentro deles. Assim, a utilização do algoritmo *K-means* forneceu uma configuração estatisticamente consistente para a formação de quatro grupos ou *clusters*. A análise desses agrupamentos indicou que o grupo 3 seria aquele que poderia conter a maior quantidade de exoplanetas capazes de abrigar algum tipo de vida. Essa conclusão é baseada principalmente na análise do ESI. Ficou claro também, que outros fatores são importantes na investigação das condições de habitabilidade planetária como, por exemplo, a idade estelar, o tipo da estrela hospedeira, a densidade do planeta, a temperatura superficial do planeta e seu período orbital.

CONTRIBUIÇÕES DOS AUTORES

P.S.L.G. contribuiu com a conceituação, análise formal e redação – rascunho original. L.B. contribuiu com a conceituação, supervisão, recursos, curadoria e análise de dados, metodologia e redação – revisão e edição. Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

REFERÊNCIAS

- BERNARDES, L. Machine Learning: uma ferramenta para o estudo de exoplanetas que podem apresentar atmosferas biogênicas. Monografia (Especialização em Data Science e Analytics) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, São Paulo, 2023. p. 30.
- DORN, C.; NOACK, L.; ROZEL, A. Outgassing on stagnant-lid super-Earths. *Astronomy & Astrophysics*, v. 614, p. A18, 2018.
- FÁVERO, L. P.; BELFIORE, P. Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata. Rio de Janeiro: Elsevier, 2017.
- HARTIGAN, P.; WONG, M. A. A K-means clustering algorithm: Algorithm AS 1366. *Applied Statistics*, v. 28, p. 126–130, 1979.

MANFREDI, G. M.; AREIAS, H. S.; PUPO, N. L.; SANTOS, I. F.; VALIO, A. Classification of exoplanets with data mining techniques. *Boletim da Sociedade Astronômica Brasileira*, v. 31, n. 1, p. 21–22, 2020.

NASA. NASA Exoplanet Archive: a service of NASA Exoplanet Science Institute. Disponível em: <https://exoplanetarchive.ipac.caltech.edu/index.html>. Acesso em: 23 jul. 2024.

NOACK, L.; BREUER, D. Plate tectonics on rocky exoplanets: influence of initial conditions and mantle rheology. *Planetary and Space Science*, v. 98, p. 41–49, 2014.

SCHULZE-MAKUCH, D.; MÉNDEZ, A.; FAIRÉN, A. G.; VON PARIS, P.; TURSE, C.; BOYER, G.; DÁVILA, A. F.; RESENDES DE SOUSA ANTÔNIO, M.; IRWIN, L. N.; CATLING, D. A two-tiered approach to assess the habitability of exoplanets. *Astrobiology*, v. 11, n. 10, p. 1041–1052, 2011.

ZENG, L.; SASSELOV, D. D.; JACOBSEN, S. B. Mass-radius relation for rocky planets based on PREM. *The Astrophysical Journal*, v. 819, p. 127, 2016.