Observatório Nacional

DISSERTAÇÃO DE MESTRADO

CARACTERIZAÇÃO DE ESTRELAS HOSPEDEIRAS DE EXOPLANETAS USANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA EM DADOS DOS LEVANTAMENTOS J-PLUS E S-PLUS

ÍCARO MEIDEM

RIO DE JANEIRO
2025

Ministério da Ciência, Tecnologia, Inovações e Comunicações
Observatório Nacional
Programa de Pós-Graduação

Dissertação de Mestrado

CARACTERIZAÇÃO DE ESTRELAS HOSPEDEIRAS DE EXOPLANETAS USANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA EM DADOS DOS LEVANTAMENTOS J-PLUS E S-PLUS

por

Ícaro Meidem

Dissertação submetida ao Corpo Docente do Programa de Pós-graduação em Astronomia do Observatório Nacional, como parte dos requisitos necessários para a obtenção do Grau de Mestre em Astronomia.

Orientador: Dr. Marcelo Borges Fernandes

Rio de Janeiro, RJ – Brasil
Março de 2025

M499 Meidem, Ícaro
Caracterização de Estrelas Hospedeiras de Exoplanetas usando Técnicas de Aprendizagem de Máquina em Dados dos Levantamentos J-PLUS e S-PLUS [Rio de Janeiro] 2025.
xix, 153 p. 29,7 cm:

Dissertação (mestrado) - Observatório Nacional - Rio de Janeiro, 2025.

1. Aprendizado de Máquina. 2. *Random Forest*. 3. *XGBoost*. 4. Parâmetros estelares. 5. Exoplanetas. I. Observatório Nacional. II. Título.

CDU 000.000.000

“CARACTERIZAÇÃO DE ESTRELAS HOSPEDEIRAS DE EXOPLANETAS USANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA EM DADOS DOS LEVANTAMENTOS J-PLUS E S-PLUS”

ÍCARO MEIDEM

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ASTRONOMIA DO OBSERVATÓRIO NACIONAL COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ASTRONOMIA.

Aprovada por:

Dr. Marcelo Borges Fernandes – Observatório Nacional (ON/MCTI)
(Orientador)

Dra. Adriana B. M. Valio – Universidade Presbiteriana Mackenzie (UPM)

Dr. Hélio D. Perottoni – Observatório Nacional (ON/MCTI)

RIO DE JANEIRO, RJ – BRASIL
11 DE MARÇO DE 2025

À todos aqueles que
(des)acreditaram.

Agradecimentos

Agradeço, primeiramente, aos meus pais. Mais do que especial ao meu pai José Marcos que nunca desistiu de mim, que sempre me apoiou nas minhas decisões, minha escolha de futuro, por ser meu pai, meu amigo, meu companheiro e por ser o maior exemplo de homem guerreiro e bondoso que eu já conheci. Eu prometo sempre lhe orgulhar e nunca decepcioná-lo! À minha mãe Katia Cilene pela ajuda para me manter no Rio de Janeiro e por ser minha mãe. Tenho muito orgulho de ser filho de vocês.

Agradeço a minha namorada Natália, imensamente por ser essa mulher maravilhosa na minha vida, por nunca me abandonar, mesmo com meus erros, por estar presente mesmo na distância e por sempre estar disposta em me ajudar no que diz respeito a tudo na minha vida.

Agradecimento especial ao meu orientador Marcelo Borges Fernandes, Lethycia de Carvalho e Vinícius Cordeiro que me ajudaram nos meus primeiros passos no *machine learning* e nesse projeto como um todo. Sem eles esse projeto não sairia do papel.

Agradeço aos professores do Observatório Nacional pelos conhecimentos passados durantes as disciplinas e conversas nos corredores.

Agradeço aos funcionários do ON pelo apoio e serviços prestados à instituição. Principalmente para Giane que é uma mãe para nós no prédio, para Mari e Ana que não deixam faltar café na copa pra gente. Muitíssimo obrigado!

Agradeço a cada um dos meus amigos que fiz no ON. Alessandro, Anthony (famoso Toninho), Maria Eduarda, Marie, Raquel e Vinícius Bessa (o Besser) pelas resenhas, cafezinhos na copa, pelo vôlei fim de semana, apoios e conselhos, vou levá-los para a vida.

Agradecimento especial, de coração, aos meus amigos de sala do ON. Romualdo, que me ajudou em um momento difícil na cidade do Rio, e mais do que especial, ao Patrick e Vinícius Sanches, colegas de casa que são como irmãos. Agradeço pelas suas ajudas foram essenciais para que eu continuasse no ON. Muitíssimo obrigado por tudo!

À banca examinadora por aceitar meu convite, por contribuir para nosso trabalho e à minha formação pessoal e profissional.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, quem agradeço a bolsa de mestrado concedida.

"Não me sigam
Que eu também estou perdido.
Ou façam ou descubram o próprio
caminho...
Façam como eu: Inventem!
Ou melhor, não façam como eu:
Inventem!".

Belchior (1946 - 2017)

Ícaro Meidem

CARACTERIZAÇÃO DE ESTRELAS HOSPEDEIRAS DE EXOPLANETAS USANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA EM DADOS DOS LEVANTAMENTOS J-PLUS E S-PLUS

RESUMO

Desde a descoberta do primeiro exoplaneta em 1995, orbitando uma estrela do tipo solar, mais de 5.867 exoplanetas foram identificados, segundo o NASA Exoplanet Archive. Sabendo que os parâmetros planetários dependem diretamente dos parâmetros estelares, este trabalho teve como objetivo aprimorar a caracterização de estrelas hospedeiras de exoplanetas observadas pelos levantamentos fotométricos J-PLUS e S-PLUS, utilizando técnicas de aprendizado de máquina, com foco nos algoritmos *Random Forest* e *XGBoost*. Foram utilizados nos treinamentos dos modelos, dados fotométricos dos 12 filtros ópticos do J-PLUS e S-PLUS em conjunto com os levantamentos LAMOST e APOGEE, para prever os parâmetros como temperatura efetiva (T_{ef}), gravidade superficial (log g) e metalicidade ([Fe/H]). Os modelos foram aplicados nas estrelas hospedeiras presentes nos catálogos das missões Kepler e TESS e do Espectrógrafo HARPS. Os resultados encontrados destacaram a eficácia do *Random Forest*, que apresentou maior precisão do que o *XGBoost*, principalmente para T_{ef} e log g. A significativa redução nos erros, quando comparados aos valores da literatura, reafirma a viabilidade de combinar aprendizado de máquina com um sistema fotométrico que combina filtros de banda larga, intermediária e estreita para obter estimativas precisas de parâmetros estelares. Além da previsão dos parâmetros, foram derivadas propriedades fundamentais das estrelas, como luminosidade, magnitude absoluta, magnitude bolométrica, massas e raios estelares. Esses resultados foram utilizados para caracterizar exoplanetas que possuem valores de trânsito planetário presentes nos catálogos do Kepler e do TESS em campos comuns com os do J-PLUS e S-PLUS. Este estudo demonstra que as técnicas de aprendizado de máquina não apenas melhoram a caracterização das estrelas hospedeiras, mas também abrem novas possibilidades para identificar exoplanetas e outros objetos de interesse, como binárias eclipsantes e anãs marrons. Por fim, os resultados encontrados são promissores para análises futuras, especialmente com a aplicação desta metodologia em outros levantamentos, como o J-PAS, que possui um maior número de filtros e potencial para resultados ainda mais precisos.

Ícaro Meidem

CHARACTERIZATION OF EXOPLANET HOST STARS USING MACHINE LEARNING TECHNIQUES ON J-PLUS AND S-PLUS SURVEY DATA

ABSTRACT

Since the discovery of the first exoplanet in 1995 orbiting a Sun-like star, more than 5,867 exoplanets have been identified, according to the NASA Exoplanet Archive. Knowing that planetary parameters depend directly on stellar parameters, this work aimed to improve the characterization of exoplanet host stars observed by the J-PLUS and S-PLUS photometric surveys using machine learning techniques, with a focus on the Random Forest and XGBoost algorithms. The training of the models used photometric data from the 12 optical filters of J-PLUS and S-PLUS combined with data from the LAMOST and APOGEE surveys to predict stellar parameters such as effective temperature (T_{eff}), surface gravity (log g), and metallicity ([Fe/H]). The models were applied to host stars listed in the catalogs of the Kepler and TESS missions and the HARPS spectrograph. The results highlighted the effectiveness of the Random Forest algorithm, which showed greater accuracy than XGBoost, particularly for T_{eff} and log g. The significant reduction in errors, when compared to literature values, reaffirms the viability of combining machine learning with a photometric system that integrates broad, intermediate, and narrow-band filters to obtain precise estimates of stellar parameters. In addition to predicting these parameters, fundamental stellar properties such as luminosity, absolute magnitude, bolometric magnitude, mass, and radius were derived. These results were used to characterize exoplanets with transit values available in the Kepler and TESS catalogs in fields overlapping with J-PLUS and S-PLUS. This study demonstrates that machine learning techniques not only enhance the characterization of host stars but also open new possibilities for identifying exoplanets and other objects of interest, such as eclipsing binaries and brown dwarfs. Finally, the findings are promising for future analyses, especially with the application of this methodology to other surveys like J-PAS, which features a larger number of filters and the potential for even more accurate results.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

> Abri asas confiantes no espaço e eleveime em direção ao infinito, (...) eu vi que o sol era só outra estrela e que as estrelas eram outros sois. Cada um deles acompanhado por outras terras como a nossa, a revelação dessa imensidão foi como se apaixonar.
>
> *Visão de Giordano Bruno (1548–1600) transcrita da Série Cosmos: A Spacetime Odyssey (2014).*

A indagação da existência de múltiplos mundos não é recente, os primeiros registros revelam que esse pensamento vem desde entre os séculos V e III a.C., nos quais os atomistas[1] já especulavam sobre a possibilidade de outros mundos em sistemas planetários distintos e distantes do nosso (Warren, 2004). Na Idade Moderna, Giordano Bruno (1548-1600) foi pioneiro ao sugerir publicamente que as estrelas eram sóis distantes, com seus próprios planetas e possivelmente capazes de abrigar vida. Em 17 de fevereiro de 1600, no Campo das Flores, em Roma, Bruno foi condenado à morte pela Inquisição, sacrificando sua vida em defesa de suas ideias sobre a existência de mundos infinitos.

As primeiras observações científicas com o propósito específico de detectar exoplanetas (também chamados de planetas extrassolares) aconteceram no final da década de 80, utilizando como técnica principal o método de Velocidade Radial, para encontrar possíveis exoplanetas nas estrelas *Chi1 Ori* e *Gamma Cep* (Campbell *et al.*, 1988). Porém, esses objetos só foram definitivamente confirmados em 1992, quando foram detectados

[1] O pensamento filosófico no qual postula que a matéria é composta por partículas indivisíveis chamadas átomos, que se combinam de diversas maneiras para formar todas as substâncias e fenômenos observados na natureza.

2 exoplanetas orbitando o pulsar PSR 1257 + 12, denominados como PSR 1257+12B e PSR 1257+12C Wolszczan & Frail (1992). O primeiro com pelo menos 2,8 $M_{\oplus}$(massas terrestres) e o segundo, com 3,4 $M_{\oplus}$ (Wolszczan & Frail, 1992). Um terceiro, denominado como PSR 1257+12 A, foi descoberto mais tarde em 1994 por Wolszczan (1994), com massa de pelo menos 2 vezes a massa da Terra.

Alguns anos depois, em 1995, foi descoberto o primeiro exoplaneta orbitando uma estrela do tipo solar, a 51 Pegasi, sendo ele chamado de 51 Pegasi-b (Mayor & Queloz, 1995). Esse planeta é um gigante gasoso pertencente à classe conhecida como Jupiters Quentes (*Hot Jupiters*)[2]. Essa descoberta rendeu o Prêmio Nobel de Física de 2019 aos seus descobridores (Nobel Prize Outreach AB, 2019).

Desde então, inúmeros exoplanetas foram descobertos, revelando uma diversidade impressionante de tamanhos, composições e órbitas. Esse crescente número de descobertas tornou evidente a alta probabilidade desses objetos serem tão comuns quanto as estrelas no universo. Atualmente, já são 5.867 exoplanetas confirmados[3].

1.1 Técnicas de Detecção de Exoplanetas

Diversas técnicas são utilizadas para detectar exoplanetas, cada uma com métodos específicos e características próprias. A primeira delas é através da Astrometria (Quirrenbach, 2010), que consiste em medir a posição de uma estrela em relação ao fundo estelar, observando deslocamentos causados pela perturbação gravitacional de um planeta. Essa técnica, no entanto, é limitada a estrelas próximas da Terra devido à dificuldade de medir deslocamentos muito pequenos a grandes distâncias.

Outra técnica importante é o Imageamento Direto (Traub & Oppenheimer, 2010), que, como o nome sugere, registra imagens diretas do exoplaneta. Essa abordagem é extremamente útil para identificar exoplanetas, especialmente gigantes gasosos, mas enfrenta como principal desafio o brilho excessivo da estrela hospedeira, que frequentemente ofusca o planeta, já que eles estão muito próximos entre si no campo de visão. Além disso, há a técnica de Microlentes Gravitacionais (Gaudi, 2010), que detecta exoplanetas por meio do efeito de lente gravitacional gerado por um objeto massivo, que passa em frente a outra estrela. Nesse caso, a luz da estrela que está atrás é amplificada primeiro pela estrela que passa na frente e atua como lente e, posteriormente, é a vez do exoplaneta atuar com uma outra lente, amplificando novamente a luz da estrela que está atrás. A principal limitação dessa técnica é a raridade e a brevidade dos eventos observados.

A Cronometria de Tempo de Pulsar (ou *Pulsar Time Variations*) foi a primeira técnica

[2]Constituem uma classe de exoplanetas com massas comparáveis ou superiores à de Júpiter. Sua característica distinta é a órbita extremamente próxima de suas estrelas hospedeiras, em comparação com os planetas gigantes do nosso Sistema Solar. Esses planetas têm períodos orbitais curtos, resultando em altas temperaturas em suas atmosferas devido à proximidade com as estrelas.

[3]Recuperado de https://exoplanetarchive.ipac.caltech.edu/ em 10 abr. de 2025.

que permitiu a detectar um exoplaneta (Wolszczan & Frail, 1992) e consiste na medição do tempo de chegada dos pulsos emitidos por pulsares, utilizando radiotelescópios. Pulsares são estrelas de nêutrons em rotação, que emitem pulsos de rádio com extrema regularidade na direção do seu eixo magnético. Essa radiação é detectada sempre que o feixe estiver apontando na direção da Terra, geralmente com períodos da ordem de milissegundos. Devido a essa regularidade, mesmo desvios sutis nos tempos de chegada dos pulsos podem indicar alterações no movimento da estrela e esse movimento orbital provoca variações periódicas nos pulsos recebidos na Terra, resultado da mudança na velocidade radial do pulsar em relação ao observador. Com isso, a partir dessas variações temporais, é possível inferir os parâmetros orbitais do planeta que orbita essa estrela, incluindo sua massa mínima e período orbital.

As técnicas mais consolidadas e amplamente utilizadas na detecção e caracterização de exoplanetas são a de Trânsito Planetário e a de Velocidade Radial, que juntas correspondem a cerca de 94% do número total de detecções, de acordo com o NASA Exoplanet Archive (as cores roxa e verde na Figura 1.1). A técnica de trânsito consiste em observar a variação de brilho da estrela quando um planeta passa na linha de visada entre o observador e a estrela. Já a técnica de velocidade radial mede, por meio da espectroscopia, o movimento da estrela causado pela influência gravitacional do planeta através de deslocamentos na posição de linhas espectrais. Essas duas técnicas, quando combinadas, permitem inferir parâmetros como raio, período orbital, massa, densidade do planeta, entre outros. Por serem as mais relevantes para o trabalho, ambas serão exploradas com mais detalhes nas seções 1.1.1 e 1.1.2.

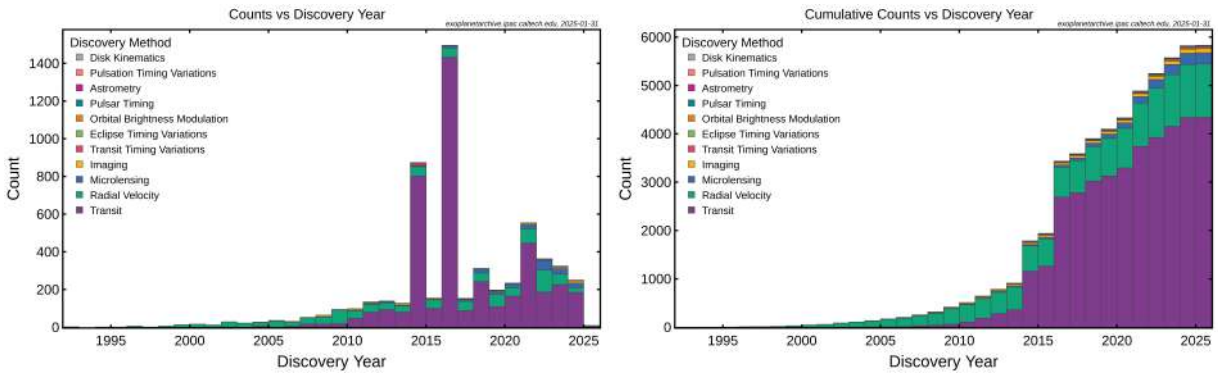

Figura 1.1: Número de exoplanetas detectados por ano de acordo com o método utilizado. À esquerda: Número de detecções realizadas por cada técnica por ano, representadas por cores diferentes (ver a legenda na imagem). À direita: Número de detecções acumuladas ao longo dos anos. Note que as principais técnicas utilizadas são a de Trânsito e a de Velocidade Radial, respectivamente. Fonte: NASA Exoplanet Archive, https://exoplanetarchive.ipac.caltech.edu/exoplanetplots/.

1.1.1 Método de Trânsito Planetário

A técnica de trânsito planetário é a principal técnica utilizada na detecção de exoplanetas e outros sistemas planetários. Sozinha representa cerca de 74% do total de detecções com 4.360 objetos já confirmados[4]. O método consiste no monitoramento do brilho da estrela quando um planeta passa em frente a ela, bloqueando parte do seu brilho temporariamente.

Este método tem suas origens nos estudos realizados entre os séculos XVIII e XIX, durante os trânsitos dos planetas Mercúrio e Vênus pelo disco solar. O objetivo desses estudos era determinar a distância entre a Terra e o Sol por meio de trigonometria. De maneira geral, as estrelas que não sofrem obscuração por trânsito apresentam um brilho relativamente constante. No entanto, isso não ocorre em casos específicos, como nas estrelas variáveis, que apresentam variações em seu brilho intrínseco ao longo do tempo. As oscilações de brilho observadas durante o trânsito se manifestam como mínimos na curva de luz da estrela (Perryman, 2018). Além disso, é importante destacar que nem todos os exoplanetas apresentam trânsitos, uma vez que a geometria favorável depende do alinhamento do plano orbital com a linha de visada.

Podemos ilustrar a situação através da Figura 1.2 que mostra, fora de escala, como acontece o trânsito.

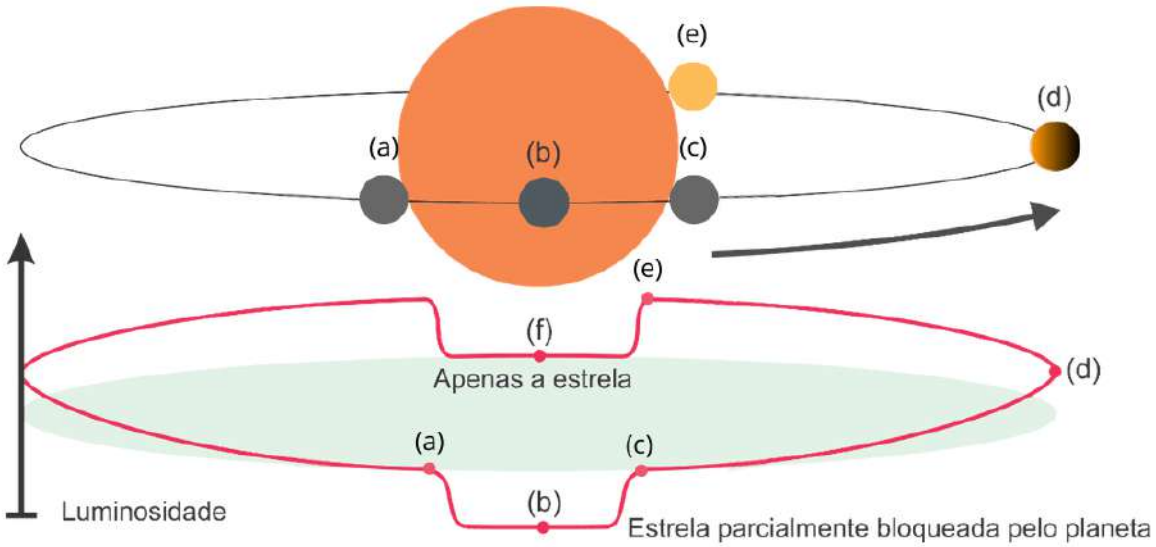

Figura 1.2: Esquema fora de escala ilustrando a duração do trânsito do planeta através do disco estelar e o comportamento do brilho da estrela hospedeira. As letras representam as posições do exoplaneta na órbita em torno de sua estrela hospedeira. a) é a posição de início do trânsito, b) o ponto médio do trânsito, c) o fim do trânsito, o ponto d) a estrela começa a iluminar a face voltada para o observador e o ponto e) é o inicio do eclipse secundário. Fonte: Meidem (2022). Adaptado pelo autor.

[4]Recuperado de https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html em 10 abr. de 2025.

O eclipse primário é caracterizado no momento que o planeta (esfera menor) se encontra na posição (b) da Figura 1.2, no qual o brilho da estrela hospedeira (esfera laranja maior) atinge seu mínimo. As posições (a) e (c) representam o início e fim do trânsito, respectivamente. A medida que o planeta vai se movendo da posição (c) até a (d) o seu brilho refletido vai aumentando, ate finalmente chegar na posição (e) onde é o inicio do eclipse secundário, no qual, é possível medir o espectro térmico do planeta. Para a análise da atmosfera são obtidos espectros de transmissão nos pontos (a) e (c). Por fim, na posição (f) o planeta não pode ser mais visto, portanto o brilho registrado é apenas da hospedeira, ocorrendo assim o eclipse secundário.

O método de trânsito não serve apenas para detectar exoplanetas, mas também para identificar anãs marrons e outras estrelas companheiras (caracterizando um sistema binário ou múltiplo eclipsante). Como são objetos maiores, eles podem caracterizar uma queda mais acentuada no brilho da estrela. Podem ocorrer variações do brilho que são causadas por efeitos sistemáticos e não necessariamente por um objeto real, que são denominados como falsos trânsitos ou falso-positivos. Esses efeitos podem ser causados por píxels mortos (aqueles que não apresentam nenhuma contagem). Além dos falso-positivos, há também os falso-negativos nos quais são estrelas que possuem objetos em órbita, mas sua inclinação orbital não permite que ele passe diante da estrela na linha de visada do observador. Isso faz parecer que não existe nenhum objeto em órbita (e em dado momento, em trânsito).

O ponto de mínimo dessa curva, ou seja, o local onde ocorre a maior diminuição no fluxo sinalizado como o ponto (b) na Figura 1.2, representa a profundidade máxima do trânsito (δ). Em geral, essa profundidade máxima está associada ao instante em que o objeto se encontra mais próximo do centro do disco estelar, região de maior fluxo observado. Nas bordas da estrela, o fluxo vindo dela diminui, um fenômeno explicado por dois fatores principais: profundidade óptica e temperatura. Esse efeito é denominado escurecimento de borda (em inglês: *limb darkening*).

De acordo com Nogueira (2020), em um cenário ideal, a forma da curva de luz durante um trânsito planetário se assemelharia a um trapézio isósceles, como ilustrado na Figura 1.3, onde o brilho observado da estrela diminuiria e aumentaria de forma linear à medida que o planeta ingressa e sai do disco estelar aparente, apresentando um segmento central relativamente plano enquanto o planeta estiver completamente à frente da estrela. No entanto, o autor também aponta que, na prática, essa forma idealizada sofre alterações, especialmente em faixas espectrais mais azuladas, devido ao efeito conhecido como escurecimento de borda. Esse fenômeno ocorre porque, nas bordas do disco estelar, a radiação detectada provém de camadas atmosféricas mais externas e frias da estrela, ao passo que, no centro do disco, a mesma profundidade óptica revela camadas mais internas e quentes (Nogueira, 2020).

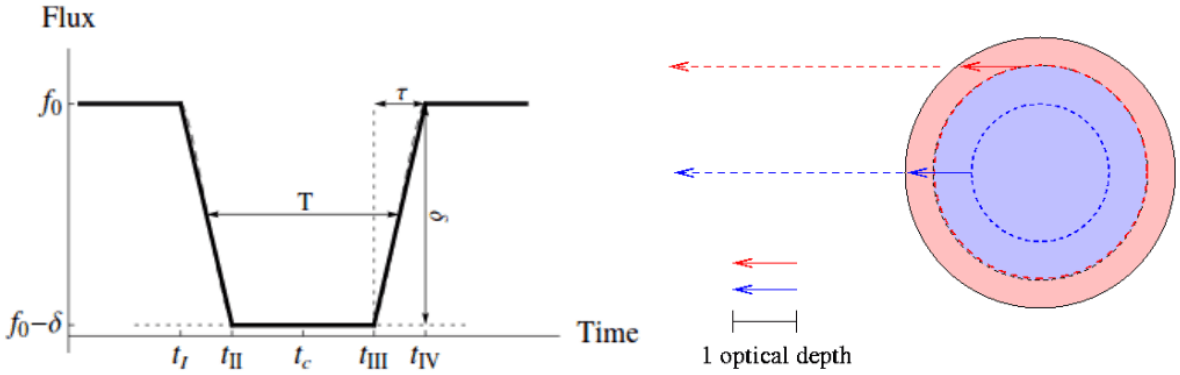

Figura 1.3: Esquerda: Curva de luz ideal sem o efeito de escurecimento de borda. Direita: Nas bordas do disco estelar, a profundidade óptica corresponde a camadas mais superficiais e frias, emitindo menor fluxo. No centro, a mesma profundidade óptica se refere a camadas mais profundas e quentes, resultando em maior fluxo. Fonte: Nogueira (2020)

Assim, o observador vê um menor fluxo proveniente das bordas em relação ao centro, caracterizando o escurecimento do limbo. Este efeito contribui para moldar os trânsitos observados, arredondando a curva de luz e aproximando seu formato ao de um "U" vistos em curvas de luz reais.

1.1.2 Método de Velocidade Radial

A técnica de velocidade radial é a segunda técnica mais bem sucedida utilizada na caracterização e detecção de exoplanetas e foi a primeira a ser utilizada para esse fim. Ela representa 19% com 1.112 objetos confirmados[5].

Considerando que tanto a estrela quanto o planeta orbitam um centro de massa comum (indicado na Figura 1.4 como x), e assumindo que o CM é fixo em relação ao observador, à medida que o sistema órbita esse centro, a estrela se aproxima e se afasta dele. Esse movimento causa, graças ao efeito Doppler, um deslocamento do comprimento de onda das linhas espectrais presentes no espectro estelar, resultando em um desvio para o azul (*Blueshift*), quando a estrela se aproxima e um desvio para o vermelho (*Redshift*), quando ela se afasta.

[5]Recuperado de https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html em 12 de mar. de 2025.

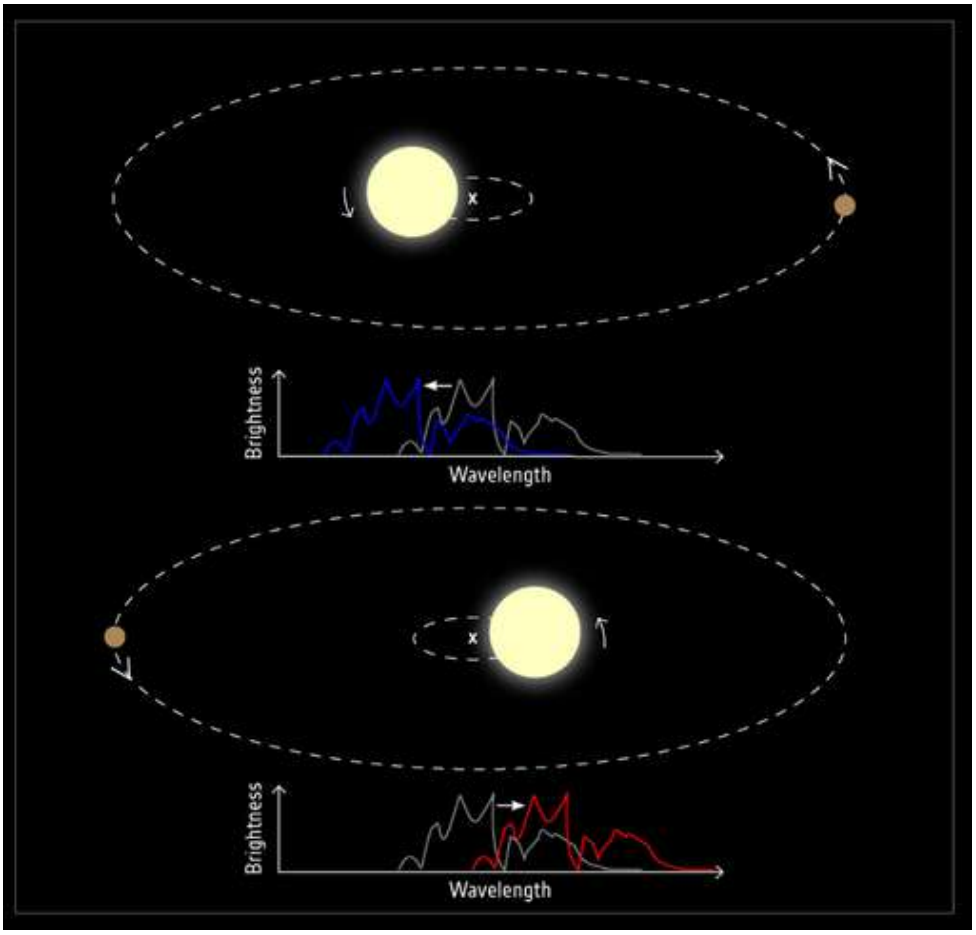

Figura 1.4: Esquema fora de escala ilustrando o comportamento da estrela e o deslocamento de seu espectro devido ao movimento ao redor do centro de massa do sistema estrela-planeta (representado pelo x). O espectro em cinza representa o espectro com as linhas espectrais nos comprimentos de onda de laboratório, o espectro azul representa o deslocamento do espectro para o azul com a estrela se aproximando do observador e o espectro vermelho representa o deslocamento para o vermelho com a estrela se afastando. Créditos: ESA, 2019. Adaptado pelo autor.

A análise pode ser extrapolada para os casos em que o centro de massa se move com relação ao observador e as oscilações provocadas no espectro (deslocamento do comprimento de onda das linhas espectrais) constituem indícios de que a estrela possui um corpo orbitando-a cuja massa é significativa, visto que é capaz de produzir oscilações mensuráveis na posição da estrela hospedeira (Lovis & Fischer, 2010).

A técnica baseia-se na detecção da velocidade da movimentação da estrela em torno do CM, sendo particularmente eficaz para gigantes gasosos orbitando muito próximos de sua estrela hospedeira, conhecidos como Jupiters Quentes. No entanto, planetas com massas inferiores à de Júpiter, aproximadamente $9,5 \times 10^{-4} M_{\odot}$, apresentam maior dificuldade de detecção por meio do método de velocidade radial, uma vez que provocam oscilações muito pequenas em sua estrela hospedeira (Perryman, 2000; Válio, 2009).

Soto (2020) alerta que a velocidade de rotação da estrela, assim como fenômenos

estelares intrínsecos, como manchas solares, pulsações e variações causadas por contrações, podem prejudicar as medições de velocidade radial. Essas características podem não apenas mascarar a presença de um planeta, mas também gerar sinais falsos que imitam a existência de um corpo em órbita da estrela. Se a rotação for muito alta, ocorre um alargamento das linhas de absorção, dificultando a medição, especialmente em sistemas onde a velocidade radial é da ordem de apenas alguns metros por segundo.

Além disso, outra limitação que pode prejudicar as medidas, mencionada por Soto (2020), está relacionada ao tipo espectral e à atividade da estrela. Essa atividade refere-se a fenômenos como manchas estelares, flares, oscilações e variações cromosféricas, que podem induzir sinais periódicos no espectro da estrela, gerando perturbações na medição da velocidade radial. Esses sinais podem apresentar amplitudes comparáveis às causadas por um planeta, tornando difícil distinguir entre variações causadas por fenômenos estelares intrínsecos e aquelas provocadas por um corpo em órbita. Como muitas dessas atividades possuem periodicidade similar à rotação estelar, elas podem mascarar a presença de um exoplaneta ou até mesmo simular sua existência.

1.2 Dependência entre Parâmetros Planetários e Estelares

Para compreender as características dos exoplanetas, é fundamental determinar com precisão os parâmetros físicos das estrelas hospedeiras, como temperatura, luminosidade, massa e raio. Esses parâmetros são essenciais para calcularmos as propriedades do planeta, como seu raio e sua massa.

1.2.1 Raio Estelar e Raio Planetário

A partir da profundidade de trânsito, inferida através do método de trânsito planetário, é possível determinar o raio do planeta (ou corpo que orbita a estrela hospedeira). É importante atentar-se que alguns trânsitos presentes nos dados da Missão Kepler (Borucki, 2016) ou em outros levantamentos não foram observados de maneira contínua. Sem observações contínuas, a profundidade registrada pode não corresponder ao ponto de maior redução no fluxo estelar, prejudicando a precisão dos cálculos do raio planetário. Portanto, o que pode ser inferido durante as medições é a profundidade observada (Q), que pode ser exemplificada na Figura 1.5.

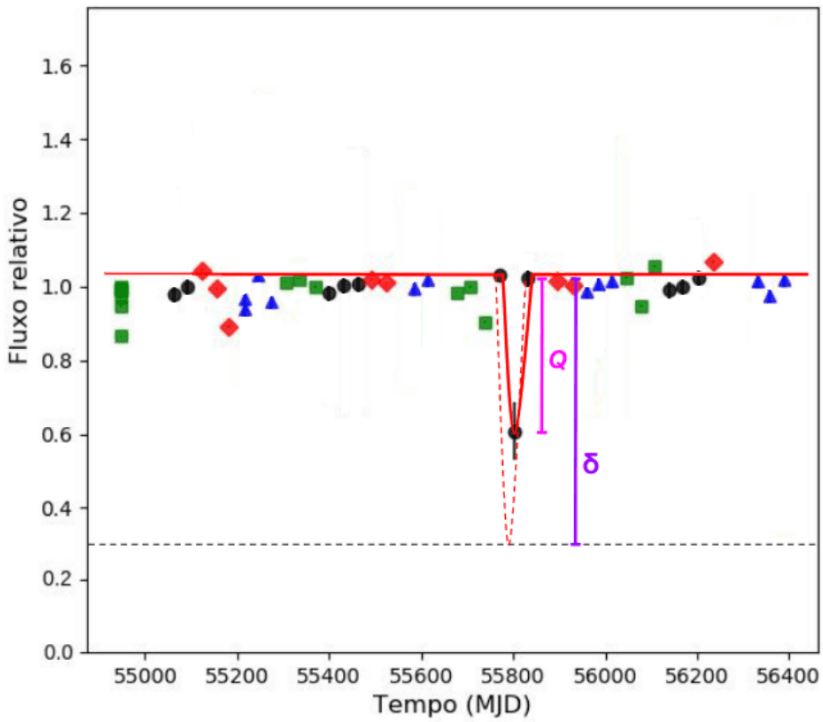

Figura 1.5: Curva de luz da estrela KIC 8561192 com diferentes orientações do telescópio Kepler. Os quadrados verdes, triângulos azuis, círculos pretos e losangos vermelhos correspondem às quatro orientações da missão Kepler (ver a Seção 1.2.1 de Carvalho (2022)). A curva sólida em vermelho representa a aproximação da curva de luz calculada por Nogueira (2020), enquanto a curva pontilhada em vermelho refere-se à curva de luz contínua dessa estrela, conforme o Catálogo de Binárias Eclipsantes do Kepler (Kirk *et al.*, 2016). A profundidade máxima observada do trânsito (Q) é destacada, como sendo uma variável representativa, já que $Q \leqslant \delta$. Como a profundidade real do trânsito não pode ser menor que Q, esta variável corresponde ao valor mínimo da profundidade real. Fonte: Carvalho (2022).

Na Figura 1.5 os quadrados verdes, triângulos azuis, círculos pretos e losangos vermelhos correspondem às 4 diferentes orientações do telescópio Kepler, as quais não serão tratadas neste trabalho[6]. A curva sólida em vermelho representa a aproximação da curva de luz calculada por Nogueira (2020) e a curva pontilhada em vermelho se refere a curva de luz contínua desta estrela, segundo o Catálogo de Binárias Eclipsantes do Kepler (Kirk *et al.*, 2016). A profundidade observada, que pode ser chamada de profundidade máxima observada do trânsito (Q), é uma alusão ao valor que ela representa. Considerando que $Q \leqslant \delta$ e que a profundidade real do trânsito não pode ser menor que Q, essa variável representará o valor mínimo da profundidade, conforme evidenciado por Carvalho (2022) e Nogueira (2020) em seus trabalhos.

Para calcular Q é necessário utilizar a Equação 1.1, onde o F é o fluxo:

[6]Consulte os trabalhos de Carvalho (2022) e Nogueira (2020) para melhor entendimento sobre essas orientações do Telescópio Espacial Kepler.

$$Q = \left(\frac{F_A + F_D}{2}\right) - F_T \tag{1.1}$$

Onde os prefixos indicam os fluxos relativos observados imediatamente antes (F_A) e logo após o trânsito (F_D) e o fluxo relativo mínimo associado ao possível trânsito (F_T).

Se o trânsito for observado de forma ideal, ou seja, medindo momentos antes, durante e após o eclipse, a profundidade Q será igual à profundidade real do trânsito (δ). É de extrema importância conhecer Q quando se quer conhecer o raio do objeto (R_p) causador do trânsito, neste caso, um exoplaneta. E tão importante quanto é conhecer o raio da estrela ($R_\star$) no qual ele orbita. Para isso, este trabalho relata a relevância de adquirir com precisão um valor confiável para o raio estelar.

Para entender-se melhor essa importância e a relação entre os parâmetros, a Equação 1.2 mostra a relação existente entre a profundidade do trânsito, o raio do planeta e o raio da estrela hospedeira, da qual é notório que, para dois objetos que provocam um mesmo Q, quanto maior $R_\star$, maior R_p. Vide:

$$\frac{\Delta F}{F} = Q = \left(\frac{R_p}{R_\star}\right)^2 \rightarrow R_p = Q^{1/2} R_\star \tag{1.2}$$

As incertezas de R_p podem ser propagadas da seguinte forma:

$$\sigma_{R_p} = \sqrt{\left(\frac{Q^{-1/2}\sigma_Q R_\star}{2}\right)^2 + (Q^{1/2}\sigma_{R_\star})^2} \tag{1.3}$$

Se a profundidade do trânsito for a ideal, no caso, a real (δ), pode-se reescrever a Equação 1.2 e encontrar o raio do planeta (R_p) causador da queda do fluxo da seguinte forma:

$$R_p = \delta^{1/2} R_\star \tag{1.4}$$

A partir da Lei de Stefan-Boltzmann, onde a luminosidade da estrela depende da temperatura (T_{ef}) e do raio estelar ($R_\star$). Temos:

$$L = 4\pi R_\star^2 \sigma T_{ef}^4 \rightarrow R_\star = \sqrt{\frac{L}{4\pi\sigma T_{ef}^4}} \tag{1.5}$$

onde σ é a constante de Stefan-Boltzmann, que tem o valor de $5,6697 \times 10^{-5}$ erg cm^{-2} s^{-1} K^4. Sabendo sua luminosidade (L) e temperatura (T_{ef}), podemos calcular seu raio estelar ($R_\star$), logo, é possível calcular o raio planetário (R_p).

1.2.2 Massa Estelar e Massa Planetária

Há algumas formas de estimar a massa de uma estrela. Como por exemplo, através do diagrama de Hertzsprung-Russell (HR), das trilhas evolutivas (en. *evolutionary tracks*) e de isócronas comparando suas propriedades observacionais com modelos teóricos. O diagrama HR relaciona luminosidade e temperatura, as trilhas evolutivas mostram como as estrelas evoluem ao longo do tempo para diferentes massas, e as isócronas representam estrelas de diferentes massas que têm a mesma idade, sendo úteis para a estimativa da idade de aglomerados estelares. E também, através de sua luminosidade, que pode ser determinada a partir da magnitude aparente, da correção bolométrica e da distância.

A massa do planeta (M_p), depende diretamente da massa da estrela $M_\star$ e para obter-la, é preciso determinar a variação da velocidade radial. Para isso, é necessário partir da dinâmica Kepleriana, considerando órbitas elípticas, com o centro de massa localizado em um dos focos e que tanto a energia quanto o momento angular permanecem constantes durante o movimento. Como a velocidade radial de um corpo, visto da Terra, varia com o movimento orbital, é comumente utilizada a semi-amplitude da variação da velocidade radial (K). Considerando a Terceira Lei de Kepler e como descrito em Lovis & Fischer (2010) e Murray & Correia (2010)), temos:

$$K = \sqrt{\frac{G}{1-e^2}} m_2 \sin i (m_1 + m_2)^{-1/2} a^{-1/2} \tag{1.6}$$

Reescrevendo em unidades utilizadas na astronomia, tem-se:

$$K = \frac{28.4329 m s^{-1}}{\sqrt{1-e^2}} \frac{m_2 \sin i}{M_{Jup}} \left(\frac{m_1 + m_2}{M_\odot}\right)^{-2/3} \left(\frac{a}{1\mathrm{AU}}\right)^{-1/2} \tag{1.7}$$

onde m_1 é a massa da estrela (em $M_\odot$), m_2 é a massa do corpo que orbita a estrela (companheira ou planeta, em M_{Jup}), a é o semi-eixo maior da órbita desse corpo, em AU, i é o ângulo de inclinação do plano orbital com relação a linha de visada, e e é a excentricidade da orbita. Reescrevendo, a Equação 1.7 em termos do período orbital do planeta (P), em anos, e considerando que a massa da estrela (m_1 ou $M_\star$) é muito maior do que a massa do planeta (m_2 ou M_p), tem-se que $m_1 + m_2 = m_1 = M_\star$. Com isso:

$$K = \frac{28.4329 m s^{-1}}{\sqrt{1-e^2}} \frac{M_p \sin i}{M_{Jup}} \left(\frac{M_\star}{M_\odot}\right)^{-2/3} \left(\frac{P}{1\mathrm{yr}}\right)^{-1/3} \tag{1.8}$$

Na Figura 1.6 podemos observar a variação da semi amplitude da velocidade radial medida para a estrela 51 Peg, que é orbitada por um planeta. A curva foi ajustada para um período de 4,23 dias, obtida por Mayor & Queloz (1995). O sinal detectado é resultado da presença de um companheiro orbital com massa mínima de 0,47 M_{Jup}, marcando a primeira observação de um exoplaneta em órbita de uma estrela de tipo solar.

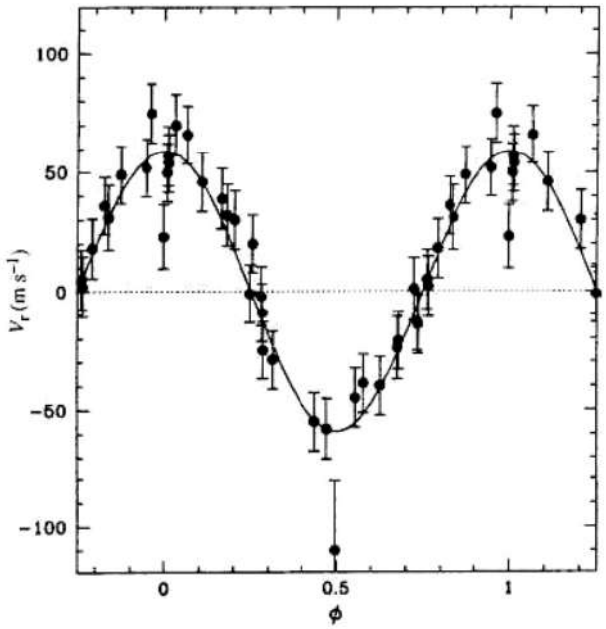

Figura 1.6: Curva original de semi amplitude da velocidade radial da estrela 51 Peg, ajustada para um período de 4,23 dias, obtida por Mayor & Queloz (1995).

Para obter M_p, é necessário conhecer a inclinação i do plano orbital do exoplaneta. Em sistemas onde se consegue observar o trânsito do planeta, i pode ser aproximadamente igual a 90°, e $M_p \sin i$ será aproximadamente M_p. Em outros casos, a massa mínima é a melhor estimativa que se pode obter com a técnica de velocidade radial. Esse método requer observações precisas ao longo de várias órbitas para obter o valor de K, e também uma estimativa precisa de $M_\star$, geralmente baseada em modelos de evolução estelar e observações astrométricas da estrela.

Portanto isolando a massa mínima em 1.7, tem-se:

$$M_p = M_p \sin i = \frac{K M_\star^{2/3} P^{1/3} \sqrt{1-e^2}}{28,4329} \tag{1.9}$$

Onde M_p é dado em M_{Jup}.

Fazendo a propagação de incertezas, levando em consideração as incertezas de todos os parâmetros, tem-se que:

$$\sigma M_p = \sqrt{\left(\frac{\partial(M_p)}{\partial K}\sigma_K\right)^2 + \left(\frac{\partial(M_p)}{\partial M_\star}\sigma_{M_\star}\right)^2 + \left(\frac{\partial(M_p)}{\partial P}\sigma_P\right)^2 + \left(\frac{\partial(M_p)}{\partial e}\sigma_e\right)^2} \tag{1.10}$$

onde:

$$\frac{\partial(M_p)}{\partial K} = \frac{M_\star^{2/3} P^{1/3}\sqrt{1-e^2}}{28,4329}, \quad \frac{\partial(M_p)}{\partial M_\star} = \frac{2K M_\star^{-1/3} P^{1/3}\sqrt{1-e^2}}{3 \times 28,4329}, \tag{1.11}$$

$$\frac{\partial(M_p)}{\partial P} = \frac{K M_\star^{2/3} P^{-2/3}\sqrt{1-e^2}}{3 \times 28,4329}, \quad \frac{\partial(M_p)}{\partial_e} = \frac{K M_\star^{2/3} P^{1/3} e}{28,4329\sqrt{1-e^2}} \tag{1.12}$$

Sendo σ_K, $\sigma_{M_\star}$, σ_P e σ_e os erros da semi amplitude da velocidade radial, da massa da estrela, do período orbital do planeta e da excentricidade, respectivamente. Utilizando os métodos de trânsito e o de velocidade radial combinados, é possível inferir com maiores precisões a massa, o raio e o período orbital do planeta. A massa de um planeta desempenha um papel crucial em sua capacidade de manter uma atmosfera, uma vez que massas muito baixas resultam em uma gravidade demasiadamente fraca. Em planetas desprovidos de atmosfera, a evolução de formas de vida complexas na superfície é improvável, e a impossibilidade da existência de água em estado líquido. Por outro lado, planetas excessivamente massivos podem apresentar gravidade tão intensa que inviabiliza a vida tal como a conhecemos, podendo sobrecarregar organismos com pressões excessivas.

Conhecendo a massa e o raio do planeta, é possível calcular sua densidade através da massa do planeta e do seu volume (V_p), considerando o planeta como uma distribuição esférica:

$$\rho = \frac{M_p}{V_p} = \frac{3M_p}{4\pi R_p^3} \tag{1.13}$$

1.2.3 Outros Parâmetros Planetários

Diversos outros parâmetros planetários são importantes para avaliar a habitabilidade de exoplanetas e merecem atenção. A distância orbital, que pode ser determinada pelo Método de Trânsito, pelo Método de Velocidade radial e também pela 3ª Lei de Kepler, desempenha um papel essencial nesse quesito. Esta distância é um dos fatores determinante na capacidade do planeta de sustentar vida, pois se o planeta estiver muito próximo da estrela, as condições extremas de calor inviabilizam processos bioquímicos fundamentais causando a evaporação de qualquer água líquida em sua superfície, enquanto a atmosfera seria dissipada pela fotoevaporação causada pela radiação UV e raio-X da estrela. Por outro lado, quando o planeta está muito afastado da estrela, torna-se excessivamente frio para manter a água em estado líquido, resultando novamente na inviabilidade de vida complexa em sua superfície.

Entender os limites para a distância orbital é crucial na determinação da chamada "zona habitável", faixa orbital em que a temperatura superficial permite a existência de água líquida (entre 0° a 100° C), um dos componentes mais vitais para a busca por ambientes propícios à vida como a conhecemos (ver a Figura 1.7). Portanto, a análise desses parâmetros é fundamental para a identificação de exoplanetas potencialmente habitáveis e para a compreensão da diversidade de condições que podem existir em sistemas estelares além do nosso.

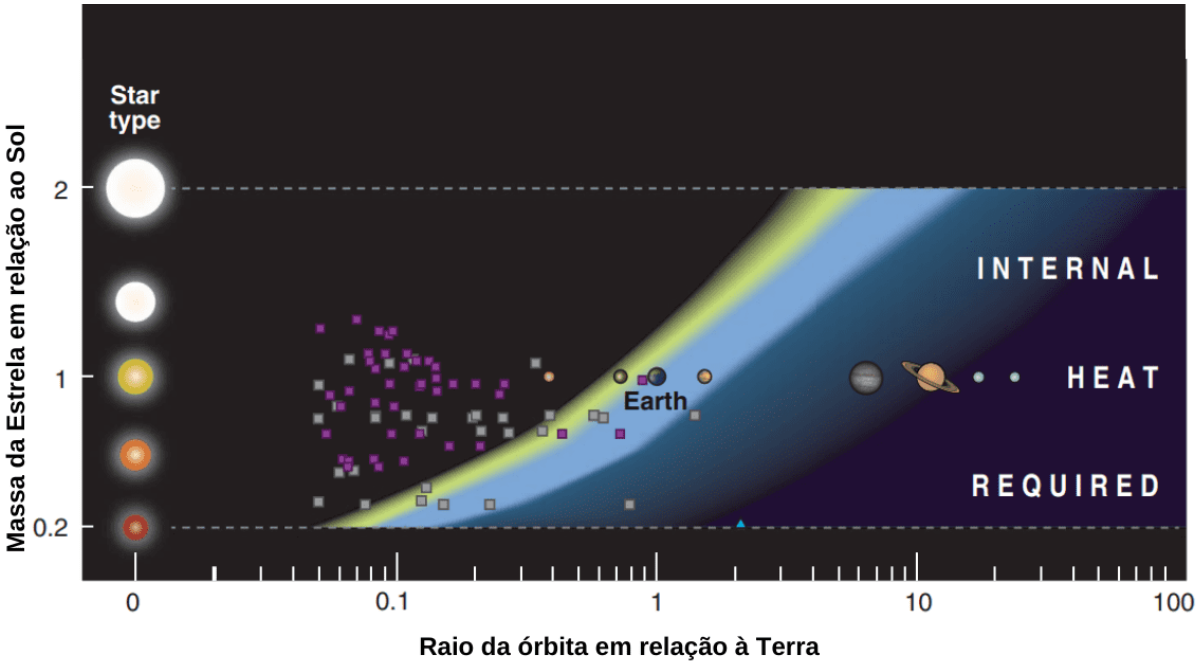

Figura 1.7: Zona habitável (região em azul claro) para diferentes estrelas de diferentes massas. Fonte: Seager (2013). Adaptado pelo autor.

Além desses citados, outros parâmetros estelares desempenham um papel significativo no entendimento das características dos planetas no sistema. Um exemplo notável é a correlação entre o raio planetário e a metalicidade da estrela hospedeira, que foi observada por Buchhave *et al.* (2012) e Ghezzi *et al.* (2021). Nesses estudos, os autores investigaram como a distribuição de metalicidade no disco fino da Galáxia, especialmente na vizinhança solar, influencia as propriedades de sistemas planetários. Ao comparar as estrelas hospedeiras de exoplanetas com diferentes tamanhos planetários, identificaram uma transição significativa entre sub-Netunos e sub-Saturnos, associada a variações nas metalicidades estelares, ajudando a elucidar uma região anteriormente pouco compreendida da arquitetura planetária.

Ghezzi *et al.* (2021) observaram que as distribuições de metalicidade tornam-se cada vez mais distintas à medida que o raio do maior planeta nos sistemas aumenta, especialmente para planetas com $R_p > 2,7R_\oplus$. Vide a Figura 1.8.

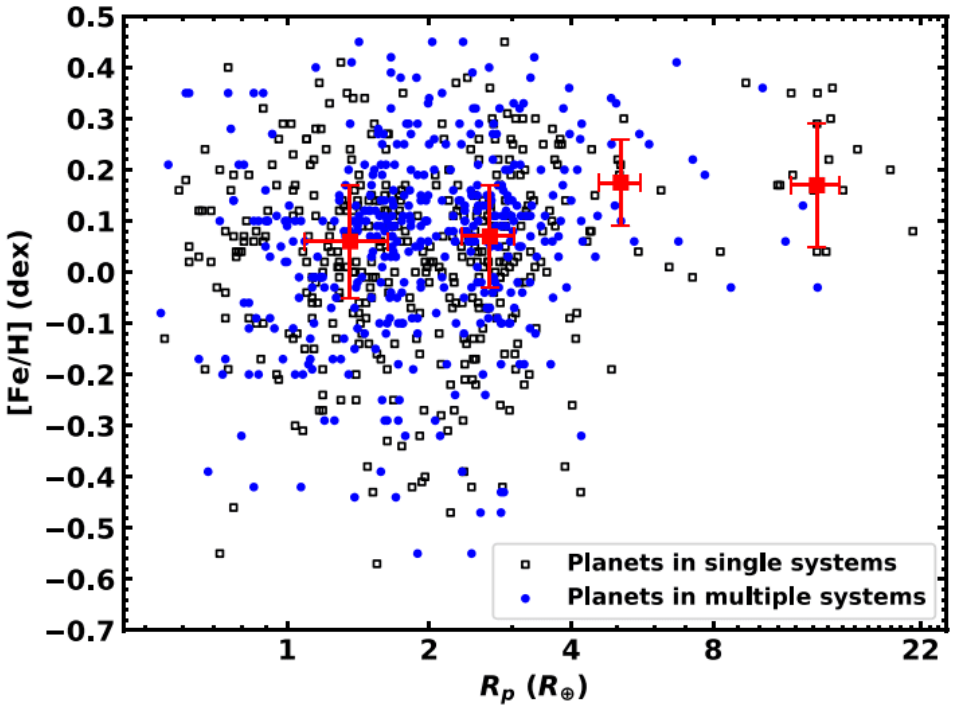

Figura 1.8: Metalicidade estelar como função dos raios planetários para a amostra de 804 planetas. Os quadrados pretos abertos mostram planetas em sistemas únicos e os círculos azuis preenchidos representam planetas em sistemas múltiplos. As barras vermelhas representam os erros nos valores de R_p medianos definidos pelos autores. Fonte: Ghezzi *et al.* (2021).

Estudos realizados por Dong *et al.* (2018), Mulders *et al.* (2016), Petigura *et al.* (2018), Wilson *et al.* (2018) e Wilson *et al.* (2022) indicam uma correlação entre a metalicidade estelar e o período orbital dos planetas. Em particular, esses estudos mostram que estrelas com maior metalicidade tendem a hospedar planetas de curto período, como os Jupiteres Quentes e Netunos Quentes, que apresentam uma maior frequência em torno dessas estrelas. Isso pode ser observado na Figura 1.9, que ilustra essa relação.

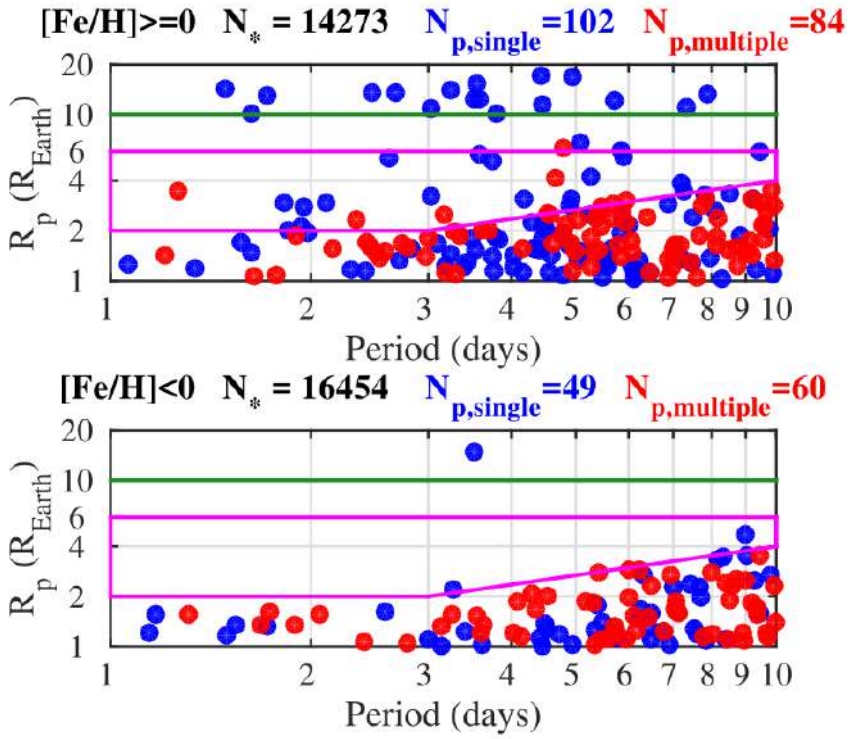

Figura 1.9: Distribuição período-raio para candidatos a planetas de curto período, com estrelas hospedeiras ricas em metais (parte superior) e pobres em metais (parte inferior). Os planetas em sistemas com um único trânsito estão representados por círculos azuis, enquanto os planetas em sistemas com múltiplos trânsitos estão indicados por círculos vermelhos. O número de planetas e estrelas (incluindo os alvos não-trânsitos) é mostrado no topo de cada painel (N_*). A linha horizontal verde escura ($R_p = 10\ R_\oplus$) denota o limite empírico inferior dos *hot Jupiters*, e as linhas magentas marcam os limites empíricos dos *Hot Neptunes*. Fonte: Dong *et al.* (2018).

Além disso, a metalicidade estelar também pode influenciar tanto a ocorrência desses planetas quanto suas características, como o raio e a distribuição de tamanhos. Os planetas ao redor de estrelas mais ricas em metais tendem a ter raios maiores do que os planetas ao redor de estrelas pobres em metais. Isso sugere que a metalicidade pode influenciar o tamanho e a composição dos planetas.

Na presente seção, foi possível examinar diversas situações em que as análises obtêm benefícios significativos a partir de medidas mais precisas das estrelas hospedeiras. A seguir, apresentamos a motivação científica deste trabalho e delineamos nossas propostas para a realização de medições mais confiáveis dos parâmetros estelares para melhor caracterização dos exoplanetas.

1.3 Motivação Científica e Objetivos

Como discutido na Seção 1.2, os parâmetros planetários estão intimamente relacionados aos parâmetros estelares. Assim, antes de proceder com a análise das curvas de

luz e dos espectros estelares, é crucial inferir os parâmetros estelares com um alto grau de confiança. Isso garante incertezas significativamente menores do que as relatadas na literatura, maximizando a precisão para o maior número possível de objetos.

Para avaliar a habitabilidade de exoplanetas, sua composição química e as características de sua atmosfera, é fundamental ter um conhecimento detalhado sobre sua estrela hospedeira. As propriedades estelares, como temperatura, luminosidade, composição química e idade, afetam diretamente o ambiente planetário e, consequentemente, a capacidade de um planeta sustentar vida (vide a Figura 1.10). Além disso, variações nos parâmetros estelares podem influenciar a detecção e interpretação de sinais planetários, como a assinatura espectral de gases atmosféricos, reforçando a necessidade de uma caracterização estelar precisa e com baixa incerteza.

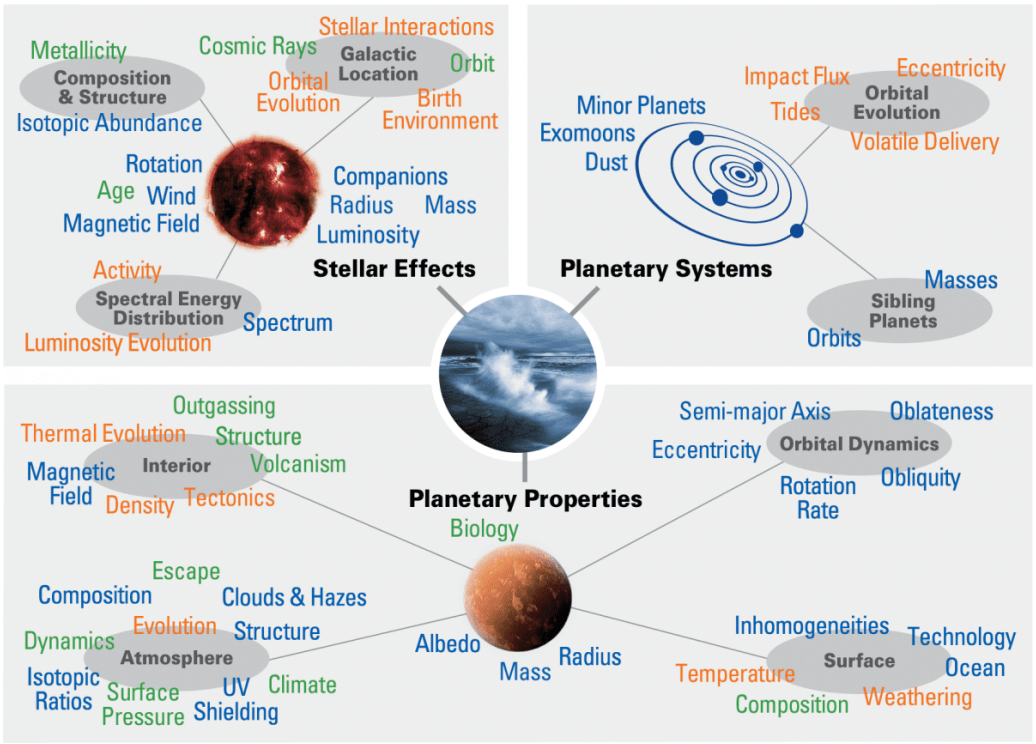

Figura 1.10: Esquema ilustrando como a habitabilidade e as propriedades planetárias dependem dos parâmetros estelares, destacando a importância de uma caracterização precisa da estrela hospedeira para compreender a composição química, a atmosfera e a evolução do planeta. Fonte: Meadows & Barnes (2018).

Para isso, esse trabalho utilizou-se de técnicas de Aprendizado de Máquina (em inglês: *Machine Learning* (ML)) para a determinação de parâmetros estelares. Baron (2019) destaca que técnicas de aprendizado de máquina podem prever parâmetros físicos de forma eficaz, desde que haja uma amostra de treinamento adequada. Além disso, von Marttens *et al.* (2022) ressaltaram que os algoritmos de ML aprendem e aprimoram seu desempenho automaticamente com base na experiência, sem necessidade de programação explícita, sendo capazes de encontrar soluções para sistemas complexos onde métodos

analíticos tradicionais são inviáveis.

Com base em um conjunto de treinamento, um algoritmo de Aprendizado de Máquina aprende a identificar padrões e relações presentes nos dados, o que o torna capaz de estimar valores para novos exemplos que apresentem características semelhantes às da amostra usada no treinamento. Este processo de treinamento é fundamental para que o modelo generalize suas previsões a partir das informações aprendidas, tornando-o eficaz na identificação e análise de dados semelhantes.

No contexto deste trabalho, duas técnicas de Aprendizado de Máquina foram utilizadas: *Random Forest* (RF) e *eXtreme Gradient Boosting* (*XGBoost* ou XGB). Ambas são baseadas em árvores de decisão, onde o modelo divide os dados em ramificações utilizando regras de *if-then* para fazer previsões. Essas técnicas identificam padrões a partir das magnitudes em estrelas com parâmetros físicos específicos, proporcionando uma maior precisão nas previsões. Tanto o RF quanto o XGB constroem múltiplas árvores de decisão e combinam suas previsões para aumentar a precisão e a robustez do modelo. Essas técnicas serão melhor descritas no Capítulo 3.

Dito isso, este trabalho tem como objetivo geral a determinação de parâmetros estelares, tais como temperatura efetiva (T_{ef}), gravidade superficial (log g) e metalicidade ([Fe/H]) com uma precisão mais alta do que na literatura para estrelas hospedeiras de exoplanetas que foram observadas pelos grandes levantamentos fotométricos J-PLUS, S-PLUS, e futuramente, o J-PAS aplicando técnicas de ML. Com a caracterização precisa desses parâmetros, essenciais para orientar análises mais aprofundadas, busca-se melhorar a caracterização dos exoplanetas identificados ao redor dessas estrelas. Dessa forma, os resultados encontrados e que serão apresentados nas seções a seguir, servirão como base para análises mais aprofundadas por meio de validação e acompanhamento destes objetos.

Esta dissertação está organizada da seguinte forma: o Capítulo 2 descreve, com mais detalhes, os principais levantamentos fotométricos utilizados neste trabalho. O Capítulo 3 apresenta uma seção dedicada à metodologia, com uma explicação de como funcionam as técnicas de ML que foram utilizadas, o treinamento dos modelos e suas aplicações nos levantamentos que buscam exoplanetas. O Capítulo 4 aborda os resultados das análises e discussões sobre eles. O Capítulo 5 contém as conclusões acerca dos resultados encontrados e as perspectivas futuras do projeto.

Capítulo 2

Levantamentos Astronômicos Utilizados

Nesse Capítulo, descreveremos todos os levantamentos de dados, cujos os dados foram utilizados na construção e aplicação dos modelos de aprendizado de máquina desenvolvidos neste projeto de pesquisa de mestrado.

Nas Seções 2.1 e 2.2 descreveremos com detalhes os levantamentos fotométricos principais utilizados neste trabalho. E na Seção 2.3, descreveremos os levantamentos auxiliares que nos forneceram parâmetros estelares para o treinamento e testes dos modelos de ML. E por fim, na Seção 2.4, descreveremos com mais detalhes os levantamentos que buscam exoplanetas nos quais foram aplicados os modelos de ML na determinação dos parâmetros das estrelas hospedeiras.

2.1 Javalambre Photometric Local Universe Survey (J-PLUS)

O *Javalambre Photometric Local Universe Survey* (J-PLUS; Cenarro *et al.*, 2019) é um levantamento fotométrico de 8500 graus quadrados, cobrindo o halo da Galáxia, visto do hemisfério norte. Teve as operações iniciadas em 2018 a partir do Observatório Astrofísico de Javalambre (OAJ) localizado em Teruel, Espanha. Os dois telescópios no OAJ são o *Javalambre Survey Telescope* (JST/T250) e o *Javalambre Auxiliary Survey Telescope* (JAST/T80).

O JST/T250 é um telescópio de 2,55 m de diâmetro, com campo de visão (FoV) de 3 graus no céu, especificamente projetado para grandes levantamentos, como o *Javalambre Physics of the Accelerating Universe Astrophysical Survey* (J-PAS; Benitez *et al.*, 2014). O JAST80 é um telescópio de 83 cm de diâmetro com campo de visão de 2 graus, inicialmente desenvolvido para efetuar as calibrações essenciais ao J-PAS, foi posteriormente designado para conduzir as observações do J-PLUS. Além disso, conta com a T80Cam, que é uma câmera de grande campo equipada com um CCD de alta eficiência com 9200x9200 pixels, cada um com 10 μm de tamanho. A leitura é feita por 16 portas simultaneamente,

permitindo tempos de leitura de 12 segundos com baixo ruído, instalado no foco Cassegrain do JAST80.

Cenarro *et al.* (2019) citam que um dos principais objetivos do J-PLUS é a determinação precisa de Distribuições Espectral de Energia (SEDs; do inglês *Spectral Energy Distributions*) de estrelas da Via Láctea e de galáxias próximas. Portanto, é evidente que o conjunto de filtros do J-PLUS deve cobrir tanto o contínuo óptico quanto as linhas espectrais mais proeminentes. Logo, seguindo essa abordagem, o conjunto de filtros do J-PLUS é composto pelos 12 filtros dispostos na Tabela 2.1 com suas respectivas curvas de transmissão na Figura 2.1.

Filtro	λ_c (Å)	FWHM (Å)
u	3485	508
$J0378$	3785	168
$J0395$	3950	100
$J0410$	4100	200
$J0430$	4300	200
g	4803	1409
$J0515$	5150	200
r	6254	1388
$J0660$	6600	145
i	7668	1535
$J0861$	8610	400
z	9114	1409

Tabela 2.1: Conjunto de filtros do J-PLUS com seus respectivos valores de comprimento de onda central (λ_c) e largura de banda (FWHM). Fonte: Cenarro *et al.* (2019). **Nota:** O filtro u também pode ser chamado de *uJAVA* para não confundir com a nomenclatura de outros filtros.

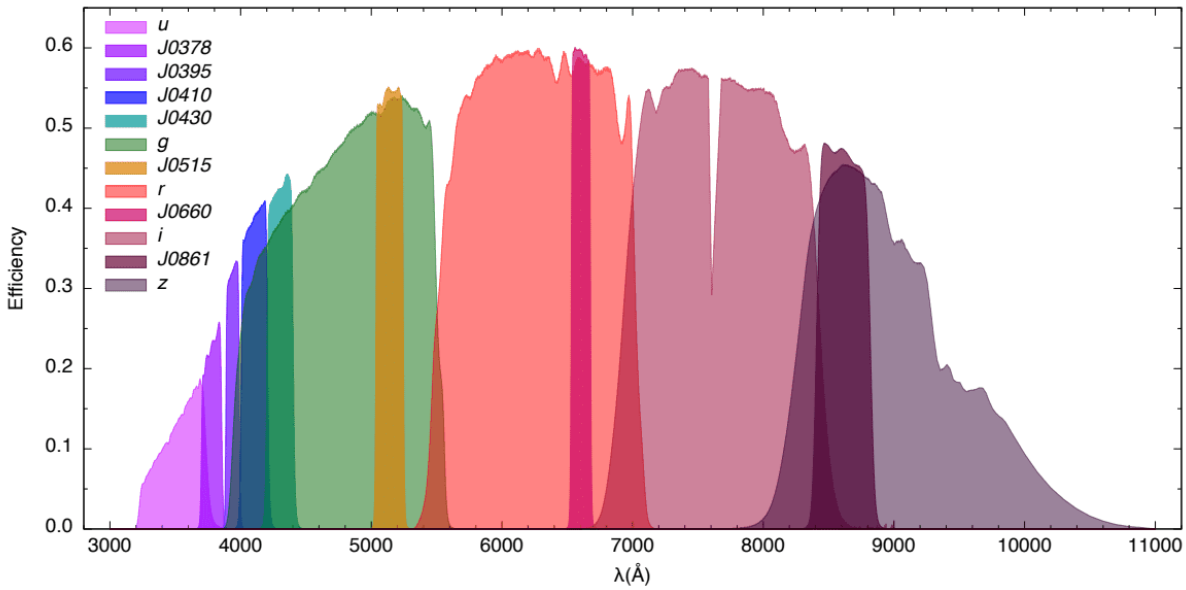

Figura 2.1: Curvas de transmissão para o conjunto de 12 filtros do J-PLUS. Fonte: Cenarro *et al.* (2019).

O conjunto de filtros do J-PLUS inclui quatro filtros de banda larga (g, r, i e z) que também fazem parte do sistema fotométrico do *Sloan Digital Sky Survey* (SDSS; Fukugita *et al.*, 1996). Esses filtros foram incorporadas ao J-PLUS para garantir compatibilidade com levantamentos anteriores e possibilitar comparações diretas com dados do SDSS. Além desses, há seis filtros de banda intermediária que estão centrados em regiões espectrais cruciais e foram construídos especificamente para o projeto. Entre eles, temos o filtro *u*, ou *uJAVA*, localizado na região azul da Série de Balmer, de 3700 a 4000 Å. Temos também os filtros *J*0395 para as linhas H e K do CaII, o *J*0410 para Hδ, o *J*0430 para a banda G, o *J*0515 para o tripleto de Mg*b* e o *J*0861 para o tripleto de Ca II. Os filtros de banda estreita incluem dois filtros: o filtro *J*0378, no qual é sensível às linhas de emissão do O II e o *J*0660, que é sensível à linha de Hα. Os filtros *uJAVA*, *J*0378 e *J*0660 são filtros compartilhados com o J-PAS e Cenarro *et al.* (2019) os consideram como um valor adicional para o procedimento geral de calibração do J-PAS, facilitando correções dos chamados *Photometric Zero Points* (ZPs)[7].

Neste trabalho utilizamos a terceira divulgação de dados do J-PLUS, (*Data Release* 3; DR3), lançada em julho de 2022. Os dados foram coletados entre novembro de 2015 e fevereiro de 2022 pelo telescópio JAST80, cobrindo uma área total de 3.192 graus quadrados resultando em 1.642 campos observados espalhados no céu do hemisfério norte com aproximadamente 338,4 milhões de objetos detectados (ver a Figura 2.2).

[7]Vide Varela *et al.* (2014) para melhor entendimento do uso desses ZPs na calibração do J-PAS.

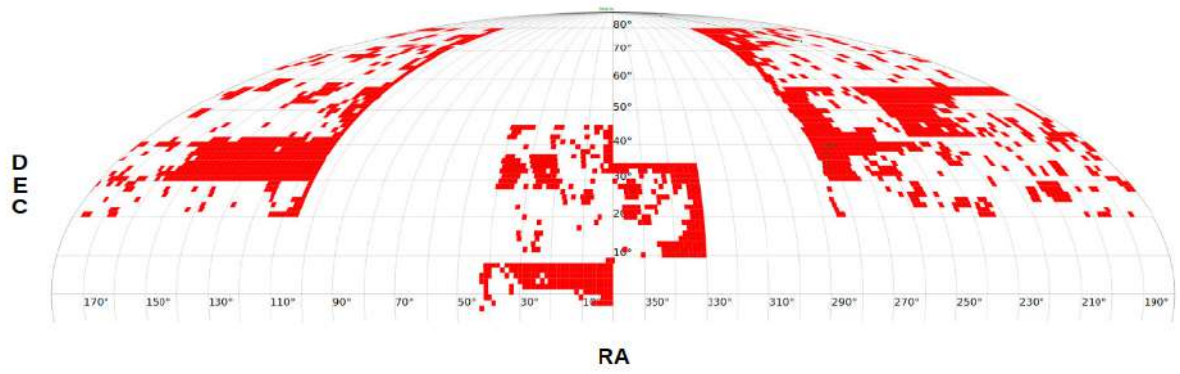

Figura 2.2: Distribuição dos campos observados pelo J-PLUS no DR3, em 1642 campos observados (sinalizados como quadrados/retângulos vermelhos na imagem. O eixo horizontal representa a ascensão reta (α) e o vertical representa a declinação (δ). Fonte: https://www.j-plus.es/datareleases/data_release_dr3.

Ao final, totalizando aproximadamente ~373GB de dados disponibilizados publicamente para a comunidade científica. Essas e outras informações estão disponíveis no site oficial do projeto[8].

2.2 Southern Photometric Local Universe Survey (S-PLUS)

O projeto *Southern Photometric Local Universe Survey* (S-PLUS; Mendes de Oliveira *et al.*, 2019) consiste em um levantamento do céu em 12 bandas ópticas que iniciou suas operações 2016. Utiliza-se uma réplica do telescópio, câmera e filtros do J-PLUS para mapear uma área equivalente do céu meridional, na mesma região espectral. O S-PLUS está conduzindo observações a partir do Observatório Interamericano Cerro Tololo (CTIO), localizado no Chile.

O S-PLUS pretende cobrir uma área de 9300 graus quadrados do céu no hemisfério sul, utilizando como telescópio principal o T80-South (T80S) que possui uma abertura de 80cm de diâmetro e com campo de visão (FoV) de 2 graus no céu, no qual foi otimizado para operar roboticamente. Além disso, é equipado com um CCD de alta eficiência com resolução de 9232x9216 pixels com cada pixel de 10μm de tamanho, a T80Cam-S.

O sistema fotométrico do S-PLUS é o mesmo do J-PLUS (ver a Figura 2.1), no entanto, os filtros utilizados possuem pequenas diferenças, pois é extremamente complexo desenvolver filtros ópticos idênticos. Variações nos materiais e nos processos de fabricação podem resultar em respostas espectrais distintas. Por isso, os valores apresentados na Tabela 2.2 são diferentes dos da Tabela 2.1. Nessa tabela, listamos o comprimento de onda central e a largura dos filtros do S-PLUS.

[8] https://www.j-plus.es/datareleases/releases.

Filtro	λ_c (Å)	FWHM (Å)
u	3577	352
$J0378$	3771	151
$J0395$	3941	103
$J0410$	4094	201
$J0430$	4292	201
g	4292	1545
$J0515$	5133	207
r	6275	1465
$J0660$	6614	147
i	7402	1506
$J0861$	8611	408
z	8882	1182

Tabela 2.2: Conjunto de filtros do sistema *Javalambre* do S-PLUS com seus respectivos valores de comprimento de onda central (λ_c) e largura de banda (FWHM). Fonte: S-PLUS Cloud.

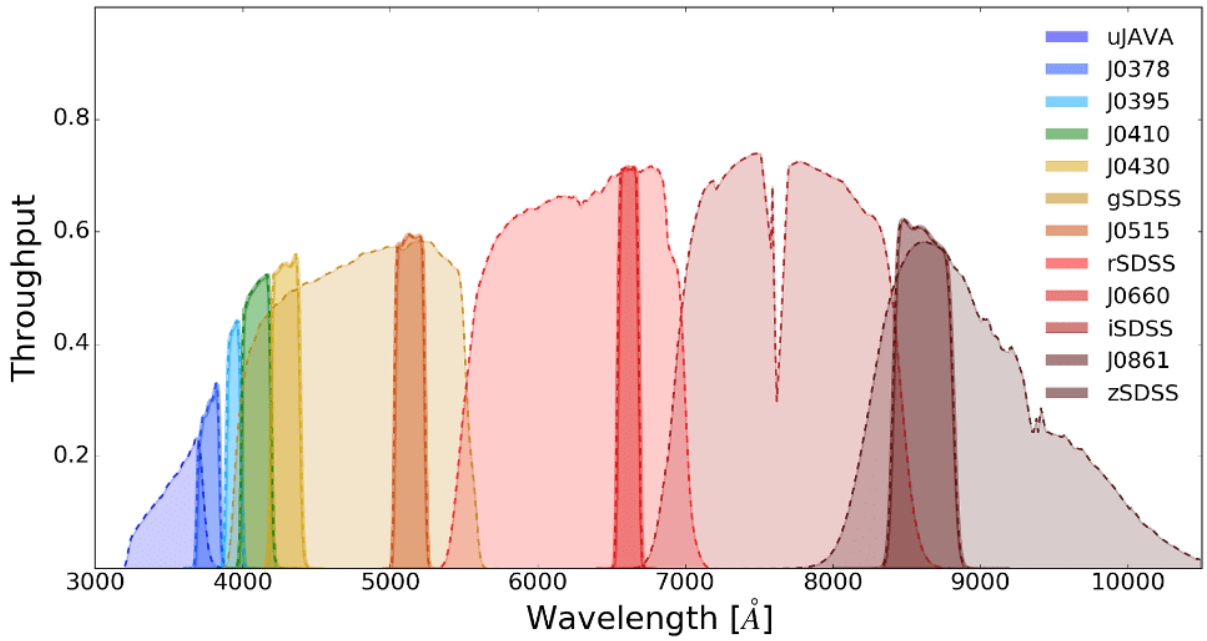

Figura 2.3: Curvas de transmissão dos 12 filtros do S-PLUS. Fonte: Mendes de Oliveira *et al.* (2019).

Diferentemente do levantamento J-PLUS, o levantamento S-PLUS é subdividido em cinco sub-levantamentos, buscando expandir a importância e a utilidade do projeto para diferentes interesses científicos da comunidade astronômica (Mendes de Oliveira *et al.*, 2019). São eles:

- *Main Survey* (MS): Consiste no levantamento principal, cobrindo uma área de 8000 graus quadrados. O MS realiza observações em um único período para cada campo e filtro, sob condições fotométricas e com um *seeing* entre 0,82 e 2,02. A estratégia do MS é semelhante à do J-PLUS, e a expectativa é que ambos os levantamentos possam ser combinados, cobrindo até 16000 graus quadrados no céu.

- *Ultra-Short Survey* (USS): O USS tem o objetivo de cobrir a mesma área que o MS, mas com 1,6s de tempo de exposição, resultando em limites de saturação que permitem a detecção de objetos mais brilhantes. Este levantamento concentra-se na busca por estrelas de baixa metalicidade brilhantes.

- *Variability Fields Survey* (VFS): Realiza observações repetidas em alguns campos já observados pelo MS, com o propósito de monitorar eventos de variabilidade, como asteroides, supernovas (SNe), variáveis cataclísmicas, binárias eclipsantes, pulsares e núcleo ativos de galaxias (AGNs).

- *Galactic Survey* (GS): Consiste em uma região de 1300 graus quadrados no plano galáctico, incluindo áreas do bojo e do disco. Além disso, o GS possibilita o estudo de estrelas variáveis e aglomerados estelares abertos.

- *Marble Field Survey* (MFS): O MFS visa observar campos selecionados específicos com a maior frequência possível quando o *seeing* é muito alto para a realização do MS. Alguns dos objetos de interesse incluem a galáxia M83, a Pequena Nuvem de Magalhães, o aglomerado Hydra, entre outros.

A área de cobertura de cada subdivisão está mostrada na Figura 2.4:

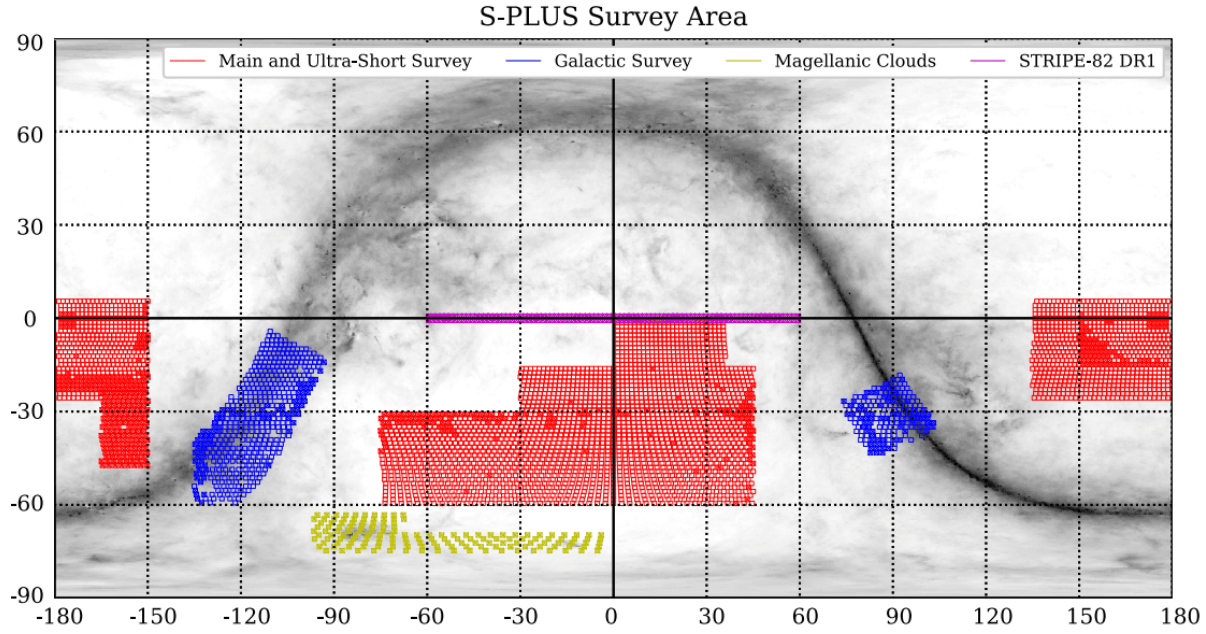

Figura 2.4: Diagrama em coordenadas equatoriais mostrando a cobertura de três dos cinco sublevantamentos do S-PLUS plotados sobre o mapa de extinção de Schlegel *et al.* (1998). A malha vermelha representam o MS e o USS, que compartilham a mesma área. A malha azul representa os campos galácticos e a malha amarela destaca a área das Nuvens de Magalhães, que também estão incluídas no MS. Os quadrados preenchidos de cada malha representa as áreas que já foram observadas até março de 2019. E por fim, a cor magenta representa a área da Stripe82. Fonte: Mendes de Oliveira *et al.* (2019)

Na Figura 2.4, a malha vermelha representam o MS e o USS, que compartilham a mesma área. A malha azul representa os campos galácticos e a malha amarela destaca a área das Nuvens de Magalhães, que também estão incluídas no MS. Os quadrados preenchidos de cada malha representam as áreas que já foram observadas até março de 2019. E por fim, a cor magenta representa a área da Stripe82[9] contida na primeira liberação pública de dados, o *Data Release 1* (DR1), no qual também faz parte do MS.

Neste trabalho, será utilizada a quinta liberação interna de dados (iDR5)[10] que foi liberada em agosto de 2024. O iDR5 consiste em observações de agosto de 2016 a maio de 2023. De acordo com a equipe do S-PLUS, durante esse intervalo houve duas interrupções significativas. A primeira entre abril e novembro de 2017, devido a uma série de problemas técnicos que foram identificados e corrigidos durante a fase de verificação científica. E a segunda entre março e outubro de 2020, devido a restrições externas causadas pela pandemia de Covid-19 que forçaram a interrupção das observações.

No total, o iDR5 inclui observações de 2491 campos totalizando cerca de 4592,2 graus quadrados de área de cobertura do céu do hemisfério sul (Figura 2.5). Diferentemente

[9]Refere-se à região localizada ao longo do Equador Celeste na região do Hemisfério Galáctico Sul. Corresponde a um campo de aproximadamente 270 graus quadrados com ascensão reta entre -50 a 59 graus e declinação entre -1,25 a 1,25 graus. Mais informações em: https://classic.sdss.org/legacy/stripe82.php

[10]Disponível apenas para membros da colaboração.

do DRs 1, 2 e 3 e 4, o iDR5 inclui dados inéditos do bojo da Galáxia e alguns campos adicionais de alta latitude galáctica.

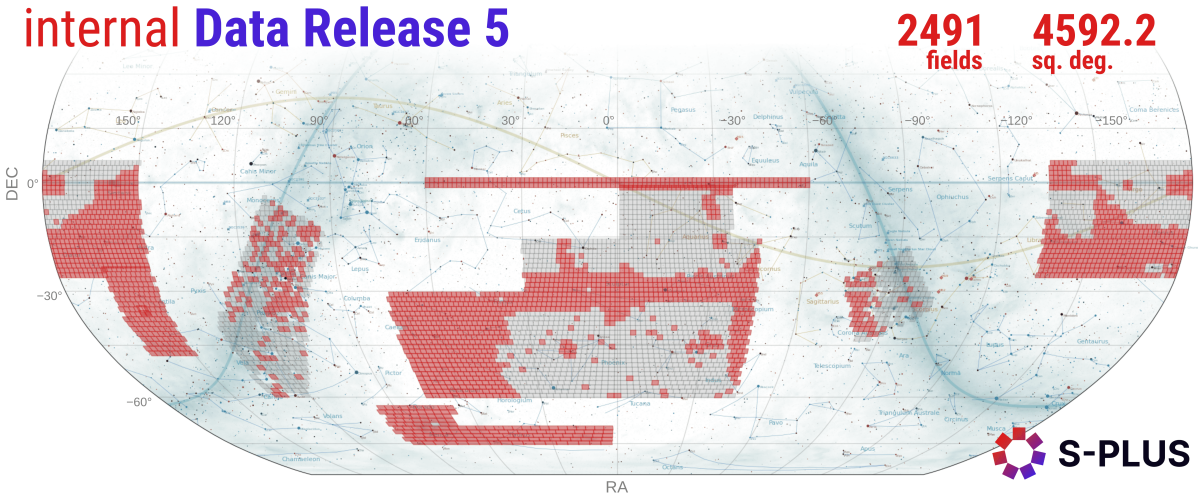

Figura 2.5: Diagrama em coordenadas equatoriais sobre o mapa de extinção de Schlegel *et al.* (1998) mostrando os campos observados (quadrados vermelhos) que foram liberados no iDR5 e os quadrados cinzas referem-se aos campos que ainda serão observados e liberados em próximos DR's. Fonte: S-PLUS Cloud.

Nesta divulgação dos dados, a calibração fotométrica é baseada em espectros do Gaia, garantindo uma calibração uniforme para todo o conjunto de dados. A metodologia empregada no iDR5 será descrita em um próximo artigo da colaboração, no qual serão explicadas a calibração, as observações, a fotometria e a redução dos dados em detalhe.

As magnitudes dos 12 filtros fornecidos pelos levantamentos J-PLUS e S-PLUS junto com as suas combinações de cores serão essenciais para o treinamento dos modelos de ML realizado neste trabalho. Elas funcionarão como *features* (ou variáveis explicativas), a partir das quais os modelos serão capazes de prever parâmetros estelares como temperatura efetiva, gravidade superficial e metalicidade, utilizando a metodologia já aplicada em estudos anteriores. No Capítulo 3 será explicado com mais detalhes como essas *features* serão definidas e implementadas no treinamento e teste dos modelos de ML.

2.3 Levantamentos Auxiliares

Como o J-PLUS e o S-PLUS não estimam parâmetros físicos como temperatura efetiva (T_{ef}), gravidade superficial (log g), metalicidade ([Fe/H]), os quais são os parâmetros de interesse deste trabalho, foi necessário recorrer a bancos de dados de outros levantamentos que tenham esses parâmetros calculados.

Para a escolha desses levantamentos, foi levado em consideração o estudo de Carvalho (2022) e Cordeiro da Silva (2023), no qual os autores realizaram testes utilizando técnicas

de ML, tendo o J-PLUS e S-PLUS como levantamentos principais em diferentes bases de dados. São eles:

2.3.1 LAMOST

O *Large Sky Area Multi-Object Fiber Spectroscopic Telescope* (LAMOST, Zhao *et al.*, 2012) teve seu início em 2011 e está localizado na Estação Xinglong do Observatório Astronômico Nacional da China. É composto por dois componentes principais: o *LAMOST ExtraGAlactic Survey* (LEGAS) e o *LAMOST Experiment for Galactic Understanding and Exploration* (LEGUE). O LEGAS foca na exploração de galáxias, enquanto o LEGUE se dedica ao estudo da estrutura estelar da Via Láctea.

Seu sistema óptico é composto por 3 componentes principais que estão alinhados ao longo da longitude terrestre (plano meridiano). São eles: um espelho corretor asférico ativo[11] na extremidade norte, um espelho primário esférico na extremidade sul e uma superfície focal equipada com fibras ópticas entre eles, formando uma inclinação de 25° em relação ao plano horizontal dos outros espelhos (ver a Figura 1 de Yan *et al.* (2022) para mais detalhes sobre a montagem).

O espelho primário tem dimensões de 6,67 x 6,05 metros e é composto por 37 espelhos de formato hexagonal. Já o espelho secundário possui dimensões de 5,72 x 4,40 metros e é composto por 24 espelhos de formato hexagonal. A superfície focal abriga até 4.000 fibras ópticas, possibilitando a coleta de luz de objetos celestes distantes e pouco luminosos, alcançando magnitudes de até 20,5 (Yan *et al.*, 2022). E, abaixo da superfície focal, localiza-se os 16 espectrógrafos com 32 CCD's. Este design inovador, que combina uma grande abertura com um amplo campo de visão, permite ao LAMOST obter várias dezenas de milhares de espectros por noite (Yan *et al.*, 2022).

Após realizar os testes, Carvalho (2022) concluiu que o LAMOST foi o levantamento que apresentou a amostra com maior quantidade de objetos. Outros levantamentos utilizados, como por exemplo o TESS, não apresentaram consistência nos dados de log g devido a correlação muito baixa entre as *features*. Já o GALAH não possui uma amostra suficientemente grande para o treinamento dos modelos de ML, e o SEGUE apresenta uma quantidade de objetos muito inferior. Além disso, outros autores como Yang *et al.* (2022) e Cordeiro da Silva (2023), também destacam o LAMOST como a opção mais promissora para este tipo de análise.

Com base nisso, foi escolhido o LAMOST como levantamento auxiliar para este trabalho. O levantamento está em sua décima terceira liberação de dados (DR13), porém está disponível publicamente somente até o DR10, dito isso foi selecionada a amostra de baixa resolução das estrelas AFGK do DR10 versão v2.0 cuja a área de cobertura pode ser conferida na Figura 2.6.

[11] É um componente óptico projetado para corrigir aberrações em sistemas de telescópios e otimizar a qualidade da imagem capturada.

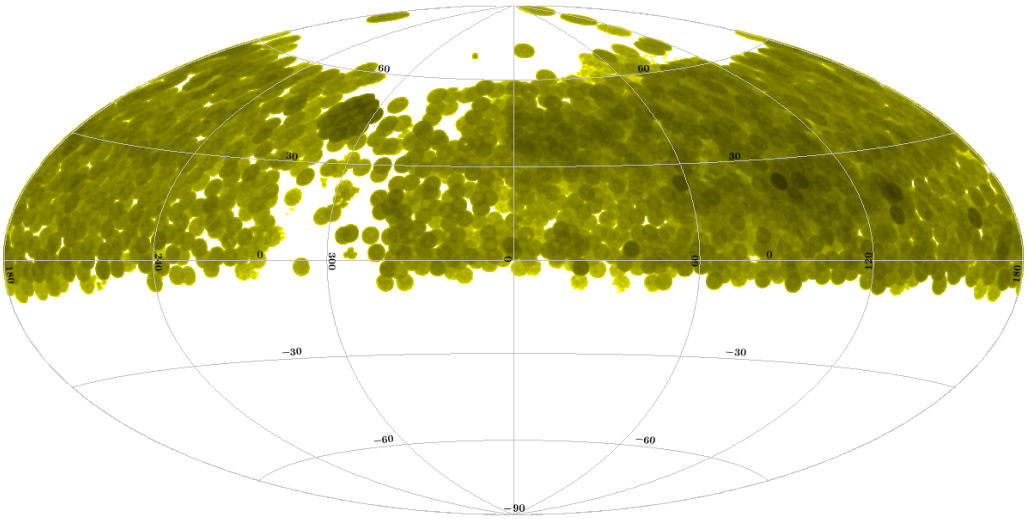

Figura 2.6: Área de cobertura do LAMOST DR10 v2.0 baixa resolução para estrelas AFGK no referencial ICRS. Site oficial do LAMOST.

Esse *data release* está disponível publicamente em seu site oficial[12] e apresenta parâmetros estelares para 7.450.303 estrelas. Mais informações detalhadas do levantamento podem ser encontradas em seu site oficial disponível em http://www.lamost.org/.

2.3.2 APOGEE

Considerando que o LAMOST está focado no hemisfério norte devido a sua localização geográfica, seria necessário para este projeto dados do hemisfério sul, portanto identificamos um levantamento que forneça os parâmetros estelares de interesse e que tenha uma quantidade significativa de objetos no HS. Para isso, foi feito uso também do *Apache Point Observatory Galactic Evolution Experiment* (APOGEE; Majewski, 2016).

O APOGEE é um projeto de espectroscopia no infravermelho próximo (banda H) que integra a terceira e quarta fases do SDSS. Tem como objetivo principal explorar a evolução e a estrutura da Via Láctea por meio da análise de estrelas. O projeto foi dividido em duas gerações: APOGEE-1 e APOGEE-2. A primeira geração, APOGEE-1 (região em ciano na Figura 2.7), operou entre 2011 e 2014, utilizando espectroscopia de alta resolução (R ∼22.500) e e alta razão sinal-ruído (> 100). O objetivo inicial do levantamento é observar 100.000 estrelas gigantes vermelhas. No entanto, o projeto conseguiu examinar mais de 150.000 estrelas ao longo do bojo, barra, disco e halo galáctico, fornecendo dados detalhados sobre velocidades radiais, parâmetros físicos e abundâncias químicas. Dados que são cruciais para entender a evolução dinâmica e química da Via-Láctea.

A segunda fase do projeto, conhecida como APOGEE-2, expande a investigação para estudar o registro arqueológico da Via Láctea, analisando os movimentos orbitais e as composições química de centenas de milhares de estrelas para entender com melhor pre-

[12]Disponível em: http://www.lamost.org/dr10/v2.0/.

cisão a história de formação e evolução da nossa Galáxia (Majewski, 2016). Utilizando também espectroscopia no infravermelho próximo, o APOGEE-2 cria um mapeamento detalhado dos padrões dinâmicos e químicos das estrelas da nossa Galáxia. Os dados coletados no hemisfério norte é chamado de APOGEE-2N (região em azul da Figura 2.7) e os dados do hemisfério sul é chamado de APOGEE-2S (região em vermelho da Figura 2.7). Além de continuar o estudo das estrelas tardias, o APOGEE-2 também explora estrelas jovens, regiões de formação estelar, estrelas variáveis, e estrelas em aglomerados e galáxias satélites.

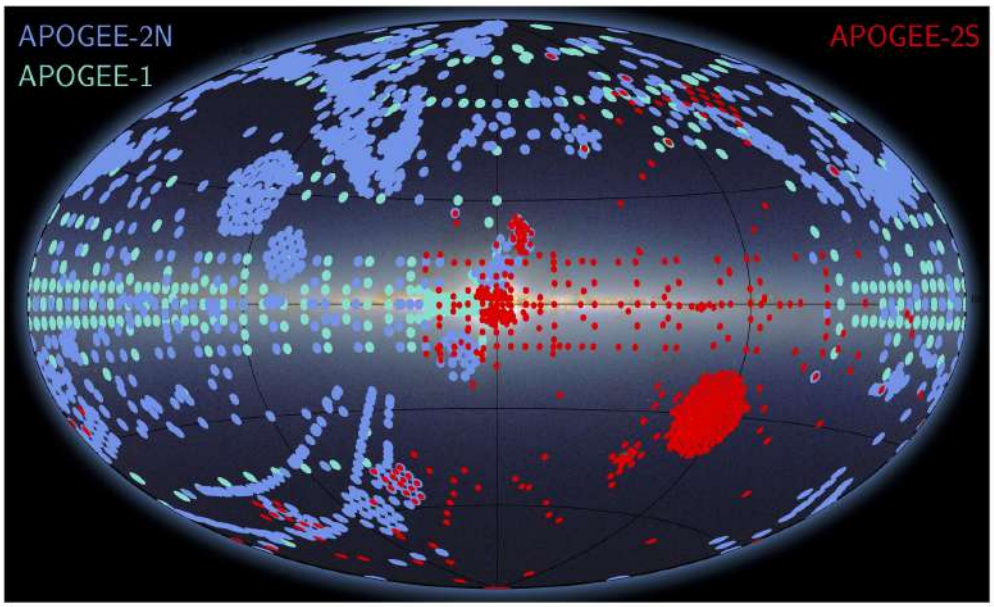

Figura 2.7: Área de cobertura do APOGEE SDSS-IV DR17. Em ciano temos os campos observados pela primeira geração do levantamento (APOGEE-1), enquanto os campos azul e vermelho representam a segunda geração APOGEE-2N e APOGEE-2S, respectivamente. Fonte: https://www.sdss4.org/surveys/apogee-2/.

Além disso, tanto o APOGEE-1 quanto o APOGEE-2 têm contribuído significativamente para o avanço das questões fundamentais da astrofísica, complementando, por exemplo, os dados obtidos pelos levantamentos de exoplanetas Kepler e CoRoT. Eles têm sido fundamentais para a realização de testes de cosmologia em pequena escala, abordando a formação das galáxias e a distribuição da matéria escura (Majewski, 2016).

A primeira geração de dados apareceu pela primeira vez no DR13 do SDSS-III, enquanto a segunda geração (que inclui a primeira) apareceu pela primeira vez no DR17 do SDSS-IV. Ambos estão disponíveis publicamente. Portanto, para esse trabalho foi utilizado a última liberação pública de dados, o DR17, que pode ser encontrada no Vizier (III/286/catalog) ou no site oficial do SDSS[13].

[13] https://www.sdss4.org/dr17.

2.3.3 Gaia

Após a seleção das bases de dados de parâmetros estelares de interesse, seria necessário obter informações astrométricas, como a paralaxe (que permite calcular a distância). A melhor fonte para esses dados é o catálogo da missão espacial Gaia (Collaboration *et al.*, 2016). O Gaia é uma missão da Agência Espacial Européia (sigla em inglês: ESA) que visa mapear com alta precisão a posição e os movimentos próprios de quase dois bilhões de objetos na Via Láctea, construindo um mapa tridimensional da Galáxia.

A sua instrumentação inclui dois telescópios ópticos equipados com três instrumentos de medições astrométricas. Esses instrumentos são utilizados para determinar com precisão a localização das estrelas e suas velocidades, além de obter um espectro para análise detalhada, utilizado na determinação de parâmetros como temperatura, composição química, dentre outros.

De acordo com o Gaia Archive a sonda encontra-se a 1.600.342 km de distância da Terra e está "estacionada" no ponto conhecido como Ponto Lagrange 2[14]. Durante sua missão, a espaçonave gira lentamente, permitindo que os dois telescópios cubram toda a esfera celeste e além disso, durante todo o ano, os telescópios observam cada região pelo menos 14 vezes, assim, obtendo melhor precisão.

Para este trabalho, selecionamos a terceira liberação de dados do Gaia, o DR3, que foi disponibilizada pela colaboração em junho de 2022 e pode ser acessada atráves do site oficial Gaia Archive[15]. Nele, encontramos no total, parâmetros de 1.811.709.771 fontes.

2.4 Levantamentos que Buscam Exoplanetas

Para este trabalho, foram escolhidos levantamentos que utilizaram as técnicas de detecção por Trânsito Planetário e Variação de Velocidade Radial, pois, como dito anteriormente, representam a maior parte das detecções de exoplanetas disponíveis na literatura. Os levantamentos selecionados para Trânsito foram o Kepler e o TESS e para Velocidade Radial foi selecionado os catálogo do Espectrógrafo HARPS. Eles serão melhor descritos a seguir.

2.4.1 Missão Kepler

A Missão Kepler é uma das missões mais bem sucedidas da NASA na busca de exoplanetas pela técnica de trânsito planetário. Lançado em 6 de março de 2009, representou um marco na busca por exoplanetas ao combinar técnicas inovadoras de medição do brilho estelar com a maior câmera digital já projetada para observações espaciais até aquele

[14] Um ponto de Lagrange é uma localização no espaço onde as forças gravitacionais de dois corpos celestes, como a Terra e o Sol, se equilibram, permitindo que um terceiro objeto, como um satélite, permaneça em uma posição estável relativa a esses dois corpos.

[15] https://www.cosmos.esa.int/web/gaia/dr3.

momento (NASA Science Mission, 2017). Equipado com um telescópio com espelho de 1,4 metros de diâmetro, um fotometro de 0,95m de abertura com um sistema de 42 CCDs observando cerca de 4,5 milhões de estrelas, das quais 150 mil foram monitoradas continuamente em uma região entre as constelações de Cisne e Lira cobrindo uma área de aproximadamente 105 graus quadrados (vide a Figura 2.8). Essa abordagem permitiu à missão identificar planetas do tamanho da Terra na zona habitável de suas estrelas hospedeiras, possibilitando o primeiro grande levantamento de planetas na nossa Galáxia.

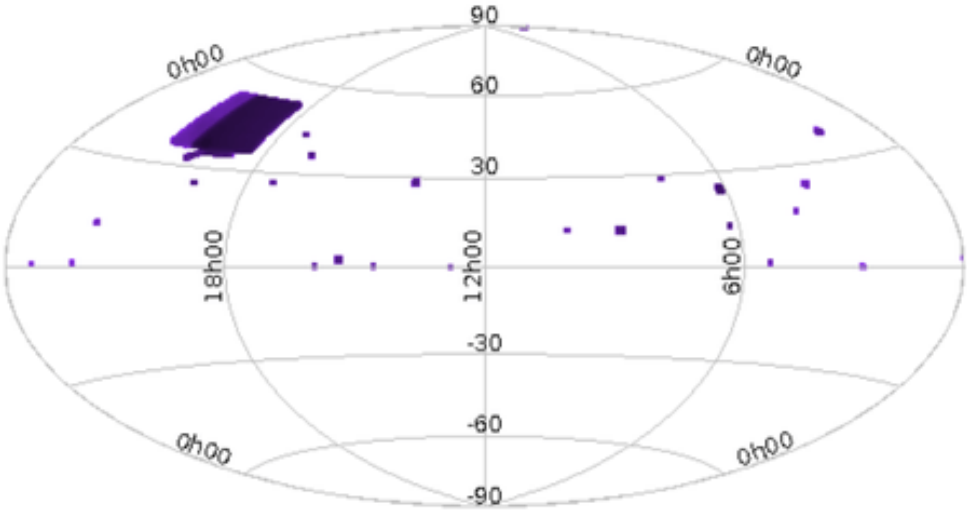

Figura 2.8: Área de cobertura do Kepler Input Catalog plotado no sistema ICRS com a região em quadrangular de Cisne e Lira em destaque. Os outros pontos roxos são campos observados na segunda campanha do Kepler (K2) que estão presentes no catálogo do KIC. Fonte dos dados: `https://archive.stsci.edu/missions-and-data/kepler`.

Durante seus primeiros anos de operação, a missão Kepler descobriu não apenas planetas isolados, mas também sistemas planetários complexos. Em 2010, os cientistas da missão anunciaram o sistema Kepler-9, o primeiro a ter múltiplos planetas confirmados em torno de uma mesma estrela do tipo solar (Holman *et al.*, 2010). Em 2011, foi a vez do sistema Kepler-11, que revelou seis planetas, alguns comparáveis em tamanho a Urano e Netuno (Lissauer *et al.*, 2011). Um dos achados mais intrigantes foi a prevalência de planetas de tamanho intermediário entre a Terra e Netuno, um tipo inexistente no Sistema Solar, mas comum na Galáxia. Além disso, Kepler mostrou que sistemas planetários densos, com muitos planetas orbitando próximos de suas estrelas, são mais comuns do que se imaginava.

A missão principal do Kepler durou até maio de 2013 devido à perda de dois dos quatro giroscópios disponíveis na espaçonave. Para se manter fixo e estável, o telescópio necessitava de pelo menos três funcionando perfeitamente para manter a precisão original da missão. Nos anos seguintes, entre 2014 e 2018, o telescópio iniciou uma fase estendida, conhecida como missão K2. Durante essa fase, o Kepler observou diferentes campos ao redor do plano da eclíptica, adotando uma estratégia que minimizava o torque causado

pela pressão do vento solar. Isso reduzia a deriva do apontamento, permitindo que os dois giroscópios restantes e os propulsores mantivessem o controle efetivo da espaçonave. Assim, o telescópio podia observar um campo de visão com estabilidade por cerca de 80 dias, período denominado como uma campanha de observação.

O legado da missão Kepler vai além da descoberta de exoplanetas. Seus dados permitiram avanços na compreensão das propriedades das estrelas e sua influência sobre os planetas que as orbitam. Mesmo após o esgotamento de combustível em 2018, a missão continuou a entregar dados científicos valiosos, que complementam os esforços do *Transiting Exoplanet Survey Satellite* (TESS; Ricker, 2015), lançado em 2018. Com a análise dos dados da missão Kepler, estima-se que 20% a 50% das estrelas visíveis a olho nu podem ter planetas rochosos, semelhantes à Terra, na zona habitável (NASA Science Mission, 2017). De acordo com o NASA Exoplanet Archive somente com dados da primeira etapa das observações foi possível confirmar 2778 exoplanetas, enquanto a segunda etapa, o K2, descobriu 548 e ainda há cerca de 2950 candidatos à exoplanetas no total das duas campanhas.

O catálogo do Kepler está disponível publicamente em *Mikulski Archive for Space Telescopes* (MAST)[16].

2.4.2 TESS

O TESS é uma missão dedicada à realização de um levantamento de trânsitos planetários em todo o céu. Lançado em abril de 2018, o principal objetivo do TESS é detectar planetas do tamanho da Terra que orbitam estrelas brilhantes próximas, adequadas para observações complementares que permitam determinar suas massas e composições atmosféricas. Durante sua missão principal de dois anos, o TESS realizou fotometria de alta precisão em mais de 200.000 estrelas, com um intervalo de leitura de aproximadamente dois minutos. As imagens completas dos campos de visão, chamadas de Full Frame Images (FFIs), foram obtidas a cada 30 minutos, permitindo uma análise fotométrica detalhada de qualquer alvo dentro de um campo (FOV) de 24 x 96 graus em diferentes setores do céu (vide a Figura 2.9).

[16] https://archive.stsci.edu/missions-and-data/kepler.

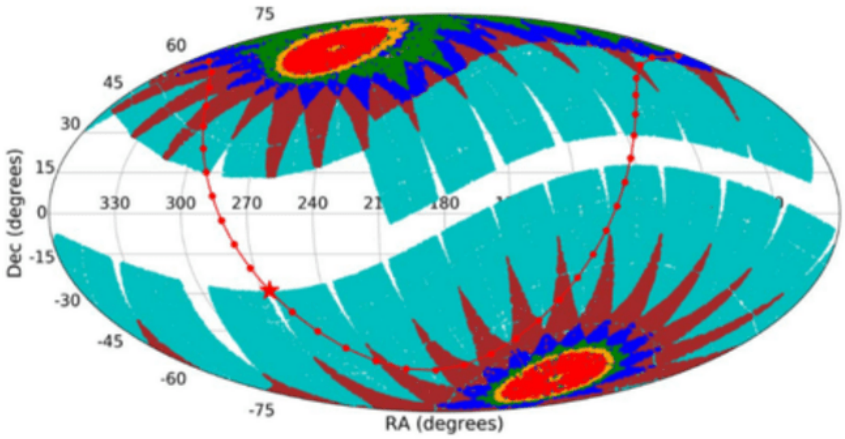

Figura 2.9: Área de cobertura da missão TESS colorida por setor. A curva vermelha em forma de "U" mostra o plano galáctico, com a posição do centro galáctico mostrada pela estrela vermelha. Cada cor representa um setor que foi observado por 27 dias cada. O setor 12 (em vermelho vivo), localizados nos polos, foi observada pelo TESS de forma contínua durante 351 dias. Fonte: Guerrero *et al.* (2021).

A missão TESS é planejada para realizar pesquisas fotométricas sequenciais de setores do céu com duração de 27,4 dias cada. Durante o primeiro ano, o telescópio observou o hemisfério sul da eclíptica, seguido pelo hemisfério norte no segundo ano. As observações incluem leituras completas de todos os detectores das câmeras do TESS, bem como o armazenamento de "postage stamps" (pequenas áreas de interesse nos detectores) de cerca de 200.000 a 400.000 estrelas pré-selecionadas (Guerrero *et al.*, 2021). Essas estrelas são escolhidas com base em sua alta prioridade para a busca de trânsitos planetários. Esse formato de operação é semelhante ao da missão Kepler, mas com uma abordagem expandida para cobrir uma área muito maior do céu. Até o momento, de acordo com o NASA Exoplanet Archive, foram confirmados 613[17] exoplanetas pela missão TESS com ainda outros 4809[18] candidatos a serem confirmados. Informações mais completas sobre a missão e o telescópio encontram-se em seu site oficial.

O TESS Input Catalog (TIC) atualmente na versão 8.2 (TIC v8.2: Paegert *et al.* (2021)), desempenha um papel fundamental na seleção de alvos para a missão. Este catálogo inclui fontes luminosas ópticas em todo o céu, baseando-se em dados disponíveis de catálogos de campo amplo. O TIC fornece parâmetros estelares essenciais para a avaliação de sinais de trânsito utilizados para calcular os parâmetros dos planetas detectados (Paegert *et al.*, 2021).

O catálogo está disponível publicamente e pode ser acessado pelo portal do MAST[19].

[17]Segundo o NASA Exoplanet Archive em 16 mar. 2025. Fonte: https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html.

[18]Segundo o NASA Exoplanet Archive em 16 mar. 2025. Fonte: https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html.

[19]https://archive.stsci.edu/missions-and-data/tess.

2.4.3 Espectrógrafo HARPS

O *High Accuracy Radial velocity Planet Searcher* (HARPS; Barbieri, 2023) é um espectrógrafo de alta precisão desenvolvido pelo Observatório Europeu do Sul (ESO) e instalado no telescópio de 3,6 metros em La Silla, no Chile. Este instrumento foi projetado para medir velocidades radiais com precisão da ordem de 1 m/s. Para alcançar tal precisão, o instrumento utiliza uma grade de difração do tipo *Echelle*, que é composta por ranhuras dispostas de maneira a induzir a difração da luz em altas ordens. Essa configuração é projetada para operar eficientemente com feixes de luz que incidem sob ângulos elevados, permitindo a separação precisa de comprimentos de onda e a obtenção de espectros de alta resolução. O sistema opera em um câmara a vácuo para minimizar desvios espectrais causados por variações de temperatura e pressão atmosférica, garantindo resultados consistentes mesmo em condições adversas.

A rede do HARPS é alimentada por um par fibras ópticas, onde cada uma cobre cerca 1" de céu (Barbieri, 2023). Uma das fibras coleta a luz da estrela, enquanto a outra é empregada para registrar simultaneamente o espectro de uma lâmpada de referência Th-Ar ou o brilho do céu de fundo. O instrumento é equipado com um mosaico de dois CCDs, totalizando 4000 x 4000 pixels de 15 mícrons cada, e conta com um sistema de calibração em tempo real com um intervalo espectral de 378 nm a 691 nm. Em sua configuração padrão, conhecida como *Simultaneous Thorium Reference Method*, o HARPS pode atingir um erro de determinação de velocidade radial de aproximadamente 0,90 m/s para estrelas de tipo espectral G2V com magnitude visual V=6 (Barbieri, 2023). Essa precisão é especialmente valiosa para estudos de estrelas frias, com rotações moderadas e sistemas planetários.

Além disso o HARPS possui um pipeline automatizado que fornece aos astrônomos, em tempo quase real, espectros extraídos e calibrados em comprimento de onda. Quando o método de referência simultânea é aplicado, o pipeline também entrega velocidades radiais precisas, relativas ao baricentro do Sistema Solar, para estrelas de tipo tardio com velocidades radiais previamente conhecidas. Desde seu início em operação, em outubro de 2003, o HARPS se consolidou como uma ferramenta essencial na busca e caracterização de exoplanetas, revolucionando nossa compreensão sobre a diversidade de sistemas planetários (Barbieri, 2023).

Em 2023, o ESO disponibilizou a primeira versão do catálogo, que inclui observações realizadas entre junho de 2003 e junho de 2023, extraídas de seu próprio arquivo científico. Para garantir a qualidade e a precisão dos dados, foram excluídas observações realizadas com configurações polarimétricas (Curto *et al.*, 2012). O catálogo contém 289.843 observações realizadas por 327 pesquisadores diferentes em 630 programas. As observações estão concentradas principalmente em campos onde outros levantamentos acompanham trânsitos de exoplanetas na nossa Galáxia e concentradas nas Nuvens de Magalhães. O

catálogo está disponível publicamente no *site* oficial[20].

[20] https://www.eso.org/qi/

Capítulo 3

Metodologia

A metodologia deste projeto pode ser dividida em três grandes etapas, facilitando o entendimento do processo até a obtenção dos resultados esperados: (1) Seleção e preparação da amostra, (2) Treinamento e teste de algoritmos de Aprendizado de Máquina para a construção do modelo, e (3) Aplicação do modelo em levantamentos voltados para a busca de exoplanetas.

3.1 Seleção da Amostra

Como mencionado anteriormente no Capítulo 2 no caso do J-PLUS DR3, os dados selecionados seguiram a metodologia apresentada e testada por Carvalho (2022). Selecionamos, portanto, as estrelas que foram observadas com uma abertura de 6" em todos os 12 filtros do levantamento e que possuem uma probabilidade de pelo menos 90% de serem classificadas como estrelas (prob_star $\geqslant$ 0,9). Ao final, isso resultou em uma amostra com 2.048.514 objetos.

Para o levantamento S-PLUS iDR5 foram selecionadas as estrelas que foram observadas em todos os 12 filtros e que tenham pelo menos 90% de possibilidade de serem estrelas (CLASS $\geqslant$ 0,9). No que diz respeito a melhor abertura a ser utilizada a equipe do S-PLUS Cloud recomenda fortemente o uso de fotometrias diferentes, uma para regiões mais densas como o disco galáctico e aglomerados globulares e outra para o halo. Para a região do disco e aglomerados é recomendável utilizar a fotometria *Point-Spread Function* (PSF). A fotometria PSF mede o brilho de um objeto celeste ajustando um modelo da função de espalhamento da luz. Ela leva em consideração a forma com que a luz de uma estrela ou galáxia é distribuída no detector, permitindo uma medição mais precisa em regiões com alta densidade estelar, no caso do S-PLUS, obtidas usando DOphot (Alonso-García *et al.*, 2012).

Para a região do halo recomenda-se o uso da fotometria PSTotal, que consiste em uma medida fotométrica que visa refletir com precisão a magnitude total de uma fonte pontual, como estrelas ou quasares, enquanto aproveita a alta relação sinal-ruído (S/N)

de uma abertura muito pequena (Herpich *et al.*, 2024). A fotometria é derivada ajustando as medições obtidas em aberturas maiores para compensar a diferença na captura de luz devido ao tamanho da abertura. Esse ajuste é realizado utilizando a curva de crescimento da magnitude, que descreve como a magnitude muda com o aumento da abertura.

Devido ao baixo número de campos observados utilizando a fotometria PSF, este trabalho utilizou somente os dados obtidos com a fotometria PSTotal, porque combina a alta precisão da medição com a abertura restrita e a confiabilidade da função de espalhamento (PSF), assegurando uma estimativa robusta da magnitude total da fonte (Herpich *et al.*, 2024). Os objetos selecionados foram observados em todos os 12 filtros do levantamento e que possuem uma probabilidade de pelo menos 90% de serem classificadas como estrelas (CLASS_STAR $\geqslant$ 0,9). Ao final, resultou em uma amostra com 2.592.643 objetos.

Além disso, todos os valores de magnitude devem possuir valor para correção de extinção interestelar calculado.

3.1.1 Cruzamento entre os Levantamentos Principais e os Auxiliares

Antes de iniciar o processo de treinamento e teste do algoritmo de ML, é necessário reunir as amostras das estrelas em comum nos campos dos levantamentos principais com os levantamentos auxiliares. Essa técnica amplamente utilizada na análise de dados é conhecida em inglês como *crossmatching*, um cruzamento entre dados comuns presentes em Tabelas diferentes, sendo ascensão reta (RA) e declinação (DEC) como referências, resultando em uma única amostra consolidada ao final.

Para realizar o cruzamento de amostras de forma eficiente, foi utilizado o software TOPCAT v.4.8.3[21]. Nele há inúmeras ferramentas que permitem a manipulação e análise de dados astronômicos em grandes levantamentos de forma rápida. E com uma dessas ferramentas foi possível identificar as correspondências exatas entre as estrelas nos diferentes levantamentos, garantindo a precisão necessária para a criação de uma amostra única composta dos parâmetros de magnitude dos levantamentos principais e dos parâmetros de interesse dos levantamentos auxiliares. Sendo assim, temos as seguintes amostras:

- Dados do J-PLUS DR3 em campos em comum com o GAIA DR3 e o LAMOST DR10 v2.0, resultando em uma amostra de 314.999 objetos;
- Dados do J-PLUS DR3 em campos em comum com o GAIA DR3 e o APOGEE DR17 SDSS-IV, resultando em uma amostra de 11.797 objetos.
- Dados do S-PLUS iDR5 em campos em comum com o GAIA DR3 e o LAMOST DR10 v2.0, resultando em uma amostra de 110.430 objetos;

[21] Documentação completa em: https://www.star.bris.ac.uk/~mbt/topcat/.

- Dados do S-PLUS iDR5 em campos em comum com o GAIA DR3 e o APOGEE DR17 SDSS-IV, resultando em uma amostra de 20.447 objetos.

Após a definição dos objetos que seriam utilizados nas amostras, as variáveis (*features*) consideradas como entrada para os modelos foram compostas pelas 12 magnitudes descritas no capitulo 2, já com as correções de extinção aplicadas e suas combinações de cores.

3.1.2 Preparação das Amostras de Dados

Antes de iniciar a etapa de treinamento e teste dos modelos, é preciso preparar as amostras de dados para se chegar em modelos bons e precisos. Seguindo a recomendação de Cordeiro da Silva (2023) e após realizado alguns testes, foi definida uma filtragem em alguns parâmetros específicos fornecidos pelos levantamentos auxiliares utilizados neste trabalho. Para as amostras com dados do levantamento LAMOST, a filtragem utilizada foi a seguinte:

- Razão sinal-ruído nas faixas espectrais que correspondem aos filtros g, i e z tem que ser maior ou igual a 10 ([snri, snrg, snrz] $\geqslant$ 10);
- Razão sinal-ruído maior ou igual a 20 na faixa espectral que corresponde ao filtro r ([snrr] $\geqslant$ 20).
- Os erros inferiores a 300 K para T_{ef}, inferiores a 0,4 dex para log g e 0,4 dex para [Fe/H].

Essa abordagem permite que os usuários avaliem a qualidade do espectro em regiões que são relevantes para cada estudo. Para as amostras com dados do levantamento APOGEE, é levemente diferente, pois o levantamento não fornece o sinal ruído específico para diferentes faixas específicas do espectro, porém fornece a média do sinal ruído combinado por pixel. Portanto neste caso a filtragem utilizada foi:

- A média da razão sinal-ruído maior ou igual a 20 ([SNR] $\geqslant$ 20).

Além dos levantamentos LAMOST e APOGEE, foram definidas filtragens para os parâmetros fornecidos pelo Gaia. Segue:

- Erro de Peso Unitário Renormalizado menor ou igual a 1,4 ([RUWE[22]] $\leqslant$ 1,4);

[22]É uma métrica utilizada para avaliar a qualidade dos ajustes astrométricos, no qual é verificado se as observações de um objeto estão bem ajustadas ao modelo astrométrico, sendo que um valor próximo de 1 indica um bom ajuste, enquanto valores mais altos podem sugerir que o objeto tem um comportamento astrométrico incomum e/ou outros desvios que podem indicar erros ou binaridade.

- Paralaxe dividida pelo seu erro padrão menor ou igual a 5 ([RPlx[23]] $\leqslant 5$).

Outro ponto importante a ser mencionado é que o treinamento dos modelos foi realizado utilizando a magnitude absoluta. Com a magnitude absoluta e consequentemente a distância, é possível eliminar interpretações errôneas causadas pela degenerescência das magnitudes aparentes. Os levantamentos J-PLUS e S-PLUS fornecem apenas as magnitudes aparentes (que foram corrigidas pela extinção interestelar). Portanto, foi necessário calcular a magnitude absoluta em cada filtro utilizando a Equação 3.1, também conhecida como módulo de distância:

$$M = m_0 - 5\log_{10}(d) + 5 \tag{3.1}$$

Onde M é a magnitude absoluta em cada filtro, m_0 é a magnitude aparente fornecida pelos levantamentos J-PLUS e S-PLUS nos 12 filtros corrigidas a extinção interestelar, e d é a distância que é calculada por Bailer-Jones *et al.* (2021). Neste trabalho, esse cálculo foi feito utilizando o pacote *astropack*[24].

E seguindo a recomendação de Carvalho (2022), após serem preparadas conforme os filtros citados anteriormente, os dados foram separados em amostras que chamamos de menos restrita e restrita. A amostra menos restrita, consiste em uma amostra que contém somente objetos com erro de magnitude em todos os filtros J-PLUS e S-PLUS menor ou igual a 0,2 (e_mag $\leqslant$ 0,2). Enquanto a amostra restrita, consiste nos dados de estrelas com erro de magnitude menor ou igual a 0,1 (e_mag $\leqslant$ 0,1).

A quantidade específica de objetos após a preparação das amostras é apresentada conforme a Tabela 3.1, a seguir:

Amostra	**Restrita**	**Menos Restrita**
J-PLUS DR3 + GAIA DR3 + LAMOST DR10	255.039	278.879
J-PLUS DR3 + GAIA DR3 + APOGEE DR17 SDSS-IV	72.036	76.035
S-PLUS iDR5 + GAIA DR3 + LAMOST DR10	2.724	4.782
S-PLUS iDR5 + GAIA DR3 + APOGEE DR17 SDSS-IV	8.123	9.125

Tabela 3.1: Quantidade de objetos em cada amostra de entrada para o treinamento após a preparação.

Para cada treino, como será descrito na próxima seção, as amostras foram separadas, conforme o padrão dos algoritmos de ML utilizados neste trabalho, em 75% para treinamento e 25% para teste e validação. Esses 25%, mesmo sendo parte da mesma amostra

[23]Também conhecida como significância da paralaxe e indica a confiabilidade da medição da paralaxe. Um valor de pelo menos 5 é geralmente considerado confiável para determinar distâncias estelares.

[24]Documentação completa em: https://github.com/cordeirossauro/astropack/tree/main/src/astropack. Desenvolvido por Cordeiro da Silva (2023) em seu trabalho.

inicial, o algoritmo não a conhece, com isso sendo possível fazer um teste e uma validação consistente dos modelos após o treino. +

3.2 Treinamento e Teste de Algoritmos de Aprendizado de Máquina

Antes de abordarmos a segunda grande etapa do trabalho, apresentaremos uma breve explicação sobre os algoritmos de Aprendizado de Máquina (do inglês: ML), além de detalhar as técnicas específicas que foram utilizadas neste trabalho: *Random Forest* (Breiman, 2001) e *XGBoost* (Chen & Guestrin, 2016). Por fim, apresentaremos a descrição do processo de treinamento e teste desses algoritmos.

3.2.1 Aprendizagem de Máquina (*Machine Learning*)

O uso de técnicas de ML tem se tornado cada vez mais comum em diversas áreas da ciência, incluindo a astronomia, devido à sua capacidade de lidar com grandes volumes de dados e encontrar padrões complexos (Baron, 2019). Diante desse rápido avanço, os astrônomos têm criado ferramentas automatizadas para detectar, caracterizar e classificar objetos nos vastos e complexos conjuntos de dados coletados por diversos observatórios e satélites (Baron, 2019; Yao & Liu, 2013).

A ML é uma subárea da inteligência artificial que desenvolve algoritmos capazes de aprender e realizar previsões ou decisões com base em dados oferecidos à eles sem que haja uma programação explícita para tal. Esses algoritmos identificam padrões em grandes volumes de informações, permitindo que eles realizem previsões sobre novos dados de maneira eficiente.

Além disso, esses algoritmos podem ser classificados em duas categorias principais: modelos supervisionados e modelos não supervisionados. Nos modelos supervisionados, o algoritmo aprende a partir de dados rotulados (as chamadas *features*), ou seja, o conjunto de treinamento inclui as respostas corretas (variáveis-alvo), permitindo que o modelo faça previsões para novos dados com base nesse aprendizado. Já os modelos não supervisionados trabalham sem rótulos, buscando identificar padrões ou agrupamentos de forma autônoma, permitindo uma exploração mais livre dos dados, o que pode ser útil em cenários onde as informações completas não estão disponíveis.

Neste trabalho, foram exploradas duas das técnicas bem populares no campo da ML: *Random Forest* e *XGBoost*. Ambas são técnicas de aprendizado supervisionado e baseadas em árvores de decisões (explicação melhor na Seção 3.2.1.1), o que significa que treinamos os algoritmos em um conjunto de dados rotulados (*features*) e, em seguida, esses modelos treinados foram utilizados para prever parâmetros estelares em novos conjuntos de dados.

3.2.1.1 Árvores de Decisão

Árvores de decisão são modelos amplamente utilizados em tarefas de classificação e regressão, que representam decisões em uma estrutura hierárquica semelhante a uma árvore. Cada nó interno corresponde a uma condição baseada em uma das características do conjunto de dados, enquanto os nós terminais, ou folhas, fornecem o resultado da previsão. Durante o treinamento, o algoritmo busca as divisões mais informativas com base em critérios como a impureza de Gini[25] ou o ganho de informação, permitindo separar eficientemente os dados (Baron, 2019). A construção da árvore é feita de forma recursiva, até que todas as amostras estejam corretamente agrupadas nos nós finais ou até que um critério de parada seja atendido.

Os modelos de aprendizado baseados em árvores de decisão utilizam um processo de divisão sequencial dos dados, onde cada divisão ou "ramificação" é guiada por regras do tipo *if-then*. Essas regras avaliam uma variável de entrada em comparação com um valor de corte, o que direciona o dado para uma de suas ramificações. Esse processo é repetido em diferentes nós da árvore, formando várias ramificações, até que os dados sejam classificados em uma categoria ou uma previsão seja feita para um valor numérico, como em problemas de regressão.

Por outro lado, as *features* são informações inseridas pelo usuário nos modelos supervisionados e representam as características que o modelo utiliza para fazer previsões. Elas são essenciais para o aprendizado do modelo, pois cada *feature* fornece uma parte da informação necessária para que o algoritmo identifique padrões e faça inferências a partir dos dados. A escolha e a qualidade dessas *features* influenciam diretamente o desempenho do modelo, tornando fundamental a seleção criteriosa das variáveis mais relevantes para a tarefa em questão.

Baron (2019) cita que uma das principais vantagens das árvores de decisão é sua capacidade de lidar com grandes volumes de dados e um grande número de características, mantendo a simplicidade na interpretação dos resultados. Elas podem identificar a importância relativa de diferentes características na previsão, permitindo que os astrônomos compreendam quais parâmetros estelares são mais relevantes para as suas análises. Além disso, a construção recursiva das árvores permite que o modelo ajuste-se a padrões complexos nos dados, o que é particularmente útil em contextos como a astronomia, onde os dados podem ser altamente variáveis e multidimensionais (Baron, 2019).

3.2.2 *Random Forest*

Como mencionado anteriormente, a técnica de *Random Forest* (RF), ou Floresta Aleatória, é um modelo supervisionado baseado em árvores de decisão. O RF pertence à categoria dos métodos *ensemble*, que combinam múltiplos modelos — no caso, diversas

[25] Veja Yao & Liu (2013), Baron (2019) e Cordeiro da Silva (2023) para melhor explicação.

árvores de decisão — para formar um único modelo preditivo, resultando em uma melhora do desempenho em comparação com o desempenho de cada algoritmo supervisionado individualmente. Além disso, podem ser combinados com diferentes algoritmos supervisionados ou informações de um único algoritmo treinado em diferentes subconjuntos do conjunto de treinamento (Baron, 2019).

Por sua vez, os métodos de ensemble são divididos em duas categorias principais: métodos de *Bagging* e métodos de *Boosting*. Nos métodos de *Bagging* (também chamados de Métodos de Média), como o RF, os estimadores base (neste caso, árvores de decisão) são treinados de forma independente, e a previsão final é obtida pela média (ou votação majoritária) das previsões de cada árvore. Esse procedimento adiciona uma camada de aleatoriedade, daí o *Random*, o que melhora a robustez e a generalização do modelo ao reduzir a correlação entre as árvores.

Além disso, quando um novo objeto é introduzido, ele é propagado por todas as árvores da floresta, e a *feature* atribuída será aquele que for mais frequente entre as árvores. Embora uma única árvore tenha uma tendência a sobreajustar (*Overfitting*) os dados de treinamento, o uso de múltiplas árvores combinadas em uma floresta permite que o modelo se generalize melhor para novos dados, resultando em um desempenho superior (Breiman, 2001). O sobreajuste ocorre quando um modelo aprende padrões muito específicos dos dados de treinamento, perdendo a capacidade de generalizar para novos dados, ou seja, em termos mais práticos o modelo "decorou" os padrões iniciais impossibilitando um novo aprendizado e uma nova aplicação em uma base de dados diferentes.

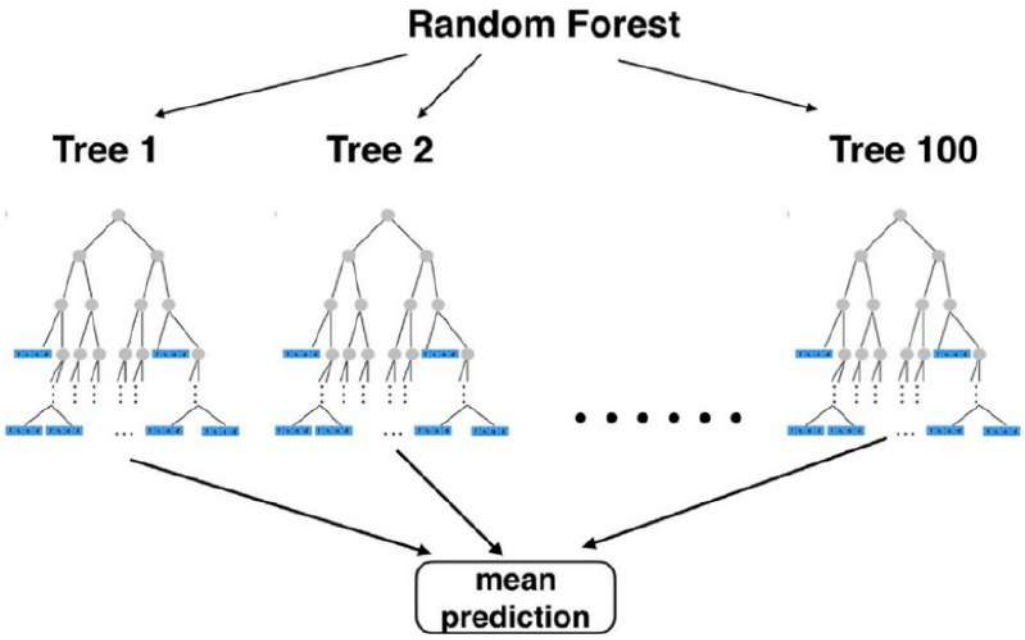

Figura 3.1: Esquema representando a hierarquização das árvores no *Random Forest*. As árvores são compostas por nós internos (representados por círculos cinza), nós terminais (retângulos azuis) e arestas (linhas cinza) que as conectam. Fonte: Nedjati-Gilani *et al.* (2017).

Na Figura 3.1, é apresentada a hierarquização das árvores no algoritmo. As árvores são compostas por nós internos (representados por círculos cinza), nós terminais (retângulos azuis) e arestas (linhas cinza) que as conectam. Cada nó interno possui uma borda de entrada e duas bordas de saída e armazena um teste que, com base nos dados, define qual borda de saída será seguida. O resultado final da previsão é obtido pela média ponderada das estimativas fornecidas por cada árvore de decisão individual.

O RF é um algoritmo disponível gratuitamente e de código aberto na biblioteca *Scikit-Learn*[26] para *Python*. Neste trabalho, foi utilizada a versão 1.5.0 da *Scikit-Learn*, integrada à versão 3.12.4 do *Python*. Além disso, o pacote *Pandas* v2.2.2 foi empregado para a importação e manipulação das amostras, permitindo o ajuste e treinamento do modelo. Para mais informações, exemplos de *scripts* e detalhes de configuração, consulte a documentação oficial da biblioteca *Scikit-Learn*.

Já os métodos de *Boosting*, que é o caso do *XGBoost*, serão melhor explicados na Seção 3.2.3 a seguir.

3.2.3 *XGBoost*

O *eXtreme Gradient Boosting*, ou simplesmente *XGBoost* (XGB), é um algoritmo baseado no método de *boosting* que ganhou notoriedade em meados de 2015, quando ele foi amplamente reconhecido em vários desafios envolvendo ML (Chen & Guestrin, 2016). De acordo com o autor das 29 soluções que foram vencedoras nos desafios que foram publicados no blog Kaggle [27], 11 delas apresentaram soluções envolvendo XGB. Além disso, em várias outras competições, como por exemplo as que foram realizadas durante a KDD 2015[28], onde todas as equipes que ficaram entre as 10 melhores, usaram o XGB (Chen & Guestrin, 2016). Esses resultados evidenciam a eficácia do algoritmo e sua ampla aceitação na área de ML.

O algoritmo é uma implementação otimizada da técnica de *Gradient Boosting* chamada de *eXtreme Gradient Boosting*, cujo foco está em melhorar o desempenho por meio do aprendizado em um espaço funcional, utilizando pseudo-resíduos em vez de resíduos convencionais. Diferente dos resíduos tradicionais, que são a diferença direta entre a predição e o valor real, os pseudo-resíduos são calculados a partir do gradiente da função de perda, ou seja, são uma aproximação dos resíduos que indicam a direção em que o modelo deve ser ajustado para reduzir o erro. Essa modificação permite que o modelo aprenda de

[26] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.htAM.

[27] Kaggle é um grande repositório usado por pesquisadores, cientistas, professores e entusiastas de ML onde compartilham, testam a resistência e mantém-se atualizados sobre todas as técnicas e tecnologias de ML mais recentes. Fonte: https://www.kaggle.com/.

[28] A *Knowledge Discovery and Data Mining 2015* foi uma conferência realizada em 2015 na cidade de Sydney, Austrália, reunindo pesquisadores e profissionais de mineração e análise de *big data*. Fonte: https://www.kdd.org/kdd2015/.

forma mais eficaz e eficiente durante o processo de treinamento (Friedman, 2002; Hastie *et al.*, 2009).

O algoritmo cria um conjunto de modelos fracos, normalmente árvores de decisão simples, que, quando combinados, formam um conjunto robusto de predição (Friedman, 2002; Hastie *et al.*, 2009). Quando árvores de decisão são usadas como aprendizes fracos, o método resultante é chamado de gradient-boosted trees, que geralmente supera o Random Forest em desempenho (Hastie *et al.*, 2009).

Diferentemente do RF, que combina as previsões de várias árvores independentes em um processo paralelo, o XGB segue uma abordagem sequencial. Em vez de treinar todas as árvores ao mesmo tempo, o algoritmo ajusta cada árvore para corrigir os erros cometidos pelas anteriores, permitindo que o modelo se ajuste progressivamente aos dados, identificando padrões que os primeiros modelos não conseguiram capturar.

Cada nova árvore é treinada para minimizar os erros residuais do modelo anterior, tornando o aprendizado mais direcionado e eficaz. Dessa forma, ele reduz iterativamente os erros de predição, melhorando a precisão geral do modelo (veja a Figura 3.2).

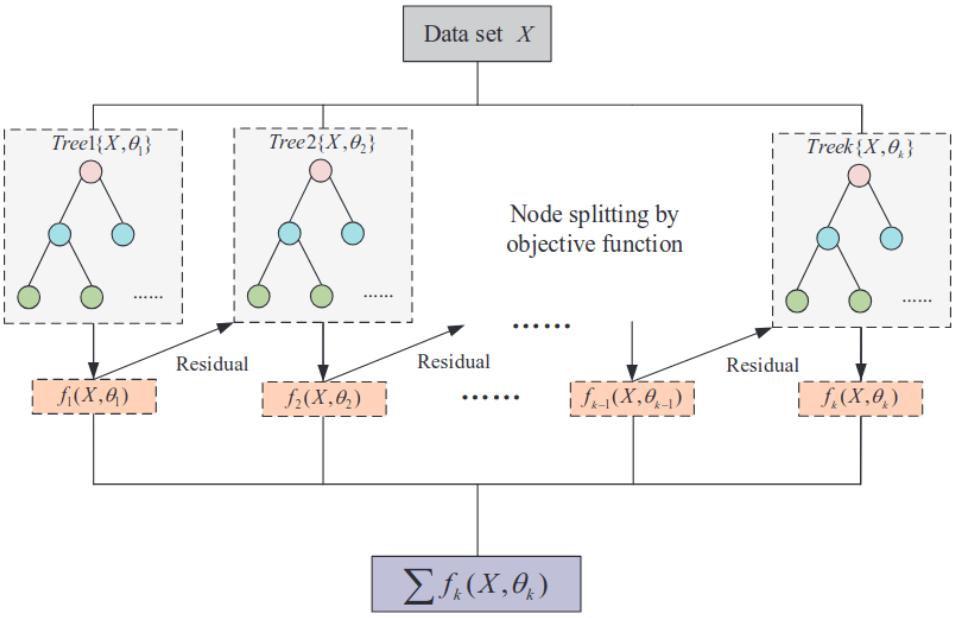

Figura 3.2: Esquema de sequenciamento do *XGBoost* na previsão de modelos. Os retângulos tracejados em cinza mais claro representam cada árvore, conforme esquematizado na Figura 3.1. Os retângulos tracejados em laranja mostram os resíduos resultantes de cada árvore, que são direcionados para a próxima árvore com o objetivo de minimizar os erros até a previsão final, indicada pelo modelo no retângulo em cinza mais escuro. No processo de aprendizagem, X representa as características ou variáveis de entrada (a matriz de dados), e θ representa os parâmetros aprendidos pelo modelo durante o treinamento, como os pesos nos nós das árvores. Fonte: Chen & Guestrin (2016).

No contexto de predição de parâmetros estelares, utilizando a temperatura efetiva, como exemplo, o XGB utiliza um processo iterativo para ajustar as previsões. Suponha

que a primeira árvore faça uma previsão inicial da T_{ef} de uma estrela, por exemplo, como sendo 5000 K, enquanto o valor real observado é 5200 K. O resíduo dessa previsão seria de 200 K. A segunda árvore, então, se concentra em prever e corrigir esse resíduo de 200 K, ajustando a previsão para 5100 K, 5120K e assim por diante, aproximando-se ainda mais do valor real. Árvores subsequentes continuam a corrigir os erros acumulados das previsões anteriores, ajustando os resíduos e melhorando continuamente as previsões.

Cada árvore não faz uma previsão independente, esse processo permite que o modelo aprenda de maneira sequencial e corrija suas deficiências, resultando em previsões mais precisas dos parâmetros das estrelas. Esse procedimento evita que o modelo caia em armadilhas comuns de *overfiting* e tende a resultar em um desempenho superior, especialmente em conjuntos de dados complexos e com alta variabilidade.

Além disso, o algoritmo permite várias otimizações, como a regularização para evitar o sobreajuste, a paralelização de partes do processo de aprendizado e o controle preciso da complexidade do modelo. Isso faz com que, em comparação com outros métodos de *boosting*, como o *AdaBoost*, o XGB atinja um equilíbrio mais eficiente entre viés e variância[29], oferecendo modelos mais robustos e previsões mais precisas (Chen & Guestrin, 2016; Friedman, 2001; Hastie *et al.*, 2009).

O XGB também é um algoritmo gratuito e de código aberto. Sua instalação é feita normalmente usando o *pip install* em qualquer plataforma Linux. Neste trabalho foi usada a versão 2.0.3, e sua documentação pode ser consultada em seu site oficial[30] para informações mais detalhadas de instalação e configuração.

3.2.4 Hiperparâmetros

Diferentemente dos parâmetros, que são aprendidos a partir dos dados durante o treinamento, os hiperparâmetros devem ser definidos antes do início do processo de aprendizado. Eles são variáveis de configuração que controlam o comportamento do treinamento e o funcionamento dos modelos de ML (Yang & Shami, 2020). A escolha dos valores para os hiperparâmetros tem um impacto direto na qualidade do modelo, influenciando sua capacidade de aprendizado e sua habilidade de generalizar para dados não vistos anteriormente (Yang & Shami, 2020).

Eles podem incluir configurações como a profundidade máxima de uma árvore de decisão, a taxa de aprendizado em algoritmos de *boosting*, o número de *features* a serem consideradas e entre outros. A escolha adequada dos hiperparâmetros pode melhorar significativamente o desempenho de um modelo, para isso é necessário fazer uma busca dos valores que melhor se encaixam com base nos conjuntos de dados disponíveis.

[29] O viés refere-se ao erro sistemático do modelo (quando o modelo é muito simples e perde a precisão), enquanto a variância diz respeito à sensibilidade do modelo às flutuações nos dados de treinamento (quando o modelo é muito complexo e acaba "memorizando" os dados).

[30] https://xgboost.readthedocs.io/en/stable/index.htAM.

Existem inúmeras formas de encontrar os melhores hiperparâmetros, e de acordo com Bergstra & Bengio (2012), as mais utilizadas são a busca em grade e a busca aleatória. Em ambas as abordagens, cada hiperparâmetro é tratado como uma dimensão de um hiperespaço, onde cada ponto representa uma possível combinação de valores de hiperparâmetros para o algoritmo.

Na busca em grade (*GridSearch*), o hiperespaço é explorado de forma sistemática, ao longo de uma grade predefinida de valores (vide a Figura 3.3). São selecionados conjuntos de valores para cada hiperparâmetro, e todas as combinações possíveis entre esses valores são testadas, escolhendo-se a melhor configuração. Por outro lado, na busca aleatória (*RandomSearch*), o espaço de hiperparâmetros é explorado de maneira estocástica (conforme mostrado na Figura 3.3), onde um número pré-definido de pontos é sorteado aleatoriamente dentro dos intervalos de valores possíveis para cada hiperparâmetro, e cada ponto é avaliado separadamente.

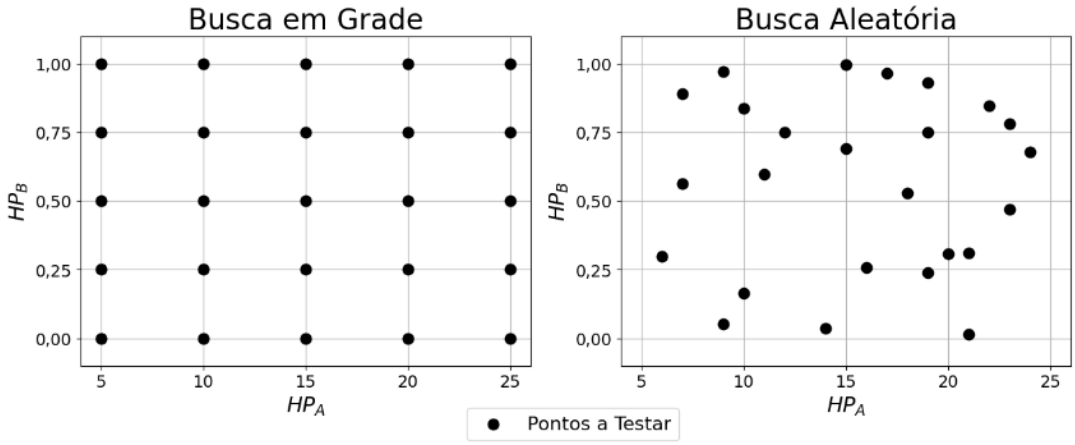

Figura 3.3: Exemplo das cobertura do hiperespaço de parâmetros. À esquerda é um exemplo de hiperespaço na busca em grade e à direita é no caso da busca aleatória, onde HP_A e HP_B são exemplos de hiperparâmetros. Fonte: Cordeiro da Silva (2023).

A busca pelos melhores hiperparâmetros é uma etapa crucial no desenvolvimento do modelo, que será aplicado na terceira grande etapa. Esses parâmetros controlam o comportamento dos modelos, como no caso dos algoritmos de ML desse trabalho, onde foi-se utilizado a busca em grade. O R^2 *Score* e o desvio mediano absoluto (*median absolute deviation*, MAD) foram utilizadas como métricas de avaliação.

O R^2 *Score* é uma métrica que avalia a qualidade da comportamento do modelo com relação aos dados, variando de 0 a 1. Quanto mais próximo de 1, melhor é o desempenho deste modelo. Ele pode ser calculado usando a seguinte fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.2}$$

onde y_i são os valores reais, $\hat{y}_i$ são os valores previstos pelo modelo, e $\bar{y}$ é a média de

todos os valores reais. A métrica mede o quão bem o modelo explica a variabilidade dos dados em relação à média.

O MAD é a mediana das diferenças absolutas entre os valores reais e os valores previstos. Ele é uma métrica robusta para medir a dispersão dos erros nas previsões, sendo menos sensível a valores atípicos em comparação às outras métricas utilizadas para avaliar modelos de ML. O MAD não considera a direção dos erros, tratando cada desvio com o mesmo peso, mas sua ênfase na mediana garante que a métrica reflita de maneira mais consistente o comportamento típico dos erros, mesmo em distribuições assimétricas ou com *outliers*[31].

Para calcular o MAD é preciso considerar os erros das previsões, que podemos também chamar de resíduos. Eles são calculados como:

$$E_i = X_{i,r} - X_{i,p} \tag{3.3}$$

onde E_i representa o erro da i-ésima previsão, $X_{i,r}$ corresponde ao valor real e $X_{i,p}$ ao valor previsto para o mesmo índice. Com os erros calculados para todos os objetos da amostra, a MAD é determinada pela seguinte Equação:

$$\text{MAD} = \text{mediana}(|E_i|) \tag{3.4}$$

O MAD é calculado como a mediana das diferenças absolutas entre todos os erros individuais.

O *GridSearch* é uma ferramenta simples de aplicar e faz parte do pacote amplamente usado neste trabalho, o *Scikit-Learn* do *Python*. Associando essa biblioteca em conjunto com a biblioteca *Pandas*, é possível importar Tabelas de dados e ajustá-las para realizar a busca em grade, o treinamento e o teste do modelo.

Para a técnica de RF, os principais hiperparâmetros ajustados foram:

- **n_estimators:** Consiste no número de árvores na floresta aleatória.
- **max_depth:** Consiste na profundidade máxima de cada árvore, que não pode ser muito profunda nem muito rasa, pois pode causar o sobreajuste (*overfiting*).
- **min_samples_split:** É o número mínimo de amostras necessárias para dividir um nó.
- **min_samples_leaf:** É o número mínimo de amostras necessárias em um nó folha.
- **max_features:** Número de *features* a serem consideradas em um nó ao procurar a melhor divisão.

[31] *Outliers* são valores extremos ou atípicos em um conjunto de dados, que diferem significativamente dos demais valores e podem distorcer análises estatísticas.

- **bootstrap:** Determina se as amostras são selecionadas com ou sem substituição. Quando definido como `True`, cada árvore é treinada em uma amostra aleatória com substituição dos dados de treinamento.

- **random_state:** Controla a aleatoriedade do modelo. Definir este parâmetro garante que os mesmos resultados sejam reproduzidos em execuções diferentes do modelo.

A combinação de valores que foram implementados no *GridSearch* engloba alguns valores escolhidos e apresentados nos trabalhos de Carvalho (2022) e Cordeiro da Silva (2023). Foram escolhidos também, aqueles que garantem a reprodutibilidade do modelo em diferentes execuções, ao mesmo tempo que abrangem o maior intervalo possível de valores e combinações. Tudo isso, levando em conta o tempo e o poder de processamento disponíveis para a realização do trabalho. A inclusão de mais opções na grade resulta em um aumento no tempo de execução e na demanda por recursos computacionais, como memória RAM e CPU. Por outro lado, uma grade mais restrita reduz o tempo e o poder de processamento necessários, porém limita o número de combinações possíveis na grade. Isso pode impactar negativamente na eficácia da busca pelo melhor modelo. Após equilibrar tempo, poder de processamento e número de opções, foi possível chegar à seguinte grade de valores:

- `n_estimators`: [1, 5, 10, 25, 50, 100, 150, 200]
- `max_features`: [1, 5, 10, 20, 30, 50, 60, num_features]
- `min_samples_leaf`: [0.01, 0.05, 0.1, 1, 2, 5, 10, 20]
- `bootstrap`: [True, False]
- `max_depth`: [None, 1, 5, 10, 20, 50, 100, 150]
- `min_samples_split`: [2, 5, 10, 20, 50, 100, 150, 200]
- `random_state`: [42[32]]

O valor `num_features`, utilizado como uma das opções de `max_features`, representa o total de variáveis de entrada consideradas no modelo. Ele inclui todas as combinações possíveis de cores derivadas das magnitudes. Como a cor é definida pela diferença entre duas magnitudes, o número total de combinações possíveis a partir de 12 magnitudes é 66. Portanto, cada amostra de desenvolvimento utilizou um total de 78 variáveis (`num_features` = 78), sendo 12 magnitudes individuais e 66 combinações de cores.

Para a segunda técnica utilizada nesse trabalho, o XGB, os principais hiperparâmetros ajustados para o treinamento do modelo foram:

[32]É um valor padrão do algoritmo de RF e XGB. De acordo com a documentações, é para garantir a reprodutibilidade dos modelos.

- **n_estimators:** Consiste no número de árvores no conjunto.
- **max_depth:** Controla a profundidade máxima de cada árvore.
- **learning_rate:** Taxa de aprendizado, controlando o impacto de cada árvore no modelo.
- **subsample:** Fração de amostras utilizadas para treinar cada árvore.
- **colsample_bytree:** Fração de características a serem consideradas para cada árvore.
- **gamma:** Redução mínima na função de perda necessária para fazer uma nova partição.

Adotada a mesma estratégia utilizada para o RF. A grade de parâmetros selecionados abrange os seguintes valores:

- `n_estimators`: [1, 5, 10, 50, 100, 150]
- `max_depth`: [3, 5, 10, 20, 30, 50]
- `learning_rate`: [0.01, 0.05, 0.1, 0.2, 0.3, 0.5]
- `subsample`: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
- `colsample_bytree`: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
- `gamma`: [0, 0.01, 0.1, 0.2, 0.3, 0.5]
- `random_state`: [42]

A busca, tanto para a técnica de RF quanto para o XGB, foi aplicada aos parâmetros T_{ef}, log g e [Fe/H] reunidos na etapa anterior para todas as amostras de dados (explicada na Seção 3.1).

Para realizar as buscas dos melhores valores, os modelos foram classificados conforme apresentado na Seção 3.1.2, e feito buscas individuais para cada amostra. Assim, conhecendo os melhores hiperparâmetros para cada tipo de amostra, foi possível gerar gráficos que comparam as temperaturas, metalicidades e gravidades superficiais reais com as previstas pelo algoritmo. Essa análise permitirá avaliar o desempenho dos modelos com base nos seus R^2 *Scores*, o MAD e a quantidade de objetos da amostra de entrada.

3.3 Aplicação dos Melhores Modelos em Levantamentos que Buscam Exoplanetas

Nesta seção, será descrito como foi feita a aplicação dos modelos treinados e apresentados na seção anterior aos levantamentos que buscam exoplanetas. Cada modelo gerado resultou em um arquivo, que foi aplicado utilizando um segundo algoritmo em *Python*, com suporte da biblioteca *joblib*[33] e *astropack*. A expectativa é que essa aplicação permita uma avaliação da robustez dos modelos e a melhor caracterização de estrelas hospedeiras de exoplanetas. O novo algoritmo utiliza a modelagem obtida no processo anterior para prever os parâmetros estelares de qualquer estrela dos levantamentos que também tenha sido observada pelo J-PLUS e pelo S-PLUS, já que os melhores modelos foram definidos com base nas magnitudes das observações desses levantamentos.

Para identificar as estrelas em comum entre esses levantamentos, será novamente realizada um cruzamento utilizando ascensão reta e declinação. É essencial que as estrelas observadas atendam aos mesmos critérios das amostras de treinamento, pois os modelos foram baseados em estrelas que também seguiam esses requisitos.

Uma parte importante do processo de aplicação dos modelos é a estimativa dos erros associados aos parâmetros estelares previstos. Para isso, foi utilizado o método de Monte Carlo, que consiste em uma técnica estatística amplamente empregada para avaliar incertezas e erros em modelos preditivos. No contexto deste trabalho, o método de Monte Carlo foi aplicado gerando múltiplas amostras de dados simulados, baseadas nas distribuições de incertezas das magnitudes estelares observadas.

Em cada interação do método de Monte Carlo, as magnitudes foram alteradas de acordo com suas incertezas estimadas, gerando variações aleatórias nos valores. Os modelos treinados foram aplicados a cada uma dessas amostras simuladas, permitindo prever os parâmetros estelares em diferentes cenários. Ao final do processo, foram obtidos a média, o desvio padrão e a mediana dos erros dos parâmetros previstos. Este método é particularmente eficiente para este tipo de análise porque incorpora as incertezas dos dados de entrada de forma robusta, produzindo uma visão estatisticamente rigorosa dos erros previstos.

3.3.1 Correlação entre os Levantamentos Principais com os Levantamentos que Buscam Exoplanetas

As bases de dados das missões Kepler e TESS e do levantamento com o HARPS foram utilizadas como entrada na busca de exoplanetas ou de candidatos em potencial. Para sabermos quais objetos estão em comum com os nossos levantamentos principais, J-PLUS e S-PLUS, fizemos novamente o cruzamento (*crossmatch*) com esses levantamentos.

[33] Documentação disponível em: https://joblib.readthedocs.io/en/stable/.

Após os dados dos levantamentos serem selecionados, foi realizado o cruzamento com as amostras já preparadas (vide a Tabela 3.1) e obtivemos as seguintes amostras:

Amostra	**Restrita**	**Menos Restrita**
J-PLUS DR3 + GAIA DR3 + KIC	60.731	112.324
J-PLUS DR3 + GAIA DR3 + TIC v8.2	1.628.834	2.035.814
J-PLUS DR3 + GAIA DR3 + HARPS	0	0

Tabela 3.2: Quantidade de objetos em cada amostra de entrada para o treinamento após o cruzamento de objetos do J-PLUS com os levantamentos que buscam exoplanetas.

Amostra	**Restrita**	**Menos Restrita**
S-PLUS iDR5 + GAIA DR3 + KIC	91	101
S-PLUS iDR5 + GAIA DR3 + TIC v8.2	3.058.419	3.577.117
S-PLUS iDR5 + GAIA DR3 + HARPS	83	108

Tabela 3.3: Quantidade de objetos em cada amostra de entrada para o treinamento após o cruzamento de objetos do J-PLUS com os levantamentos que buscam exoplanetas.

Com os objetos definidos, foi aplicado os modelos, conforme descrito anteriormente e os resultados podem ser conferidos na Seção 4.4 no Capítulo 4.

3.3.2 Luminosidade, Raio e Massa das Estrelas

Com os levantamentos voltados à busca de exoplanetas definidos, foi realizado um cruzamento com os levantamentos J-PLUS e S-PLUS, e com isso, os parâmetros de T_{ef} , log g e [Fe/H] foram determinados pelos melhores modelos como explicado anteriormente.

A partir desses parâmetros, calculamos outros parâmetros fundamentais das estrelas hospedeiras, tais como luminosidade, raio e massa estelar, que são essenciais para sua caracterização completa. Essa etapa é crucial para entendermos as propriedades físicas dessas estrelas e, consequentemente, os sistemas planetários associados.

O primeiro parâmetro a ser determinado é a luminosidade ($L_\star$), que depende diretamente dos valores previstos por meio de ML. Para isso, inicia-se com o cálculo da magnitude absoluta na banda G (M_G), seguido pela magnitude bolométrica (M_{bol}). A M_{bol} considera o brilho da estrela em todos os comprimentos de onda, e não apenas em uma banda específica, como no caso de filtros fotométricos individuais. O cálculo de M_G será realizado utilizando a Equação 3.1.

Para calcular a M_{bol}, é indispensável determinar a correção bolométrica (BC). Essa correção é aplicada à magnitude absoluta na banda G (M_G) para convertê-la em M_{bol}, permitindo considerar o brilho total da estrela em todos os comprimentos de onda. Co-

nhecendo a BC de cada objeto, é possível encontrar a magnitude bolométrica M_{bol}. Para isso, utilizamos a Equação 3.5 (Jordi *et al.*, 2010).

$$BC_G = M_{bol} - M_G \rightarrow M_{bol} = M_G + BC_G \tag{3.5}$$

Para calcular BC foram consideradas as correções apresentadas no trabalho de Jordi *et al.* (2010) que fornece os dados de BC para uma amostra de estrelas observadas pelo Gaia com base em seus valores de T_{ef} , log g e [Fe/H]. Para isso, foi feito o uso das técnicas de ML usadas neste trabalho para criar um modelo que calcule a BC para estrelas do catálogos principais (J-PLUS/S-PLUS) com base nos parâmetros previstos.

O algoritmo desenvolvido utilizando RF para a previsão da correção bolométrica apresentou um R^2 de 0,9970 no desempenho geral. Enquanto o algoritmo utilizando XGB apresentou um R^2 de 0,9983. Os gráficos referentes aos treinamentos dos modelos de RF e XGB é exibido nas Figuras 3.4 e 3.5, respectivamente. Já a Tabela 3.4 apresenta as incertezas associadas às previsões de BC para cada técnica de ML.

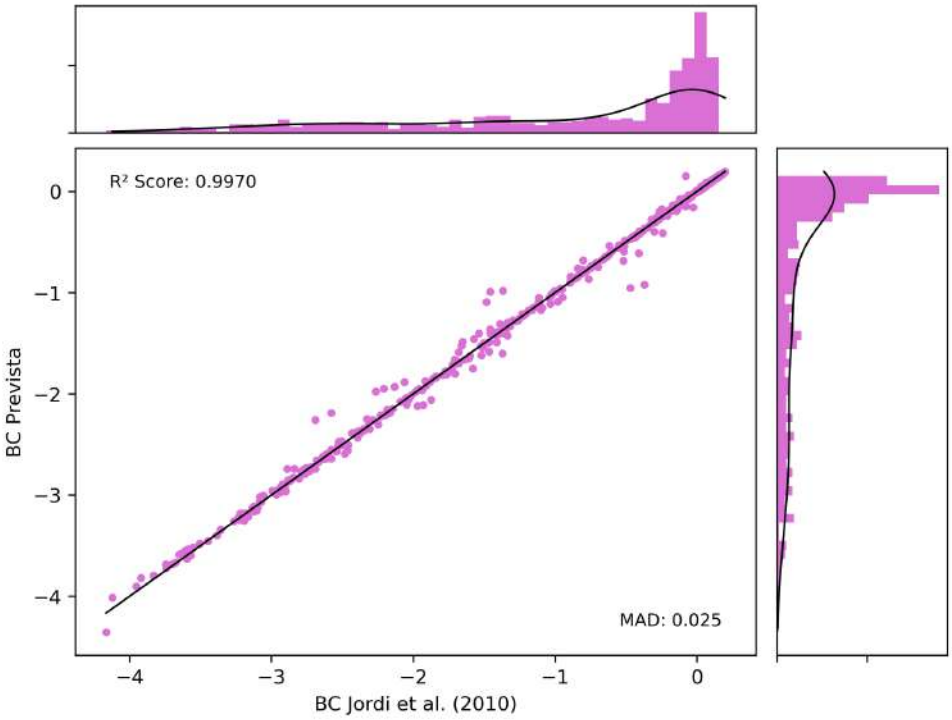

Figura 3.4: Modelagem utilizando *Random Forest* para correção bolométrica baseada nos dados de Jordi *et al.* (2010).

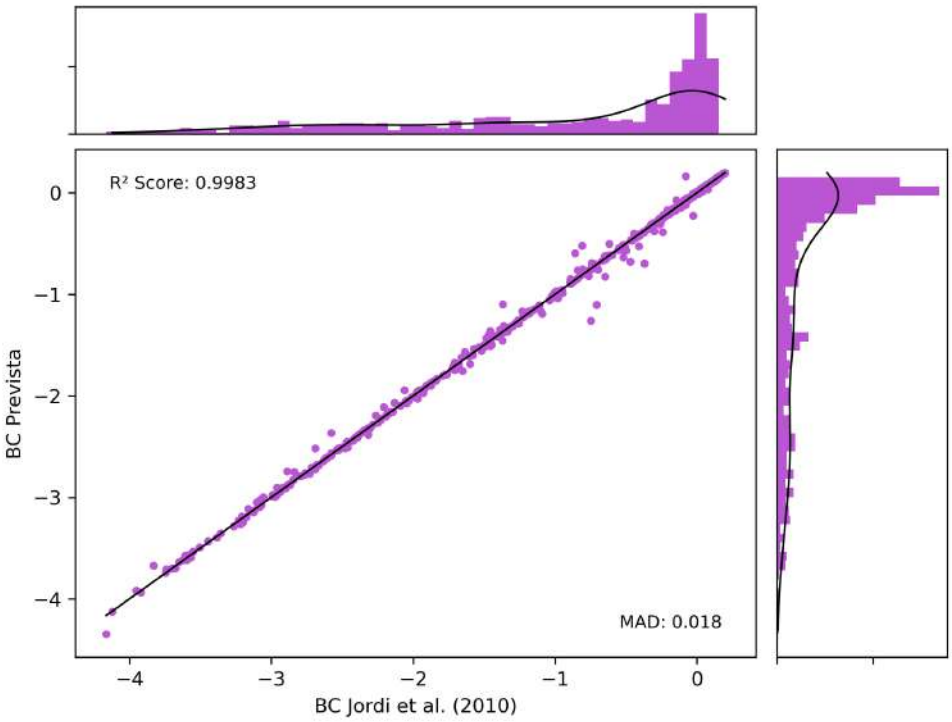

Figura 3.5: Modelagem utilizando *XGBoost* para correção bolométrica baseada nos dados de Jordi *et al.* (2010).

Métrica	**RF**	**XGB**
R^2 *Score*	0,9970	0,9983
MAD	0,0247	0,0180
Desvio padrão (σ)	0,0573	0,0430
Incerteza (3σ)	0,1719	0,1291

Tabela 3.4: Métricas para o algoritmo de correção bolométrica com *Random Forest* e com *XGBoost*.

Após alguns testes e execuções, foi identificado que o valor aceitável para se considerar a incerteza de BC prevista pelo algoritmo corresponde a até 3 vezes o do desvio padrão (3σ), o que é válido para mais de 90% dos objetos. Para determinar a BC das estrelas, foi aplicado os modelos gerados com dados de Jordi *et al.* (2010) para cada respectiva técnica, ou seja, para parâmetros obtidos com RF será usado a correção de BC obtida pela mesma técnica, assim para como no XGB.

O valor da incerteza de M_{bol} foi tomada como a incerteza em BC_G, pois muitos objetos não possuem paralaxes e distâncias confiáveis, tornando σ_d não confiável, inválido e até inexistentes em alguns casos. Com isso, os erros dos parâmetros estão subestimados. Com tudo determinado, foi possível determinar a luminosidade L com a relação demonstrada por Carroll & Ostlie (2017), onde:

$$M_{bol} = -2,5 \log \frac{L}{L_0} \rightarrow L = 10^{-0,4M_{bol}} L_0 \tag{3.6}$$

onde L_0 é a luminosidade no ponto zero com valor de $3,0128 \times 10^{28}$ W (Resolução B1 da União Astrônomica Internacional - IAU)[34]. Propagando as incertezas, é possível determinar a incerteza da luminosidade segundo a Equação 3.7:

$$\sigma_L = \sqrt{[(-0,4)10^{0,4M_{bol}} \ln(10)\sigma_{M_{bol}} L_0)]^2 + (10^{-0,4M_{bol}}\sigma_{L_0})^2} \tag{3.7}$$

Considerando zero o valor de σ_{L_0} o segundo termo na Equação 3.7 é desconsiderado. Portanto:

$$\sigma_{L_\star} = (-0,4)10^{-0,4M_{bol}} \ln(10)\sigma_{M_{bol}} L_0 \tag{3.8}$$

A partir da Equação 3.6 e com a temperatura efetiva prevista pelos algoritmos, tem-se que o raio da estrela pode ser calculado pela Equação 3.9:

$$L_\star = 4\pi R_\star^2 \sigma T_{\rm ef}^4 \rightarrow R_\star = \left(\frac{L}{4\pi\sigma T_{\rm ef}^4}\right)^{1/2} \tag{3.9}$$

onde σ é a constante de Stefan-Boltzmann, com o valor de $5,6697 \times 10^5$ erg cm^{-2} s^1 K^4. É possível, escrever em termos de unidades solares (comumente usadas na astronomia) onde:

$$\frac{L}{L_\odot} = \left(\frac{R}{R_\odot}\right)^2 \left(\frac{T_{\rm ef}}{T_\odot}\right)^4 \rightarrow \frac{R}{R_\odot} = \frac{(L/L_\odot)}{(T_{\rm ef}/T_\odot)} \tag{3.10}$$

Para propagação de erro do raio da estrela, vamos fazer uma simplificação. Vamos chamar $R_\star = R/R_\odot$ e $L_\star = L/L_\odot$, portanto a Equação 3.9 fica:

$$R_\star = L_\star^{1/2} \left(\frac{T_{\rm ef}}{T_\odot}\right)^{-2} \tag{3.11}$$

Como a Resolução B1 da IAU que considera que o valor para $T_\odot$ é de 5772 K e ela não fornece o valor de $\sigma T_\odot$, ele não será considerado na propagação das incertezas. A incerteza de $R_\star$ pode ser obtida através da Equação 3.12.

$$\sigma_{R_\star} = \sqrt{\left(\frac{1}{2} L_\star^{-1/2} \sigma_{L_\star} \left(\frac{T_{\rm ef}}{5772}\right)^{-2}\right)^2 + \left((-2) L_\star^{1/2} \left(\frac{T_{\rm ef}}{5772}\right)^{-3} \left(\frac{\sigma_{T_{\rm ef}}}{5772}\right)\right)^2} \tag{3.12}$$

Para o calculo da massa estelar ($M_\star$) foi utilizada a relação proposta por Torres *et al.* (2010). Os autores calcularam as massas de 190 estrelas pertencentes a 95 sistemas binários separados. Esses sistemas são não interagentes, ou seja, não apresentam troca de

[34] Disponível em: https://www.iau.org/static/resolutions/IAU2015_English.pdf.

matéria entre as componentes, permitindo que as estrelas evoluam de forma independente, como estrelas isoladas (do inglês, *single stars*). Além disso, o estudo se aplica à derivação de massas para estrelas isoladas na (ou após) a sequência principal com massas acima de 0,6 $M_\odot$ (Torres *et al.*, 2010).

Os autores usaram uma função polinomial que depende dos parâmetros T_{ef}, log g e [Fe/H] para calcular as massas estelares, com erros de aproximadamente 6,4%. Essa função de massa é definida pelos coeficientes de calibração a_i, cujos valores estão listados na Tabela 3.5.

i	a_i
1	1,5689 ± 0,058
2	1,3787 ± 0,029
3	0,4243 ± 0,029
4	1,139 ± 0,24
5	-0,1425 ± 0,011
6	0,01969 ± 0,0019
7	0,1010 ± 0,014

Tabela 3.5: Tabela com os coeficiente de calibração a_i e seus respectivos erros calculados por Torres *et al.* (2010). Fonte: Torres *et al.* (2010).

Com os coeficientes conhecidos, é possível determinar a massa da estrela. A massa pode ser calculada, segundo Torres *et al.* (2010), através da Equação 3.13

$$\log M_\star = a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_5(\log \mathrm{g})^2 + a_6(\log \mathrm{g})^3 + a_7[\mathrm{Fe/H}] \tag{3.13}$$

onde, de acordo com os autores, $X = \log(T_{ef})$, $M_\star$ é dado em $M_\odot$ e a incerteza de $M_\star$ ($\sigma_{M_\star}$) é:

$$\sigma_{M_\star} = \frac{6,4}{100} M_\star \tag{3.14}$$

Com essas equações, relações e com os parâmetros previstos pelos algoritmos de ML, é possível caracterizar com precisão as estrelas hospedeiras de exoplanetas. Os resultados encontrados, serão apresentados a seguir no Capítulo 4.

Capítulo 4

Resultados e Discussões

Neste Capítulo, apresentamos os resultados obtidos a partir da metodologia descrita e dos dados apresentados no Capítulo 3.

Na primeira seção deste Capítulo, discutimos os melhores hiperparâmetros otimizados pelo *GridSearch* e utilizados na construção e no teste dos modelos deste projeto, para cada técnica de ML.

Com os hiperparâmetros definidos, as Seções 4.2 e 4.3 detalham o desempenho dos modelos obtidos. Os melhores modelos são aqueles que possuem um R^2 *Score* mais próximo de 1, menor desvio mediano absoluto e maior quantidade de objetos na amostra. Em análise estatística e no treinamento de modelos de aprendizado de máquina, uma maior quantidade de objetos na amostra de treinamento tende a resultar em modelos mais confiáveis. Devido a isso, esse critério é considerado o mais relevante, superando os demais nas definições dos melhores modelos.

Na Seção 4.4, apresentamos os resultados da aplicação dos melhores modelos em levantamentos que buscam exoplanetas destacando qual a melhor técnica de ML a ser utilizada neste contexto. Também incluímos os parâmetros estelares calculados com base nos valores previstos pelos modelos. Na Seção 4.5 apresentaremos algumas comparações dos erros obtidos neste trabalho com estudos realizados pelo nosso grupo e com resultados presentes na literatura. Por fim, a Seção 4.6 apresentaremos a determinação do raio de alguns candidatos a exoplanetas e a caracterização dos mesmos com base nos modelos deste trabalho.

4.1 Otimização de Hiperparâmetros

Nesta seção apresentaremos os melhores valores de hiperparâmetros otimizados no *GridSearch*. Os hiperparâmetros são aqueles que definimos no Capítulo 3 na Seção 3.2.4.

Para facilitar a apresentação dos resultados, nomeamos cada modelo com base na amostra utilizada para o treinamento. O nome começa com o levantamento principal

(`jplus` para J-PLUS e `splus` para S-PLUS), seguido por um sufixo que indica o levantamento auxiliar (`L` para LAMOST e `A` para APOGEE). Por exemplo, o modelo denominado "`jplusL`" refere-se ao modelo treinado com estrelas do J-PLUS que têm campos em comum com o levantamento LAMOST.

Para diferenciar se é mais ou menos restrito foi inserido um sufixo "`_01`" para restrito e "`_02`" para menos restrito no nome de cada modelo, por exemplo, "`jplusL_01`". Por fim, foi inserido um sufixo "`_RF`" para modelos treinados com a técnica de RF e "`_XGB`" para aqueles treinados com a técnica *XGBoost*.

Para cada um dos três parâmetros (T_{ef}, log g e [Fe/H]) foi utilizado o *GridSearch* na otimização. Foi utilizado nas amostras restritas e também para as menos restritas, conforme a subdivisão apresentada na Seção 3.1.2. A Figura 4.1 mostra como foi essa divisão.

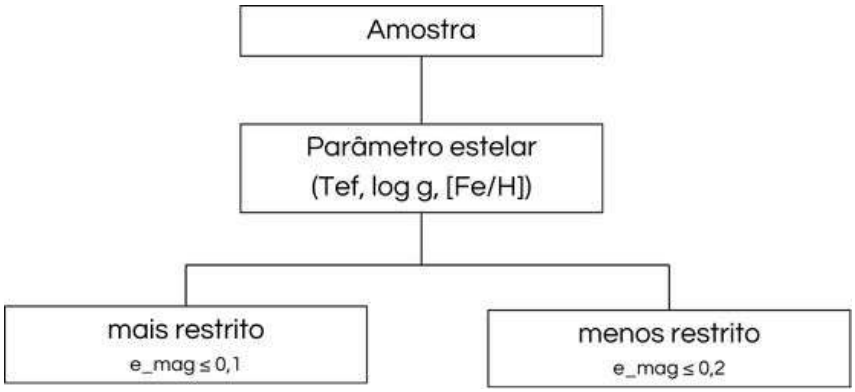

Figura 4.1: Divisão de modelos para cada técnica de ML e para cada parâmetro estelar em cada amostra apresentada na Seção 3.1.1.

A Tabela 4.1 apresenta todos os modelos com RF. Cada parâmetro estelar possui 8 modelos cada um com seus respectivos valores de hiperparâmetros. No total temos 24 modelos construídos.

T_{ef} (`_teff`)	**log g** (`_logg`)	**[Fe/H]** (`_feh`)
`jplusL_01_RF`	`jplusL_01_RF`	`jplusL_01_RF`
`jplusL_02_RF`	`jplusL_02_RF`	`jplusL_02_RF`
`splusL_01_RF`	`splusL_01_RF`	`splusL_01_RF`
`splusL_02_RF`	`splusL_02_RF`	`splusL_02_RF`
`jplusA_01_RF`	`jplusA_01_RF`	`jplusA_01_RF`
`jplusA_02_RF`	`jplusA_02_RF`	`jplusA_02_RF`
`splusA_01_RF`	`splusA_01_RF`	`splusA_01_RF`
`splusA_02_RF`	`splusA_02_RF`	`splusA_02_RF`

Tabela 4.1: Modelos com *Random Forest* para a previsão de parâmetros estelares. Cada parâmetro estelar possui 8 modelos. No total temos 24 modelos construídos.

A Tabela 4.2 apresenta todos os modelos construídos utilizando a técnica *XGBoost*. Para cada modelo, o *GridSearch* foi empregado para otimizar os valores dos hiperparâmetros de cada parâmetro estelar. Cada um dos parâmetros (T_{ef}, Log g e [Fe/H]) possui 8 modelos associados, totalizando 24 modelos desenvolvidos com XGB.

T_{ef} (`_teff`)	**log g** (`_logg`)	**[Fe/H]** (`_feh`)
`jplusL_01_XGB`	`jplusL_01_XGB`	`jplusL_01_XGB`
`jplusL_02_XGB`	`jplusL_02_XGB`	`jplusL_02_XGB`
`splusL_01_XGB`	`splusL_01_XGB`	`splusL_01_XGB`
`splusL_02_XGB`	`splusL_02_XGB`	`splusL_02_XGB`
`jplusA_01_XGB`	`jplusA_01_XGB`	`jplusA_01_XGB`
`jplusA_02_XGB`	`jplusA_02_XGB`	`jplusA_02_XGB`
`splusA_01_XGB`	`splusA_01_XGB`	`splusA_01_XGB`
`splusA_02_XGB`	`splusA_02_XGB`	`splusA_02_XGB`

Tabela 4.2: Modelos com *XGBoost* para a previsão de parâmetros estelares. Cada parâmetro estelar possui 8 modelos. No total temos 24 modelos construídos.

Apesar dos nomes parecidos, para cada parâmetro há um modelo configurado para prever somente esse parâmetro estelar. Para fins de notação e diferenciação, quando for mencionado um modelo específico para um determinado parâmetro, será adicionado um sufixo `_teff`, `_logg` ou `_feh` que representa a temperatura efetiva, a gravidade superficial e a metalicidade, respectivamente, no final dos nomes dos modelos, como por exemplo: `jplusA_RF_logg` é o modelo de previsão de log g, para a técnica de RF com dados do J-PLUS + APOGEE.

4.1.1 Hiperparâmetros Otimizados: *Random Forest*

Nesta seção, serão apresentados os hiperparâmetros utilizados no treinamento e teste dos modelos baseados na técnica de RF. Esses hiperparâmetros foram ajustados por meio de uma busca em grade (*GridSearch*), conforme detalhado no Capítulo 3, para otimizar as previsões dos parâmetros estelares T_{ef}, log g e [Fe/H].

Os modelos foram aplicados às amostras que foram apresentadas na Seção 3.1, resultando em ajustes específicos para amostras mais (`_01`) e menos restritas (`_02`).

Modelo	Hiperparâmetro	Restrito	Menos Restrito
jplusL_RF (_teff, _logg, _feh) splusL_RF (_teff, _logg, _feh)	max features	30	30
	msl	20	20
	n estimators	200	200
	bootstrap	True	True
	max depth	100	100
	mss	20	20
	random state	42	42
jplusA_RF_teff	max features	30	30
	msl	20	20
	n estimators	200	200
	bootstrap	True	True
	max depth	100	100
	mss	20	20
	random state	42	42
jplusA_RF_logg jplusA_RF_feh	max features	60	60
	msl	1	1
	n estimators	150	150
	bootstrap	False	False
	max depth	None	None
	mss	20	20
	random state	42	42
splusA_RF (_teff, _logg, _feh)	max features	50	50
	msl	5	5
	n estimators	5	5
	bootstrap	True	True
	max depth	5	5
	mss	20	20
	random state	42	42

Tabela 4.3: Valores de hiperparâmetros otimizados para os modelos restritos (_01) e menos restritos (_02) aplicados à previsão dos parâmetros estelares T_{ef}, log g e [Fe/H] com dados dos levantamentos J-PLUS DR3 e S-PLUS iDR5, com o LAMOST DR10 (com letra L) e APOGEE DR17 SDSS-IV (com letra A) aplicando a técnica de *Random Forest* (com sufixo _RF). Onde "msl" corresponde a "min samples leaf" e "mss" a "min samples split".

É possível notar que a Tabela 4.3 apresenta valores diferentes nos hiperparâmetros do modelo jplusA_RF_teff para a previsão de T_{ef} com relação aos hiperparâmetros de log g e [Fe/H] (modelos jplusA_RF_logg e jplusA_RF_feh, respectivamente), enquanto para

os outros modelos (todos aqueles utilizando o LAMOST e o modelo `splusA_RF`) são os mesmos valores de hiperparâmetros para todos os parâmetros estelares independente da amostra e do parâmetro estelar. Após alguns testes, não é possível explicar com exatidão o porque dessa diferença de valores, pois foi utilizado o mesmo *script*. O comportamento pode ser atribuído a diversos fatores, como por exemplo, a qualidade dos dados fornecidos pelo APOGEE ou filtros aplicados durante a otimização.

A análise desses hiperparâmetros é fundamental para compreender o desempenho de cada modelo e suas adaptações às diferentes características das amostras utilizadas. O impacto dessas configurações será discutido em detalhe nas próximas seções, juntamente com os resultados das previsões realizadas.

4.1.2 Hiperparâmetros Otimizados: *XGBoost*

Após a otimização realizada para a técnica de RF, foi aplicada a mesma metodologia para o XGB. Com o uso do *GridSearch*, detalhado no Capítulo 3, os hiperparâmetros foram ajustados de acordo com as características de cada amostra, visando a previsão dos parâmetros estelares T_{ef}, log g e [Fe/H].

A Tabela 4.4 apresenta os valores otimizados para a amostra restrita (`_01`) e para a amostra menos restrita (`_02`).

Modelo	Hiperparâmetro	Restrito	Menos Restrito
jplusL_XGB (_teff, _logg, _feh) splusL_XGB (_teff, _logg, _feh)	colsample by tree	0,7	0,7
	gamma	0,3	0,3
	learning rate	0,1	0,1
	max depth	10	10
	n estimators	100	100
	subsample	0,9	0,9
	random state	42	42
jplusA_XGB_teff	colsample by tree	1,0	0,6
	gamma	0,3	0,2
	learning rate	0,2	0,2
	max depth	30	10
	n estimators	150	150
	subsample	0,6	0,8
	random state	42	42
jplusA_XGB_logg	colsample by tree	1,0	0,7
	gamma	0,3	0,3
	learning rate	0,2	0,1
	max depth	30	10
	n estimators	150	100
	subsample	0,6	0,9
	random state	42	42
jplusA_XGB_feh	colsample by tree	1,0	1,0
	gamma	0,3	0,3
	learning rate	0,2	0,2
	max depth	30	30
	n estimators	150	150
	subsample	0,6	0,6
	random state	42	42

Tabela 4.4: Valores de hiperparâmetros otimizados para os modelos restritos (_01) e menos restritos (_02) aplicados à previsão dos parâmetros estelares T_{ef}, log g e [Fe/H] com dados dos levantamentos J-PLUS DR3 e S-PLUS iDR5, com o LAMOST DR10 (com letra L) e APOGEE DR17 SDSS-IV (com letra A) aplicando a técnica de *XGBoost* (com sufixo _XGB).

Modelo	Parâmetro	Restrito	menos Restrito
splusA_XGB_teff	colsample by tree	0,6	0,6
	gamma	0,4	0,4
	learning rate	0,2	0,2
	max depth	5	5
	n estimators	50	50
	subsample	0,95	0,95
	random state	42	42
splusA_XGB_logg	colsample by tree	1,0	0,6
	gamma	0,3	0,4
	learning rate	0,2	0,2
	max depth	30	5
	n estimators	150	50
	subsample	0,6	0,6
	random state	42	42
splusA_XGB_feh	colsample by tree	0,6	1.0
	gamma	0,2	0,3
	learning rate	0,2	0,2
	max depth	10	30
	n estimators	150	150
	subsample	0,8	0,6
	random state	42	42

Tabela 4.5: Valores de hiperparâmetros otimizados para os modelos restritos (_01) e menos restritos (_02) aplicados à previsão dos parâmetros estelares T_{ef}, log g e [Fe/H] com dados do levantamento S-PLUS iDR5, com o APOGEE DR17 SDSS-IV (com letra A) aplicando a técnica de *XGBoost* (com sufixo _XGB).

Podemos notar, novamente, que os valores mudam para os modelos que tem amostra do levantamento auxiliar APOGEE para cada parâmetro em específico. Alguns, para o mesmo parâmetro estelar, mudando entre si quando se usa um amostra mais ou menos restrita, como por exemplo os modelos jplusA_XGB para a previsão de T_{ef} e para a previsão de log g. Mesmo após vários testes, igualmente feitos para o RF, nos reforça a ideia de que pode ter relação com a qualidade das medidas fornecidas pelo APOGEE.

Nas seções a seguir, apresentaremos os desempenhos dos modelos e os gráficos mostrando o comportamento durante a previsão e teste dos melhores modelos, escolhidos conforme as especificações citadas anteriormente.

4.2 Desempenho dos Modelos Treinados com *Random Forest*

Nessa seção serão apresentadas as métricas alcançadas pelos modelos utilizando a técnica de *Random Forest* no treinamento, teste e construção, oferecendo uma visão detalhada sobre sua eficácia na previsão dos parâmetros estelares de interesse. Foram utilizados os hiperparâmetros otimizados pelo *GridSearch* apresentados na Seção 4.1.1. Além disso, foram utilizadas como *features* as magnitudes absolutas dos filtros calculadas conforme apresentado na Seção 3.1.2.

Nas próximas seções não serão apresentados todos os gráficos de todos os modelos gerados neste trabalho utilizando o RF, serão mostrados apenas aqueles considerados como melhores e que serão usados na determinação dos parâmetros estelares. Os outros gráficos não apresentados, podem ser consultados no Apêndice A.

4.2.1 *Random Forest* na Previsão de Temperatura Efetiva

Como mencionado anteriormente, não serão apresentadas aqui os gráficos que mostram o comportamento de cada um dos modelos treinados, serão apresentados somente as métricas alcançadas por eles durante o teste e os seus respectivos números de objetos de cada amostra, após realizados os filtros da Seção 3.1.2.

A Tabela 4.6 consiste nas métricas alcançadas pelos modelos com erro de magnitude $\leqslant$ 0,1 (restrito) e a Tabela 4.7, as métricas para os modelos com erro de magnitude $\leqslant$ 0,2 (menos restrito) e as respectivas quantidades de objetos para cada amostra ("Quant. Obj.").

Temperatura efetiva (`_teff`)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [K]**
`jplusL_01_RF`	255.039	0,9641	58,99
`splusL_01_RF`	72.036	0,9576	56,37
`jplusA_01_RF`	2.724	0,9722	37,06
`splusA_01_RF`	8.123	0,9560	48,65

Tabela 4.6: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de T_{ef} com *Random Forest*.

Temperatura efetiva (`_teff`)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [K]**
`jplusL_02_RF`	278.879	0,9714	54,98
`splusL_02_RF`	76.035	0,9594	56,69
`jplusA_02_RF`	4.782	0,9767	38,16
`splusA_02_RF`	9.125	0,9654	53,94

Tabela 4.7: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de T_{ef} com *Random Forest*.

Com relação às amostras treinadas com o levantamento J-PLUS DR3, o melhor modelo restrito para a previsão de T_{ef}, considerando apenas as métricas, é o `jplusA_01_RF_teff`, treinado e testado com estrelas em campos comuns do J-PLUS DR3 e APOGEE DR17 SDSS-IV. Esse modelo alcançou um R^2 *Score* de 0,9722 e um desvio mediano absoluto de 37,06 K. O melhor modelo menos restrito, considerando somente as métricas, consiste no modelo `jplusA_02_RF_teff`, onde também foi treinado com a amostra J-PLUS DR3 + APOGEE DR17 SDSS-IV, alcançando um R^2 *Score* de 0,9767 e um MAD de 38,16 K.

Como mencionado anteriormente, as métricas por si só não devem ser o único critério de avaliação, pois a quantidade de objetos usados no treinamento é o mais importante. Em uma análise estatística e no treinamento de modelos de aprendizado de máquina, quanto maior o número de objetos na amostra de entrada para o treinamento, mais confiável tende ser o modelo treinado.

No caso do modelo restrito para a previsão de T_{ef} usando o J-PLUS DR3 como levantamento principal, o `jplusL_01_RF_teff` passa ser o melhor, pois possui 255.039 objetos em comum com o LAMOST DR10, enquanto o modelo `jplusA_01_RF_teff` contém apenas 2.724 objetos, prejudicando assim, a confiabilidade do modelo. O modelo `jplusL_01_RF_teff` alcançou um R^2 *Score* de 0,9641 e um MAD de 58,99 K.

Para o modelo menos restrito, o melhor modelo, considerando também as condições citadas acima, no qual, a quantidade de objetos como critério principal, que apresentou melhor desempenho foi o modelo `jplusL_02_RF_teff`, treinado com 278.879 objetos, também, em comum com o LAMOST DR10, alcançando um R^2 *Score* de 0,9714 e um MAD de 54,98 K.

A Figura 4.2 ilustra a correlação entre as temperaturas previstas pelo modelo e as temperaturas fornecidas pelo LAMOST no cenário restrito (`jplusL_01_RF_teff`).

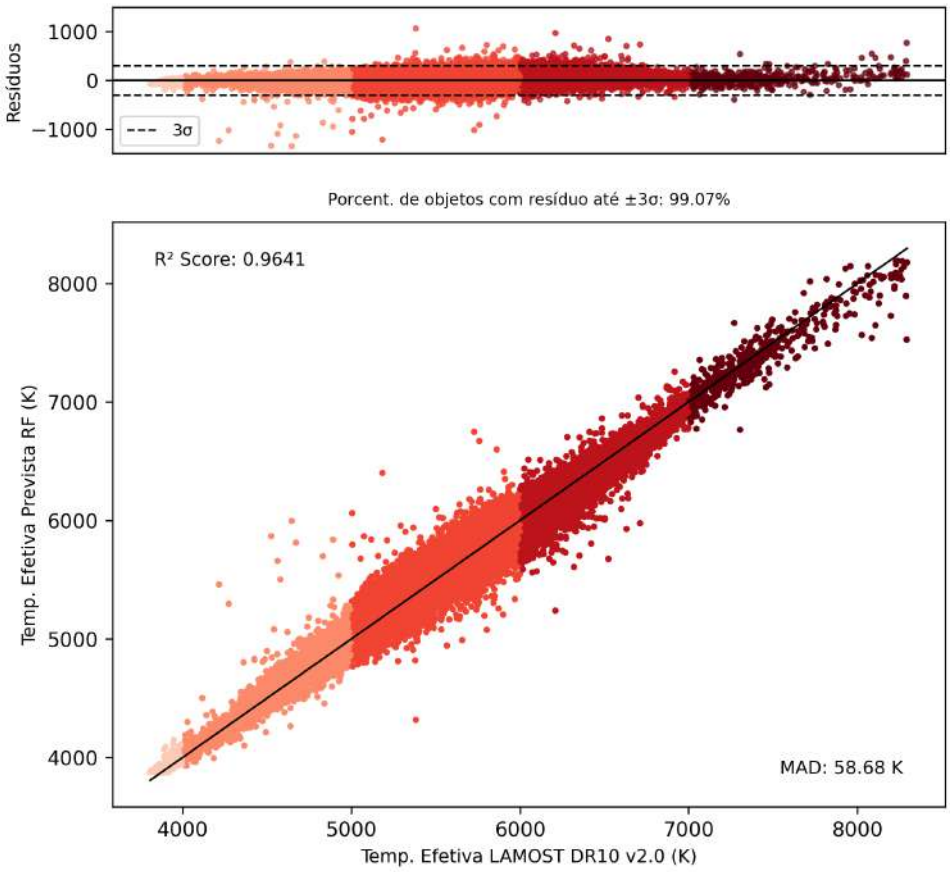

Figura 4.2: Resultados do modelo `jplusL_01_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10.

Já a Figura 4.3 ilustra a correlação entre as temperaturas previstas durante o teste pelo modelo `jplusL_02_RF_teff` e as temperaturas fornecidas pelo LAMOST, destacando o comportamento do melhor modelo menos restrito para a previsão de T_{ef}.

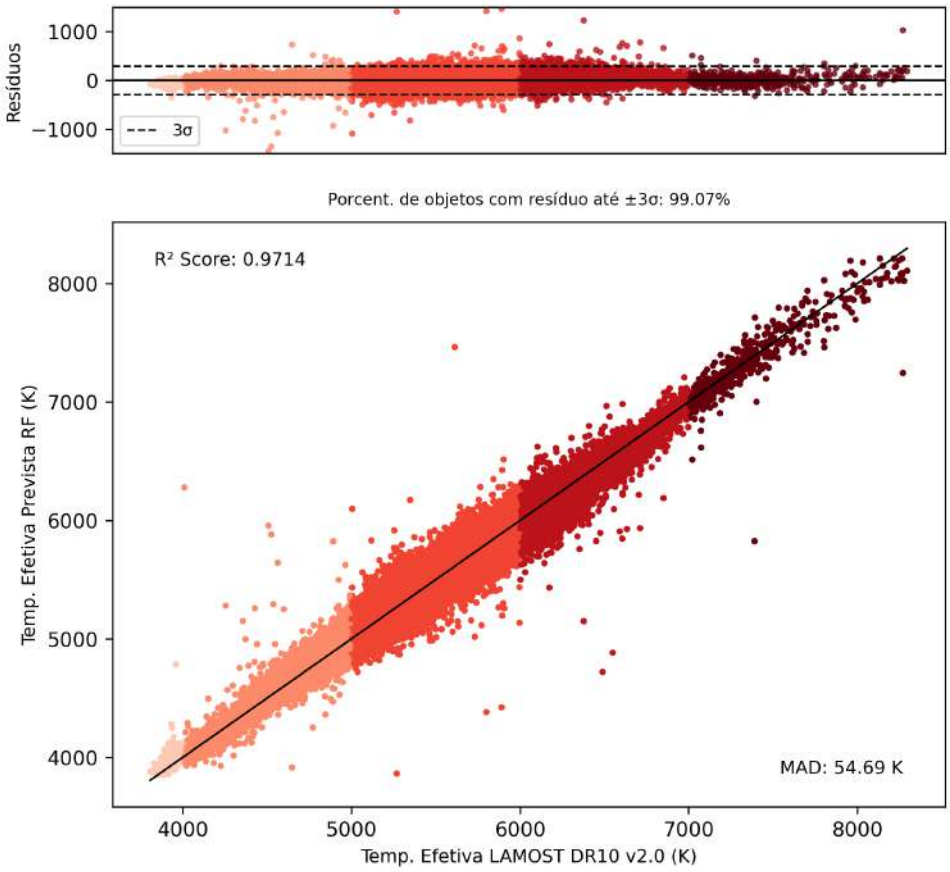

Figura 4.3: Resultados do modelo `jplusL_02_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10.

Para efeitos de visualização, as cores nas Figuras 4.2 e 4.3 (e nas próximas daqui pra frente) representam intervalos que são apenas ilustrativos. Não foram realizados cálculos, otimizações ou testes em cada intervalo específico nas cores mostradas nos gráficos. Para a temperatura foi definido um intervalo de 1000 K, variando de 3000 K até 7000 K e no final, de 7000 K até 8300 K, cobrindo toda faixa de dados das amostras de treinamento inseridas. O painel superior apresenta os resíduos das predições, definidos como a diferença entre os valores observados e os valores estimados pelo modelo, calculados pela Equação 3.3. As linhas tracejadas nesses painéis representam os limites de 3σ, utilizados para identificar possíveis *outliers* nas nossas previsões. A reta em preto nos painéis principais representa a correlação, quanto mais próximo dessa reta melhor é a correlação entre o parâmetro previsto e o parâmetro fornecido pelo levantamento auxiliar.

Considerando o levantamento principal S-PLUS, os melhores modelos, tanto o restrito quanto o menos restrito, são os modelos `splusL_01_RF_teff` e `splusL_02_RF_teff`, respectivamente. Conforme mostrado na Tabela 4.6, o modelo `splusL_01_RF_teff` obteve um R^2 *Score* de 0,9576 e um MAD de 56,37 K, utilizando uma amostra de 72.036 objetos em comum entre o S-PLUS e o LAMOST. Já a Tabela 4.7 apresenta o modelo `splusL_02_RF_teff`, que alcançou um R^2 *Score* de 0,9594 e um MAD de 56,69 K, para uma

amostra S-PLUS + LAMOST de 76.035 objetos. Ambos são considerados os melhores e os mais confiáveis modelos, levando em consideração, principalmente, a quantidade de objetos.

Replicando que foi feito com os dados do J-PLUS, gerando gráficos para visualizar o comportamento dos modelos. A Figura 4.4 mostra a correlação entre os valores previstos durante o teste pelo modelo `splusL_02_RF_teff` e os valores reais fornecidos pelo LAMOST DR10.

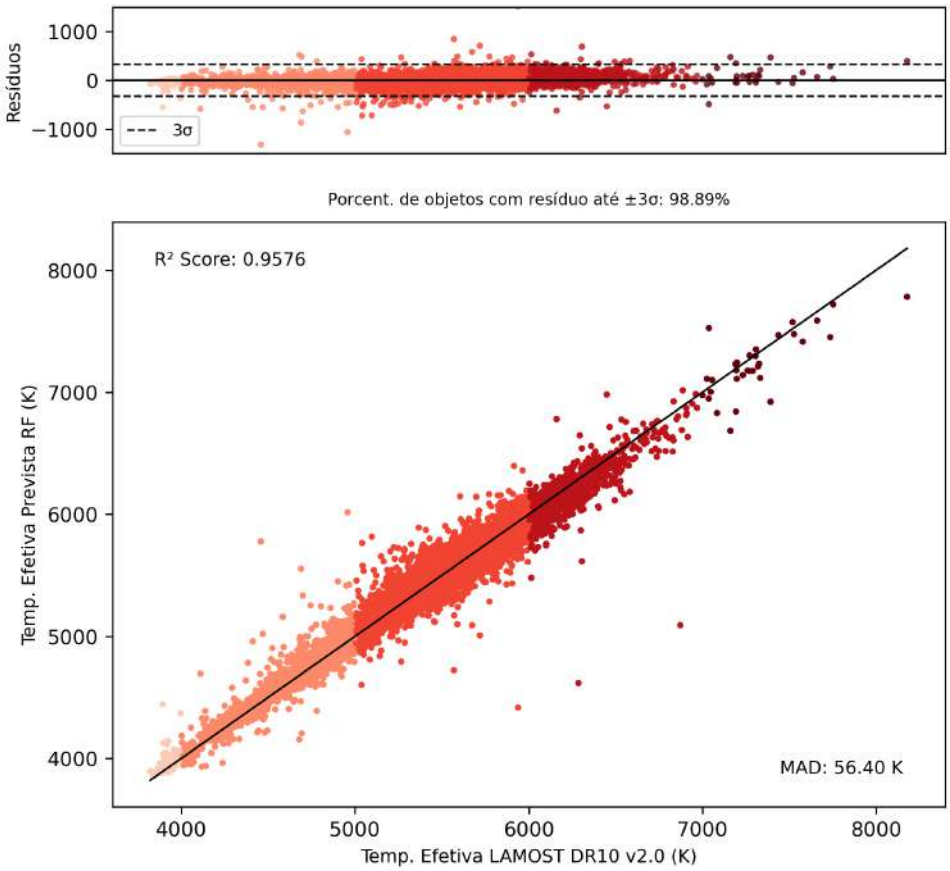

Figura 4.4: Resultados do modelo `splusL_01_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10.

Por outro lado, a Figura 4.5 ilustra a correlação entre as temperaturas previstas pelo modelo e as fornecidas pelo LAMOST, destacando o comportamento do melhor modelo menos restrito (`splusL_02_RF_teff`) para a previsão de T_{ef} com dados do S-PLUS iDR5 em comum com o levantamento LAMOST DR10.

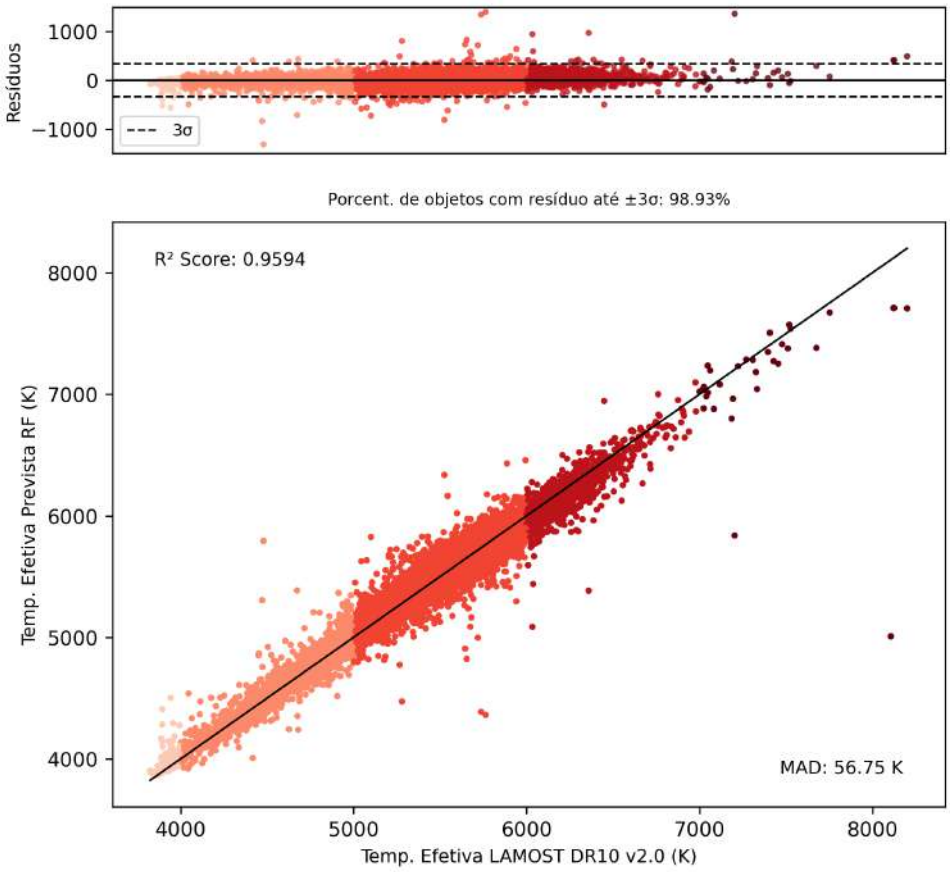

Figura 4.5: Resultados do modelo `splusL_02_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10.

Portanto, para a previsão da temperatura efetiva utilizando o RF, esses foram os melhores modelos utilizando o J-PLUS e o S-PLUS como levantamentos principais. Na Seção 4.4 apresentaremos os resultados da aplicação desses modelos e a avaliação da melhor técnica para a previsão deste parâmetro estelar.

4.2.2 *Random Forest* na Previsão da Gravidade Superficial

O mesmo procedimento feito para temperatura efetiva, foi feito para a gravidade superficial (log g). Foi aplicada a técnica de RF na previsão deste parâmetro e as métricas alcançadas pelos modelos mais e menos restritos estão apresentadas na Tabela 4.8 e 4.9, respectivamente.

Gravidade Superficial (_logg)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_01_RF	255.039	0,8793	0,07
splusL_01_RF	72.036	0,8666	0,07
jplusA_01_RF	2.724	0,9642	0,02
splusA_01_RF	8.123	0,9740	0,04

Tabela 4.8: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de log g com *Random Forest*.

Gravidade Superficial (_logg)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_02_RF	278.879	0,8809	0,07
splusL_02_RF	76.035	0,8814	0,07
jplusA_02_RF	4.782	0,9740	0,02
splusA_02_RF	9.125	0,9719	0,04

Tabela 4.9: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de log g com *Random Forest*.

Para o log g, aconteceu a mesma coisa que aconteceu para a T_{ef}. Os modelos utilizando o levantamento secundário APOGEE DR17 SDSS-IV (jplusA_01_RF_logg e jplusA_02_RF_logg) obtiveram as melhores métricas (vide as Tabelas 4.8 e 4.9). Mas analisando a coluna de quantidade de objetos, nota-se que esses modelos possuem um número reduzido de objetos na amostra de entrada para o treinamento se comparados aos modelos que usaram o LAMOST DR10, com isso, tornando-os não tão confiáveis.

Portanto, levando em consideração o número de objetos na amostra de entrada para o treinamento do modelo utilizando os objetos observados pelo J-PLUS DR10, o melhor modelo para o cenário restrito consiste no modelo jplusL_01_RF_logg, onde alcançou um R^2 *Score* de 0,8793 com um MAD de 0,07 dex para uma amostra de treinamento de 255.039 objetos comuns entre o J-PLUS DR3 e LAMOST DR10. Agora para o caso menos restrito, o melhor modelo também é aquele que foi treinado com a amostra J-PLUS + LAMOST DR10, o modelo jplusL_02_RF_logg, com 278.879 objetos de entrada, alcançando uma métrica de R^2 *Score* de 0,8809 e um MAD de 0,07 dex.

A Figura 4.6 mostra o comportamento do modelo jplusL_01_RF_logg, que destaca a correlação entre os valores de log g previstos pelo modelo e os fornecidos pelo levantamento LAMOST. Os intervalos de cores também é para efeitos de visualização, não possuem cálculos os modelos específicos para eles. Os intervalos correspondem a 1,0 dex, variando de 1,0 dex até 5,0 dex.

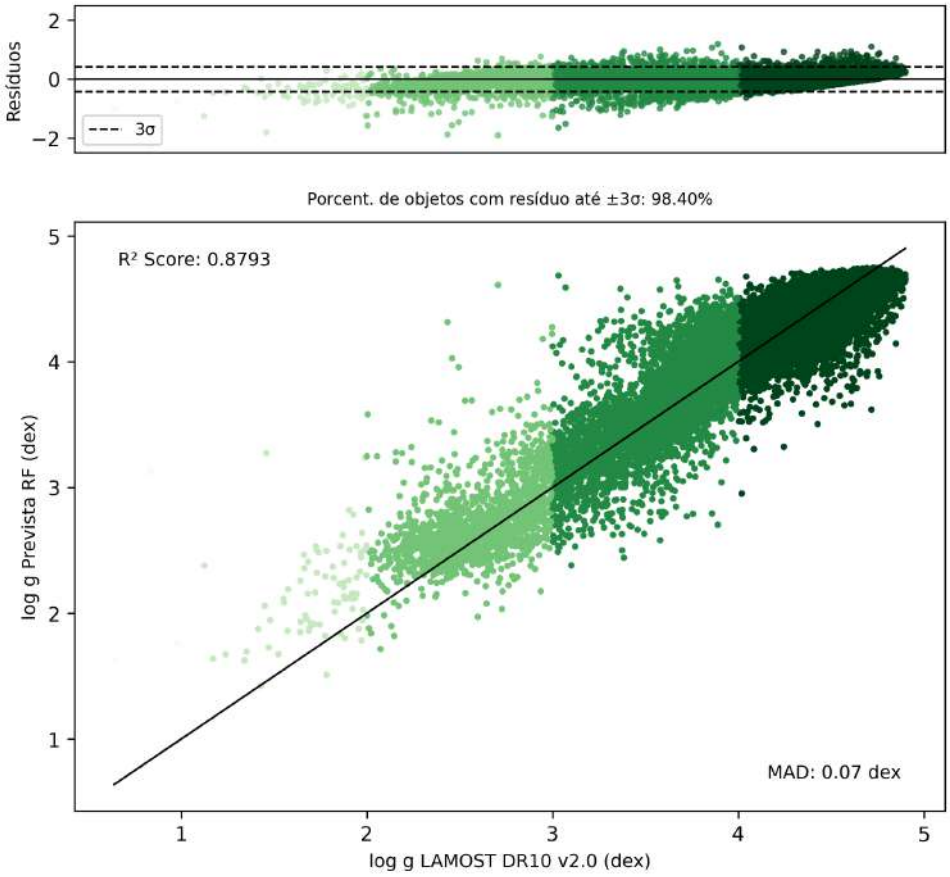

Figura 4.6: Resultados do modelo `jplusL_01_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

Por outro lado, a correlação entre os valores de gravidade superficial previstos pelo modelo `jplusL_02_RF_logg` é apresentada na Figura 4.7.

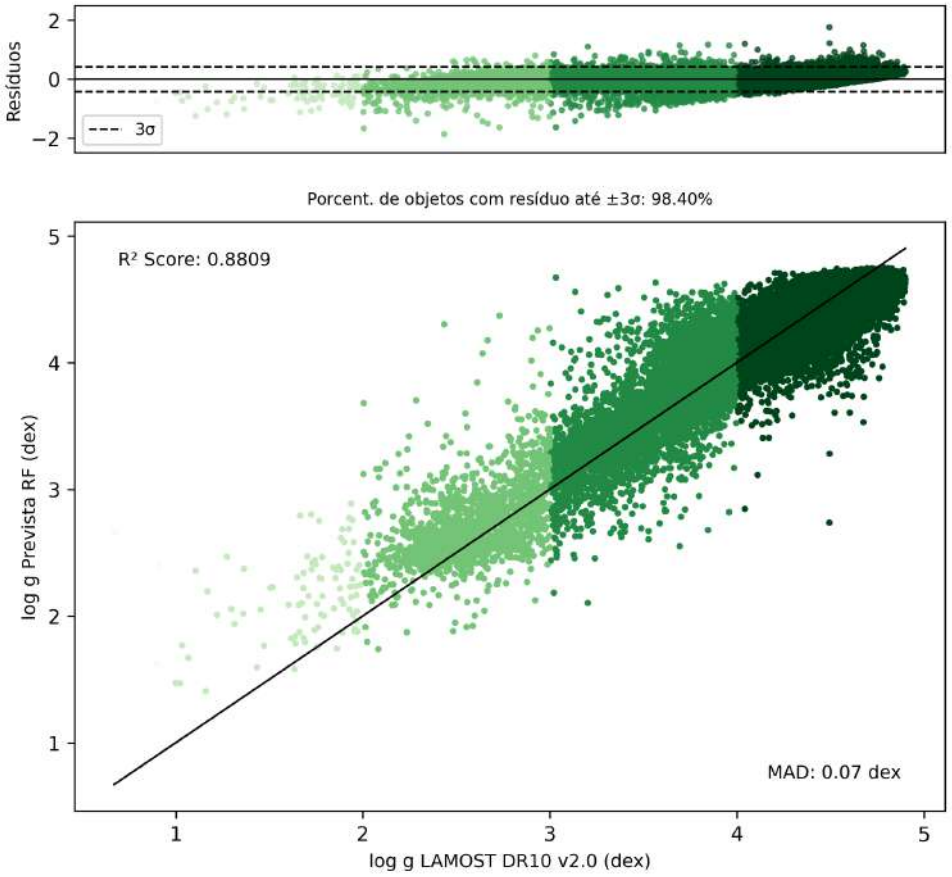

Figura 4.7: Resultados do modelo `jplusL_02_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

Fazendo o mesmo com os dados do S-PLUS, tem-se que, levando em consideração o número de objetos, o melhor e mais confiável modelo restrito foi o `splusL_01_RF_logg`, que atingiu um R^2 *Score* de 0,8666 e um MAD de 0,07 dex, com uma amostra de 72.036 objetos em comum entre os levantamentos S-PLUS e LAMOST. Já para o modelo menos restrito, o `splusL_02_RF_logg`, treinado com uma amostra maior de 76.035 objetos, obteve um R^2 *Score* de 0,8814 e um MAD de 0,07 dex.

O comportamento do modelo `splusL_01_RF_logg` pode ser visualizado na Figura 4.8, que ilustra a correlação entre os valores de log g previstos pelo modelo e os valores medidos pelo LAMOST para as estrelas em comum com o S-PLUS.

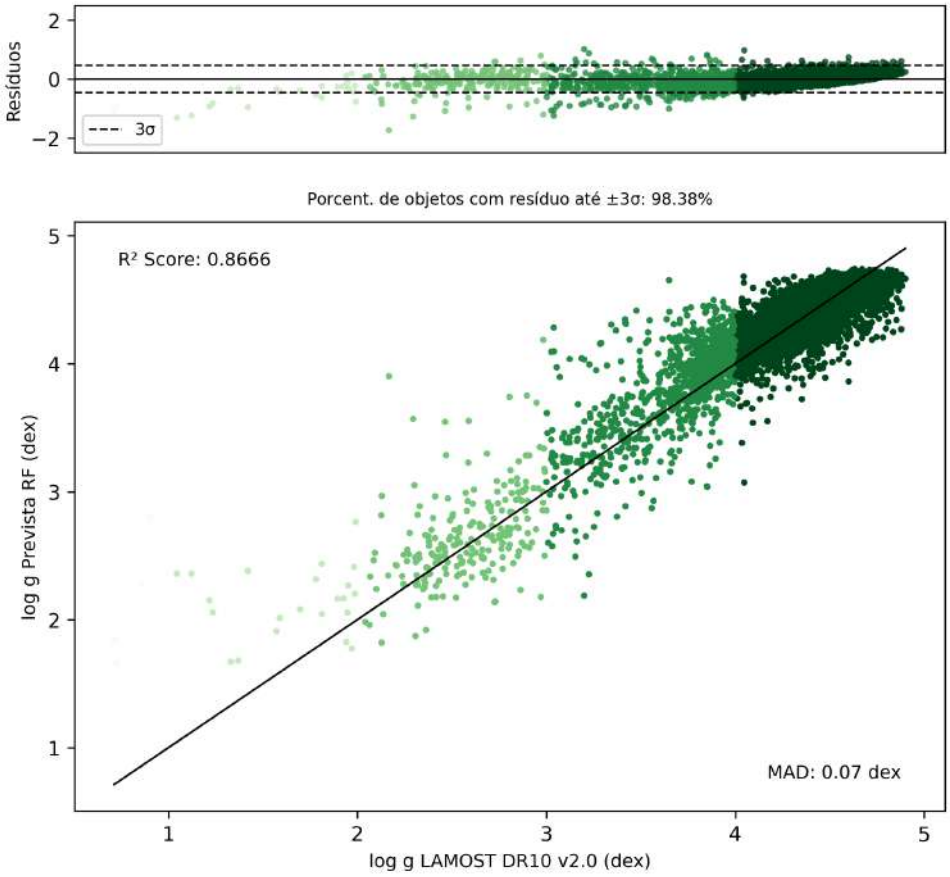

Figura 4.8: Resultados do modelo `splusL_01_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

Enquanto a Figura 4.9 apresenta a correlação obtida pelo modelo `splusL_02_RF_logg`, evidenciando a correspondência entre os valores previstos de log g e os valores medidos pelo LAMOST, com base na amostra de objetos observados simultaneamente pelo S-PLUS.

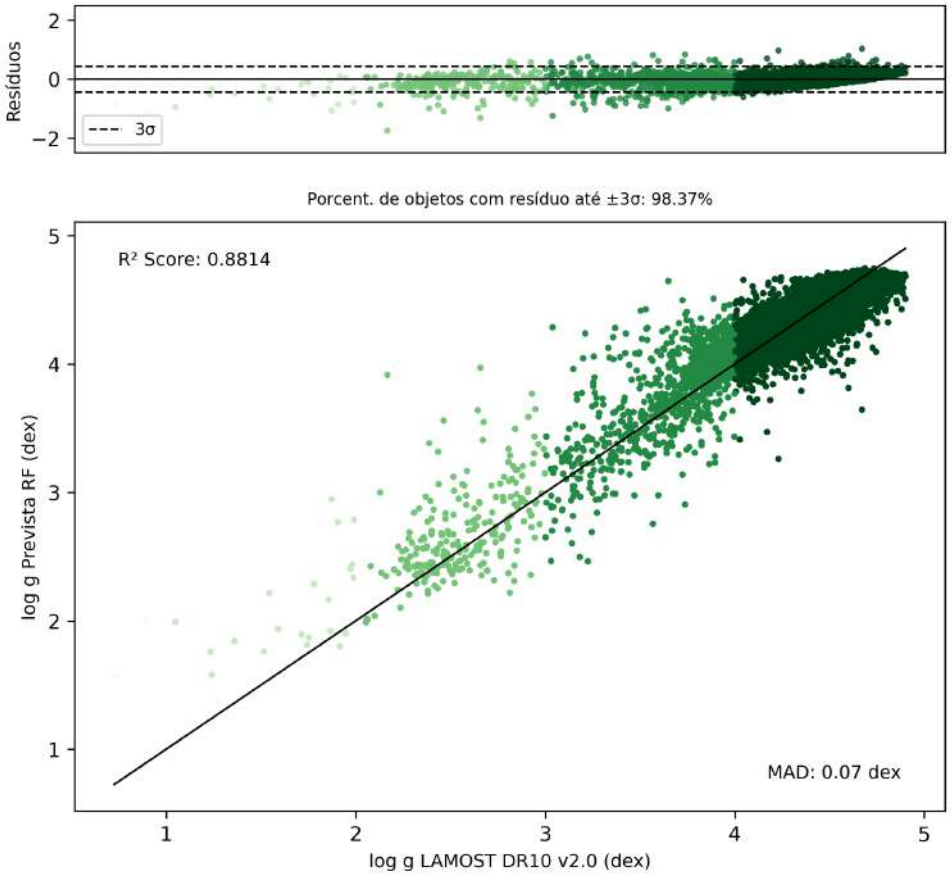

Figura 4.9: Resultados do modelo `splusL_02_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

Os resultados e discussões sobre a aplicação desses modelos nos objetos presentes nos levantamentos que buscam exoplanetas serão apresentados na Seção 4.4.

4.2.3 *Random Forest* na Previsão da Metalicidade

Para a determinação da metalicidade, também foi aplicada a técnica de RF para a previsão e teste. Nesta seção, não serão apresentados todos os gráficos gerados, mas apenas aqueles que foram considerados como os melhores modelos. As métricas obtidas pelos modelos mais e menos restritos estão detalhadas nas Tabelas 4.10 e 4.11, respectivamente.

Metalicidade (_feh)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_01_RF	255.039	0,8525	0,08
splusL_01_RF	72.036	0,8411	0,08
jplusA_01_RF	2.724	0,7887	0,07
splusA_01_RF	8.123	0,7730	0,07

Tabela 4.10: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de de [Fe/H] com *Random Forest*.

Metalicidade (_feh)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_02_RF	278.879	0,8463	0,08
splusL_02_RF	76.035	0,8420	0,08
jplusA_02_RF	4.782	0,8385	0,07
splusA_02_RF	9.125	0,7684	0,07

Tabela 4.11: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de [Fe/H] com *Random Forest*.

Diferentemente do que aconteceu com os outros parâmetros, para o parâmetro de metalicidade, durante o treinamento e testes, os modelos que apresentaram melhores desempenhos (melhores métricas) também foram aqueles que também tiveram maiores quantidades de objetos. Foram os modelos `jplusL_01_RF_feh` para o cenário mais restritivo e o `jplusL_02_RF_feh` para o caso menos restrito. Ambos são modelos treinados com objetos em comum com o levantamento LAMOST.

O modelo `jplusL_01_RF_feh` obteve um R^2 *Score* de 0,8535 e um MAD de 0,08 dex para uma amostra com 255.039 objetos na amostra de treinamento. A Figura 4.10 mostra a correlação entre os valores de [Fe/H] previstos pelo modelo e os valores fornecidos pelo levantamento LAMOST. Para a metalicidade são intervalos que correspondem a 1,0 dex, variando de -2,5 dex até 0,5 dex e no final com um intervalo de 0,5 dex, iniciando em 0,5 até 1,0 dex.

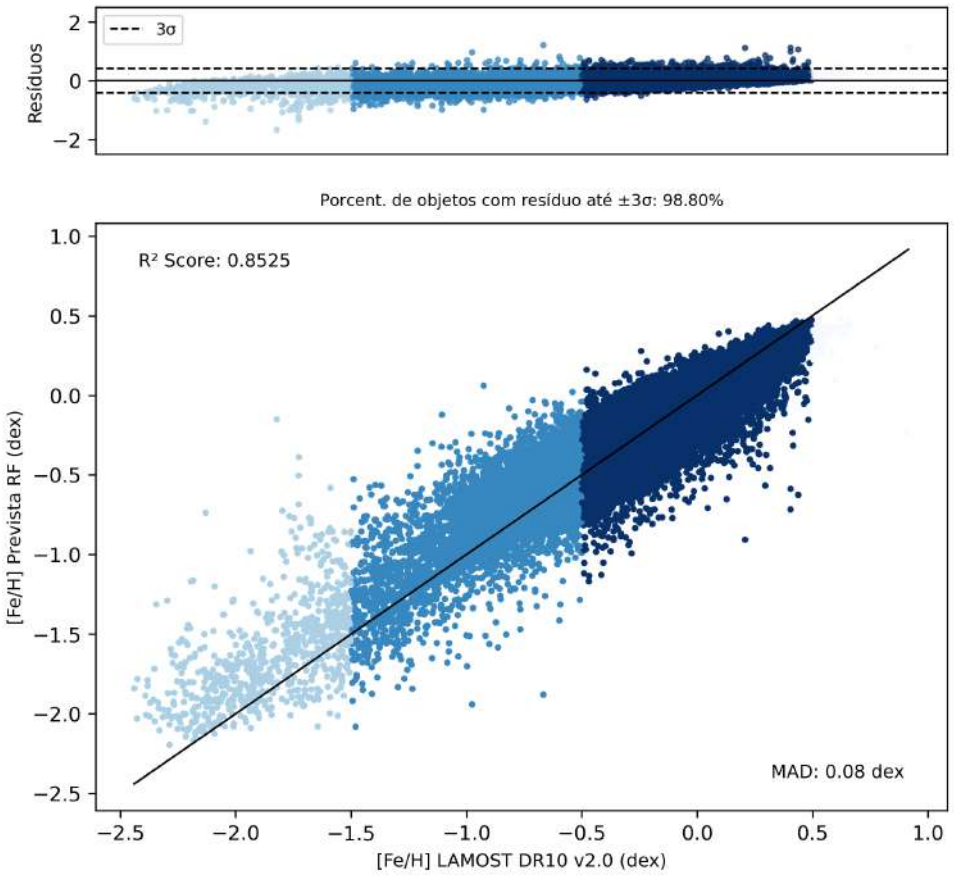

Figura 4.10: Resultados do modelo `jplusL_01_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

O modelo `jplusL_02_RF_feh` apresentou um R^2 *Score* de 0,8463 durante o teste e um MAD de 0,08 dex e uma amostra de 278.879 objetos, trazendo confiança ao modelo. A Figura 4.11 mostra a correlação entre os valores de [Fe/H] previstos pelo modelo e os valores calculados para uma amostra J-PLUS + LAMOST.

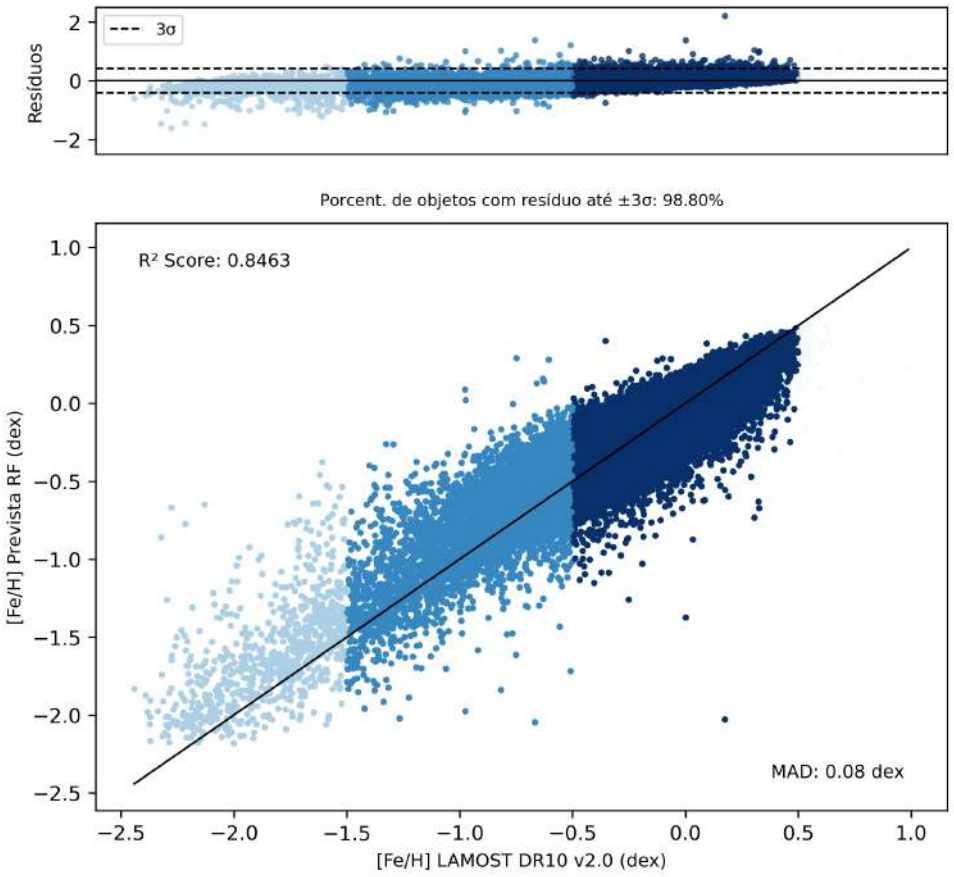

Figura 4.11: Resultados do modelo `jplusL_02_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

Para o levantamento S-PLUS, os modelos de RF que obtiveram os melhores resultados na previsão da metalicidade foram `splusL_02_RF_feh` para o cenário menos restrito e `splusL_01_RF_feh` para o restrito. Esses modelos foram treinados utilizando objetos observados tanto pelo S-PLUS iDR5 quanto pelo LAMOST DR10, garantindo uma base de comparação entre os valores reais e os previstos.

O modelo `splusL_01_RF_feh` apresentou um R^2 *Score* de 0,8411 e um MAD de 0,08 dex, utilizando uma amostra de 72.036 objetos em comum entre S-PLUS e LAMOST. Já o modelo `splusL_02_RF_feh`, que corresponde ao cenário menos restrito, atingiu um R^2 *Score* de 0,8420, com um desvio mediano absoluto de 0,08 dex para uma amostra de 76.035 objetos. Ambos os modelos demonstraram um desempenho consistente para a previsão de metalicidade.

A Figura 4.12 exibe a correlação entre os valores de [Fe/H] previstos pelo modelo restrito e as medições realizadas pelo LAMOST para uma amostra de estrelas em comum entre os levantamentos S-PLUS e LAMOST. Esse gráfico ilustra a precisão do modelo na previsão da metalicidade, destacando o bom desempenho para a amostra mais limitada.

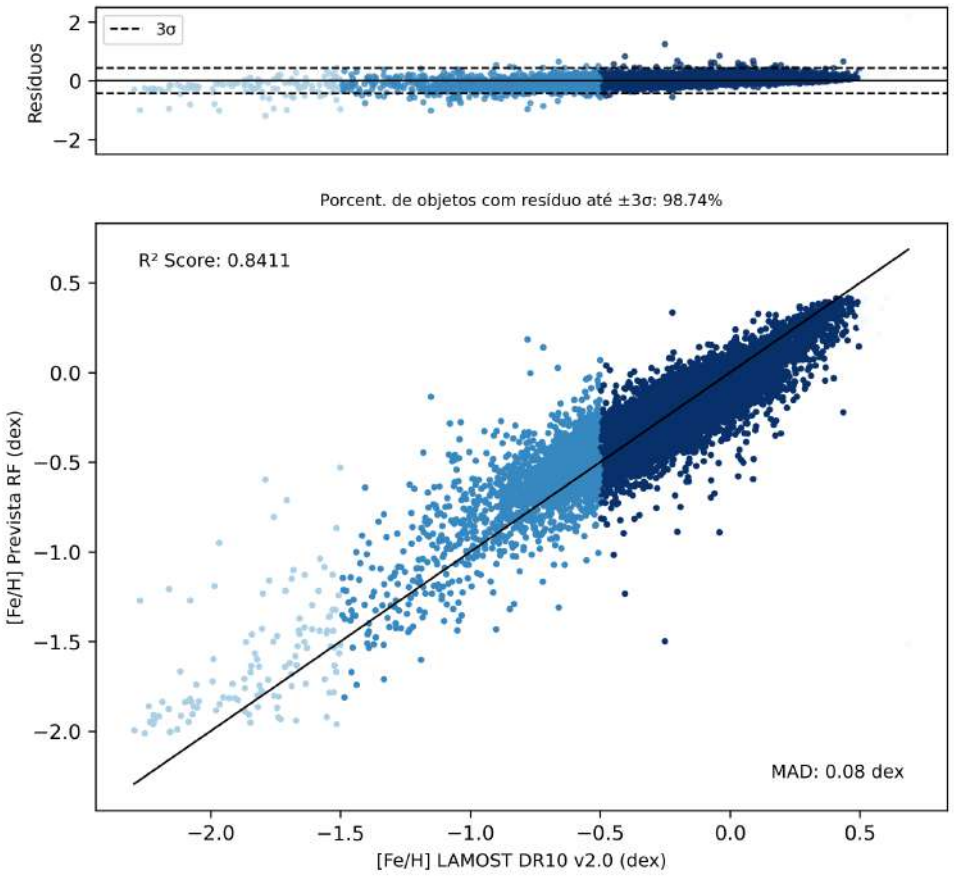

Figura 4.12: Resultados do modelo `splusL_01_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

Já a Figura 4.13 apresenta a correlação obtida para o modelo menos restrito, comparando as previsões de [Fe/H] com os valores medidos pelo LAMOST DR10. Essa Figura demonstra como o modelo se comporta ao lidar com uma amostra um pouco maior, ressaltando o comportamento do modelo nas previsões de [Fe/H].

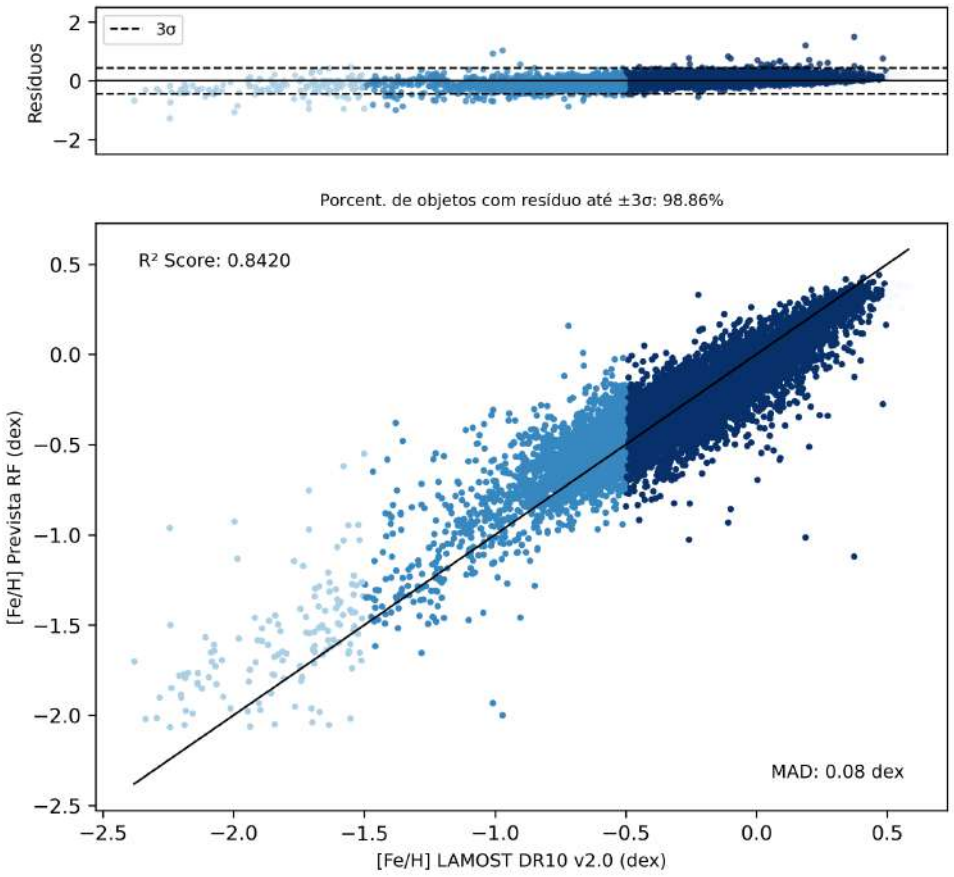

Figura 4.13: Resultados do modelo `splusL_02_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Fores* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

Com os modelos já construídos e testados, a próxima etapa envolveu a aplicação desses modelos na amostra de objetos observados nos levantamentos voltados para a busca de exoplanetas. Na Seção 4.4 apresentaremos os resultados e discussões dessa aplicação. Foram avaliadas algumas métricas que permitem uma compreensão mais aprofundada da precisão dos modelos na determinação dos parâmetros estelares utilizando o *Random Forest* como técnica de aprendizado de máquina.

4.3 Desempenho dos Modelos Treinados com *XGBoost*

Nesta seção, são discutidos os resultados obtidos com a utilização do algoritmo XGB para o treinamento e teste dos modelos, demonstrando sua eficácia na previsão dos parâmetros estelares de interesse. Assim como no caso do RF, a técnica foi aplicada para cada parâmetro estelar (T_{ef}, log g e [Fe/H]) nas mesmas subdivisões das amostras, com um modelo restrito e outro menos restrito.

Os resultados a seguir focam nos melhores desempenhos, com base nas métricas de interesse, como o R^2 *score*, desvio mediano absoluto e a quantidade de objetos. Assim como no RF, os valores dos hiperparâmetros utilizados para otimização foram obtidos através da busca em grade (*GridSearch*) e foram apresentados na Seção 4.1.2. As correlações entre os parâmetros previstos e os valores reais fornecidos pelos levantamentos auxiliares estão apresentadas nas seções a seguir.

Não serão apresentados todos os gráficos de todos os modelos gerados neste trabalho utilizando o XGB, mas apenas daqueles considerados como melhores e que foram usados na determinação dos parâmetros estelares. Os outros gráficos não apresentados, podem ser consultados no Apêndice B.

Os modelos foram nomeados de forma similar ao RF, sendo baseados na combinação de amostras principais e auxiliares (J-PLUS, S-PLUS, LAMOST e APOGEE) como explicado anteriormente.

4.3.1 *XGBoost* na Previsão da Temperatura Efetiva

Assim como foi realizado para o RF, nesta seção serão apresentados os resultados obtidos com a aplicação do XGB na previsão da temperatura efetiva (T_{ef}). O foco estará nas métricas alcançadas pelos modelos durante o teste na previsão dos parâmetros, assim como o número de objetos em cada amostra, após a aplicação dos filtros descritos na Seção 3.1.2.

A Tabela 4.12 apresenta as métricas para os modelos com erro de magnitude $\leqslant 0{,}1$ (restrito), enquanto a Tabela 4.13 exibe os resultados para os modelos com erro de magnitude $\leqslant 0{,}2$ (menos restrito), juntamente com a quantidade de objetos usados em cada amostra.

Temperatura Efetiva (`_tef`)			
Modelo	**Quant. Obj.**	***R²* Score**	**MAD [K]**
`jplusL_01_XGB`	255.039	0,9650	57,88
`splusL_01_XGB`	72.036	0,9610	55,05
`jplusA_01_XGB`	2.724	0,9678	38,49
`splusA_01_XGB`	8.123	0,9710	47,15

Tabela 4.12: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de T_{ef} com *XGBoost*.

Temperatura Efetiva (`_tef`)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [K]**
`jplusL_02_XGB`	278.879	0,9733	53,38
`splusL_02_XGB`	76.035	0,9625	56,18
`jplusA_02_XGB`	4.782	0,9712	40,19
`splusA_02_XGB`	9.125	0,9594	46,96

Tabela 4.13: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de T_{ef} com *XGBoost*.

Levando em consideração o levantamento principal J-PLUS, temos que o modelo que apresentou o melhor desempenho utilizando uma amostra restrita foi o modelo `jplusA_01_XGB_teff`. Este modelo alcançou um R^2 *Score* de 0,9678 com um MAD de 38,49.Analisando bem, é um ótimo desempenho, mas ao analisar sua confiança percebe-se que possui apenas 2.724 na amostra de treinamento, comparado ao modelo `jplusL_01_XGB_teff`, treinado com uma amostra de 255.039 objetos. Portanto, não podemos defini-lo como o melhor modelo para aplicação nos levantamentos que buscam exoplanetas.

Sendo assim, o modelo mais confiável é o modelo `jplusL_01_XGB_teff`, o qual apresentou um R^2 *Score* de 0,9650 e um desvio mediano absoluto de 57,88 K. Esse resultado indica boa correlação entre os valores previstos e os valores reais fornecidos pelo LAMOST, além de uma precisão elevada na previsão da temperatura efetiva para as estrelas dessa amostra.

Já para o modelo menos restrito, o melhor desempenho foi obtido pelo `jplusL_02_XGB_teff`, com um R^2 *Score* de 0,9733 e um MAD de 53,21 K, ainda mantendo uma precisão um pouco superior ao modelo restrito. Esse modelo trabalhou com uma quantidade maior de objetos, 278.879, indicando que a flexibilização dos filtros permitiu uma amostra mais ampla, obtendo uma correlação e precisão melhores na previsão da temperatura efetiva.

A Figura 4.14 ilustra a correlação entre os valores reais e previstos para a T_{ef} utilizando o modelo restrito treinado com a técnica de XGB. Esse gráfico destaca a capacidade do modelo `jplusL_01_XGB_teff` em prever com alta precisão os valores de T_{ef}, evidenciando a eficácia do modelo em capturar a relação entre as variáveis.

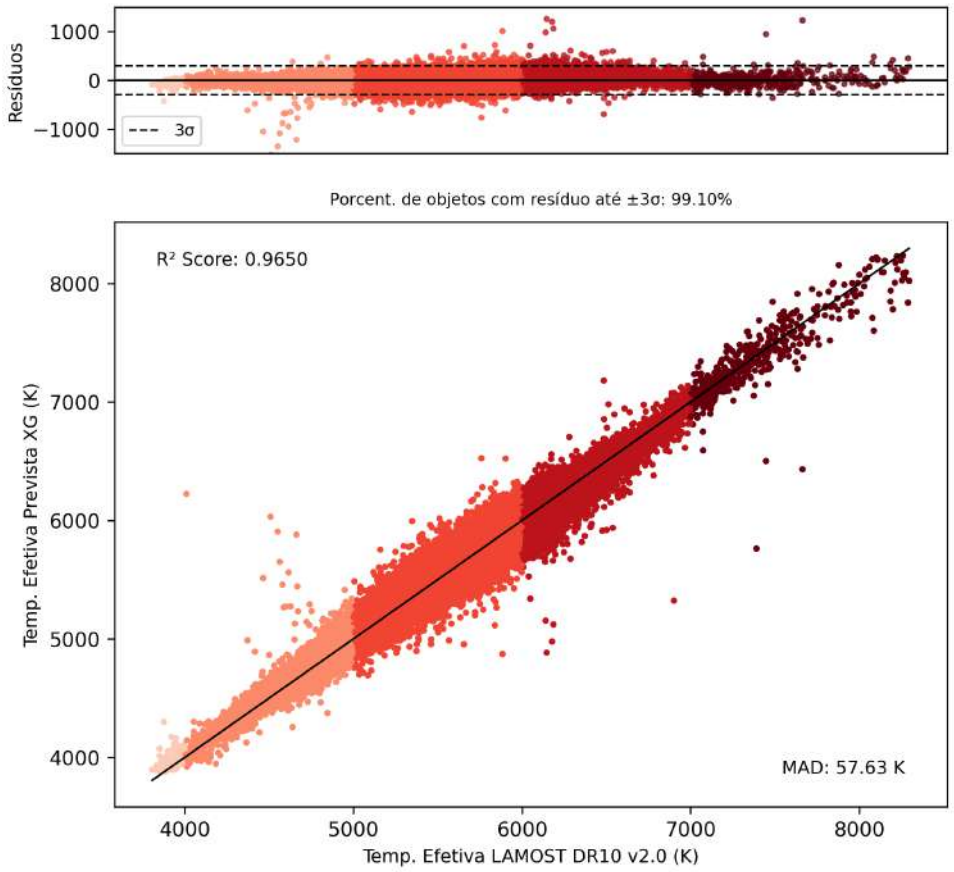

Figura 4.14: Resultados do modelo `jplusL_01_RF_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10.

Por outro lado, a Figura 4.15 apresenta a correlação para o modelo menos restrito de XGB, mostrando como o desempenho é afetado quando o conjunto de dados é menos filtrado. Embora a precisão geral se mantenha elevada, o modelo lida com uma maior variabilidade nos dados, refletindo um ligeiro aumento nos erros da previsão.

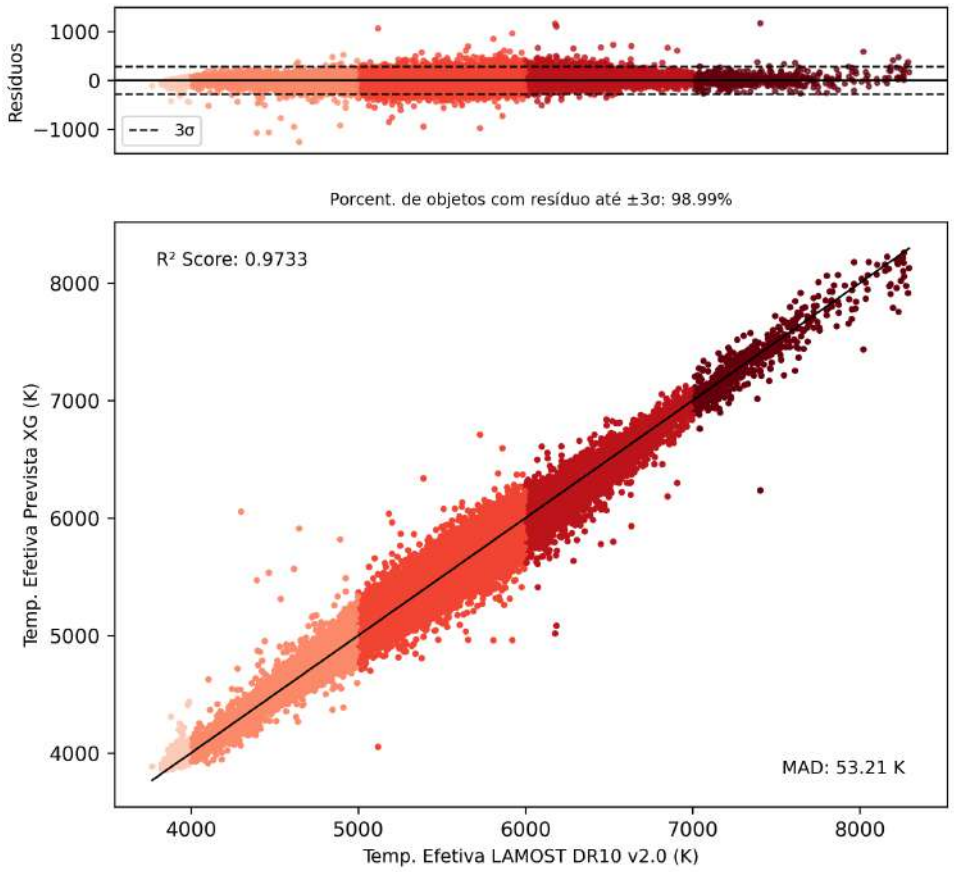

Figura 4.15: Resultados do modelo `jplusL_02_RF_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10.

No levantamento de dados S-PLUS, os modelos para a previsão de temperatura que obtiveram um bom desempenho foram os modelos `splusL_01_XGB_teff` para o cenário restrito e o `splusL_02_XGB_teff` para o cenário menos restrito. O modelo `splusL_01_XGB_teff` alcançou um R^2 *Score* de 0,9610 com um MAD de 55,05 K para uma amostra de 72.036 estrelas em comum entre S-PLUS e o LAMOST. Enquanto o modelo `splusL_02_XGB_teff` alcançou um R^2 *Score* de 0,9625 com um MAD 55,18 K para a amostra S-PLUS + LAMOST de 76.035 estrelas.

As Figura 4.16 e a Figura 4.17, a seguir, mostram a correlação entre as temperaturas previstas pelos modelos durante o teste e as temperaturas fornecidas pelo LAMOST DR10, permitindo uma visualização clara da precisão obtida.

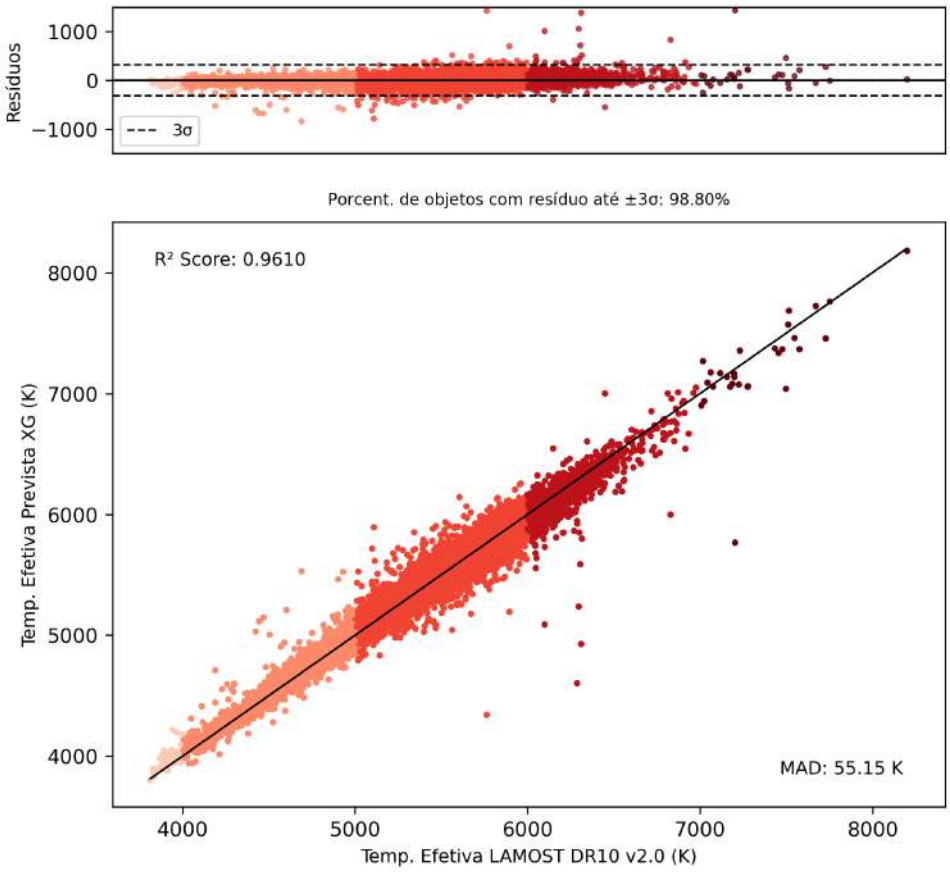

Figura 4.16: Resultados do modelo `splusL_01_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10.

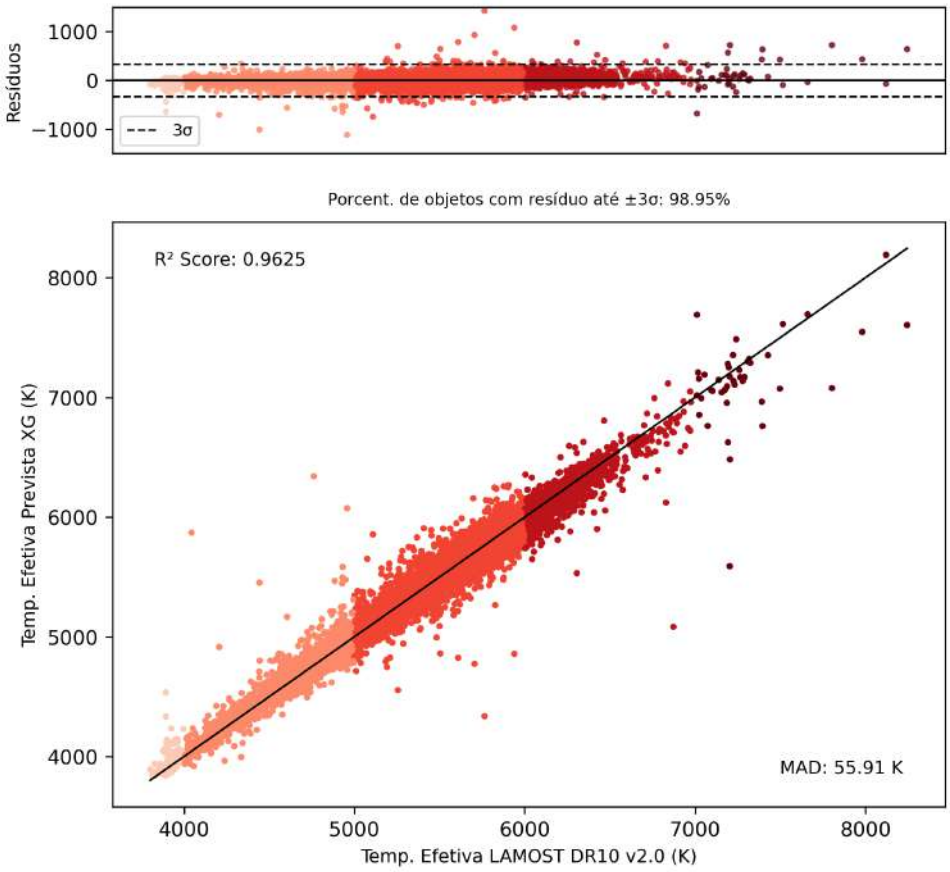

Figura 4.17: Resultados do modelo `splusL_02_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10.

Assim como foi realizado com os modelos construídos utilizando RF, os modelos treinados com XGB foram aplicados na próxima etapa para uma análise detalhada de desempenho. Após a aplicação, foi verificado qual técnica oferece os melhores resultados na previsão dos parâmetros estelares, possibilitando uma comparação eficaz entre as duas abordagens e identificando a que proporciona maior precisão e confiabilidade. Os resultados serão apresentados na Seção 4.4.

4.3.2 *XGBoost* na Previsão da Gravidade Superficial

Para o parâmetro de gravidade superficial, foi aplicada também a técnica de XGB para a previsão e teste. Assim como nas seções anteriores, não serão apresentados todos os gráficos gerados, mas apenas aqueles que representam os melhores modelos identificados. As métricas de desempenho para os modelos mais e menos restritos estão detalhadas nas Tabelas 4.14 e 4.15, respectivamente.

Gravidade Superficial (_logg)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_01_XGB	255.039	0,8802	0,07
splusL_01_XGB	72.036	0,8715	0,07
jplusA_01_XGB	2.724	0,9650	0,03
splusA_01_XGB	8.123	0,9772	0,04

Tabela 4.14: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de log g com *XGBoost*.

Gravidade Superficial (_logg)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
jplusL_02_XGB	278.879	0,8803	0,07
splusL_02_XGB	76.035	0,8778	0,07
jplusA_02_XGB	4.782	0,9724	0,03
splusA_02_XGB	9.125	0,9727	0,05

Tabela 4.15: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de log g com *XGBoost*.

Baseando-se no levantamento do J-PLUS, o melhores modelos, com base no R^2 *Score* e no MAD, igualmente como que ocorreu com o RF, são os modelos `jplusA_01_XGB_logg` e `jplusA_02_XGB_logg`, mais e menos restritos, respectivamente. Não podemos determiná-los como os melhores, pois, como dito anteriormente, possuem pouco objetos na amostra de treinamento.

Portanto, no número de objetos, é o modelo `jplusL_01_XGB_logg` é o modelo mais confiável para o cenário mais restritivo, onde alcançou um R^2 *Score* de 0,8802 com 255.039 objetos na amostra, apresentando um desvio mediano absoluto de 0,07 dex. Esses resultados indicam que este modelo teve um bom desempenho na previsão de log g, mostrando uma correlação forte entre os valores reais e previstos e um erro relativamente baixo.

Por outro lado, o melhor modelo menos restrito é o `jplusL_02_XGB_logg`, com um R^2 *Score* de 0,8803 e 278.879 objetos. O MAD para esse modelo foi de 0,07 dex, o que demonstra que este modelo apresentou um equilíbrio favorável entre precisão (MAD) e correlação (R^2 *Score*).

Ambos os modelos embora não apresentem as melhores métricas, destacam-se pelo grande número de objetos na amostra de entrada para o treinamento. Como mencionado anteriormente, quanto mais objetos disponíveis, melhor tende a ser o processo de análise e treinamento dos modelos de ML evitando o *overfitting*. Isso confere uma confiabilidade maior a esses modelos em comparação com aqueles que possuem métricas ligeiramente superiores, mas são treinados com um número reduzido de objetos. Portanto, a quantidade

significativa de objetos nos modelos `jplusL_01_XGB_logg` e `jplusL_02_XGB_logg` garante um desempenho mais sólido e confiável para a previsão de log g.

Os gráficos que ilustram o desempenho desses modelos, mostrando a correlação entre os valores previstos e reais de gravidade superficial, serão apresentados nas Figuras a seguir, onde a Figura 4.18 é para o modelo restrito e a Figura 4.19 diz respeito ao modelo menos restrito.

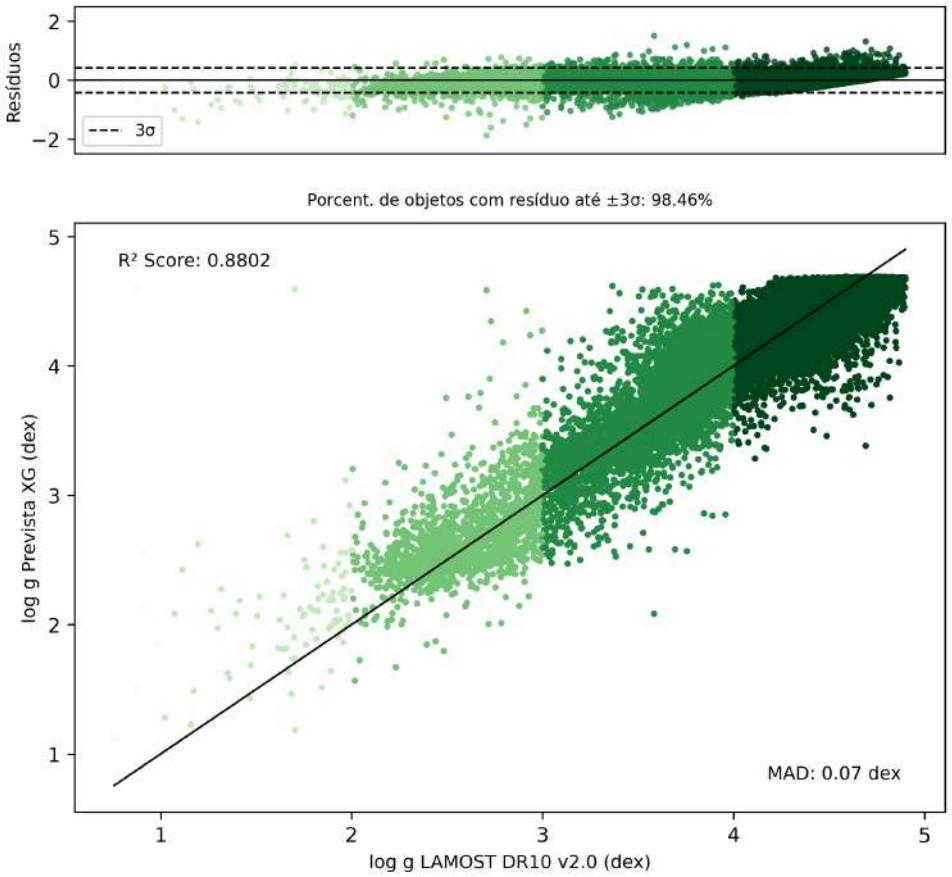

Figura 4.18: Resultados do modelo `jplusL_02_RF_logg` para a previsão de log g utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

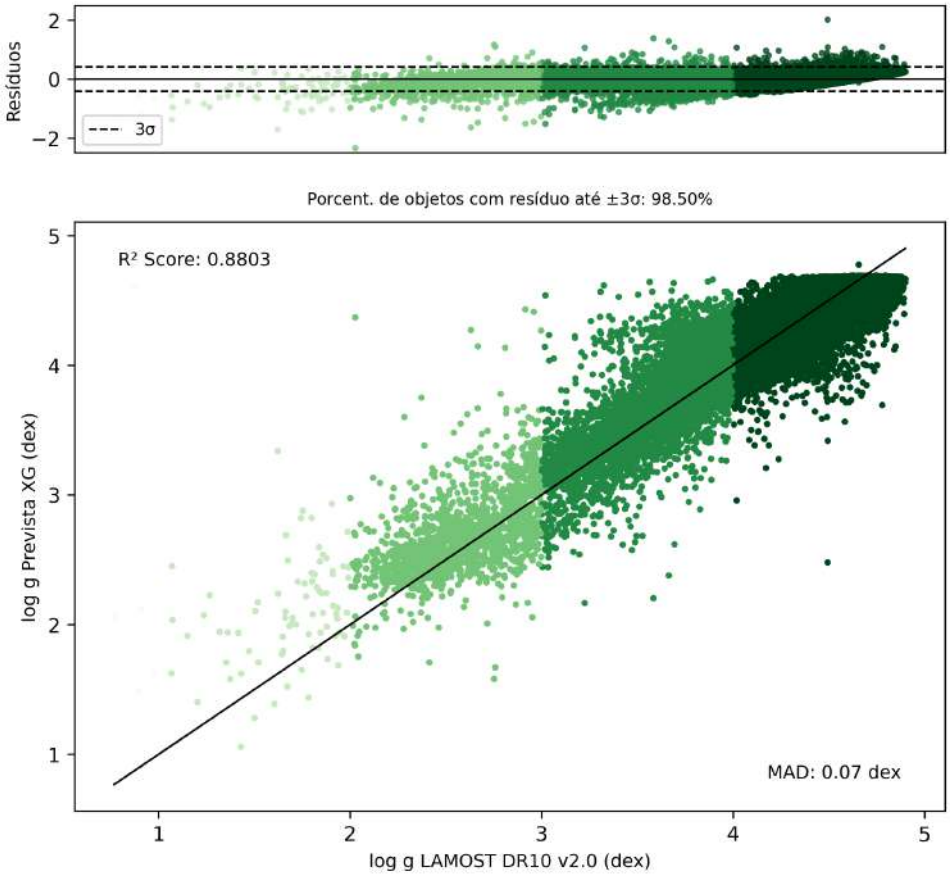

Figura 4.19: Resultados do modelo `jplusL_02_RF_logg` para a previsão de log g utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

Para o levantamento S-PLUS, tanto os modelos restritos quanto os menos restritos mais confiáveis, são aqueles que utilizaram uma quantidade maior de objetos na previsão de log g. O modelo restrito, `splusL_01_XGB_logg`, alcançou um R^2 *Score* de 0,8715, com 72.036 objetos na amostra e um desvio mediano absoluto de 0,07 dex. Já o modelo menos restrito, `splusL_02_XGB_logg`, mostrou uma leve melhoria com um R^2 *Score* de 0,8778, utilizando uma amostra maior de 76.035 objetos e mantendo o mesmo MAD de 0,07 dex.

A Figura 4.20 ilustra o comportamento do modelo `splusL_01_RF_logg`, destacando sua eficiência na previsão de log g e a consistência entre as previsões e os dados observados. Já a Figura 4.21 evidencia a correlação e o comportamento do modelo `splusL_02_RF_logg`, com um conjunto de dados mais amplo, reforçando sua capacidade de generalização nas previsões de log g.

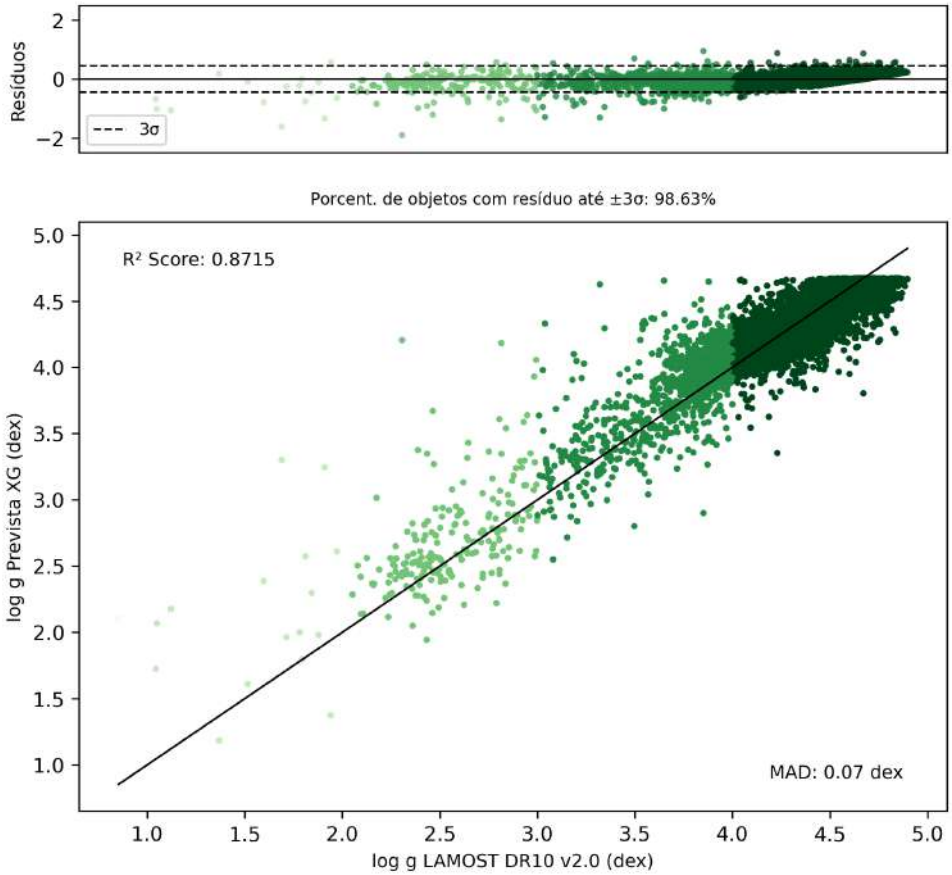

Figura 4.20: Resultados do modelo `splusL_01_RF_logg` para a previsão de log g utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

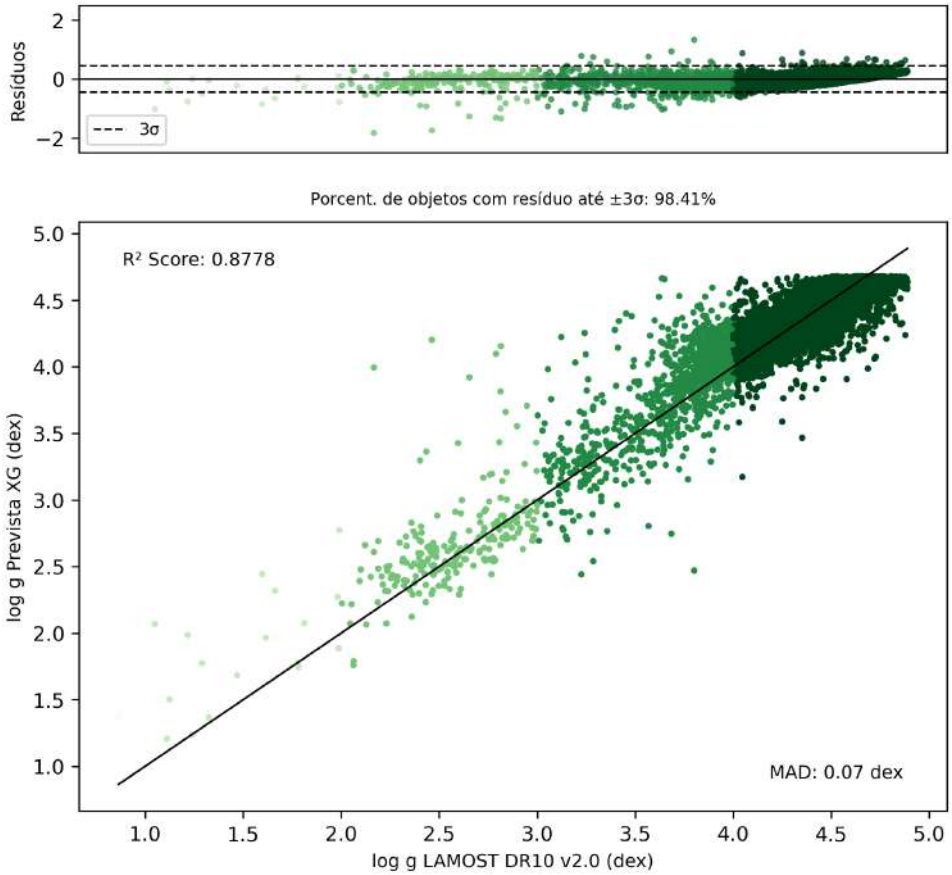

Figura 4.21: Resultados do modelo `splusL_02_RF_logg` para a previsão de log g utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de gravidade superficial de 1,0 dex, variando de 1,0 dex até 5,0 dex.

4.3.3 *XGBoost* na Previsão da Metalicidade

Para previsão da metalicidade, também foi utilizada a técnica de XGB para fazer treinamentos e testes. Assim como nas outras seções, aqui serão apresentados apenas os melhores modelos, sem incluir todos os gráficos gerados. As métricas dos modelos mais e menos restritos estão descritas nas Tabelas 4.16 e 4.17, respectivamente.

Metalicidade (`_feh`)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
`jplusL_01_XGB_feh`	255.039	0,8541	0,08
`splusL_01_XGB_feh`	72.036	0,8502	0,08
`jplusA_01_XGB_feh`	2.724	0,8045	0,08
`splusA_01_XGB_feh`	8.123	0,8033	0,07

Tabela 4.16: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo restrito para a previsão de [Fe/H] utilizando a técnica de *XGBoost*.

Metalicidade (`_feh`)			
Modelo	**Quant. Obj.**	***R² Score***	**MAD [dex]**
`jplusL_02_XGB_feh`	278.879	0,8499	0,07
`splusL_02_XGB_feh`	76.035	0,8421	0,08
`jplusA_02_XGB_feh`	4.782	0,8387	0,07
`splusA_02_XGB_feh`	9.125	0,7816	0,06

Tabela 4.17: Resultados de R^2 *Score* e desvio mediano absoluto para cada modelo menos restrito para a previsão de [Fe/H] utilizando a técnica de *XGBoost*.

O modelo restrito com o melhor desempenho foi o `jplusL_01_XGB_feh`, que apresentou um R^2 *Score* de 0,8541 e um MAD de 0,08 dex. Além disso, ele tem o maior número de objetos na amostra, 255.039 no total, o que reforça a confiança das previsões feitas no cenário restrito.

Para o caso menos restrito, o modelo `jplusL_02_XGB_feh` foi o melhor, com um R^2 *Score* de 0,8499 e um MAD de 0,07 dex. Assim como o modelo restrito, o destaque está nas métricas alcançadas e na quantidade de objetos utilizados no treinamento, 278.879 objetos no total, o que fortalece sua robustez em comparação com os demais modelos no cenário menos restrito.

A Figura 4.22 e a Figura 4.23 ilustram o comportamento dos melhores modelos encontrados, mostrando a correlação entre os valores reais e previstos para [Fe/H].

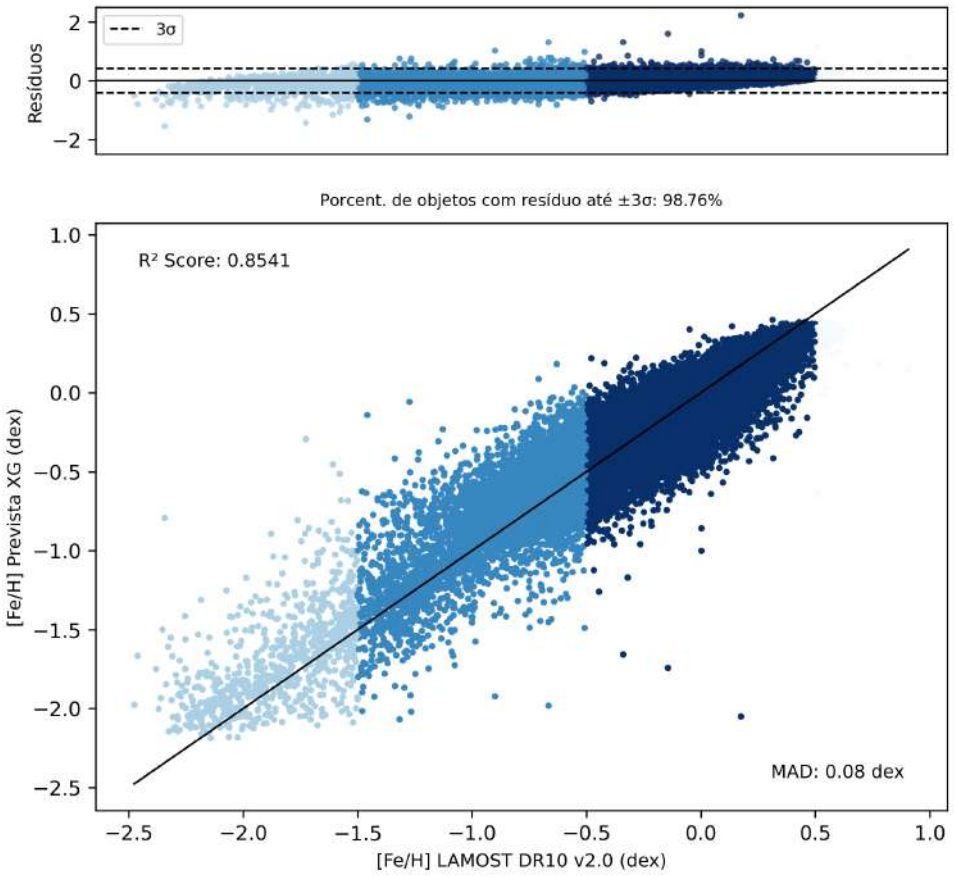

Figura 4.22: Resultados do modelo `jplusL_01_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

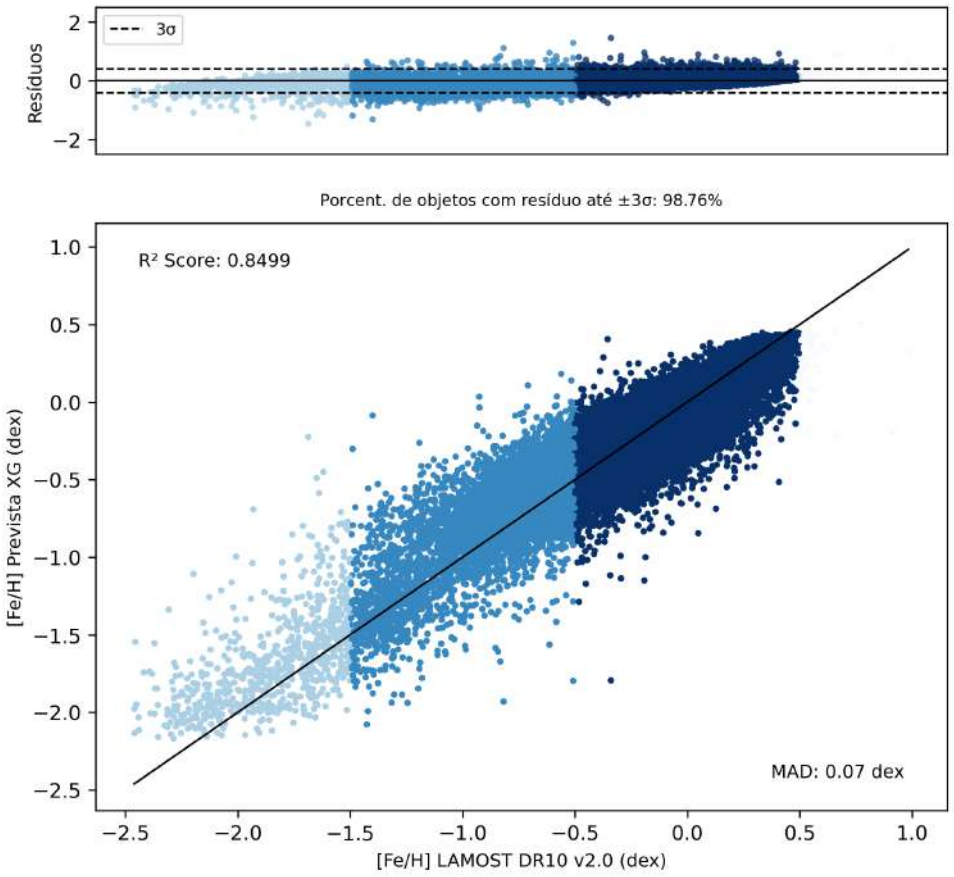

Figura 4.23: Resultados do modelo `jplusL_02_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* com objetos em comum entre o J-PLUS DR3 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

Por fim, também considerando o S-PLUS como levantamento principal, tem-se que os melhores modelos gerados com esse levantamento, também são os modelos que utilizaram o LAMOST como levantamento auxiliar. Para o cenário restrito, tem-se que aquele que apresentou as melhores métricas é o modelo `splusL_01_XGB_feh` e para o caso menos restrito consiste no modelo `splusL_02_`XGB`_feh`. Onde o `splusL_01_XGB_feh` apresentou um R^2 *Score* de 0,8502 e um MAD de 0,08 dex para uma amostra de 72.036 estrelas no treinamento e teste. Enquanto o modelo `splusL_02_XGB_feh` alcançou um R^2 *Score* de 0,8421 com um MAD de 0,08 dex.

Com os modelos já testados é possível avaliar o seus respectivos comportamentos e a correlação entre os valores previstos pelo teste e aqueles que são fornecidos pelo LAMOST. A Figura 4.24 evidencia o comportamento do modelo restrito, `splusL_01_XGB_feh`, enquanto a Figura 4.25 o comportamento do menos restrito, `splusL_02_XGB_feh`.

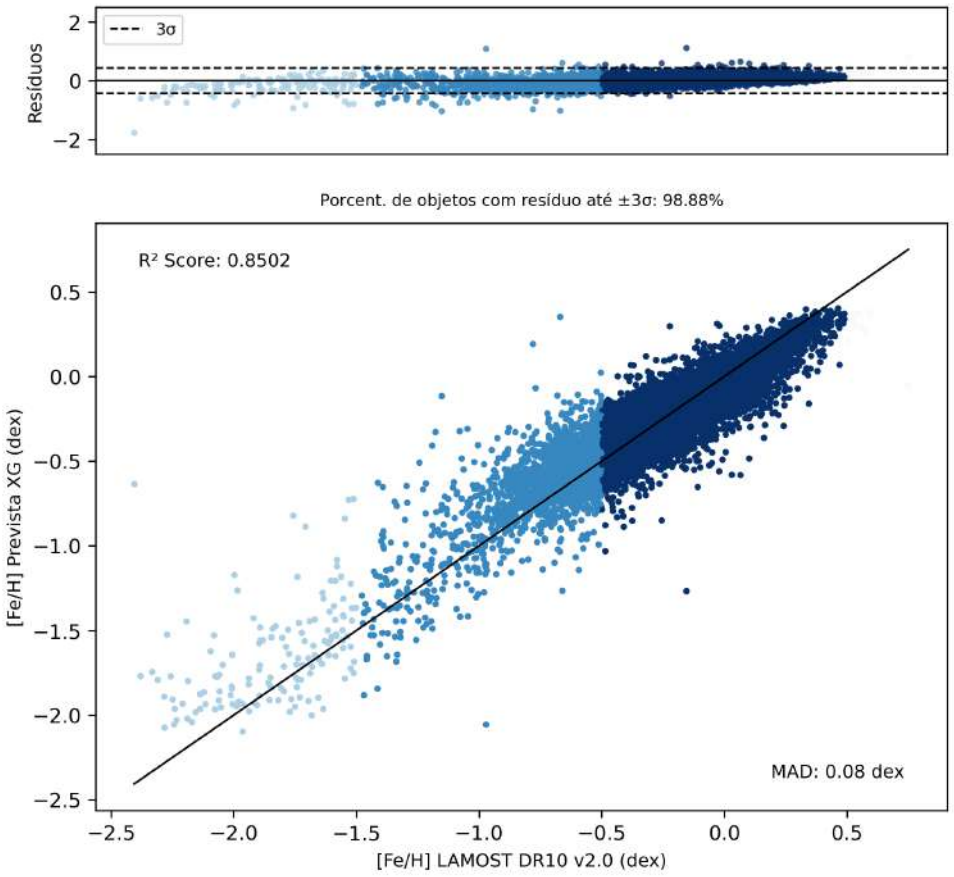

Figura 4.24: Resultados do modelo `splusL_01_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

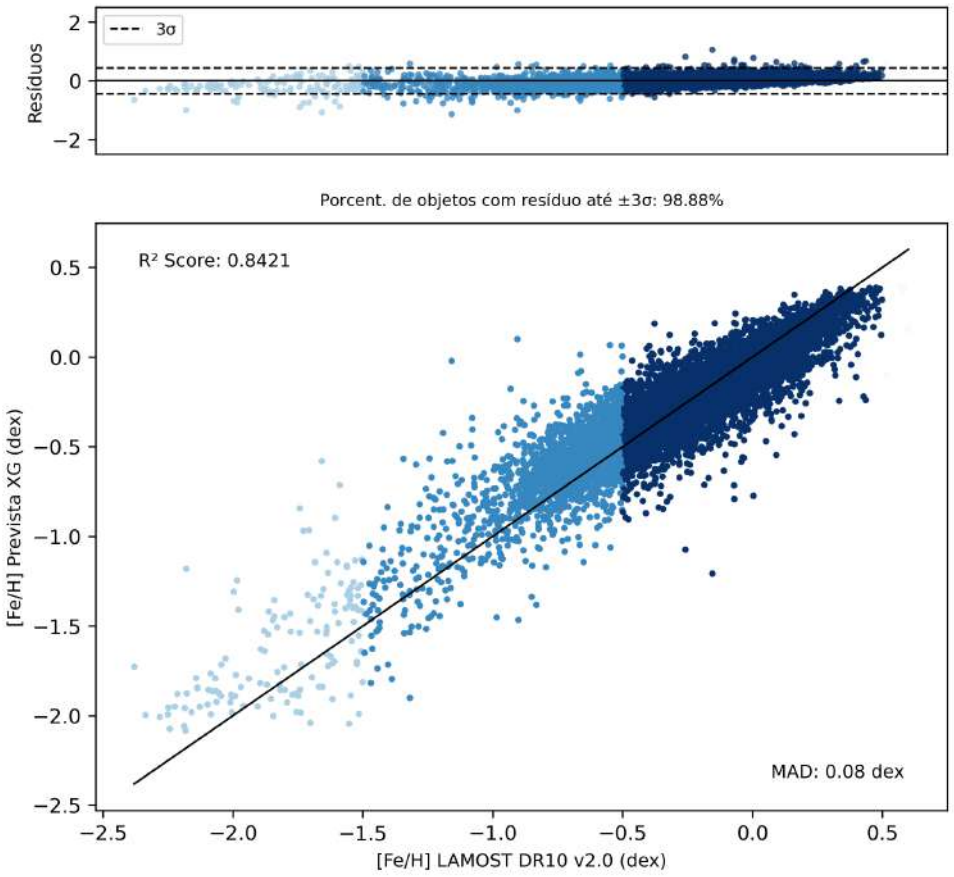

Figura 4.25: Resultados do modelo `splusL_01_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* com objetos em comum entre o S-PLUS iDR5 e o LAMOST DR10. As cores são os intervalos de metalicidade de 1,0 dex, variando de -2,5 dex até 0,5 dex.

Os modelos desenvolvidos para prever a metalicidade também serão aplicados em levantamentos voltados para a busca de exoplanetas como será mostrado na próxima seção. Após a implementação, será realizada uma análise das métricas, correlações e acurácia, a fim de compreendermos o desempenho dos modelos utilizando a técnica XGB na previsão de [Fe/H].

4.4 Determinação dos Parâmetros Estelares a partir da Aplicação dos Modelos

Nesta seção, serão apresentados os resultados obtidos a partir da aplicação dos melhores modelos treinados, conforme definidos nas Seções 4.2 e 4.3. Esses modelos foram aplicados aos levantamentos que buscam exoplanetas previamente selecionados, descritos na Seção 3.3.

Os modelos utilizados incluem as técnicas de RF e XGB, conforme detalhado nos Capítulos anteriores. A Tabela 4.18 apresenta os modelos baseados em RF, enquanto a

Tabela 4.19 lista os modelos baseados em XGB, empregados para prever cada parâmetro estelar.

T_{ef}	log g	[Fe/H]
jplusL_01_RF_teff	jplusL_01_RF_logg	jplusL_01_RF_feh
jplusL_02_RF_teff	jplusL_02_RF_logg	jplusL_02_RF_feh
splusL_01_RF_teff	splusL_01_RF_logg	splusL_01_RF_feh
splusL_02_RF_teff	splusL_02_RF_logg	splusL_02_RF_feh

Tabela 4.18: Modelos *Random Forest* que foram utilizados na determinação de parâmetros estelares de estrelas hospedeiras de exoplanetas.

T_{ef}	log g	[Fe/H]
jplusL_01_XGB_teff	jplusL_01_XGB_logg	jplusL_01_XGB_feh
jplusL_02_XGB_teff	jplusL_02_XGB_logg	jplusL_02_XGB_feh
splusL_01_XGB_teff	splusL_01_XGB_logg	splusL_01_XGB_feh
splusL_02_XGB_teff	splusL_02_XGB_logg	splusL_02_XGB_feh

Tabela 4.19: Modelos *XGBoost* que foram utilizados na determinação de parâmetros estelares de estrelas hospedeiras de exoplanetas.

Os modelos utilizaram como entrada para a previsão de temperatura efetiva (T_{ef}), gravidade superficial (log g) e metalicidade ([Fe/H]), as magnitudes absolutas calculadas a partir das magnitudes aparentes (já corrigidas pela extinção) fornecidas pelos levantamentos J-PLUS e S-PLUS.

Com os parâmetros previstos, foi possível calcular a luminosidade, raio e massa das estrelas dos levantamentos que buscam exoplanetas (Tabelas 3.2 e 3.3) utilizando as equações apresentadas na Seção 3.3.2. Devido ao tamanho das Tabelas, elas não serão apresentadas aqui. As Tabelas completas, com todas as previsões e resultados podem ser consultadas no Drive[35].

As Tabelas foram nomeadas conforme a técnica e a amostra utilizada. Por exemplo: a Tabela `kic_jplus_01_final_RF` consiste nos objetos em comum entre KIC e J-PLUS DR3 (`kic_jplus`) com parâmetros previstos pelo modelo menos restrito (`_01`) e utilizando a técnica de RF (`_RF`).

O conteúdo das Tabelas consiste em:

- Coordenadas RA e DEC dos objetos, em graus;
- Temperatura efetiva (coluna nomeada como "teff") prevista pelos nossos modelos, em K;

[35] https://linktr.ee/final_tables

- Gravidade superficial (coluna nomeada como "logg") prevista pelos nossos modelos, em dex;
- Metalicidade (coluna nomeada como "feh") prevista pelos nossos modelos, em dex;
- Seus respectivos erros previstos utilizando a técnica de Monte Carlo;
- Magnitude aparente na banda G (Gmag) e seu erro (e_Gmag) fornecidos pelo GAIA DR3;
- Distância (Dist) dos objetos calculada por Bailer-Jones *et al.* (2021) e fornecida pelo GAIA DR3, em pc;
- Correção bolométrica prevista para os objetos conforme descrito na Seção 3.3.2;
- Magnitude absoluta na banda G (Mg) calculada a partir da Equação 3.1;
- Magnitude bolométrica (Mbol) calculada a partir da Equação 3.5;
- Luminosidade da estrela (L) calculada a partir da Equação 3.6 e seu erro, em $L_{\odot}$;
- Raio da estrela (R) calculado a partir da Equação 3.11 e seu erro, em $R_{\odot}$;
- Massa da estrela (M), calculada a partir da relação proposta por Torres *et al.* (2010) e apresentada no final da Seção 3.3.2, e seu erro, em $M_{\odot}$.

Para algumas estrelas, os valores para luminosidade, massa e raio não estão disponíveis, pois durante a previsão dos modelos, eles não conseguiram prever corretamente os parâmetros estelares, exibindo valores extremamente baixos para eles. Durante o cálculo, os *scripts* interpretaram esse valores como "NaN", não realizando o cálculo desses parâmetros.

Após a previsão das incertezas avaliamos a eficácia das técnicas de RF e XGB para cada parâmetro estelar avaliando o desempenho com base na média (Me), desvio padrão (DP) e a mediana (Md) dos erros. Essas métricas foram determinadas a partir das previsões dos erros através do método de Monte Carlo (MC), com isso permitindo uma análise robusta da confiabilidade dos resultados. Nas seções a seguir discutiremos esses resultados.

4.4.1 Temperatura Efetiva

Nesta seção, apresentaremos os resultados das previsões da temperatura efetiva (T_{ef}) para os levantamentos que buscam exoplanetas.

4.4.1.1 Kepler

Esses resultados são referentes às estrelas do Kepler Input Catalog (KIC) que foram observadas em comum com o J-PLUS e o S-PLUS utilizando tando RF quanto XGB utilizando 1000 iterações na previsão dos erros por MC.

Primeiramente, com a técnica de RF, os modelos foram aplicados aos cenários mais e menos restritos, e os resultados são apresentados e na Tabela 4.20 são apresentados os modelos, quantidade de objetos de cada amostra e a estatística dos erros calculados através do MC.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_RF_teff`	60.731	36,55	20,50	33,99
`jplusL_02_RF_teff`	112.324	43,42	21,20	42,43
`splusL_01_RF_teff`	91	27,70	15,92	23,60
`splusL_02_RF_teff`	101	30,83	24,78	26,31

Tabela 4.20: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Em seguida, a técnica de XGB foi aplicada ao mesmo conjunto de dados, e as estatística dos erros são apresentadas na Tabela 4.21.

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_XGB_teff`	60.731	43,22	24,80	40,52
`jplusL_02_XGB_teff`	112.324	50,21	26,37	46,78
`splusL_01_XGB_teff`	91	37,41	20,84	33,55
`splusL_02_XGB_teff`	101	44,60	46,29	34,88

Tabela 4.21: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Ao compararmos os desempenhos das duas técnicas, o RF apresentou melhores resultados para cenários restritos, com menores valores de média global dos erros (Me) e mediana (Md) dos erros. O modelo `splusL_01_RF_teff` destacou-se como o mais consistente dentro dessa técnica, com os menores valores de Me e DP, indicando uma previsão mais estável e precisa para as estrelas do levantamento S-PLUS.

Por outro lado, o XGB apresentou resultados levemente maiores com relação ao RF, sugerindo um leve aumento de variabilidade nas previsões. Embora os valores de Me e Md tenham sido ligeiramente superiores aos do RF em alguns cenários, a estabilidade proporcionada pelo XGB o torna ainda sim bem confiável.

4.4.1.2 TESS

Diferentemente do Kepler, devido a limitações computacionais, a previsão dos erros a partir do Monte Carlo com os dados do J-PLUS e S-PLUS em comum com o TIC v8.2 foram utilizadas menos iterações, no caso 100 iterações. Sabemos que quanto mais iterações, mais confiáveis são os erros previstos, porém com essa limitação encontrada neste trabalho, que pretendemos corrigir futuramente, não foi possível inserir mais iterações no MC.

Os resultados obtidos com a técnica de RF estão listados na Tabela 4.22. Os valores foram calculados para os cenários restrito e menos restrito em ambos os levantamentos J-PLUS e S-PLUS.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_RF_teff`	1.628.834	33,06	19,90	30,55
`jplusL_02_RF_teff`	2.035.814	38,74	21,58	37,06
`splusL_01_RF_teff`	3.058.419	35,82	24,40	32,25
`splusL_02_RF_teff`	3.577.117	42,09	33,18	35,26

Tabela 4.22: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *Random Forest* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Para a técnica de XGB, os resultados das previsões estão detalhados na Tabela 4.23. Assim como no RF, os valores foram obtidos para os cenários restrito e menos restrito.

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_XGB_teff`	1.628.834	39,29	21,57	37,07
`splusL_02_XGB_teff`	2.035.814	45,18	24,48	43,16
`jplusL_01_XGB_teff`	3.058.419	47,79	30,31	43,78
`splusL_02_XGB_teff`	3.577.117	52,50	36,92	46,53

Tabela 4.23: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *XGBoost* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Comparando os desempenhos das duas técnicas, o RF apresentou, em geral, menores médias e medianas dos erros em ambos os cenários. No entanto, os desvios padrão dos erros foram maiores nos modelos menos restritos, indicando uma maior variabilidade nas previsões.

Por outro lado, o XGB demonstrou valores de média e mediana dos erros superiores ao RF, com desvios padrão (DP) também mais elevados. Essa maior variabilidade sugere que o XGB não foi tão eficaz quanto o RF na previsão de T_{ef} para as estrelas presentes no TIC v8.2.

4.4.1.3 Espectrógrafo HARPS

Apresentaremos aqui os resultados das previsões para estrelas que estão nos campos de observação do J-PLUS e S-PLUS com o catálogo do espectrógrafo HARPS. Foi aplicado a mesma metodologia utilizada para os outros dois levantamentos que buscam exoplanetas. Para essa amostra de objetos, também foram utilizadas 1000 iterações na previsão dos erros de T_{ef} das medidas com o Monte Carlo.

Ao realizar a correlação cruzada com dados do J-PLUS com o catálogo do HARPS notamos que não há nenhum objeto em comum, pois o J-PLUS observa no hemisfério norte e o HARPS é um instrumento do ESO (*European Southern Observatory*), observando o hemisfério sul celeste. Portanto os modelos treinados com objetos do J-PLUS não foram aplicados, com isso não possuem métricas para avaliação. Serão avaliados somente a aplicação dos modelos treinados com dados do S-PLUS para as duas técnicas de ML.

Nas Tabela 4.24 e 4.25 é possível conferir as incertezas alcançadas após a previsão por MC.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_RF_teff`	0	-	-	-
`jplusL_02_RF_teff`	0	-	-	-
`splusL_01_RF_teff`	83	8,67	11,39	5,67
`splusL_02_RF_teff`	108	13,21	22,87	7,04

Tabela 4.24: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrográfo HARPS, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Para o mesmo conjunto de dados, aplicando a técnica de XGB, alcançamos os seguintes resultados apresentados na Tabela 4.25.

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [K]**	**DP [K]**	**Md [K]**
`jplusL_01_XGB_teff`	0	-	-	-
`jplusL_02_XGB_teff`	0	-	-	-
`splusL_01_XGB_teff`	83	14,88	13,81	12,76
`splusL_02_XGB_teff`	108	19,61	26,36	13,42

Tabela 4.25: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para T_{ef} com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrógrafo HARPS, Me é a média global dos erros (em K), DP é o desvio padrão dos erros (em K) e Md é a mediana dos erros (em K).

Ao avaliar os resultados obtidos para os dados do HARPS, observa-se que a técnica de RF apresentou erros médios menores nos dois cenários aplicados (`splusL_01_RF_tef` e `splusL_02_RF_teff`). A média dos erros foi de 8,67 K para o cenário restrito e 13,21 K para o menos restrito. Além disso, o desvio padrão também foi consideravelmente mais baixo no modelo `splusL_01_RF_tef`, indicando menor variabilidade nos erros previstos. Esses resultados sugerem maior precisão e estabilidade da técnica de RF, especialmente nos cenários restritos.

Por outro lado, a técnica de *XGBoost* apresentou valores ligeiramente maiores de média e desvio padrão em comparação com o RF. No modelo restrito (`splusL_01_XGB_teff`), a média dos erros foi de 14,88 K, com um DP de 13,81 K, enquanto o cenário menos restrito (`splusL_02_XGB_teff`) apresentou valores de Me e DP mais elevados, indicando maior variabilidade nas previsões.

4.4.2 Gravidade Superficial

Nesta seção, apresentaremos os resultados das previsões da gravidade superficial (log g) para os levantamentos que buscam exoplanetas, onde foi aplicada a mesma metodologia para T_{ef}. Os resultados aqui apresentados são, também, para os dados do Kepler, TESS e HARPS.

4.4.2.1 Kepler

Apresentaremos aqui os resultados referentes às estrelas em comum com o J-PLUS e o S-PLUS e com o levantamento Kepler. Da mesma forma que foi feito para T_{ef}, para a previsão utilizando MC foram utilizadas 1000 iterações.

Com a técnica RF obtivemos os seguintes resultados pós aplicação:

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_logg`	60.731	0,03	0,02	0,02
`jplusL_02_RF_logg`	112.324	0,03	0,02	0,03
`splusL_01_RF_logg`	91	0,02	0,01	0,02
`splusL_02_RF_logg`	101	0,02	0,01	0,02

Tabela 4.26: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

Para o XGB, temos os seguintes resultados para as incertezas:

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_logg`	60.731	0,04	0,04	0,03
`jplusL_02_XGB_logg`	112.324	0,04	0,04	0,03
`splusL_01_XGB_logg`	91	0,02	0,01	0,02
`splusL_02_XGB_logg`	101	0,02	0,01	0,02

Tabela 4.27: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

A análise dos resultados para a previsão da gravidade superficial (log g) nos dados do Kepler revela diferenças importantes entre as técnicas de RF e XGB. A técnica de RF apresentou desempenhos ligeiramente melhores nos cenários restritos, especialmente para

os modelos com relação ao levantamento J-PLUS. Os modelos em relação ao S-PLUS, para ambas as técnicas, apresentaram o mesmo desempenho, ou seja, os mesmos valores de média global dos erros (Me), desvio padrão (DP) e mediana (Md), indicando maior precisão e estabilidade nas previsões.

Por outro lado, a técnica de XGB apresentou resultados inferiores, com Me e DP mais elevadas em comparação ao RF, nos modelos aplicados ao levantamento J-PLUS. As previsões com do XGB não superou o RF em nenhuma métrica analisada, demonstrando um desempenho geral menos eficaz para este parâmetro utilizando o J-PLUS.

Para o S-PLUS independe de qualquer técnica de ML visto que apresentaram o mesmo desempenho, levando em consideração as métricas apresentadas. Seria o caso, de utilizar outros meios de avaliação não apresentados aqui, como por exemplo, tempo de execução, consumo de processamento e energia e entre outros.

4.4.2.2 TESS

Foi utilizada também a mesma metodologia para a previsão da gravidade superficial (log g) para os dados do TESS (100 iterações no Monte Carlo). Os resultados obtidos com a técnica de RF estão listados na Tabela 4.28, enquanto os valores para a técnica de XGB são apresentados na Tabela 4.29. Ambas as técnicas foram aplicadas em cenários restritos e menos restritos para os levantamentos J-PLUS e S-PLUS.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_logg`	1.628.834	0,02	0,01	0,02
`jplusL_02_RF_logg`	2.035.814	0,02	0,01	0,04
`splusL_01_RF_logg`	3.058.419	0,02	0,02	0,02
`splusL_02_RF_logg`	3.577.117	0,03	0,02	0,02

Tabela 4.28: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *Random Forest* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_logg`	1.628.834	0,03	0,03	0,02
`splusL_02_XGB_logg`	2.035.814	0,03	0,03	0,02
`jplusL_01_XGB_logg`	3.058.419	0,03	0,03	0,03
`splusL_02_XGB_logg`	3.577.117	0,03	0,03	0,03

Tabela 4.29: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *XGBoost* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

Analisando os resultados apresentados nas Tabelas para o RF, os modelos aplicados aos levantamentos J-PLUS e S-PLUS apresentaram valores similares de média global dos erros (Me), desvio padrão (DP) e mediana (Md). No entanto, observou-se um aumento no desvio padrão nos cenários menos restritos utilizando o J-PLUS, o que sugere uma menor estabilidade, mas não prejudicando o seu desempenho.

Para a técnica de XGB, os resultados foram ligeiramente inferiores em comparação ao RF. Embora os valores de Me tenham sido consistentes entre os modelos, os desvios padrão foram levemente mais elevados. Além disso, não houve uma redução significativa nos valores de Md, exceto para o modelo utilizando o J-PLUS no cenário menos restrito, reforçando assim que a técnica não conseguiu superar o desempenho do RF.

4.4.2.3 Espectrógrafo HARPS

De forma análoga ao que foi realizado para os dados do TESS, foi aplicada a mesma metodologia para a previsão da gravidade superficial (log g) nas estrelas presentes nos dados do espectrógrafo HARPS. Considerando o tamanho da amostra e a necessidade de uma boa análise, o método de Monte Carlo foi utilizado com 1000 iteração para os dados.

Os resultados obtidos com a técnica de RF para o espectro HARPS estão listados na Tabela 4.30, enquanto os valores para a técnica de XGB são apresentados na Tabela 4.31. Ambas as técnicas foram implementadas considerando cenários restritos e menos restritos, com o intuito de analisar a precisão das previsões para o espectro HARPS em diferentes condições.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_logg`	0	-	-	-
`jplusL_02_RF_logg`	0	-	-	-
`splusL_01_RF_logg`	83	0,01	0,01	0,002
`splusL_02_RF_logg`	108	0,01	0,01	0,003

Tabela 4.30: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrográfo HARPS, Me é a média global dos erros (em dex), DP é o desvio padrão dos erros (em dex) e Md é a mediana dos erros (em dex).

Para o mesmo conjunto de dados, aplicando a técnica de XGB, alcançamos os seguintes resultados apresentados na Tabela 4.25.

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_logg`	0	-	-	-
`jplusL_02_XGB_logg`	0	-	-	-
`splusL_01_XGB_logg`	83	0,01	0,01	0,003
`splusL_02_XGB_logg`	108	0,01	0,01	0,02

Tabela 4.31: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrógrafo HARPS, Me é a média global dos erros (em dex), DP é o desvio padrão dos erros (em dex) e Md é a mediana dos erros (em dex).

Em termos de precisão, os modelos tanto utilizando RF quanto XGB apresentaram erros médios (Me), desvios padrão (DP) e medianas (Md) dos erros iguais, independente do levantamento principal utilizado (J-PLUS/S-PLUS). Considerando que o objetivo deste trabalho é obter boas previsões de log g, para esse levantamento independe de qual técnica utilizar para esse tipo de análise, pois ambas ofereceram previsões confiáveis e precisas.

4.4.3 Metalicidade

Nesta seção, apresentaremos os resultados das previsões da metalicidade ([Fe/H]) para os levantamentos de exoplanetas, utilizando as duas técnicas de RF e XGB. As previsões foram realizadas de maneira similar àquelas para a temperatura e gravidade, com a biblioteca *astropack*, que integra outras bibliotecas do *Python*.

O objetivo é analisar a precisão das previsões para as estrelas dos levantamentos

Kepler, TESS e HARPS, em cenários mais e menos restritos, e avaliar qual das duas técnicas oferece o melhor desempenho.

4.4.3.1 Kepler

Nesta seção, apresentaremos os resultados das incertezas de metalicidade ([Fe/H]) para os dados obtidos pelo levantamento Kepler previstas pelo Monte Carlo. Com a técnica de RF, os modelos foram aplicados aos cenários mais e menos restritos, e os resultados são apresentados na Tabela 4.32.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_feh`	60.731	0,07	0,04	0,07
`jplusL_02_RF_feh`	112.324	0,08	0,04	0,08
`splusL_01_RF_feh`	91	0,05	0,03	0,05
`splusL_02_RF_feh`	101	0,05	0,03	0,05

Tabela 4.32: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para [Fe/H] com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

E para as previsões usando o XGB, a Tabela 4.33 apresenta os seguintes resultados:

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_feh`	60.731	0,09	0,05	0,08
`jplusL_02_XGB_feh`	112.324	0,09	0,05	0,10
`splusL_01_XGB_feh`	91	0,06	0,03	0,06
`splusL_02_XGB_feh`	101	0,06	0,03	0,06

Tabela 4.33: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para [Fe/H] com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o KIC, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

Com base nas incertezas apresentados nas Tabelas 4.32 e 4.33, para os modelos baseados em *Random Forest*, observa-se que a média dos erros variou entre 0,05 dex e 0,08 dex, enquanto o desvio padrão esteve consistentemente em torno de 0,03 dex a 0,04 dex, com a mediana alinhada aos valores médios. Este comportamento indica boa estabilidade e precisão para as previsões, especialmente nos modelos restritos (`_01`), onde a média e a mediana são menores, sugerindo menor dispersão dos erros.

Por outro lado, os modelos com XGB apresentaram valores de Me ligeiramente mais elevados, variando entre 0,06 dex e 0,09 dex, com desvios padrão (DP) em torno de 0,03 dex a 0,05 dex. Embora os resultados para os modelos restritos (`_01`) e menos restritos (`_02`) sejam próximos, os valores ligeiramente superiores em média e mediana dos erros indicam que, mais uma vez, o RF apresenta desempenho levemente superior em termos de precisão.

4.4.3.2 TESS

Nesta seção, apresentaremos os resultados das previsões da metalicidade ([Fe/H]) para o levantamento TESS. A Tabela 4.34 apresenta os resultados para as previsões com o RF enquanto a Tabela 4.35 apresenta os resultando com previsões com o XGB.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_feh`	1.628.834	0,06	0,04	0,06
`jplusL_02_RF_feh`	2.035.814	0,07	0,04	0,06
`splusL_01_RF_feh`	3.058.419	0,06	0,04	0,05
`splusL_02_RF_feh`	3.577.117	0,06	0,04	0,06

Tabela 4.34: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *Random Forest* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_feh`	1.628.834	0,07	0,04	0,07
`splusL_02_XGB_feh`	2.035.814	0,08	0,05	0,07
`jplusL_01_XGB_feh`	3.058.419	0,07	0,04	0,06
`splusL_02_XGB_feh`	3.577.117	0,07	0,04	0,08

Tabela 4.35: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para log g com *XGBoost* utilizando 100 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o TIC v8.2, Me é a média global dos erros (dex), DP é o desvio padrão dos erros (dex) e Md é a mediana dos erros (dex).

Para a técnica de RF, os resultados indicam uma maior estabilidade nos erros previstos, com valores de média dos erros, desvio padrão, e mediana muito próximos (ou iguais) entre os modelos restritos (`_01`) e menos restritos (`_02`). Em particular, os modelos aplicados utlizando o levantamento J-PLUS demonstraram ligeira superioridade, com valores de Md relativamente baixos (0,05-0,06 dex), indicando precisão nas previsões.

Por outro lado, a técnica de XGB apresentou desempenhos similares ao RF em relação aos valores de Me e Md nos levantamentos restritos. No entanto, nos cenários menos restritos, observou-se um aumento considerável no Md, que aumento 0,01 a 0,02 dex. Esse comportamento sugere uma ligeira instabilidade na técnica com relação a identificação de padrões, porém ainda sim, apresentou melhores métricas.

4.4.3.3 Espectrógrafo HARPS

Nesta seção, apresentaremos os resultados das previsões das incertezas para a metalicidade ([Fe/H]) das estrelas em comum com o espectrógrafo HARPS com os levantamentos J-PLUS e S-PLUS realizadas com 1000 iterações com o método de Monte Carlo.

As Tabelas 4.36 e 4.37 apresentam os resultados para o RF e para o XGB, respectivamente.

Random Forest				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_RF_feh`	0	-	-	-
`jplusL_02_RF_feh`	0	-	-	-
`splusL_01_RF_feh`	83	0,01	0,02	0,01
`splusL_02_RF_feh`	108	0,01	0,03	0,01

Tabela 4.36: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para [Fe/H] com *Random Forest* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrográfo HARPS, Me é a média global dos erros (em dex), DP é o desvio padrão dos erros (em dex) e Md é a mediana dos erros (em dex).

Para o mesmo conjunto de dados, aplicando a técnica de XGB, alcançamos os seguintes resultados apresentados na Tabela 4.25.

XGBoost				
Modelo Aplicado	**Quant. Obj.**	**Me [dex]**	**DP [dex]**	**Md [dex]**
`jplusL_01_XGB_feh`	0	-	-	-
`jplusL_02_XGB_feh`	0	-	-	-
`splusL_01_XGB_feh`	83	0,01	0,01	0,01
`splusL_02_XGB_feh`	108	0,02	0,02	0,01

Tabela 4.37: Resultados das previsões utilizando os modelos mais (`_01`) e menos (`_02`) restritos para [Fe/H] com *XGBoost* utilizando 1000 iterações. Quant. Obj. é a quantidade de objetos em comum J-PLUS e S-PLUS com o catálogo do espectrógrafo HARPS, Me é a média global dos erros (em dex), DP é o desvio padrão dos erros (em dex) e Md é a mediana dos erros (em dex).

Para a técnica de RF, o modelo restrito (`splusL_01_RF_feh`) apresentou uma média dos erros de 0,01 dex, com um desvio padrão de 0,02 dex e uma mediana de 0,01 dex. Já no modelo menos restrito (`splusL_02_RF_feh`), observou-se uma diminuição da Me para 0,01 dex, mas com um aumento no DP (0,02 dex) e uma Md de 0,01 dex, indicando maior variabilidade nos erros.

Por outro lado, a técnica de XGB apresentou desempenho pouco superior em termos de precisão e estabilidade. O modelo restrito (`splusL_01_XGB_feh`) obteve uma Me de 0,01 dex, com DP e Md iguais a 0,01 dex. Já o modelo menos restrito (`splusL_02_XGB_feh`) apresentou uma Me de 0,02 dex, mantendo um DP e uma Md consistentes de 0,01 dex. Esses valores indicam que, mesmo nos cenários menos restritos, o XGB conseguiu manter previsões estáveis, com menor variabilidade nos erros.

4.5 Comparações com a Literatura

Para avaliar a eficácia dos modelos desenvolvidos neste trabalho, comparamos nossos resultados com aqueles obtidos por outros trabalhos publicados na literatura. O primeiro trabalho analisado foi realizado por Carvalho (2022), que determinou parâmetros estelares de uma amostra restrita contendo 29.164 estrelas e uma amostra menos restrita com 44.483 estrelas do Kepler, também observadas pelo J-PLUS. A autora utilizou a técnica de RF como principal abordagem de aprendizado de máquina. Este estudo serviu como ponto de partida para o desenvolvimento deste trabalho, adotando uma metodologia similar em várias etapas.

Na Tabela 4.38, apresentamos uma comparação dos erros médios absolutos calculados por Carvalho (2022) com os erros globais médios obtidos neste estudo para a amostra restrita. Já na Tabela 4.39, apresentamos os mesmos dados para a amostra menos restrita, para o mesmo levantamento, o Kepler.

Restrito			
Trabalho	**T_{ef} [K]**	**log g [dex]**	**[Fe/H] [dex]**
Carvalho (2022)	70,00	0,07	0,10
Me. RF	36,55	0,03	0,07
Me. XGB	43,22	0,04	0,09

Tabela 4.38: Comparação com as incertezas alcançadas por Carvalho (2022) com este trabalho para estrelas em comum entre Kepler e J-PLUS no cenário restrito. Carvalho (2022) utilizou modelos baseados na técnica de *Random Forest* enquanto este trabalho utilizou de *XGBoost* e também de *Random Forest*. "Me. RF" e "Me. XGB" consiste na média das incertezas com *Random Forest* e *XGBoost*, respectivamente.

Menos Restrito			
Trabalho	**T_{ef} [K]**	**log g [dex]**	**[Fe/H] [dex]**
Carvalho (2022)	58,00	0,08	0,10
Me. RF	43,42	0,03	0,08
Me. XGB	50,21	0,04	0,09

Tabela 4.39: Comparação com as incertezas alcançadas por Carvalho (2022) com este trabalho para estrelas em comum entre Kepler e J-PLUS no cenário menos restrito. Carvalho (2022) utilizou modelos baseados na técnica de *Random Forest* enquanto este trabalho utilizou de *XGBoost* e também de *Random Forest*. "Me. RF" e "Me. XGB" consiste na média das incertezas com *Random Forest* e *XGBoost*, respectivamente.

Observamos que as incertezas medidas neste trabalho são significativamente menores do que as obtidas por Carvalho (2022). Para a amostra restrita, a incerteza média em T_{ef} foi reduzida de 70 K para 36,55 K com o RF e de 70 K para 43,22 K com o XGB. Já o log g diminuiu de 0,07 dex para 0,03 dex utilizando RF, enquanto utilizando o XGB diminuiu para 0,04 dex. E, para [Fe/H] utilizando RF teve uma boa redução, de 0,10 dex para 0,07 dex e utilizando XGB como técnica de ML teve uma sútil diminuição de 0,01 dex apenas. Resultados semelhantes foram observados para a amostra menos restrita, com uma redução de 58 K para 43,42 K em T_{ef}, de 0,08 dex para 0,03 dex em log g e de 0,10 para 0,08 dex na [Fe/H] com o RF. Enquanto para o XGB, a redução em T_{ef} foi de 70 K para 50,21, em log g foi de 0,08 para 0,04 e para [Fe/H] foi uma pequena redução de 0,10 para 0,09.

A redução das incertezas alcançada neste trabalho em relação a Carvalho (2022) era esperada, dado o uso de dados mais mais precisos fornecidos pelo LAMOST e também devido à quantidade de objetos, no caso deste trabalho, foi maior comparado com da autora. Enquanto Carvalho (2022) utilizou magnitudes (corrigidas pela extinção) do J-PLUS DR2 e dados do LAMOST DR8, este estudo empregou magnitudes provenientes do J-PLUS DR3 e parâmetros estelares do catálogo espectroscópico mais recente, LAMOST DR10, disponibilizado em 2024. Essas reduções destacam o impacto positivo da combinação de dados mais recentes, mais número de objetos na amostra de treinamento e técnicas avançadas de ML, como XGB e Monte Carlo na previsão dos erros, que demonstraram ser eficazes em capturar padrões mais sutis nos dados.

Com relação exclusivamente ao levantamento J-PLUS, foi possível comparar os resultados deste trabalho com os estudos feitos por Whitten *et al.* (2019), Yang *et al.* (2022) e Galarza *et al.* (2022), este último tendo usado o mesmo critério de erros de magnitude usado nos nossos modelos restritos, ou seja, erros de magnitude menor que 0,1.

Parâm.	**MAD J-PLUS RF**	**MAD J-PLUS XGB**	**MAD S-PLUS RF**	**MAD S-PLUS XGB**
T_{ef}	58,99	57,88	56,37	55,05
log g	0,07	0,07	0,07	0,07
[Fe/H]	0,08	0,08	0,08	0,08

Tabela 4.40: Desvio mediano absoluto (MAD) alcançados pelos nossos modelos restritos utilizando dados J-PLUS/S-PLUS com o LAMOST DR10 utilizando as técnicas de *Random Forest* e *XGBoost* na previsão de cada parâmetro estelar. T_{ef} é dado em K, log g e [Fe/H] é dado em dex.

Parâm.	**MAD J-PLUS RF**	**MAD J-PLUS XGB**	**MAD S-PLUS RF**	**MAD S-PLUS XGB**
T_{ef}	54,98	53,88	56,37	55,05
log g	0,07	0,07	0,07	0,07
[Fe/H]	0,08	0,07	0,08	0,08

Tabela 4.41: Desvio mediano absoluto (MAD) alcançados pelos nossos modelos menos restritos utilizando dados J-PLUS/S-PLUS com o LAMOST DR10 utilizando as técnicas de *Random Forest* e *XGBoost* na previsão de cada parâmetro estelar. T_{ef} é dado em K, log g e [Fe/H] é dado em dex.

Comparando com Whitten *et al.* (2019) onde os autores utilizaram dados do J-PLUS DR1 e desenvolveram algoritmos baseados em redes neurais e magnitudes sintéticas calculadas a partir dos espectros dos levantamentos BOSS, SEGUE e Legacy para prever a temperatura efetiva e metalicidade de estrelas pobres em metais. Para a T_{ef}, os autores alcançaram um desvio padrão de $\sim$ 91 K para estrelas com temperatura entre 4500 K até 8500 K, enquanto os nossos modelos, calculando para a mesma faixa de temperatura, alcançaram um MAD de 60,24 K utilizando RF e 58,91 K com XGB. Para a metalicidade, Whitten *et al.* (2019) alcançaram um desvio padrão de $\sim$ 0,2 dex para estrelas com metalicidade menores que -0,5. Para a mesma faixa de metalicidade, nossos modelos alcançaram um MAD de 0,1 dex com RF e XGB. Embora utilizem métricas de erro diferentes, no caso desvio padrão ao invés de desvio mediano absoluto, os modelos desenvolvidos neste trabalho apresentaram um desempenho melhor para a temperatura e similar para a metalicidade aos obtidos por Whitten *et al.* (2019).

Yang *et al.* (2022) utilizaram dados do J-PLUS DR1 e LAMOST DR5 para prever os parâmetros estelares e abundâncias, para mais de 2 milhões de estrelas utilizando redes neurais. Para a T_{ef}, os autores alcançaram uma precisão de $\sim$ 55 K, para log g de $\sim$ 0,15 dex e para [Fe/H] de $\sim$ 0,07 dex. Os valores encontrados pelos autores são similares com os nossos na previsão da T_{ef} e [Fe/H] (vide a Tabela 4.40) tendo uma performance equiparável com diferentes técnicas de ML. Para log g, os erros reportados pelos autores

são muito maiores quando comparados aos nossos.

Galarza *et al.* (2022) utilizaram de dados do J-PLUS DR2, LAMOST DR5 e do SEGUE SDSS-II para treinar algoritmos de aprendizado de máquina, também baseados em árvores de decisão, para a previsão dos parâmetros de estrelas de baixa metalicidade. Os autores em seu trabalho reportaram incertezas de $T_{ef} \sim 41$ K e [Fe/H] $\sim$ 0,09 dex que são resultados equiparáveis, exceto para log g, onde do autor é maior que o nosso. Os autores utilizaram intervalos similares aos utilizados nesse trabalho para treinamento e teste de ML.

Um estudo que é interessante comparar é com o de Yuan *et al.* (2023). Esses autores utilizaram dados do J-PAS (miniJPAS; Benitez *et al.* (2014)) para prever parâmetros estelares a partir do ajuste da distribuição espectral de energia (SED) em comparação com dados do LAMOST DR7. Para T_{ef}, os autores reportaram uma incerteza típica inferior a 150 K, valor consideravelmente maior do que as incertezas encontradas em nosso trabalho, onde alcançamos uma precisão significativamente melhor, com MAD variando entre 58,99 K (RF) e 57,88 K (XGB) (Tabela 4.40) para modelos restritos e 54,98 K (RF) 57,88 K (XGB) para modelos menos restritos (Tabela 4.41). Para [Fe/H], os autores demonstraram que é possível estimar esse parâmetro com uma precisão melhor que 0,1 dex a partir da fotometria do miniJPAS. Nossos resultados são melhores comparados aos dos autores, com MAD variando entre 0,08 dex (RF) e 0,08 dex (XGB) (Tabela 4.40) para modelos restritos e 0,08 dex (RF) e 0,07 dex (XGB) para modelos menos restritos (Tabela 4.41). Isso indica que técnicas baseadas em aprendizado de máquina aplicadas aos dados do J-PLUS e S-PLUS podem alcançar precisão superiores à abordagem utilizada por Yuan *et al.* (2023). Outro ponto a ser destacado é o fato de que a amostra utilizada pelos autores é uma amostra incipiente, pois o J-PAS é um projeto que ainda está no início, portanto amostras futuras podem melhorar o desempenho aqui mostrado.

Outro estudo com o qual podemos comparar os nossos resultados, é com o de Cordeiro da Silva (2023). O autor utilizou o RF para determinar os mesmos parâmetros deste trabalho com o objetivo de identificar novas subanãs quentes com dados do J-PLUS e S-PLUS. O autor avaliou os desempenhos dos modelos utilizando a mesma métrica utilizada por nós, o desvio mediano absoluto (MAD). Com isso, podemos fazer a comparação entre os dois trabalhos, permitindo uma análise mais detalhada da eficácia do nosso estudo.

Vale ressaltar que Cordeiro da Silva (2023) utilizou em seu trabalho a magnitude aparente como *feature* no treinamento e testes de seus modelos para a previsão de T_{ef}, log g e [Fe/H]. Ele também realizou previsões e teste de um modelo que utiliza magnitude absoluta, mas somente para a obtenção de log g (vide a Seção 4.3.3 de Cordeiro da Silva (2023)). Além disso, o autor utilizou para treino e teste, dados com erros de magnitude igual aos nossos modelos restritos, e_mag $\leqslant$ 0,1.

Considerando o J-PLUS, Cordeiro da Silva (2023) reportou em seu trabalho um MAD de 59 K para T_{ef}, 0,08 dex para log g e 0,09 dex para [Fe/H] utilizando magnitude aparente.

Utilizando magnitude absoluta, o autor reportou um MAD de 0,06 dex. Comparando com a Tabela 4.40 nossos modelos tiveram melhores performances na previsão de T_{ef} e log g, independente da técnica utilizada e uma performance similar na previsão de [Fe/H]. Com relação a magnitude absoluta, utilizada na previsão de log g, o modelo do autor teve uma performance relativamente melhor quando comparado ao nosso, reduzindo 0,01 dex. Com relação ao S-PLUS, Cordeiro da Silva (2023) encontrou 51 K para T_{ef}, 0,09 dex para log g e 0,08 dex para [Fe/H] e, para magnitude absoluta, encontrou um MAD de 0,06 dex para log g, que são incertezas melhores comparadas às nossas.

É possível também fazer uma comparação entre as incertezas médias encontradas neste trabalho através do Monte Carlo, com aqueles que são fornecidos por levantamentos que buscam exoplanetas, por exemplo o catálogo TIC v8.2 (Paegert *et al.*, 2021). Ele foi desenvolvido a partir do levantamento TESS, fornece parâmetros estelares e suas incertezas com base em observações espectroscópicas. Esses dados são utilizados na busca e caracterização de exoplanetas, sendo fundamentais para a criação do catálogo TOI (TESS Objects of Interest; Guerrero *et al.* (2021))[36], que lista os objetos de interesse identificados pelo TESS, juntamente com parâmetros de potenciais exoplanetas associados. Para isso, efetuamos uma correlação cruzada a fim de identificar as estrelas em comum entre os levantamentos J-PLUS, S-PLUS e o TIC v8.2, permitindo uma análise detalhada dos resultados.

Ao realizar o cruzamento entre o J-PLUS e o TIC v8.2, obtivemos 1.628.834 estrelas com e_mag $\leqslant$ 0,1 e 2.035.814 com e_mag $\leqslant$ 0,2. Para o S-PLUS, os números correspondentes foram 3.058.419 estrelas com e_mag $\leqslant$ 0,1 e 3.577.117 com e_mag $\leqslant$ 0,2. No entanto, o TIC v8.2 não forneceu valores dos parâmetros para todos os objetos. Por isso, realizamos uma contagem dos valores ausentes antes de calcular as médias. Para as 1.628.834 estrelas com e_mag $\leqslant$ 0,1 observadas tanto pelo J-PLUS quanto pelo TESS, 927 não possuem valores para a temperatura efetiva, 125.914 não possuem valores para a gravidade superficial e 108.653 não possuem valores para a metalicidade. Para estrelas com e_mag $\leqslant$ 0,2, das 2.035.814 estrelas, 1.097 não possuem valores para a temperatura efetiva, 126.023 não possuem valores para a gravidade superficial e 115.468 não possuem valores para a metalicidade.

Para as estrelas observadas pelo S-PLUS e pelo TESS que tiveram valores previstos pelos nossos modelos, com e_mag $\leqslant$ 0,1, das 3.058.419 estrelas, 622 não possuem valores de temperatura, 53.335 não possuem valores de gravidade e 48.336 não possuem valores de metalicidade na base de dados do TIC v8.2. Enquanto para as estrelas com e_mag $\leqslant$ 0,2, das 3.577.117 estrelas, 4.217 não possuem valores para a temperatura efetiva, 72.797 não possuem valores para a gravidade superficial e 74.947 não possuem valores para a metalicidade.

Com os objetos que possuem os valores dos parâmetros no TIC, comparamos com os

[36] https://exoplanetarchive.ipac.caltech.edu/docs/API_TOI_columns.html

resultados obtidos pelos nossos modelos. Vide a Tabela 4.42 para os objetos em comum com o J-PLUS e a Tabela 4.43 com relação ao S-PLUS:

Parâm.	**Cenário**	**Me. RF**	**Me. XGB**	**TIC v8.2**
T_{ef}	Restrito	33,06	39,29	122,44
	Menos Restrito	38,74	45,19	123,42
log g	Restrito	0,03	0,03	0,08
	Menos Restrito	0,03	0,03	0,07
[Fe/H]	Restrito	0,07	0,07	0,10
	Menos Restrito	0,07	0,08	0,10

Tabela 4.42: Comparação com as incertezas obtidas neste trabalho com as fornecidas pelo TIC v8.2 para as mesmas estrelas em comum entre J-PLUS e TESS. T_{ef} é dado em K, log g e [Fe/H] é dado em dex. "Me. RF" e "Me. XGB" consiste na média das incertezas com *Random Forest* e *XGBoost*, respectivamente.

Parâm.	**Cenário**	**Me. RF**	**Me. XGB**	**TIC v8.2**
T_{ef}	Restrito	35,82	47,79	124,15
	Menos Restrito	42,09	52,50	124,73
log g	Restrito	0,03	0,03	0,08
	Menos Restrito	0,03	0,03	0,07
[Fe/H]	Restrito	0,06	0,07	0,09
	Menos Restrito	0,06	0,07	0,09

Tabela 4.43: Comparação com as incertezas obtidas neste trabalho com as fornecidas pelo TIC v8.2 para mesmas estrelas em comum entre S-PLUS e TESS. T_{ef} é dado em K, log g e [Fe/H] é dado em dex. "Me. RF" e "Me. XGB" consiste na média das incertezas com *Random Forest* e *XGBoost*, respectivamente.

A análise das incertezas para o J-PLUS, apresentados na Tabela 4.42 indicam que, de maneira geral, as incertezas previstas pelos modelos RF e XGB também são mais baixas que as fornecidas pelo TIC v8.2. No cenário restrito, o modelo RF apresenta uma incerteza para T_{ef} de 33,06 K, em contraste com os 122,44 K do TIC v8.2, e para log g, os valores de 0,03 dex são significativamente mais baixos do que os 0,08 dex do TIC v8.2. Para a metalicidade, a diferença é de 0,03 dex. Esses resultados demonstram o bom desempenho dos modelos de aprendizado de máquina no contexto do J-PLUS, proporcionando estimativas com menores incertezas, o que pode ser crucial para estudos mais detalhados de exoplanetas e seus parâmetros estelares.

No caso do S-PLUS, conforme apresentado na Tabela 4.43, as incertezas nos parâmetros de temperatura efetiva, gravidade superficial e metalicidade obtidas com o modelo RF e XGB são em média mais baixas quando comparadas com as fornecidas pelo TIC

v8.2. O modelo RF, especialmente no cenário restrito, apresenta incertezas de T_{ef} de 35,82 K e log g de 0,03 dex, ambos inferiores aos valores obtidos pelo TIC v8.2 (124,15 K para T_{ef} e 0,08 dex para log g). De maneira similar, a metalicidade apresenta valores melhores, com uma diferença de 0,03 dex para o parâmetro [Fe/H]. Isso sugere que os modelos RF e XGB tendem a fornecer incertezas mais precisas para as estrelas observadas no S-PLUS, quando comparados com os valores relatados pelo TIC v8.2.

Em termos gerais, as incertezas desempenham um papel fundamental na avaliação da qualidade dos modelos de previsão. A comparação entre os parâmetros previstos por nossos modelos (RF e XGB) e os valores fornecidos pelo TIC v8.2, como mostrado nas Tabelas 4.42 e 4.43, evidencia a importância de se considerar os erros associados.

O TIC v8.2 calcula os erros previstos a partir das incertezas das medidas fotométricas e astrométricas disponíveis, propagando essas incertezas na estimativa dos parâmetros estelares. Para isso, o catálogo utiliza relações empíricas calibradas com amostras bem caracterizadas, além de combinar informações de diferentes catálogos, como Gaia e 2MASS, ajustando modelos estelares para inferir os parâmetros físicos (Guerrero *et al.*, 2021). Isso implica que os erros reportados podem ser subestimados, o que é uma limitação importante a ser considerada. Como alertado por Guerrero *et al.* (2021), esses erros podem não ser confiáveis, especialmente quando os dados do TIC v8.2 não estão disponíveis para todos os objetos.

Molina-Jorquera *et al.* (2024) utilizaram técnicas de aprendizado de máquina, especificamente Redes Neurais Artificiais (ANN, do inglês *Artificial Neural Networks*), para estimar a metalicidade de estrelas gigantes a partir da fotometria do S-PLUS DR3, treinando seus modelos com dados espectroscópicos do APOGEE SDSS-IV DR17. Embora nossos melhores modelos tenham sido treinados com dados do LAMOST DR10, devido à maior quantidade de objetos disponíveis, a comparação com nossos modelos baseados no APOGEE SDSS-IV DR17 é válida para avaliar a eficácia das diferentes técnicas de aprendizado de máquina empregadas neste trabalho.

Molina-Jorquera *et al.* (2024) reportaram um desvio padrão de 0,07 dex para a metalicidade, enquanto os nossos modelos, treinados com dados de magnitudes mais recentes do S-PLUS, o iDR5, obtiveram a mesma incerteza de 0,07 dex tanto no cenário restrito quanto no menos restrito, utilizando tanto RF quanto XGB (RF: Tabela 4.10 e Tabela 4.11; XGB: Tabela 4.16 e Tabela 4.17).

Apesar das diferenças entre as abordagens metodológicas — Redes Neurais no estudo de Molina-Jorquera *et al.* (2024) e RF/XGB neste trabalho — podem alcançar alta precisão na estimativa de metalicidade estelar. Além disso, a qualidade dos levantamentos fotométricos utilizados, como J-PLUS e S-PLUS, se mostrou adequada para a derivação de parâmetros estelares com incertezas reduzidas. Dessa forma, este estudo reforça que a combinação de aprendizado de máquina com dados fotométricos é uma abordagem robusta e confiável para a caracterização estelar.

4.6 Caracterização de Objetos de Interesse

Nesta seção, apresentamos os resultados da caracterização de alguns exoplanetas cujas estrelas hospedeiras estão na nossa amostra de objetos observados pelo J-PLUS ou S-PLUS. Conforme discutido na Seção 1.1.1, o raio de um planeta em trânsito pode ser determinado a partir do raio de sua estrela hospedeira, desde que a profundidade do trânsito seja conhecida, conforme descrito na Equação 1.2. Essa relação permite inferir parâmetros fundamentais sobre o exoplaneta a partir das observações fotométricas de sua estrela.

Para determinar o tipo de objeto que orbita a estrela, utilizamos a descrição de Petigura *et al.* (2018), que classifica os objetos com base em seus raios. De acordo com os autores é considerado como planeta ou anã marrom, os objetos com raios menores que 24 $R_{\oplus}$ (2,14 R_{Jup}). Objetos com raios maiores que esse limite foram classificados como binárias eclipsantes. Não é possível determinar a natureza dos objetos, como por exemplo no caso de anãs marrons, somente pelo raio, sendo preciso calcular a massa com outras técnicas de detecção, o que não foi feito por nós.

Neste trabalho, não calculamos diretamente a profundidade do trânsito dos objetos analisados, mas utilizamos valores disponíveis na literatura. Nogueira (2020), por exemplo, conduziu um estudo detalhado das curvas de luz obtidas pela missão Kepler, permitindo a inferência dos raios mínimos dos objetos que causam possíveis trânsitos. Com esses trânsitos, foi possível classificar os objetos como planetas ou sistemas binários eclipsantes. A análise resultou na identificação de 230 estrelas com 236 possíveis trânsitos, dos quais 31 foram classificados como candidatos a planetas ou anãs marrons, 178 como binárias eclipsantes e 27 objetos permaneceram sem classificação devido à falta de determinação do raio (Nogueira, 2020).

Dessas 230 estrelas que tiveram trânsitos calculados por Nogueira (2020), 7 delas estão presentes no campo do J-PLUS DR3, sendo possível a determinação de seus parâmetros pelos nossos modelos e, consequentemente, o calculo do raio do objeto que as orbitam. Carvalho (2022), em sua dissertação, também caracterizou esses mesmos objetos, determinando o raio da estrela e o raio do planeta. Para 6 dessas estrelas foi possível determinar os parâmetros, através dos modelos menos restritos enquanto somente para uma delas, determinamos através do modelo restrito. A classificação proposta pela autora se manteve, com exceção de somente um objeto, que foi reclassificado pelos nossos modelos.

Esse objeto que foi reclassificado, é o KIC 7661441 que teve o raio estelar calculado pelo KIC e utilizado por Nogueira (2020), como sendo $R_{\star}^{Nog} = 1,04\ R_{\odot}$, enquanto Carvalho (2022) determinou um raio de $R_{\star}^{Carv} = 1,03 \pm 0,03\ R_{\odot}$ utilizando RF. Já o nosso modelo menos restrito com RF, calculou um $R_{\star}^{RF} = 0,73 \pm 0,04\ R_{\odot}$ e com XGB foi de $R_{\star}^{XGB} = 0,74 \pm 0,05\ R_{\odot}$, o que representa um raio estelar 29% menor do que os valores calculados pelos outros autores.

Carvalho (2022) obteve o raio do objeto que orbita a estrela como sendo $R_p^{Carv} = 3,00 \pm 0,04$ R_{Jup}, enquanto Nogueira (2020) determinou o valor de $R_p^{Nog} = 3,03 \pm 0,07$ R_{Jup}, ambos classificando o objeto como uma binária eclipsante. No entanto, ao calcular o raio do planeta (R_p^{RF} e R_p^{XGB}) a partir da Equação 1.2, utilizando o raio da estrela calculado por nós, obtemos os seguintes valores: $R_p^{RF} = 2,11 \pm 0,27$ R_{Jup} e $R_p^{XGB} = 2,15 \pm 0,27$ R_{Jup} (utilizando os modelos em RF e XGB, respectivamente). Isso resulta em uma mudança na classificação, agora para exoplaneta ou uma anã marrom (no caso do XGB considerando a margem de erro). Este objeto teve trânsito identificado por Nogueira (2020), com profundidade Q igual a 8,91% de queda no fluxo, com um erro de 22,33% do valor de Q. Este valor de profundidade é muito alto, mesmo para planetas supergigantes, portanto é necessário estudos posteriores para confirmar a sua natureza.

É importante observar que as incertezas associadas ao raio do planeta (R_p) propostas pelos autores em seus trabalhos são muito baixas. Ao realizar alguns testes, foi possível verificar que ambos os estudos não consideraram a incerteza do raio da estrela ao calcular as incertezas do raio do planeta. Esse ponto é relevante, pois a ausência desse fator pode comprometer a precisão e a confiabilidade dos valores de incerteza apresentados.

O Kepler e o TESS fornecem, no Exoplanet Archive, dois catálogos chamados Kepler Objects of Interest (KOI)[37] e TESS Objects of Interest (TOI), que contêm informações sobre objetos observados pelo KIC e pelo TIC, respectivamente, e que apresentaram trânsitos detectados, mas que não tiveram o objeto eclipsante determinado. O catálogo do Kepler inclui 1982 objetos, enquanto o TESS conta com 7358 objetos de interesse. Como determinamos os parâmetros das estrelas observadas pelo J-PLUS e S-PLUS, em comum com o KIC e o TIC, podemos classificar alguns objetos de acordo com a descrição de Petigura *et al.* (2018).

Além de fornecerem os valores da profundidade de trânsito, os catálogos também apresentam o raio do planeta (R_p) e seu erro, em unidades de $R_\oplus$ (raios terrestres). Com isso, comparamos os valores de raio do planeta fornecidos pelos catálogos com os valores calculados pelos nossos modelos e fazemos uma classificação com base nos nossos raios.

Ao cruzarmos o J-PLUS DR3 com o KOI, encontramos que 7 estrelas tiveram seus parâmetros previstos pelos modelos restritos e 16 pelos nossos modelos menos restritos. Ao calcular o raio (R_p, tanto com os parâmetros calculados por RF quanto com os calcularmos por XGB) e realizarmos a classificação, nenhum dos objetos foi reclassificado. Portanto, não discutiremos individualmente cada caso. Com relação ao S-PLUS, o cruzamento não resultou em objetos em comum.

Com relação ao J-PLUS e TOI, o cruzamento resultou em 44 objetos em comum com e_mag $\leqslant$ 0,1 e 55 objetos com e_mag $\leqslant$ 0,2, e não houve uma mudança de classificação. Com relação ao S-PLUS, ao realizarmos o cruzamento com o TOI, obtivemos 168 objetos com e_mag $\leqslant$ 0,1 e 194 com e_mag $\leqslant$ 0,2. Entre eles, 7 objetos tiveram sua classificação

[37] https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_KOI_docs.html

alterada com base nos raios planetários calculados a partir dos resultados dos nossos modelos. A Tabela 4.44 e a Tabela 4.45 consistem nos valores calculados a partir dos parâmetros previstos pelos modelos restritos e menos restrito, respectivamente, para esses 7 objetos.

Restrito					
TOI ID	$\mathbf{R_\star^{RF}}$ **[$R_\odot$]**	$\mathbf{R_\star^{XGB}}$ **[$R_\odot$]**	$\mathbf{R_p^{RF}}$ **[$R_\oplus$]**	$\mathbf{R_p^{XGB}}$ **[$R_\oplus$]**	$\mathbf{R_p^{TOI}}$ **[$R_\oplus$]**
147.01	2,11 ± 0,13	2,16 ± 0,13	21,50 ± 1,28	21,97 ± 1,31	35,41 ± 4,85
360.01	1,60 ± 0,10	1,62 ± 0,10	12,63 ± 0,77	12,78 ± 0,78	24,31 ± 7,87
917.01	2,44 ± 0,15	2,52 ± 0,15	20,67 ± 1,23	21,30 ± 1,28	24,16 ± 2,01
1032.01	4,71 ± 0,28	4,82 ± 0,29	30,45 ± 1,81	31,17 ± 1,85	14,77 ± 0,75
2442.01	0,97 ± 0,06	0,99 ± 0,06	38,43 ± 2,29	39,42 ± 2,35	10,23 ± 0,52
3083.01	1,60 ± 0,10	1,65 ± 0,10	14,47 ± 0,87	14,84 ± 0,89	NA

Tabela 4.44: Raio das estrelas em comum entre S-PLUS e TOI, com parâmetros previstos pelos modelos restritos. $R_\star^{RF}$ e $R_\star^{XGB}$ são os raios das estrelas calculados a partir de parâmetros previstos por *Random Forest* e *XGBoost* em raios solares, respectivamente. R_p^{RF} e R_p^{XGB} são os raios dos planetas calculados a partir de parâmetros previstos por *Random Forest* e *XGBoost* em raios terrestres, respectivamente, e R_p^{TOI} são os raios dos planetas fornecidos pelo TOI, em raios terrestres.

Menos Restrito					
TOI ID	$\mathbf{R_\star^{RF}}$ **[$R_\odot$]**	$\mathbf{R_\star^{XGB}}$ **[$R_\odot$]**	$\mathbf{R_p^{RF}}$ **[$R_\oplus$]**	$\mathbf{R_p^{XGB}}$ **[$R_\oplus$]**	$\mathbf{R_p^{TOI}}$ **[$R_\oplus$]**
2664.01	0,99 ± 0,06	1,02 ± 0,06	8,51 ± 0,51	8,82 ± 0,52	24,45 ± 3,13

Tabela 4.45: Raio da estrela em comum entre S-PLUS e TOI, com parâmetros previstos pelo modelo menos restrito. $R\star^{RF}$ e $R\star^{XGB}$ são os raios da estrela calculados a partir de parâmetros previstos por *Random Forest* e *XGBoost*, em raios solares, respectivamente. R_p^{RF} e R_p^{XGB} são os raios do planeta calculados a partir de parâmetros previstos por *Random Forest* e *XGBoost*, em raios terrestres, respectivamente. R_p^{TOI} é o raio do planeta fornecido pelo TOI, em raios terrestres.

Os valores de profundidade de trânsito observado (Q), para esses objetos, utilizados nos cálculos dos raios são:

TOI ID	Q [ppm]	Q_err [ppm]
147.01	8643,1	89,25
360.01	5180	1684,47
917.01	5967,22	454,16
1032.01	3491,82	95,25
2442.01	131000	0,70
2664.01	6193	10789,20
3083.01	6780	2912,70

Tabela 4.46: Profundidade de trânsito fornecido pelo TOI para os objetos em campo comum S-PLUS + KOI e sua respectiva incerteza em ppm.

Conforme os raios planetários apresentados nas Tabelas 4.44 e 4.45 podemos classificar os objetos da seguinte forma:

- TOI 147.01: A classificação inicial deste objeto com base no raio do planeta do TOI, que é superior a 24 $R_\oplus$ (35,41 $\pm$ 4,85 $R_\oplus$), indicando que ele seria parte de um sistema binário eclipsante. Ao calcular o raio do planeta a partir dos modelos de RF (21,50 $\pm$ 1,28 $R_\oplus$) e XGB (21,97 $\pm$ 1,31 $R_\oplus$), observamos que o valor obtido é inferior ao limite de 24 $R_\oplus$. Portanto, embora a classificação inicial sugira uma binária eclipsante, os cálculos do raio do objeto indica é um exoplaneta ou anã marrom.

- TOI 360.01: Com um raio do planeta fornecido pelo TOI de 24,31 $\pm$ 7,87 $R_\oplus$, o objeto é inicialmente classificado como exoplaneta devido ao valor ligeiramente acima de 24 $R_\oplus$ (dentro da margem de erro). No entanto, os cálculos com os modelos RF (12,63 $\pm$ 0,77 $R_\oplus$) e XGB (12,63 $\pm$ 0,76 $R_\oplus$) indicam um valor consideravelmente menor para o raio, confirmando sua classificação como exoplaneta.

- TOI 917.01: A classificação inicial, com o raio do TOI de 24,16 $\pm$ 2,01 $R_\oplus$, sugere que o objeto é como um exoplaneta, considerando a margem de erro. Os cálculos baseados nos modelos com RF (20,67 $\pm$ 1,23 $R_\oplus$) e com XGB (21,30 $\pm$ 1,28 $R_\oplus$) resultam em raios planetários menores, reforçando a classificação como exoplaneta ou anã marrom.

- TOI 1032.01: O raio fornecido pelo TOI (14,77 $\pm$ 0,75 $R_\oplus$) indica uma classificação inicial como exoplaneta. No entanto, os valores calculados com os modelos de RF (30,45 $\pm$ 1,81 $R_\oplus$) e de XGB (31,17 $\pm$ 1,85 $R_\oplus$) são consideravelmente maiores, superando o limite de 24 $R_\oplus$ e, portanto, o objeto é classificado como uma binária eclipsante.

- TOI 2442.01: A classificação inicial, com o raio fornecido pelo TOI de 10,23 $\pm$ 0,52 $R_\oplus$, sugere que o objeto seria um exoplaneta. No entanto, os raios planetários

calculados pelos modelos de RF (38,43 ± 2,29 $R_\oplus$) e de XGB (39,42 ± 2,35 $R_\oplus$) são muito maiores classificando este objeto como parte de sistema binário eclipsante.

- TOI 3083.01: O raio fornecido pelo TOI não está disponível ou não foi calculado, mas os raios calculados pelos modelos com RF (14,47 ± 0,87 $R_\oplus$) e com XGB (14,84 ± 0,89 $R_\oplus$) são condizentes com um exoplaneta ou anã marrom.

- TOI 2664.01: Foi o único objeto que aparece somente nos modelos menos restrito. Com o raio do TOI (24,45 ± 3,13 $R_\oplus$) já superior a 24 $R_\oplus$, a classificação inicial é de binária eclipsante. Com os nossos modelos com o RF, o raio calculado foi de de 8,51 ± 0,51 $R_\oplus$ e com XGB foi de 8,82 ± 0,52 $R_\oplus$, com isso podemos concluir que se trata de um exoplaneta ou anã marrom.

Vale ressaltar que para uma classificação mais precisa é necessário calcular a massa mínima do objeto orbitante, o que não foi feito pelos nossos modelos. As tabelas finais, com todos os parâmetros das estrelas e os parâmetros planetários de todos os objetos classificados e reclassificados encontram-se no Drive para consulta[38].

[38]S-PLUS + TOI: https://linktr.ee/final_tables.

Capítulo 5

Considerações Finais e Perspectivas Futuras

Os resultados obtidos ao longo deste trabalho demonstram que os levantamentos J-PLUS e S-PLUS, com seus sistemas de filtros fotométricos, são fontes de dados promissoras para a construção de modelos de aprendizado de máquina voltados para a determinação de parâmetros estelares, como temperatura efetiva (T_{ef}), gravidade superficial (log g) e metalicidade ([Fe/H]).

Em comparação com a literatura, foi possível observar uma redução significativa nos erros das previsões realizadas, principalmente ao utilizar modelos restritos baseados nos dados do LAMOST (Zhao *et al.*, 2012). Isso reafirma a qualidade deste levantamento, que, devido à sua vasta quantidade de dados e precisão nos parâmetros estelares, é uma das melhores referências disponíveis atualmente. Essa superioridade também foi corroborada por outros estudos, como os de Dong *et al.* (2018), Carvalho (2022) e Cordeiro da Silva (2023).

Os modelos de aprendizado de máquina empregados neste trabalho, RF e XGB, demonstraram ser altamente eficazes na previsão dos parâmetros de interesse de grandes amostras de dados, bem como na caracterização de estrelas hospedeiras de exoplanetas. A robustez e a qualidade dos modelos, aliada à aplicação de técnicas como o Monte Carlo, possibilitou a obtenção de resultados confiáveis, com erros significativamente menores quando comparados aos valores encontrados na literatura. Vale ressaltar que essa metodologia não substitui os métodos tradicionais para a determinação de parâmetros físicos. Na verdade, a metodologia com ML é uma ferramenta adicional para indicar futuros candidatos para estudos mais precisos com métodos tradicionais.

Além disso, os modelos apresentaram um ótimo desempenho tanto nos cenários restritos (e_mag $\leqslant$ 0,1) quanto nos menos restritos (e_mag $\leqslant$ 0,2). Para os objetos presentes em ambas as amostras, foi adotado o modelo treinado com o conjunto restrito, garantindo a melhor precisão possível. No entanto, para aqueles que estavam disponíveis apenas no cenário menos restrito, os modelos puderam ser utilizados sem prejuízos significativos,

uma vez que os valores previstos se mostraram altamente compatíveis e confiáveis.

Ao comparar as técnicas utilizadas, o RF destacou-se como a abordagem mais eficaz para prever os parâmetros estelares neste trabalho. Essa técnica apresentou, de maneira geral, menores médias globais dos erros, desvios padrão e medianas em relação ao XGB, para as estrelas presentes nos levantamentos do Kepler, TESS e HARPS em comum com os levantamentos J-PLUS e S-PLUS (vide a Seção 4.4). Mas, isso não invalida a utilização do XGB, visto que os seus resultados pós aplicação nos mostraram bem próximos aos alcançados pelo RF e bem confiáveis.

Nos dados do Kepler, para T_{ef}, a média dos erros globais com RF foi cerca de 15% menor no J-PLUS no cenário restrito e 26% menor no S-PLUS em comparação ao XGB. Em cenários menos restritos, com o RF apresentando erros 13% menores no J-PLUS e 31% menores no S-PLUS. Para log g, o RF teve um desempenho superior com médias de erros 25% menores em relação ao XGB nos cenários restritos utilizando J-PLUS. Com relação ao S-PLUS não houve diminuição, as métricas foram iguais. Nos cenários menos restritos, a vantagem permaneceu, com uma redução de cerca de 25% utilizando o J-PLUS e com o S-PLUS permaneceu a mesma. Em relação à [Fe/H], o RF obteve médias de erros até 22% menores no J-PLUS e 17% menores no S-PLUS nos cenários restritos. Nos cenários menos restritos, a diferença caiu para 11% utilizando o J-PLUS e 17% utilizando o S-PLUS, reforçando eficiência do RF.

Para as estrelas presentes no campo do levantamento TESS, os ganhos do RF foram ainda mais evidentes em cenários restritos. Em T_{ef}, a técnica apresentou erros 16% menores no J-PLUS e 25% menores no S-PLUS comparados ao XGB. Nos cenários menos restritos, a diferença foi reduzida, com melhorias de cerca de 14% no J-PLUS e 20% no S-PLUS. No caso de log g, a redução nos erros com RF foi de aproximadamente 33% no J-PLUS e S-PLUS em cenários restritos. Nos cenários menos restritos, para o J-PLUS a diferença foi mantida em torno de 33% enquanto para o S-PLUS as métricas foram as mesmas. Para [Fe/H], o RF demonstrou uma boa vantagem, com médias de erros cerca de 14% menores no J-PLUS e no S-PLUS em cenários restritos. Nos cenários menos restritos, os valores ficaram em torno de 13% e 14%, respectivamente.

Por fim, para o levantamento HARPS, a diferença entre as técnicas foi particularmente acentuada devido à qualidade e quantidade reduzida de dados. Para o J-PLUS não houve objetos em comum entre os levantamentos. Para T_{ef}, utilizando o S-PLUS, o RF apresentou erros 42% menores. Em cenários menos restritos, as diferenças diminuíram para cerca de 33%. Em log g, o RF e o XGB obtiveram os mesmo valores de média e desvio padrão, sendo melhor o RF na mediana comparada com o XGB. Para [Fe/H], não houve redução nas incertezas entre as duas técnicas em cenários restritos. Nos cenários menos restritos, a diferença ficou em torno de 50%.

Apesar dos avanços alcançados, algumas limitações computacionais restringiram o número de iterações com o método de Monte Carlo, particularmente no caso dos dados do

TESS. A utilização de mais iterações no futuro pode aprimorar a precisão dos resultados obtidos. Assim, planejamos o uso de uma infraestrutura computacional mais robusta para análises futuras.

Como outro passo futuro, pretendemos aplicar a mesma metodologia nos dados do *Javalambre-Physics of the Accelerating Universe Astrophysical Survey* (J-PAS; Benitez *et al.* (2014)). O levantamento conta com 56 filtros ópticos, o que permitirá aos algoritmos a aprendizagem de diferentes padrões que não foram aprendidos usando os filtros J-PLUS/S-PLUS, trazendo assim, maior precisão e confiança.

Outro aspecto relevante a ser destacado é que os resultados obtidos neste trabalho não se limitam apenas a estudos sobre estrelas hospedeiras de exoplanetas. Eles também podem ser aplicados a diversas outras áreas, como a evolução galáctica, a classificação espectral de estrelas, a identificação de estrelas pobres ou ricas em metais e entre outros, desde que tenham dados para treinar os modelos para prever tais parâmetros. Além disso, os modelos desenvolvidos possuem grande potencial de flexibilidade e adaptabilidade, podendo ser aprimorados e configurados para atender às necessidades específicas de cada estudo. Esses atributos reforçam sua importância como ferramentas valiosas para futuras aplicações em diferentes contextos astrofísicos.

Um artigo apresentando os resultados dos modelos desenvolvidos a partir das amostras do levantamento S-PLUS aplicados aos objetos em comum com o TESS e com o espectrógrafo HARPS e um artigo com os modelos desenvolvidos com o J-PLUS aplicados ao TESS estão em preparação, buscando aprofundar a análise e validar os resultados obtidos.

Referências Bibliográficas

Alonso-García, J., Mateo, M., Sen, B., et al., 2012, "Uncloaking globular clusters in the inner galaxy", *The Astronomical Journal*, v. 143, n. 3, pp. 70.

Bailer-Jones, C., Rybizki, J., Fouesneau, M., et al., 2021, "Estimating distances from parallaxes. V. Geometric and photogeometric distances to 1.47 billion stars in Gaia Early Data Release 3", *The Astronomical Journal*, v. 161, n. 3, pp. 147.

Barbieri, M., 2023. "ESO/HARPS Radial Velocities Catalog". Disponível em: <https://arxiv.org/abs/2312.06586>.

Baron, D., 2019, "Machine Learning in Astronomy: a practical overview", *arXiv e-prints*, art. arXiv:1904.07248. doi: 10.48550/arXiv.1904.07248.

Benitez, N., Dupke, R., Moles, M., et al., 2014, "J-PAS: the Javalambre-physics of the accelerated universe astrophysical survey", *arXiv preprint arXiv:1403.5237*.

Bergstra, J., Bengio, Y., 2012, "Random Search for Hyper-Parameter Optimization", *Journal of Machine Learning Research*, v. 13, n. 10, pp. 281–305.

Borucki, W. J., 2016, "KEPLER Mission: development and overview", *Reports on Progress in Physics*, v. 79, n. 3, pp. 036901.

Breiman, L., 2001, "Random forests", *Machine learning*, v. 45, pp. 5–32.

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*, v. 486, n. 7403, pp. 375–377.

Campbell, B., Walker, G. A., Yang, S., 1988, "A search for substellar companions to solar-type stars", *The Astrophysical Journal*, v. 331, pp. 902–921.

Carroll, B. W., Ostlie, D. A., 2017, *An introduction to modern astrophysics*. Cambridge University Press.

Carvalho, L. M., 2022, *Caracterização de estrelas fracas da missão Kepler com base em dados do J-PLUS*. Tese de Mestrado, Observatório Nacional, ON.

Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al., 2019, "J-PLUS: The javalambre photometric local universe survey", *Astronomy & Astrophysics*, v. 622, pp. A176.

Chen, T., Guestrin, C., 2016, "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, a.

Collaboration, G., others, 2016, "The gaia mission", *arXiv preprint arXiv:1609.04153*.

Cordeiro da Silva, V. E., 2023, *Identificação de Subanãs Quentes em Levantamentos Astronômicos*. Tese de Mestrado, Observatório Nacional, ON.

Curto, G. L., Pasquini, L., Manescau, A., et al., 2012, "Astronomical spectrograph calibration at the exo-earth detection limit", *Messenger*, v. 149, pp. 2–6.

Dong, S., Xie, J.-W., Zhou, J.-L., et al., 2018, "LAMOST telescope reveals that Neptunian cousins of hot Jupiters are mostly single offspring of stars that are rich in heavy elements", *Proceedings of the National Academy of Sciences*, v. 115, n. 2, pp. 266–271.

Friedman, J. H., 2001, "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, pp. 1189–1232.

Friedman, J. H., 2002, "Stochastic gradient boosting", *Computational statistics & data analysis*, v. 38, n. 4, pp. 367–378.

Fukugita, M., Ichikawa, T., Gunn, J. E., et al., 1996, "The Sloan Digital Sky Survey Photometric System", Astron. J., v. 111 (abr.), pp. 1748. doi: 10.1086/117915.

Galarza, C. A., Daflon, S., Placco, V. M., et al., 2022, "J-PLUS: Searching for very metal-poor star candidates using the SPEEM pipeline", *Astronomy & Astrophysics*, v. 657, pp. A35.

Gaudi, B. S., 2010, "Exoplanetary Microlensing", *arXiv e-prints*, art. arXiv:1002.0332. doi: 10.48550/arXiv.1002.0332.

Ghezzi, L., Martinez, C. F., Wilson, R. F., et al., 2021, "A Spectroscopic Analysis of the California-Kepler Survey Sample. II. Correlations of Stellar Metallicities with Planetary Architectures", *The Astrophysical Journal*, v. 920, n. 1, pp. 19.

Guerrero, N. M., Seager, S., Huang, C. X., et al., 2021, "The TESS objects of interest catalog from the TESS prime mission", *The Astrophysical Journal Supplement Series*, v. 254, n. 2, pp. 39.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009, *The Elements of Statistical Learning*. 2nd ed. , Springer. ISBN: 978-0-387-84857-0. Archived from the original on 2009-11-10.

Herpich, F. R., Almeida-Fernandes, F., Oliveira Schwarz, G. B., et al., 2024, "The Fourth S-PLUS Data Release: 12-filter photometry covering ∼ 3000 square degrees in the southern hemisphere", *arXiv e-prints*, art. arXiv:2407.20701. doi: 10.48550/arXiv.2407.20701.

Holman, M. J., Fabrycky, D. C., Ragozzine, D., et al., 2010, "Kepler-9: A System of Multiple Planets Transiting a Sun-Like Star, Confirmed by Timing Variations", *Science*, v. 330, n. 6000 (out.), pp. 51. doi: 10.1126/science.1195778.

Jordi, C., Gebran, M., Carrasco, J., et al., 2010, "Gaia broad band photometry", *Astronomy & Astrophysics*, v. 523, pp. A48.

Kirk, B., Conroy, K., Prša, A., et al., 2016, "Kepler eclipsing binary stars. VII. The catalog of eclipsing binaries found in the entire Kepler data set", *The Astronomical Journal*, v. 151, n. 3, pp. 68.

Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al., 2011, "A closely packed system of low-mass, low-density planets transiting Kepler-11", Nature, v. 470, n. 7332 (fev.), pp. 53–58. doi: 10.1038/nature09760.

Lovis, C., Fischer, D., 2010, "Radial Velocity Techniques for Exoplanets". In: Seager, S. (Ed.), *Exoplanets*, The University of Arizona Press Tucson, pp. 27–53.

Majewski, S., 2016, "The Apache Point Observatory Galactic Evolution Experiment (APOGEE) and its successor, APOGEE-2", *Astronomische Nachrichten*, v. 337, n. 8-9, pp. 863–870.

Mayor, M., Queloz, D., 1995, "A Jupiter-mass companion to a solar-type star", *Nature*, v. 378, pp. 355–359. doi: 10.1038/378355a0.

Meadows, V. S., Barnes, R. K., 2018, "Factors Affecting Exoplanet Habitability". In: Deeg, H., Belmonte, J. A. (Eds.), *Handbook of Exoplanets*, Springer, Cham. doi: 10.1007/978-3-319-55333-7_57. Disponível em: <https://doi.org/10.1007/978-3-319-55333-7_57>.

Meidem, Í., 2022, *Construção de uma sequência de ensino e aprendizagem com design-based research para o ensino de astronomia: foco, objetivos e compreensão do problema*. Trabalho de conclusão de curso (graduação em física licenciatura), Instituto de Física e Quimica (IFQ), Universidade Federal de Itajubá (UNIFEI).

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al., 2019, "The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters", Mon. Not, Roy. Astron. Soc., v. 489, n. 1 (out.), pp. 241–267. doi: 10.1093/mnras/stz1985.

Molina-Jorquera, F., Damke, G., Fernández-Olivares, D., et al., 2024, "Determination of metallicities of red giant stars using machine learning techniques applied to the narrow and broadband photometry of the S-PLUS survey", *Astronomy & Astrophysics*, v. 691, pp. A144.

Mulders, G. D., Pascucci, I., Apai, D., et al., 2016, "A super-solar metallicity for stars with hot rocky exoplanets", *The Astronomical Journal*, v. 152, n. 6, pp. 187.

Murray, C. D., Correia, A. C., 2010, "Keplerian orbits and dynamics of exoplanets". In: Seager, S. (Ed.), *Exoplanets*, The University of Arizona Press Tucson, pp. 15–23.

NASA Science Mission, 2017. "Kepler Mission". https://science.nasa.gov/mission/kepler/. Accesso em 20 nov. 2024.

Nedjati-Gilani, G. L., Schneider, T., Hall, M. G., et al., 2017, "Machine learning based compartment models with permeability for white matter microstructure imaging", *NeuroImage*, v. 150, pp. 119–135. doi: 10.1016/j.neuroimage.2017.02.013. Disponível em: <https://doi.org/10.1016/j.neuroimage.2017.02.013>.

Nobel Prize Outreach AB, 2019. "The Nobel Prize in Physics 2019". Recuperado de https://www.nobelprize.org/prizes/physics/2019/summary/ em 27 dez. 2023.

Nogueira, P. H. S. d. S. d. P., 2020, *Detecção de exoplanetas ao redor de estrelas fracas observadas pela missão Kepler.* Tese de Mestrado, Observatório Nacional.

Paegert, M., Stassun, K. G., Collins, K. A., et al., 2021, "TESS Input Catalog versions 8.1 and 8.2: Phantoms in the 8.0 Catalog and How to Handle Them", *arXiv e-prints*, art. arXiv:2108.04778. doi: 10.48550/arXiv.2108.04778.

Perryman, M., 2000, "Extra-solar planets", *Rep. Prog. Phys*, v. 63, pp. 1209–1272.

Perryman, M., 2018, *The exoplanet handbook.* a, Cambridge university press.

Petigura, E. A., Marcy, G. W., Winn, J. N., et al., 2018, "The California-Kepler survey. IV. Metal-rich stars host a greater diversity of planets", *The Astronomical Journal*, v. 155, n. 2, pp. 89.

Quirrenbach, A., 2010, "Astrometric Detection and Characterization of Exoplanets", *Exoplanets*, pp. 157–174.

Ricker, G. R., 2015, "The Transiting Exoplanet Survey Satellite (TESS): Discovering New Earths and Super-Earths in the Solar Neighborhood". In: *AAS/Division for Extreme Solar Systems Abstracts*, v. 47, *AAS/Division for Extreme Solar Systems Abstracts*, p. 503.01, dez.

Schlegel, D. J., Finkbeiner, D. P., Davis, M., 1998, "Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds", *The Astrophysical Journal*, v. 500, n. 2, pp. 525.

Seager, S., 2013, "Exoplanet habitability", *Science*, v. 340, n. 6132, pp. 577–581.

Soto, D. H., 2020, *Detección y caracterización de exoplanetas con telescopios espaciales*. Tese de Doutorado, Universidad de La Laguna.

Torres, G., Andersen, J., Giménez, A., 2010, "Accurate masses and radii of normal stars: modern results and applications", *The Astronomy and Astrophysics Review*, v. 18, n. 1, pp. 67–126.

Traub, W. A., Oppenheimer, B. R., 2010, *Direct imaging of exoplanets*. University of Arizona Press, Tucson.

Varela, J., Cristóbal-Hornillos, D., Cenarro, J., et al., 2014, "Statistical Challenges in the Photometric Calibration for 21st Century Cosmology: The J-PAS case", *Proceedings of the International Astronomical Union*, v. 10, n. S306, pp. 359–361.

von Marttens, R., Casarini, L., Napolitano, N. R., et al., 2022, "Inferring galaxy dark halo properties from visible matter with machine learning", *Monthly Notices of the Royal Astronomical Society*, v. 516, n. 3, pp. 3924–3943.

Válio, A., 2009, "Procuram-se planetas", *Ciência e Cultura*, v. 61, n. 4.

Warren, J., 2004, "Ancient Atomists on the Plurality of Worlds", *The Classical Quarterly*, v. 54, pp. 354–365. doi: 10.1093/clquaj/.

Whitten, D., Placco, V., Beers, T., et al., 2019, "J-PLUS: Identification of low-metallicity stars with artificial neural networks using SPHINX", *Astronomy & Astrophysics*, v. 622, pp. A182.

Wilson, R. F., Teske, J., Majewski, S. R., et al., 2018, "Elemental abundances of kepler objects of interest in APOGEE. I. Two distinct orbital period regimes inferred from host star iron abundances", *The Astronomical Journal*, v. 155, n. 2, pp. 68.

Wilson, T. G., Goffo, E., Alibert, Y., et al., 2022, "A pair of sub-Neptunes transiting the bright K-dwarf TOI-1064 characterized with CHEOPS", *Monthly Notices of the Royal Astronomical Society*, v. 511, n. 1, pp. 1043–1071.

Wolszczan, A., Frail, D., 1992, "A Planetary System Around the Millisecond Pulsar PSR1257+12", *Nature*, v. 355 (Feb), pp. 145–147.

Wolszczan, A., 1994, "Confirmation of Earth-mass planets orbiting the millisecond pulsar PSR B1257+ 12", *Science*, v. 264, n. 5158, pp. 538–542.

Yan, H., Li, H., Wang, S., et al., 2022, "Overview of the LAMOST survey in the first decade", *The Innovation*, v. 3, n. 2.

Yang, L., Yuan, H., Xiang, M., et al., 2022, "J-PLUS: Stellar parameters, C, N, Mg, Ca, and [α/Fe] abundances for two million stars from DR1", *Astronomy & Astrophysics*, v. 659, pp. A181. doi: 10.1051/0004-6361/202142724.

Yang, L., Shami, A., 2020, "On hyperparameter optimization of machine learning algorithms: Theory and practice", *Neurocomputing*, v. 415, pp. 295–316.

Yao, X., Liu, Y., 2013, "Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques: Machine Learning". In: *Springer Science + Business Media*, Springer, Boston, MA. ISBN: 978-1-4614-6940-7. doi: 10.1007/978-1-4614-6940-7_17.

Yuan, H., Yang, L., Cruz, P., et al., 2023, "The miniJPAS survey: stellar atmospheric parameters from 56 optical filters", *Monthly Notices of the Royal Astronomical Society*, v. 518, n. 2, pp. 2018–2033.

Zhao, G., Zhao, Y.-H., Chu, Y.-Q., et al., 2012, "LAMOST spectral survey—An overview", *Research in Astronomy and Astrophysics*, v. 12, n. 7, pp. 723.

Apêndice A

Desempenho dos Modelos Treinados com *Random Forest*

A.1 *Random Forest* na Previsão da Temperatura Efetiva

A.1.1 J-PLUS

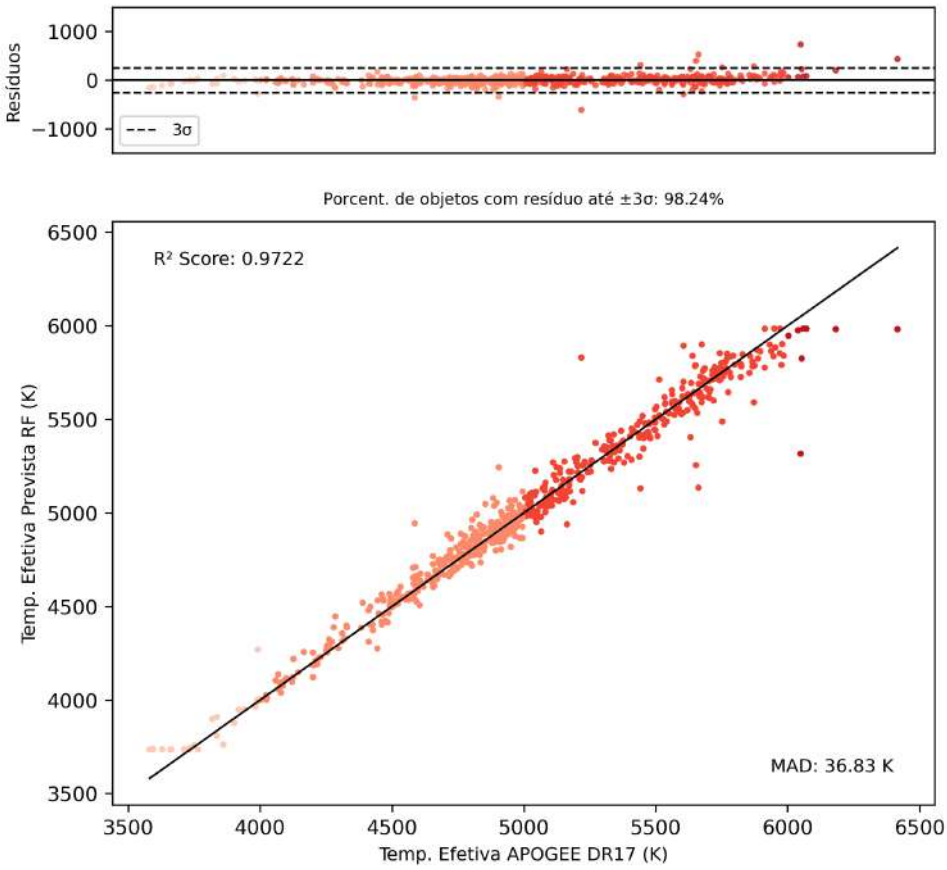

Figura A.1: Resultados do modelo `jplusA_01_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

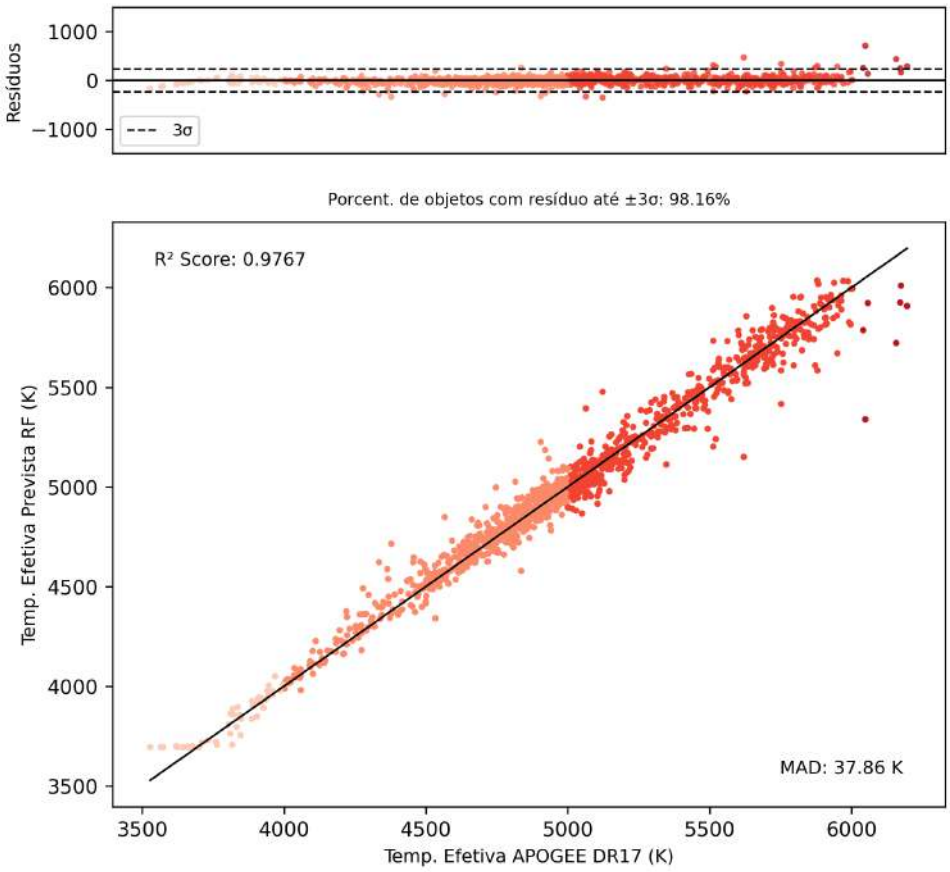

Figura A.2: Resultados do modelo `jplusA_02_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

A.1.2 S-PLUS

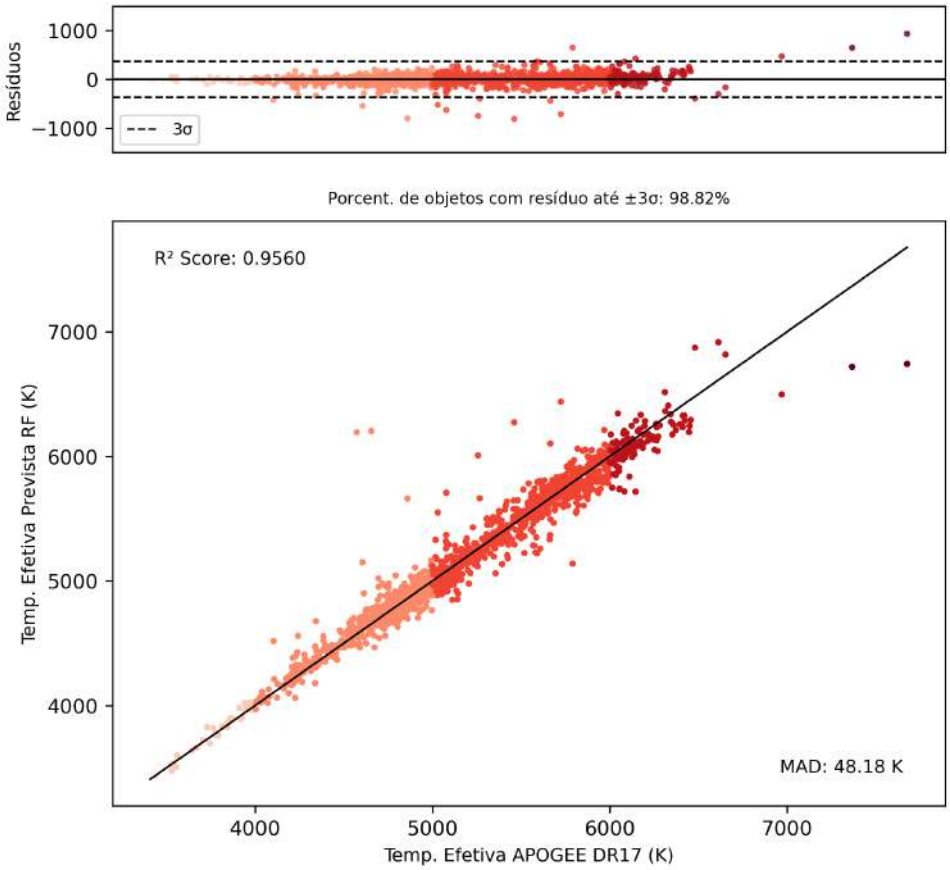

Figura A.3: Resultados do modelo `splusA_01_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

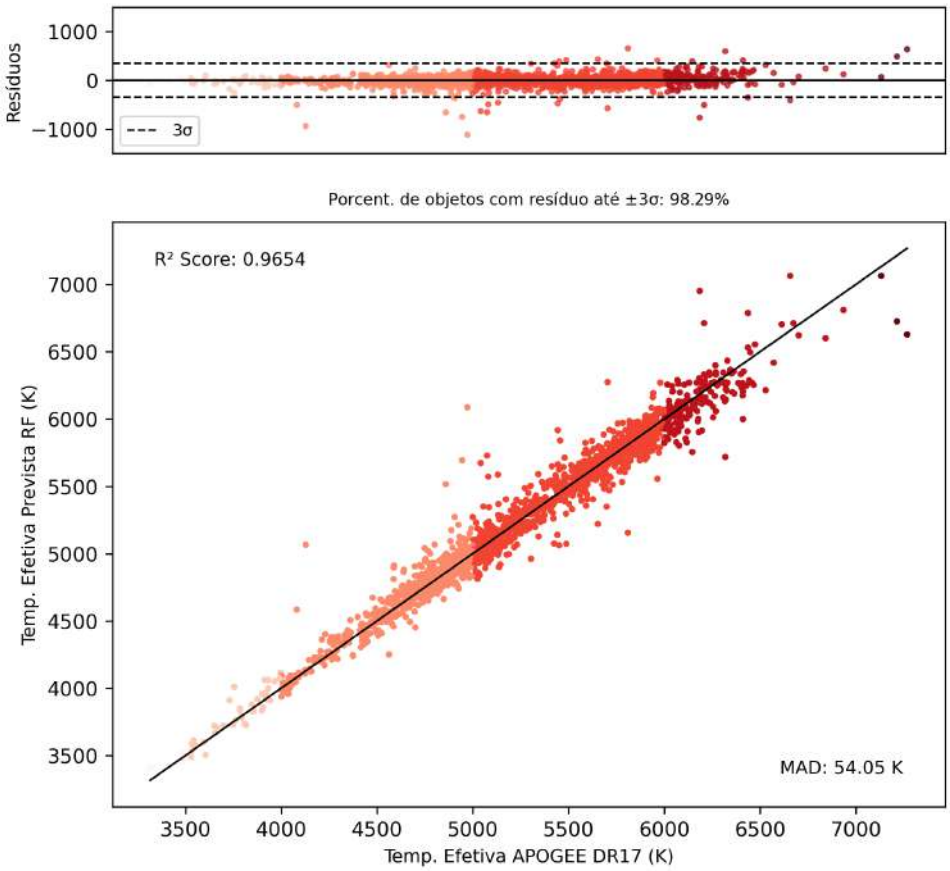

Figura A.4: Resultados do modelo `splusA_02_RF_teff` para a previsão de T_{ef} utilizando a técnica de *Random Forest* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

A.2 *Random Forest* na Previsão da Gravidade Superficial

A.2.1 J-PLUS

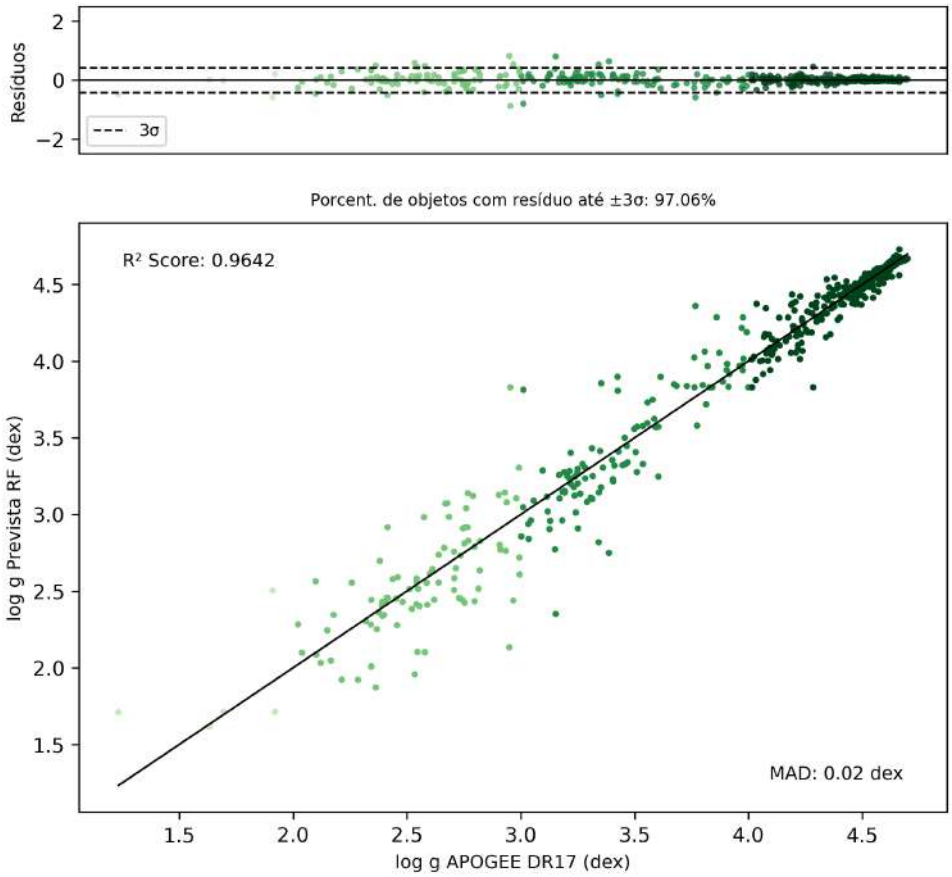

Figura A.5: Resultados do modelo `jplusA_01_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

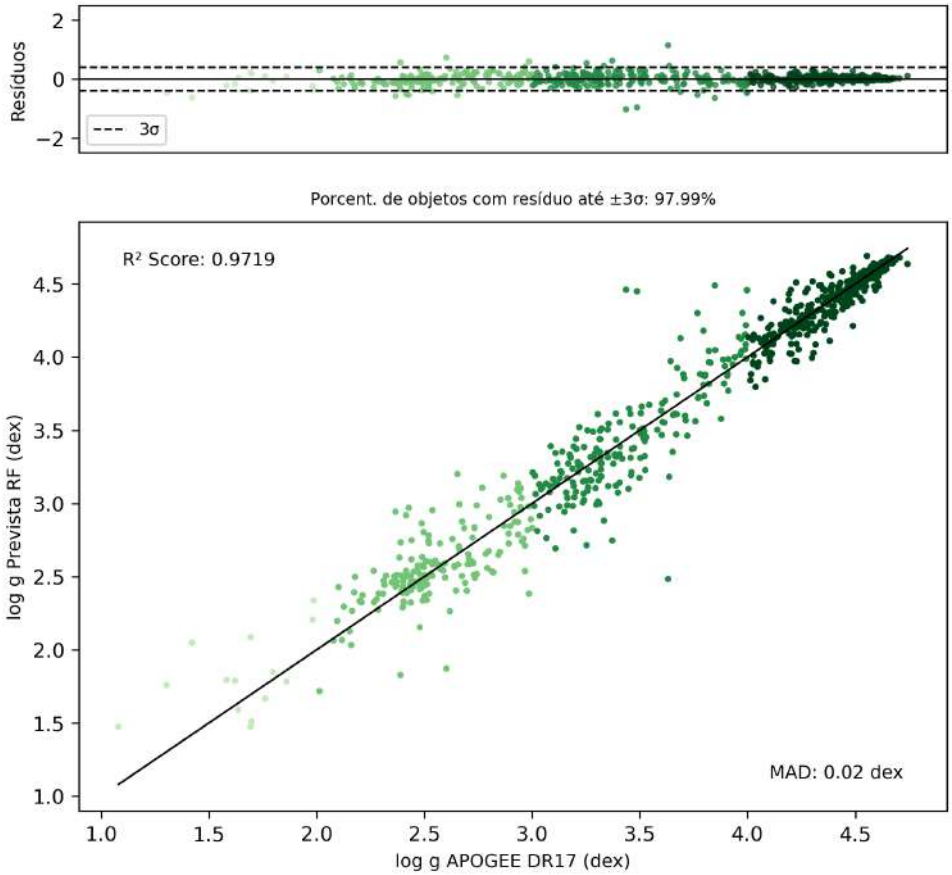

Figura A.6: Resultados do modelo `jplusA_02_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

A.2.2 S-PLUS

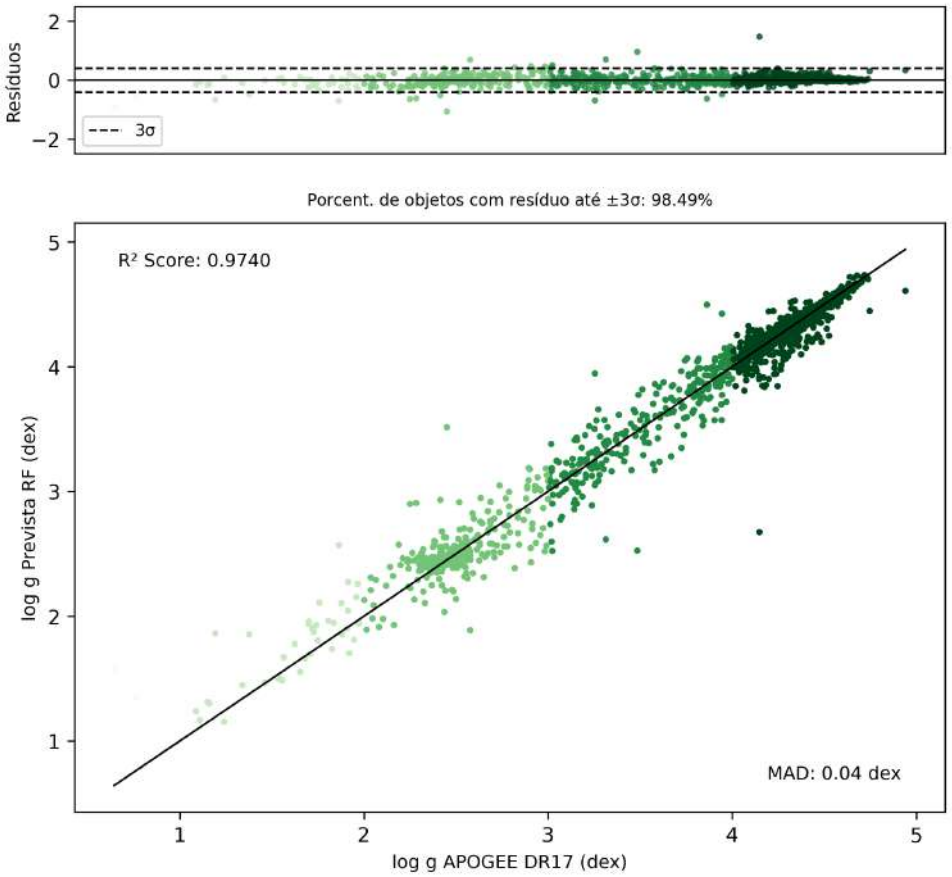

Figura A.7: Resultados do modelo `jplusA_01_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

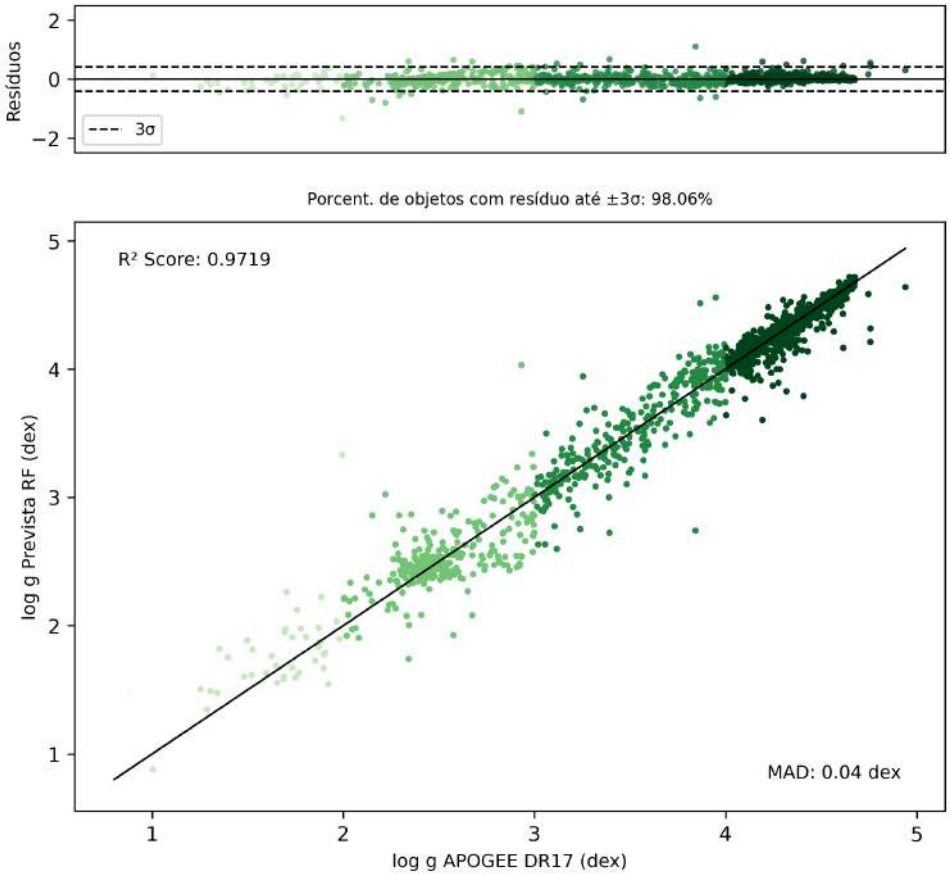

Figura A.8: Resultados do modelo `splusA_02_RF_logg` para a previsão de log g utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

A.3 *Random Forest* na Previsão da Metalicidade

A.3.1 J-PLUS

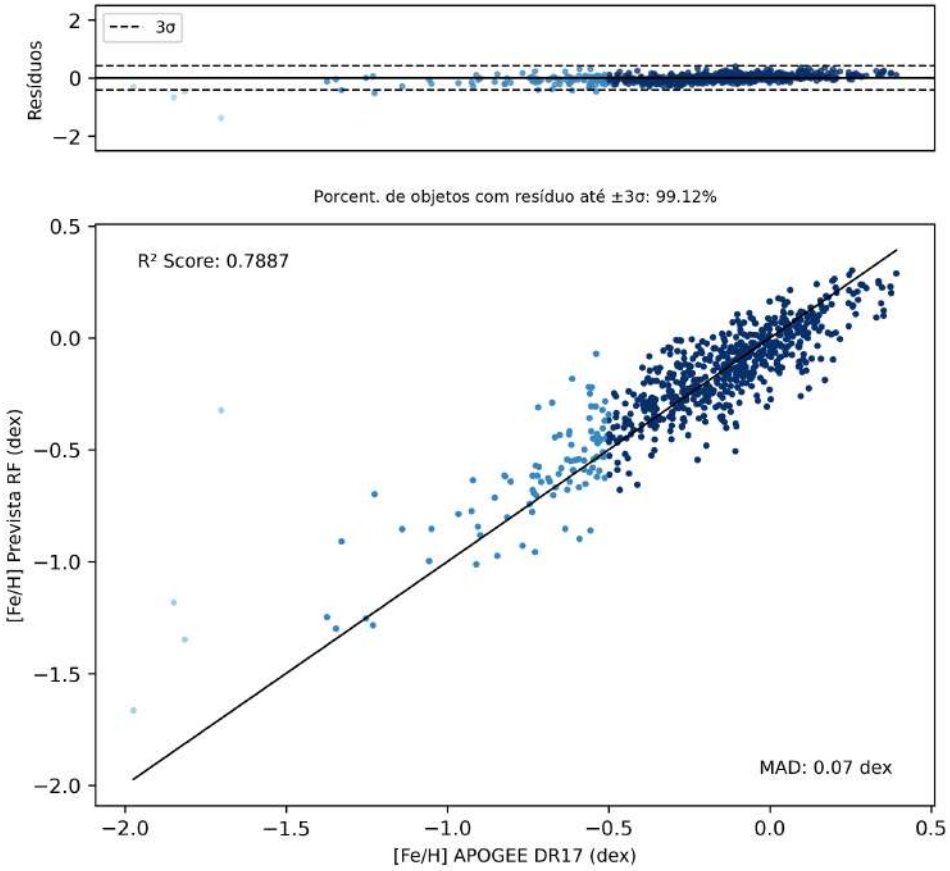

Figura A.9: Resultados do modelo `jplusA_01_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

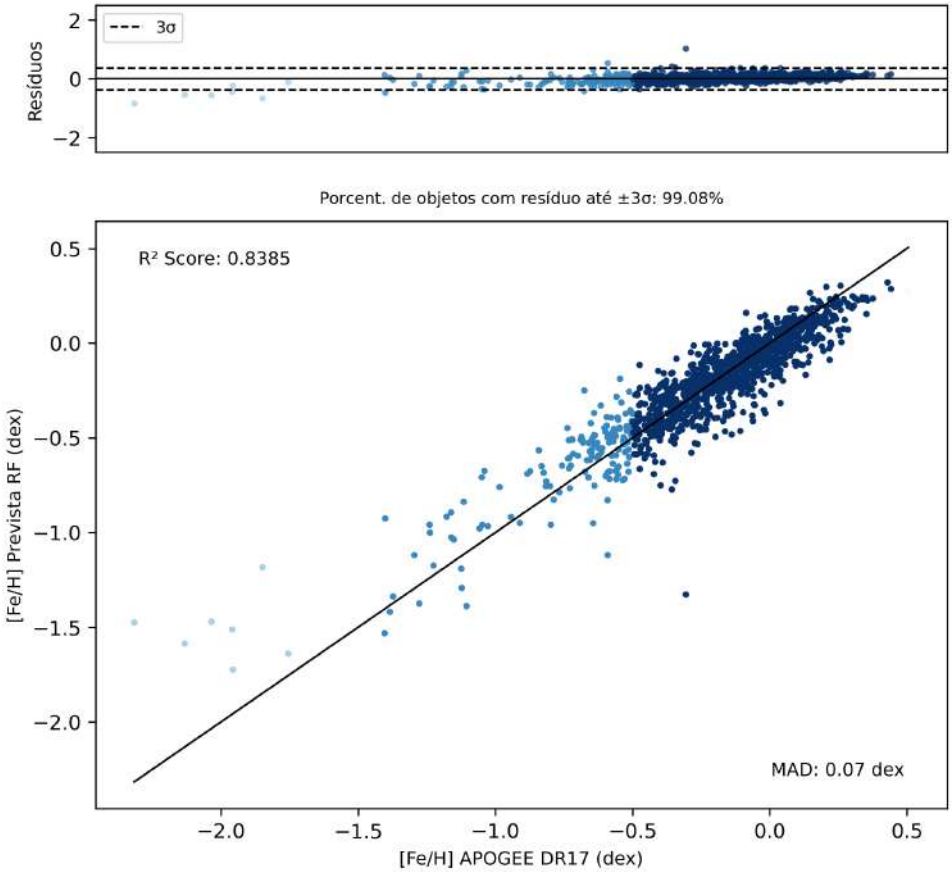

Figura A.10: Resultados do modelo `jplusA_02_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

A.3.2 S-PLUS

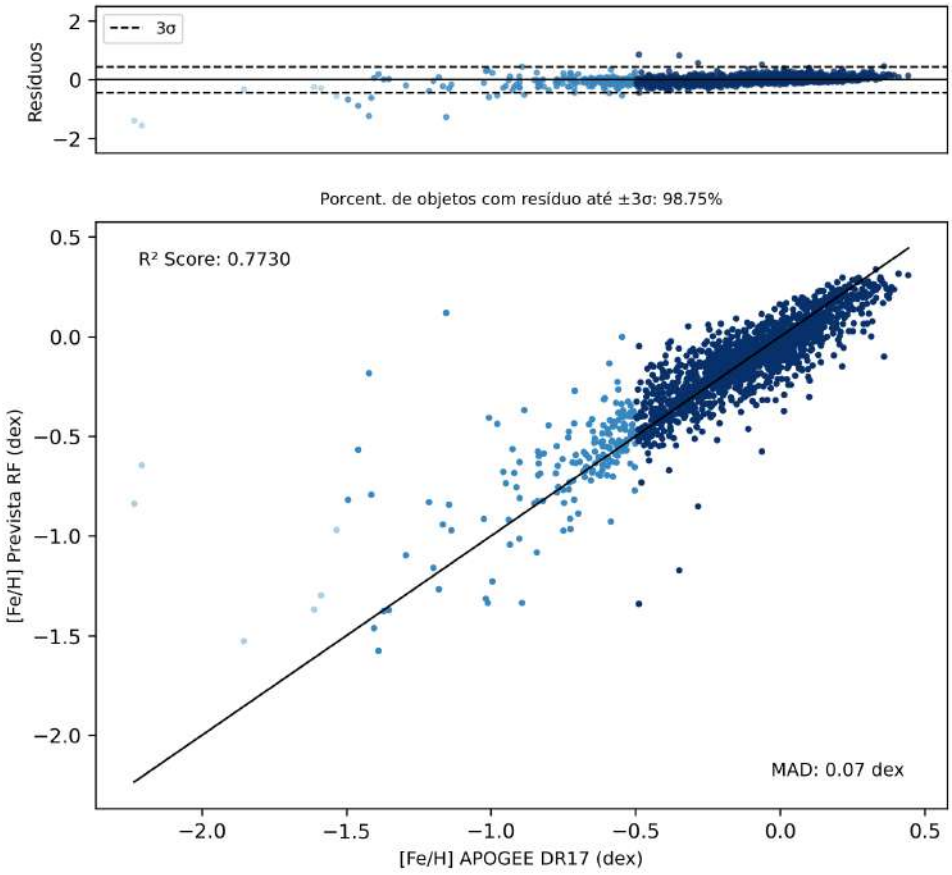

Figura A.11: Resultados do modelo `splusA_01_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Forest* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

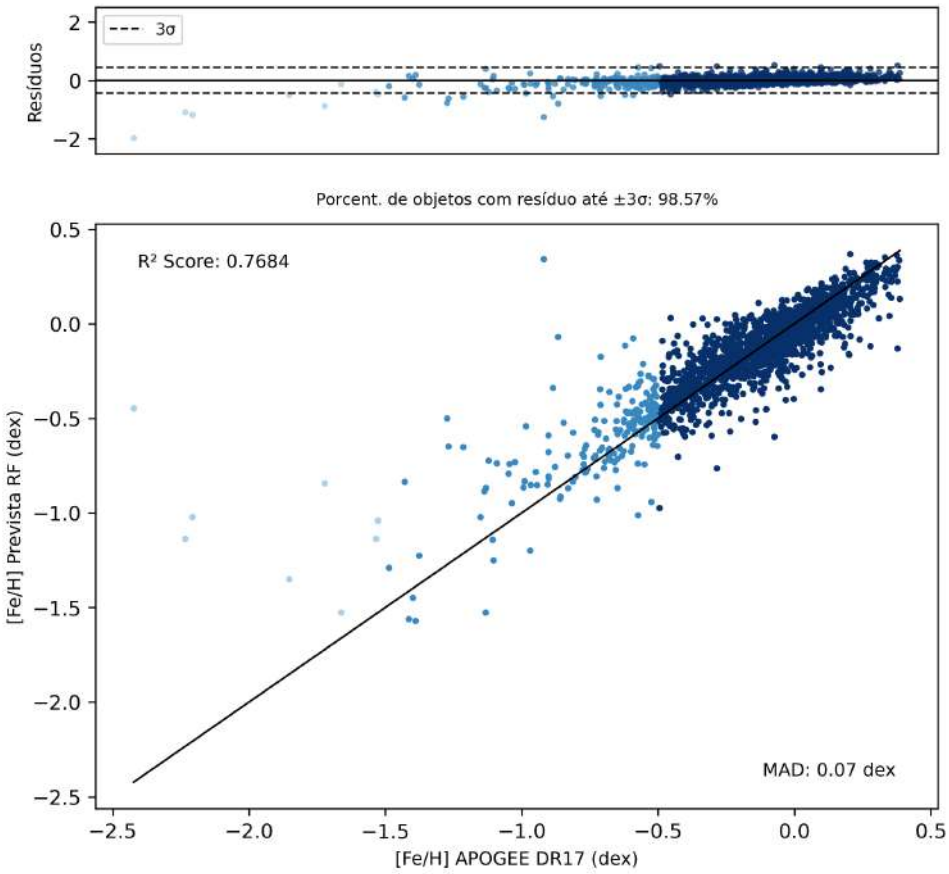

Figura A.12: Resultados do modelo `splusA_02_RF_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Fores* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

Apêndice B

Desempenho dos Modelos Treinandos com *XGBoost*

B.1 *XGBoost* na Previsão da Temperatura Efetiva

B.1.1 J-PLUS

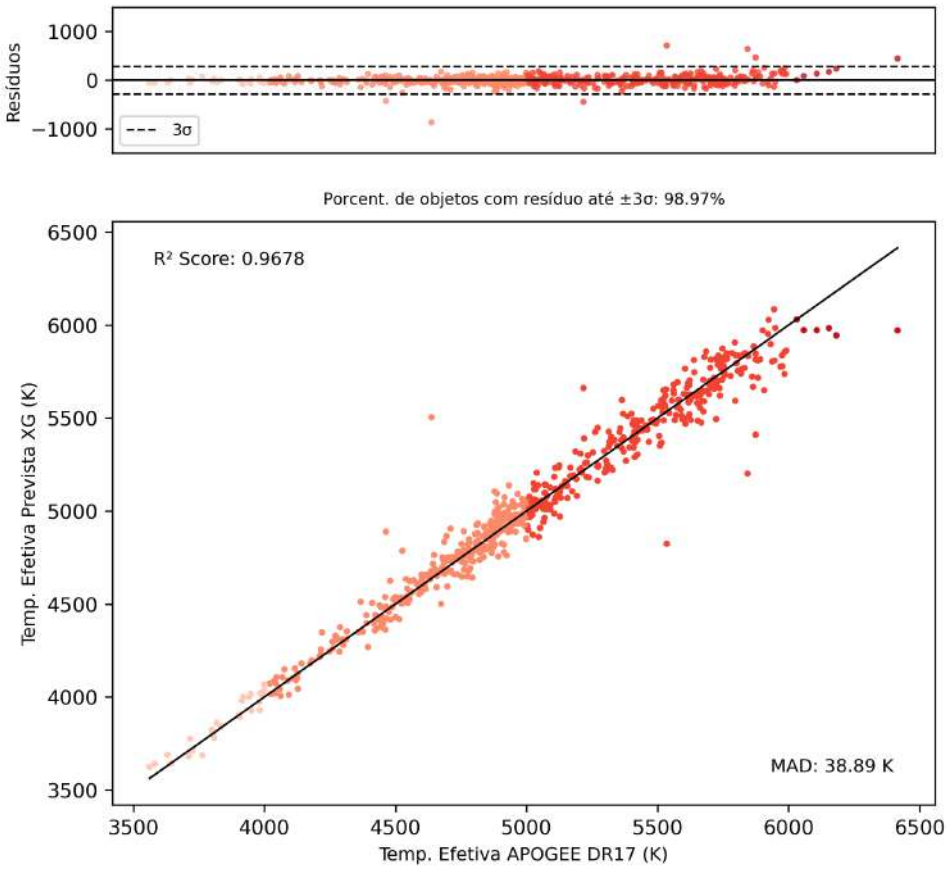

Figura B.1: Resultados do modelo `jplusA_01_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

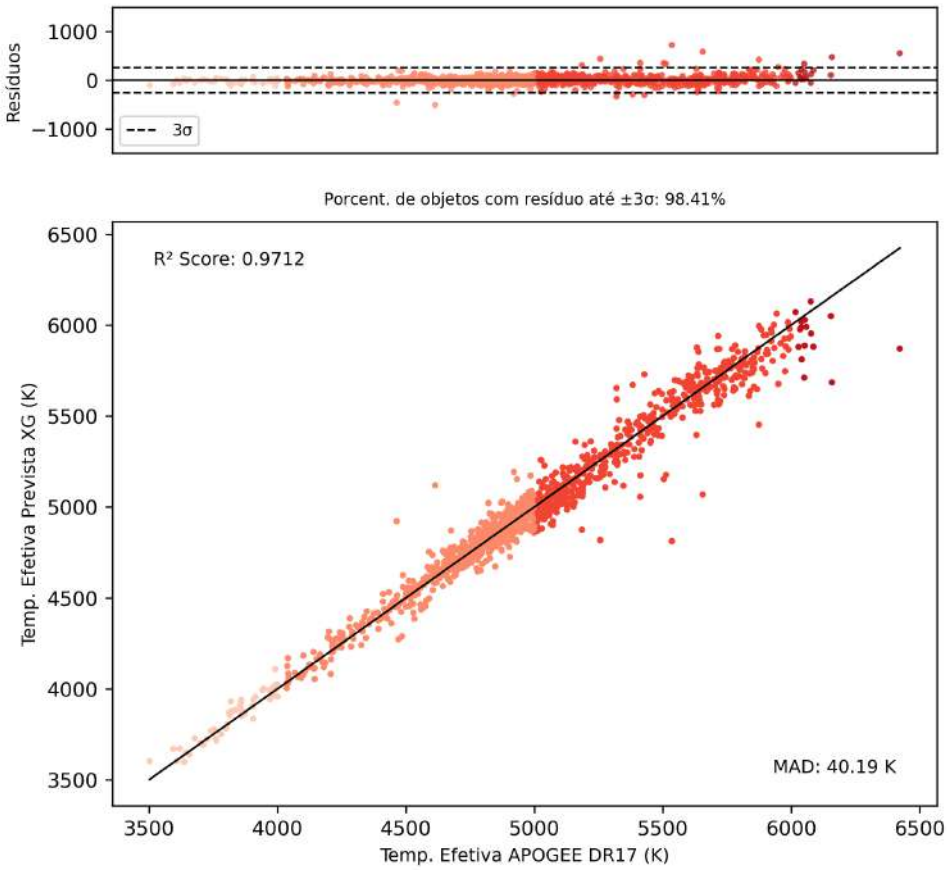

Figura B.2: Resultados do modelo `jplusA_02_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

B.1.2 S-PLUS

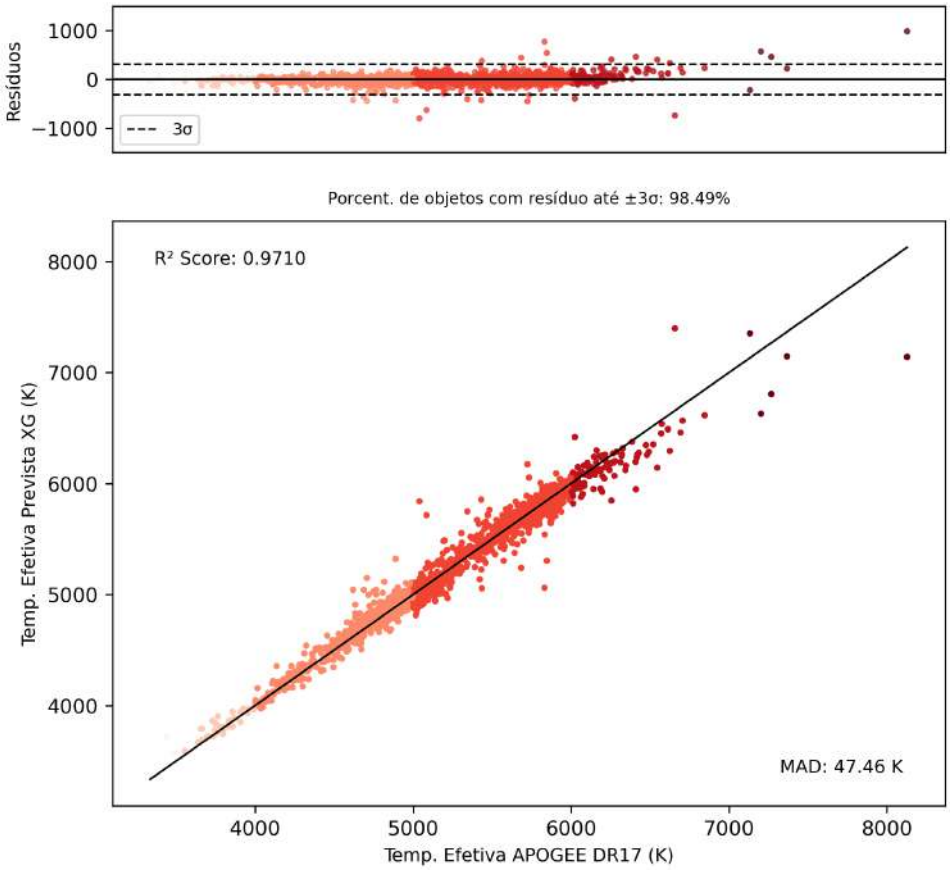

Figura B.3: Resultados do modelo `splusA_01_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

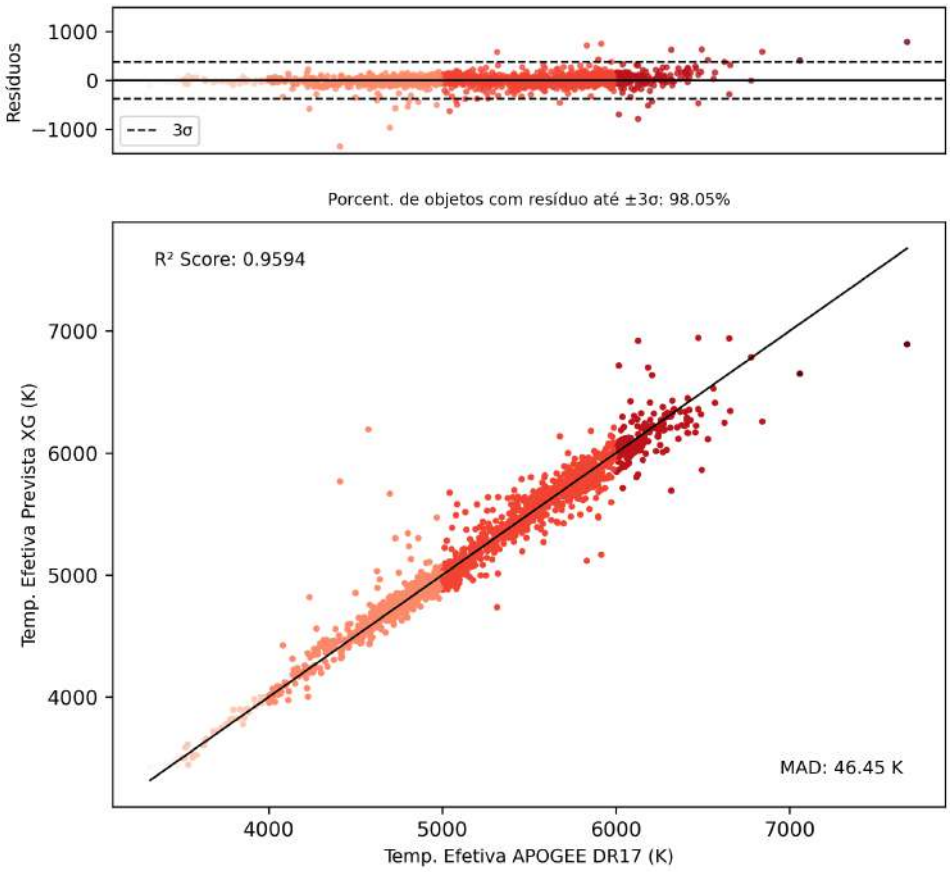

Figura B.4: Resultados do modelo `splusA_02_XGB_teff` para a previsão de T_{ef} utilizando a técnica de *XGBoost* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

B.2 *XGBoost* na Previsão da Gravidade Superficial

B.2.1 J-PLUS

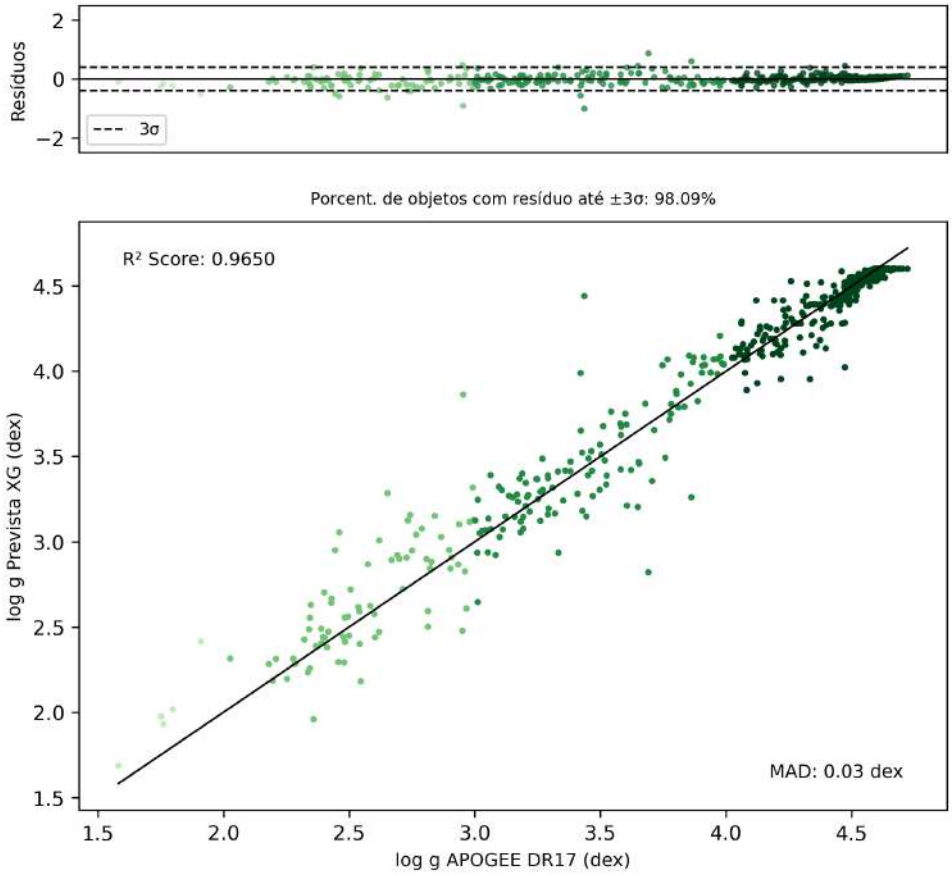

Figura B.5: Resultados do modelo `jplusA_01_XGB_logg` para a previsão de log g utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

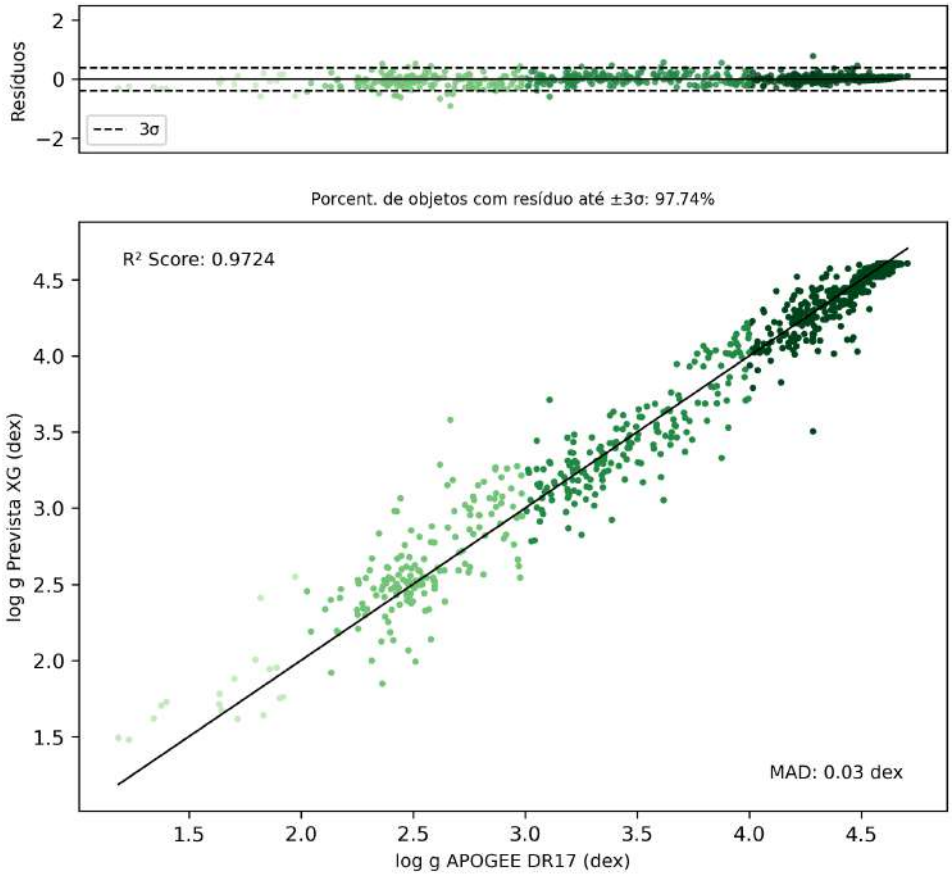

Figura B.6: Resultados do modelo `jplusA_02_XGB_logg` para a previsão de log g utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

B.2.2 S-PLUS

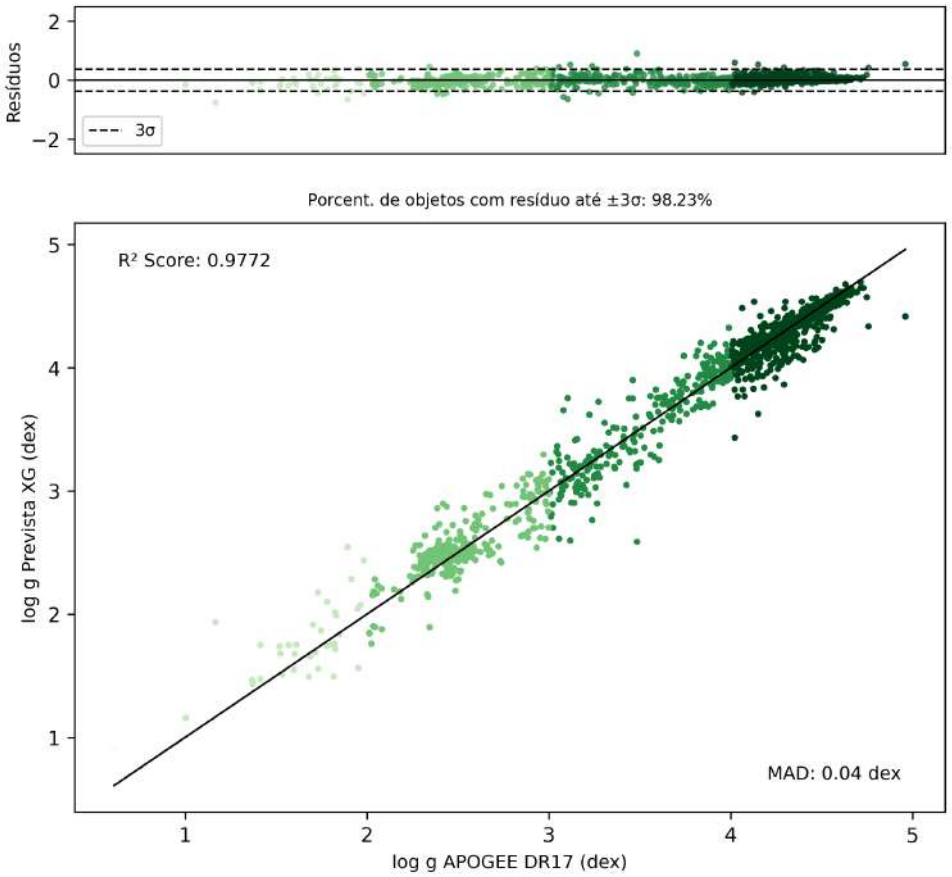

Figura B.7: Resultados do modelo `jplusA_01_XGB_logg` para a previsão de log g utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

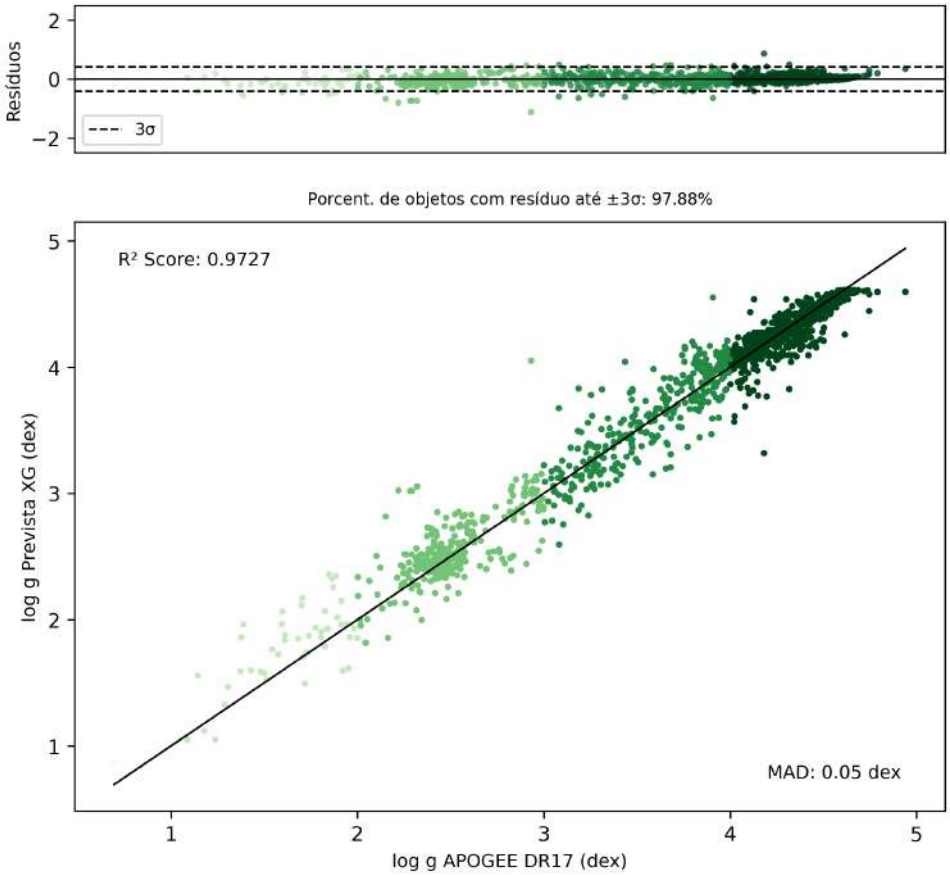

Figura B.8: Resultados do modelo `splusA_02_XGB_logg` para a previsão de log g utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

B.3 *XGBoost* na Previsão da Metalicidade

B.3.1 J-PLUS

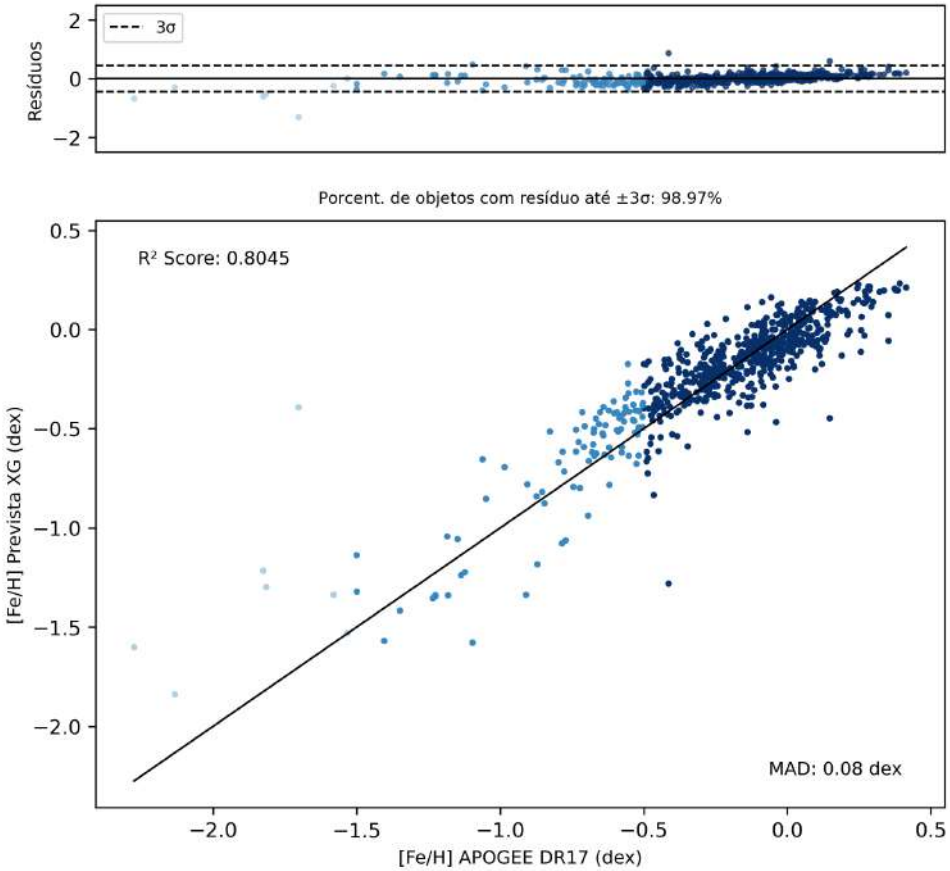

Figura B.9: Resultados do modelo `jplusA_01_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

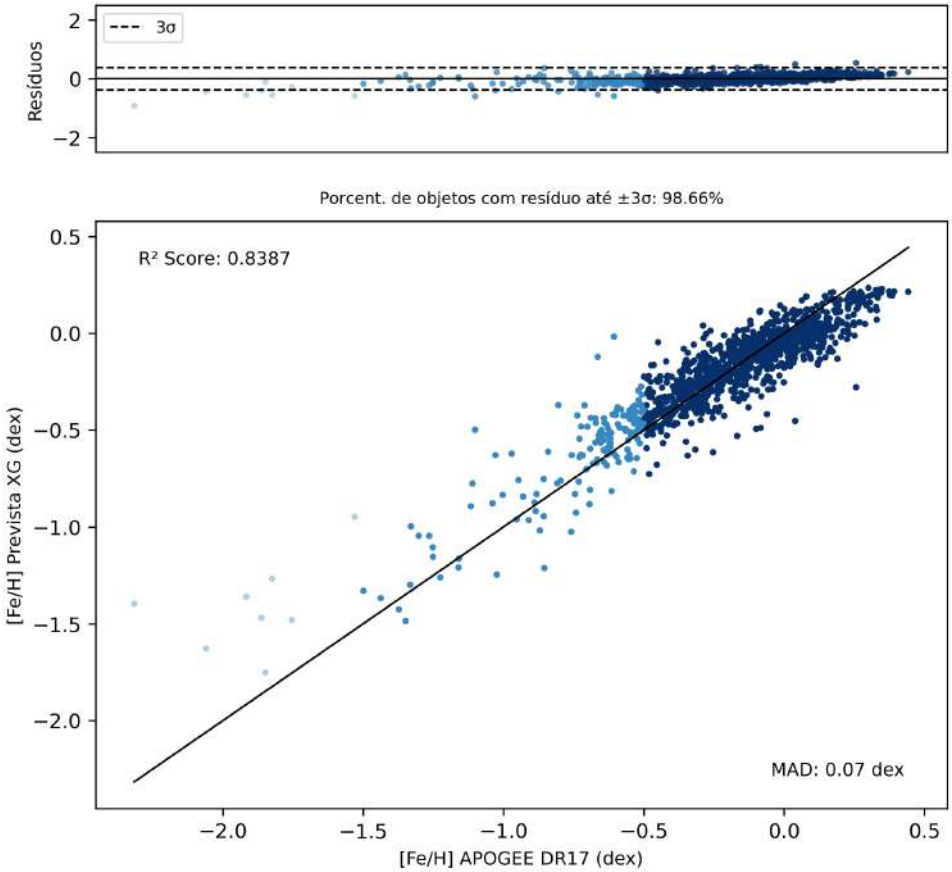

Figura B.10: Resultados do modelo `jplusA_02_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* utilizando objetos em comum entre o J-PLUS DR3 e o APOGEE DR17 SDSS-IV.

B.3.2 S-PLUS

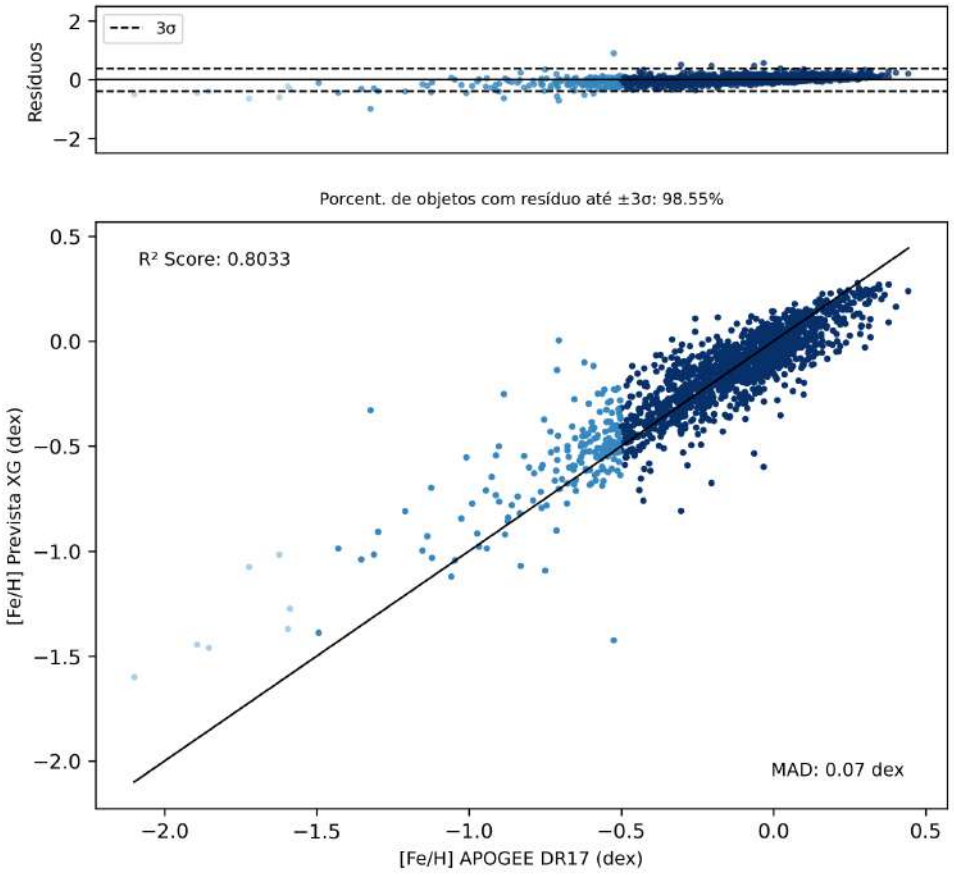

Figura B.11: Resultados do modelo `splusA_01_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *XGBoost* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.

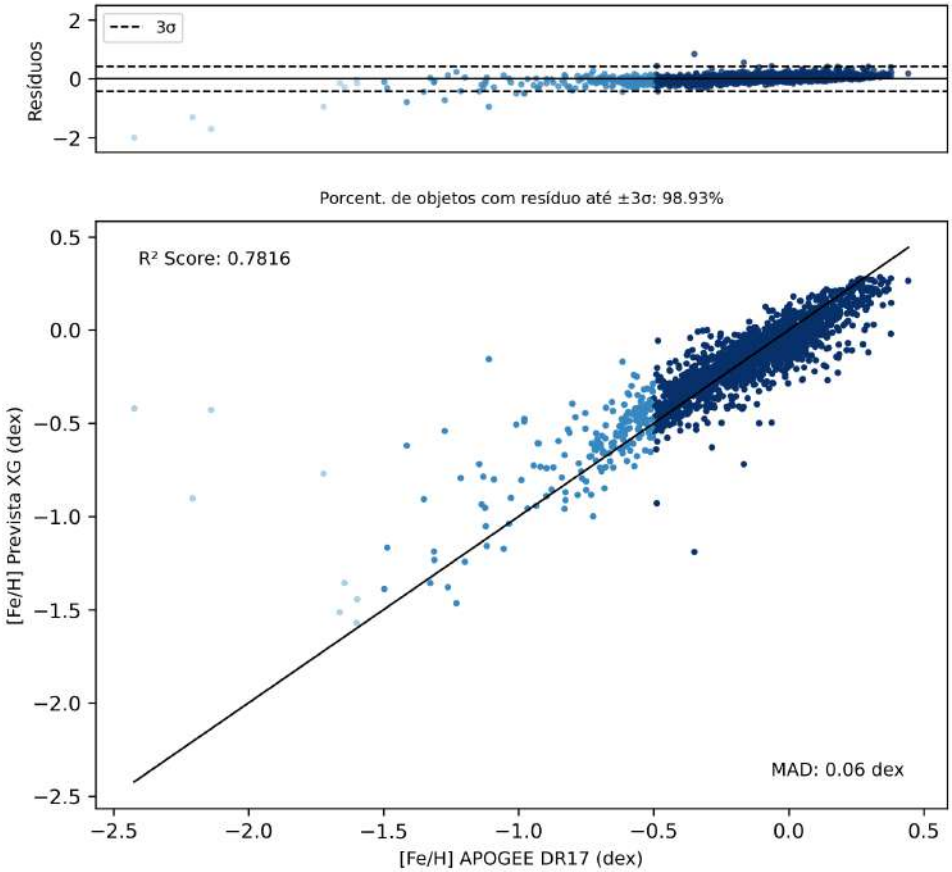

Figura B.12: Resultados do modelo `splusA_02_XGB_feh` para a previsão de [Fe/H] utilizando a técnica de *Random Fores* utilizando objetos em comum entre o S-PLUS iDR5 e o APOGEE DR17 SDSS-IV.