Observatório Nacional

DISSERTAÇÃO DE MESTRADO

CARACTERIZAÇÃO DE ESTRELAS FRACAS DA MISSÃO KEPLER COM BASE EM DADOS DO J-PLUS

LETHYCIA MARIA DE CARVALHO

RIO DE JANEIRO
2022

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

Dissertação de Mestrado

CARACTERIZAÇÃO DE ESTRELAS FRACAS DA MISSÃO KEPLER COM BASE EM DADOS DO J-PLUS

por

Lethycia Maria de Carvalho

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

Co-orientador: Dr. Luan Ghezzi F. Pinho

Rio de Janeiro, RJ – Brasil
Agosto de 2022

C837 Carvalho, Lethycia Maria de
Caracterização de estrelas fracas da missão Kepler com base em dados do J-PLUS [Rio de Janeiro] 2022.
xii, 115 p. 29,7 cm: graf. il. tab.

Dissertação (mestrado) - Observatório Nacional - Rio de Janeiro, 2022.

1. Missão Kepler. 2. Estrelas fracas. 3. Aprendizagem de Máquina. 4. Método de Trânsito. I. Observatório Nacional. II. Título.

CDU 000.000.000

“CARACTERIZAÇÃO DE ESTRELAS FRACAS DA MISSÃO KEPLER COM BASE EM DADOS DO J-PLUS”

LETHYCIA MARIA DE CARVALHO

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
(Orientador)

Dr. Luan Ghezzi F. Pinho – OV/UFRJ
(Co-orientador)

Dra. Simone Daflon dos Santos – Observatório Nacional

Dr. Raimundo Lopes de Oliveira Filho – UFS

RIO DE JANEIRO, RJ – BRASIL
30 DE AGOSTO DE 2022

Agradecimentos

Agradeço, primeiramente, aos meus orientadores, Dr. Marcelo Borges Fernandes e Dr. Luan Ghezzi Ferreira Pinho, por todo o tempo dedicado, disponibilidade, compreensão e suporte (científico e emocional).

Aos professores do Observatório Nacional, por todo o conhecimento compartilhado, dentro e fora do horário de aula.

Aos demais profissionais do Observatório Nacional, que tantas vezes facilitaram a minha vida.

Aos amigos que fiz no ON, em especial à Vinicius Cordeiro que, além de amigo, foi meu parceiro de trabalho nestes dois anos, com quem pude trocar ideias, discutir soluções e reclamar.

Aos amigos que fiz fora do ON, em especial à Arijane, minha melhor amiga de longa data, e ao meu amor, Luiz Felipe, que seguem me apoiando e tornando meus dias mais leves.

Aos membros da equipe de observação do OPD (Carlos, Marcelo, Karyne e Vinicius), com quem pude aprender bastante e compartilhar madrugadas produtivas.

À Simone e Andrés, que auxiliaram os meus primeiros passos no mundo da Aprendizagem de Máquina.

Al equipo de J-PLUS, que siempre estuvieron listos a ayudarme en todo lo que necesitara.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Lethycia Maria de Carvalho

CARACTERIZAÇÃO DE ESTRELAS FRACAS DA MISSÃO KEPLER COM BASE EM DADOS DO J-PLUS

RESUMO

Após séculos de aceitação do geocentrismo, o modelo heliocêntrico e o pluralismo cósmico fortaleceram a ideia de múltiplos mundos. Atualmente, milhares destes objetos já foram confirmados e uma fração considerável deles foi detectada pela missão Kepler, através do método de trânsito. Este método permite analisar as curvas de luz das estrelas observadas à procura de quedas temporárias de fluxo causadas pela passagem de corpos celestes em frente aos discos estelares. Entretanto, não basta apenas detectar estes corpos, é importante também poder caracterizá-los. Os parâmetros planetários dependem diretamente dos parâmetros atmosféricos de suas estrelas hospedeiras. Para isso, foi criado o Kepler Input Catalog (KIC), que é um compilado de bancos de dados menores que reúne, entre outras, informações sobre as estrelas do campo de visão da missão Kepler. Todavia, a precisão dos parâmetros do KIC tem sido discutida, visto que eles não foram obtidos de forma homogênea. Este trabalho possui o objetivo de recaracterizar estrelas da missão Kepler, unindo Aprendizagem de Máquina (Machine Learning) e as informações fotométricas do sistema de 12 filtros ópticos do Javalambre Photometric Local Universe Survey (J-PLUS) e de 4 filtros do Wide-field Infrared Survey Explorer (WISE), a fim de gerar algoritmos previsores de parâmetros estelares ($T_{\rm ef}$, log g e [Fe/H]). Além das 16 magnitudes, suas combinações em pares (cores) geram mais 120 colunas de entrada, com informações úteis para o algoritmo (totalizando 136 colunas de informação de magnitude por objeto, as *features*). Além das informações de magnitude, cada objeto da amostra utilizada para treinamento precisa possuir parâmetros físicos ($T_{\rm ef}$, log g e [Fe/H]) bem definidos na literatura. Para nossos preditores, recuperamos estes valores de um cruzamento entre LAMOST, WISE e J-PLUS, com cerca de 100 mil estrelas. Após uma cuidadosa otimização dos modelos gerados pelo algoritmo, obtivemos boas previsões para os três parâmetros atmosféricos em questão. Suas precisões foram de ≈ 70 K para $T_{\rm ef}$, $\approx 0,08$ dex para log g e $\approx 0,10$ dex para [Fe/H], na amostra de teste. Aplicamos os modelos para 44.483 estrelas da missão Kepler, também observadas pelo J-PLUS/WISE. Com os parâmetros previstos, pudemos estimar a correção bolométrica (com precisão de $\approx 0,022$ mag), e calcular as luminosidades, raios e massas destes objetos. Estas informações podem ser muito úteis para a caracterização de exoplanetas em trânsito, eventualmente detectados ao redor destas estrelas.

Lethycia Maria de Carvalho

CHARACTERIZATION OF FAINT STARS FROM THE KEPLER MISSION BASED ON J-PLUS DATA

ABSTRACT

After centuries of acceptance of geocentrism, the heliocentric model and cosmic pluralism strengthened the idea of multiple worlds. Currently, thousands of these objects have already been confirmed and a considerable fraction of them was detected by the Kepler mission, through the transit method. This method makes it possible to analyze the light curves of observed stars to search for temporary decrease in the stellar flux caused by the passage of celestial bodies in front of the stellar disks. However, it is not enough just to detect these bodies, it is also important to characterize them. The planetary parameters directly depend on the atmospheric parameters of their host stars. Based on this, the Kepler Input Catalog (KIC) was created, which is a compilation of other databases, gathering information about the stars in the Kepler mission's field of view. However, the precision of the KIC parameters has been discussed, since they were not obtained homogeneously. This work aims to recharacterize stars from the Kepler mission, combining Machine Learning and photometric information from the 12 optical filter system of the Javalambre Photometric Local Universe Survey (J-PLUS) and from 4 filters from the Wide-field Infrared Survey Explorer (WISE), in order to generate stellar parameter prediction algorithms ($T_{\rm ef}$, log g and [Fe/H]). In addition to the 16 magnitudes, their combinations in pairs (colors) create another 120 input columns, with useful information for the algorithm (totaling 136 columns of magnitude information per object, the features). In addition, each object in the sample used for training must have its physical parameters ($T_{\rm ef}$, log g and [Fe/H]) well defined in the literature. For our predictors, we retrieve these values from a crossmatch between LAMOST, WISE and J-PLUS, with about 100.000 stars. After careful optimization of the models generated by the algorithm, we obtained good predictions for these three atmospheric parameters. Their precisions were ≈ 70 K for $T_{\rm ef}$, $\approx 0,08$ dex for log g, and $\approx 0,10$ dex for [Fe/H], in the test sample. We applied the models to 44.483 stars from the Kepler mission, also observed by J-PLUS/WISE. Using the predicted parameters, we were able to estimate the bolometric correction (with a precision of $\approx 0,022$ mag), and derive luminosities, radii and masses of these objects. This information can be very useful for the characterization of transiting exoplanets, eventually detected around these stars.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Por séculos, a teoria geocêntrica de Ptolomeu (100 d.C.) sugeriu que a Terra era o centro do Universo. Apesar disso, muito tempo antes, Aristarco de Samos (320-250 a.C.) já defendia uma ideia diferente: o heliocentrismo. A teoria heliocêntrica, que destronava a Terra e afirmava que os astros, na verdade, orbitavam o Sol, foi mais tarde apoiada matematicamente por Nicolau Copérnico (1473-1543). O modelo copernicano sustentava com firmeza que o geocentrismo não tinha base lógica, mas era um pouco confuso e impreciso, já que não admitia a ideia de corpos celestes com movimentos não circulares/uniformes. Por causa disso, Copérnico foi incapaz de perceber que as órbitas eram, na verdade, elípticas. Em seu modelo, as estrelas eram fixas, além de serem tomadas como os astros mais distantes do universo, não considerando a existência dos corpos invisíveis a olho nu (Damasio, 2011).

Posteriormente, o heliocentrismo foi reforçado por Galileu Galilei (1564-1642) com a observação das luas de Júpiter. Galileu mostrou que a Terra não era o centro do universo, já que havia corpos orbitando em torno de Júpiter ao invés da Terra. Isso invalidava o pressuposto de que todos os corpos tinham que orbitar a Terra. As ideias de Copérnico também foram aprimoradas e expandidas por Johannes Kepler (1571-1630) que, por sua vez, também foram sustentadas pela física da lei da gravitação universal de Newton (1643-1727) e são aceitas até hoje (Danielson, 2001).

A ideia de múltiplos mundos também não é recente. Entre os séculos 5 e 3 a.C., 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 o responsável por propor que as estrelas eram sóis distantes, que possuíam seus próprios planetas e que podiam também abrigar vida. Essa filosofia ficou conhecida como pluralismo cósmico. Bruno também insistia que o universo era, na verdade, infinito e não possuia um centro determinado (Gatti, 2010).

Desde 1995, com a descoberta de 51 Pegasi b, o primeiro planeta fora do nosso sistema

[1] Corrente filosófica que acreditava que o mundo fisíco era composto de pequenos blocos indivisíveis e indestrutíveis, os átomos, que podiam viajar no vácuo, colidirem entre si e aglomerar-se.

planetário orbitando uma estrela de tipo solar (Mayor & Queloz, 1995), já foram identificados 5147[2] exoplanetas - e este número cresce diariamente. Isso tem ocorrido graças às diferentes técnicas de detecção, com destaque para o método de trânsito planetário, responsável por cerca de 77% de todas as detecções atuais, segundo o Nasa Exoplanet Archive[3]. Mais detalhes sobre o método de trânsito serão apresentados na Seção 1.1.

1.1 Método de trânsito

Uma das técnicas observacionais utilizadas pela astronomia para detecção de objetos ao redor de estrelas é o método de trânsito. Um trânsito consiste na passagem de um corpo em frente ao disco da estrela em torno da qual ele orbita. Essa passagem provoca uma queda, mesmo que sutil, do brilho total do sistema. Em geral, estrelas que não sofrem obscurecimento por trânsito possuem brilho relativamente constante. Isso não ocorre em casos especiais, como nas estrelas variáveis, que são objetos que possuem variação de luminosidade (brilho intrínseco) com o tempo. As oscilações de brilho detectadas durante o trânsito, se apresentam, observacionalmente, como mínimos na curva de luz da estrela (Santos & Amorim, 2017). A Figura 1.1 ilustra essa situação. Essa técnica depende das características do objeto em trânsito - por exemplo, quanto maior é este objeto, maior é a variação que ele provoca no brilho da estrela que orbita - e também da configuração da órbita, onde é necessário que o corpo transite pelo disco estelar, aproximadamente no plano da linha de visada.

No painel superior da Figura 1.1 é possível ver a estrela (esfera maior, em cinza claro) e o objeto, neste caso um exoplaneta (esfera menor), transitando pelo disco estelar em sentido anti-horário. No painel inferior, podemos ver a variação do fluxo percebido do sistema em função do tempo. Isto é o que chamamos de curva de luz. O arredondamento desta curva é causado pelo efeito de escurecimento do limbo[4]. Percebe-se que, à medida que o planeta passa em frente ao disco estelar (trânsito primário), o fluxo percebido do sistema diminui. Ele retoma o valor prévio após o planeta completar o trânsito. Essa queda no fluxo é periódica. Os demais momentos da órbita também podem provocar variações na curva de luz. Quando o planeta não está transitando pelo disco, o fluxo medido é uma combinação da energia emitida pela estrela com a luz refletida pelo planeta (emissão térmica). Quando o planeta se posiciona atrás da estrela (ocultação ou trânsito secundário), também há uma diminuição do fluxo percebido (muito menor que aquela vista no trânsito) já que não se percebe, desde a linha de visada, a luz refletida pelo planeta. Nesse caso, a única fonte de brilho é a estrela (Winn, 2010).

[2]Em 12 de agosto de 2022, segundo www.exoplanet.eu

[3]Em 12 de agosto de 2022: https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

[4]O escurecimento do limbo é um efeito óptico estelar no qual a parte central do disco parece mais brilhante do que as bordas.

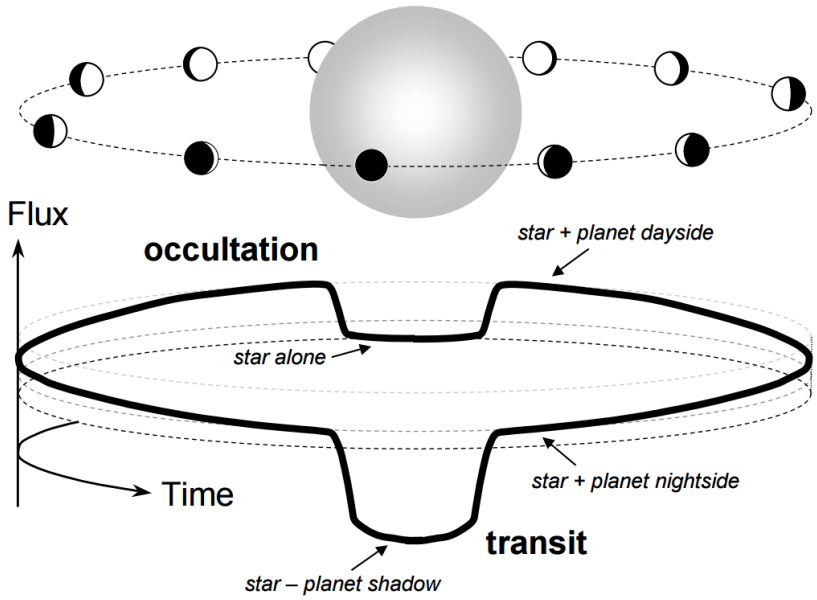


Figura 1.1: Curva de luz com evento de trânsito planetário. No painel superior, em cinza claro, temos a estrela em torno da qual orbita o objeto (estrela hospedeira). Podemos ver o objeto, neste caso um exoplaneta, transitando em sentido anti-horário. No painel inferior, podemos observar a variação do fluxo percebido do sistema em função do tempo (curva de luz). Podemos notar uma queda durante a passagem do exoplaneta em frente ao disco estelar (*transit*) e atrás (*occultation*), quando o observador não pode perceber o fluxo refletido pelo exoplaneta. Fonte: Winn (2010)

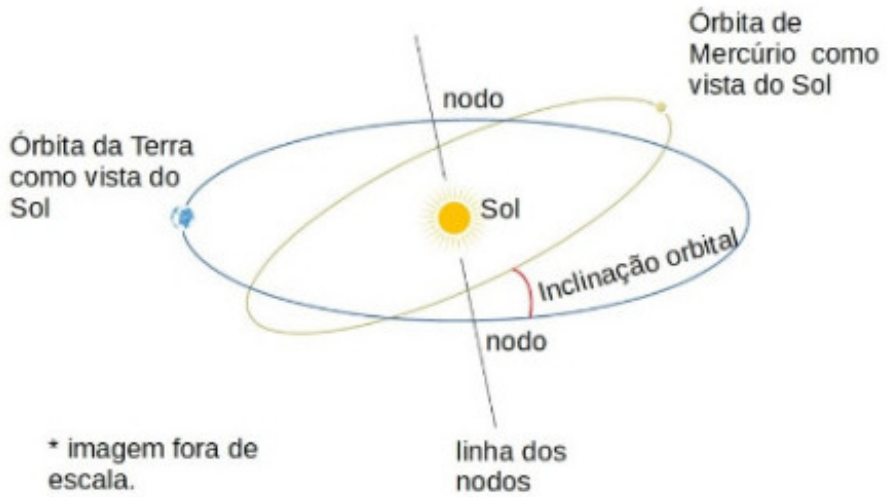


Figura 1.2: Planos orbitais de Mercúrio e Terra, ilustrando a configuração orbital necessária para observarmos, na Terra, o trânsito de Mercúrio sobre o disco solar. A linha dos nodos é definida na intersecção dos planos das duas órbitas. Registra-se um trânsito quando ambos os planetas passam simultaneamente por esta linha. Fonte: Martins et al. (2020).

O trânsito possui uma longa história de interesse para os astrônomos. Esta história se inicia sob o ponto de vista dos planetas Vênus e Mercúrio, geradores de trânsitos no disco solar que podem ser notados da Terra. Para Mercúrio, por estar mais próximo do Sol, a probabilidade de trânsito é maior: se notam cerca de 13 alinhamentos por século, o último deles tendo ocorrido em 2019 e o próximo a ocorrer em 2032. Para Vênus, detectamos de 0 a 2 trânsitos por século. Esses trânsitos não ocorrem com alta frequência porque os planetas não estão no mesmo plano na Terra. No caso de Mercúrio, sua órbita possui inclinação de $\approx 7^{\circ}$ com relação ao plano da órbita da Terra. Isso gera, portanto, apenas dois pontos (nodos) de alinhamento (vide Figura 1.2; Martins et al., 2020).

Segundo o Nasa Exoplanet Archive, o método de trânsito é responsável, desde 2011, pela maior parte das detecções anuais de exoplanetas e também, desde 2014, pela maior parte das detecções acumuladas. É possível ver mais detalhes sobre isso na Figura 1.3. Nela, pode-se ainda ver os resultados alcançados por outros métodos, com destaque também para a técnica de velocidade radial. A técnica de medição de velocidade radial, também conhecida como Espectroscopia Doppler, é um método espectroscópico para a detecção de exoplanetas e anãs marrons. Ela se baseia na observação do efeito Doppler no espectro da estrela hospedeira, que causa um deslocamento periódico das linhas espectrais.

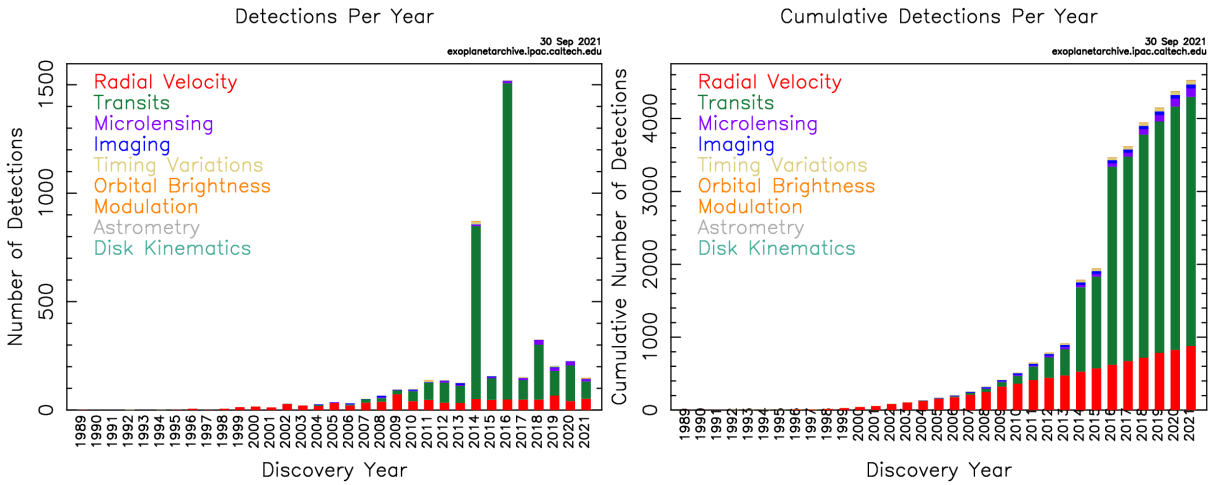


Figura 1.3: Número de exoplanetas detectados por ano, de acordo com o método utilizado, segundo o Nasa Exoplanet Archive. À esquerda, vemos o número de detecções realizadas por cada técnica (representadas por diferentes cores) por ano. À direita, vemos o número de detecções acumuladas ao longo dos anos. Observe que o método de trânsito é o principal responsável pelas detecções anuais realizadas desde 2011.

O método de trânsito não serve apenas para detectar exoplanetas, mas também para identificar os efeitos de quaisquer objetos em trânsito que provoquem variações no brilho da estrela hospedeira. Desta forma, objetos maiores podem provocar alterações ainda mais acentuadas - podemos citar objetos como anãs marrons e uma estrela secundária (caracterizando um sistema binário eclipsante). A missão Kepler, por exemplo, permitiu a detecção de mais de uma dúzia de anãs marrons (e.g. Díaz et al., 2013; Johnson et al.,

2011; Moutou et al., 2013) e 2878 binárias eclipsantes (Kirk et al., 2016). Pode ocorrer ainda de uma variação ser causada por efeitos sistemáticos e não necessariamente por um objeto real. Pode-se chamar estes casos de falsos trânsitos (ou falso-positivos). A Subseção 1.1.1 detalhará fenômenos que causam este tipo de efeito.

1.1.1 Efeitos sistemáticos

As variações nas curvas de luz de uma estrela nem sempre são produzidas por um objeto em trânsito. Algumas vezes, é possível que haja contaminações por efeitos sistemáticos. A variação pode ser causada, por exemplo, pelo deslocamento aparente da estrela, raios cósmicos e píxels mortos (aqueles que não apresentam nenhuma contagem). Vale também ressaltar que, além dos falso-positivos, temos os falso-negativos. Os falso-negativos são estrelas que possuem objetos em órbita mas que não estavam transitando o disco da estrela, na linha de visada, no momento da captura da imagem. Isso faz parecer que não existe nenhum objeto em órbita (e em dado momento, em trânsito). Esta é uma das limitações da observação não contínua.

As variações causadas pelo deslocamento aparente da estrela estão associados a aberração estelar e ao movimento próprio. A aberração estelar é um desvio aparente na posição de um astro. Ele ocorre por causa da velocidade relativa entre observador e fonte. O observador continua a se mover conforme recebe a luz do objeto e isso causa uma pequena variação angular na posição do astro. Isso faz parecer que é o objeto que se move no céu, com relação a ele (isto realmente acontece, mas não é suficiente para causar o efeito de aberração) (Crosta & Vecchiato, 2010). Para as estrelas no campo de visão do Kepler, a aberração estelar anual pode chegar a 6" ($\approx 1,5$ píxels) e é causada pelo movimento do telescópio em sua órbita. A própria rotação da Terra pode produz um efeito de aberração estelar. Podemos ver isto na Figura 1.4.

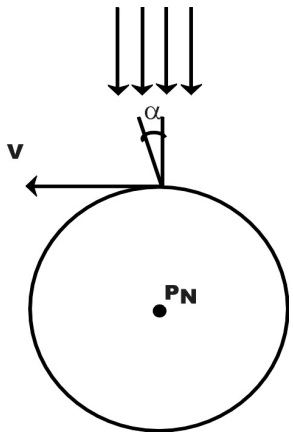


Figura 1.4: Exemplo de aberração estelar. Neste exemplo, vemos como a Terra, com velocidade v, pode produzir uma variação angular aparente (α) para um observador no equador terrestre. As setas para baixo representam a luz do objeto chegando ao observador. P_N representa o Polo Norte Terrestre. Fonte: Santiago (2005).

O movimento próprio é o deslocamento do astro com relação a objetos mais distantes e não depende da velocidade e/ou posição do observador. Para as estrelas observadas periodicamente pelo Kepler, esse deslocamento é muito sutil e oscila em milissegundos de arco ($\approx 0,1$ píxels). Este efeito pode ser detectado nas curvas de luz de uma das orientações, que apresenta uma tendência diferenciada das demais, mostrando uma variação do fluxo médio medido (Crosta & Vecchiato, 2010).

A incidência de raios cósmicos também pode causar uma impressão inicial de trânsito, já que partículas super energéticas podem afetar píxels isolados do CCD. Geralmente, isso causa um abrupto aumento no número de contagens. Neste caso, o efeito na curva de luz é facilmente identificável e descartável, já que produz um aumento e não uma queda de brilho como os trânsitos. Contudo, algumas vezes, o píxel atingido também pode sofrer uma queda abrupta no número de contagens (*Sudden Pixel Sensitivity Dropout*, SPSD), tornando-o insensível. Este caso se assemelha a uma marca de trânsito. O que o diferencia de um trânsito é que o píxel se recupera na ordem de horas a meses e não volta a ocorrer periodicamente como um trânsito (Smith et al., 2012). Para confirmar uma SPSD, precisamos de mais observações que mostrem que a curva não se recuperou e que o período é longo demais para ser um trânsito.

Também é importante ter cuidado ao analisar os dados de estrelas próximas às bordas da imagem (já que pode ter havido perda de informação dos píxels associados à detecção), com píxels saturados (que podem contaminar as contagens de píxels próximos, indicando valores altos mas artificiais nas contagens associadas às estrelas) e com píxels mortos (vide Figura 1.5).

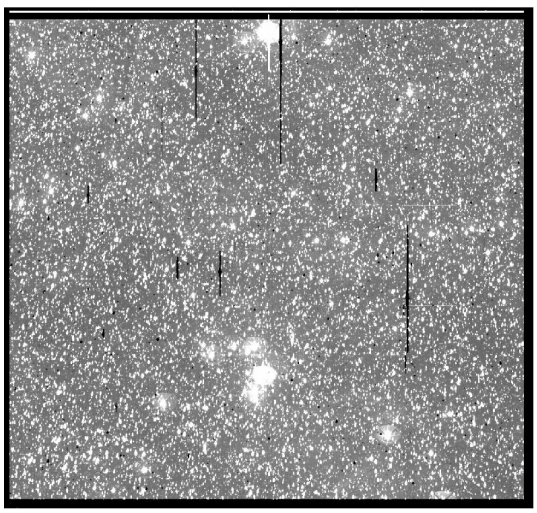


Figura 1.5: Na imagem podemos observar linhas verticais escuras que se assemelham a falhas. Elas são píxels sem contagem (ou píxels mortos), que podem ocorrer por defeitos na fabricação do equipamento ou em píxels afetados por efeitos sistemáticos, como raios cósmicos. Esta é uma imagem de saída capturada pela missão Kepler.[5]

[5]Podemos ver estes arquivos em: https://archive.stsci.edu/pub/kepler/ffi/png/ffi/

1.2 Missão Kepler

Em funcionamento de 2009 a 2013, a missão Kepler[6] consistia basicamente em um fotômetro de 0,95 m de abertura com um sistema de 42 CCDs. O equipamento foi direcionado inicialmente para uma região entre as constelações de Cisne e Lira ($\alpha \approx 19h$, $\delta \approx 40°$; vide Figura 1.6), no hemisfério norte, cobrindo uma área de $\approx$ 105 graus quadrados, observando na região espectral entre 430 e 890 nm (Borucki et al., 2010; Koch et al., 2010). A missão utilizou o método de trânsito como método de detecção de exoplanetas. Como os trânsitos duram comumente apenas algumas horas, as estrelas precisam de monitoramento contínuo, ou seja, seus brilhos devem ser medidos pelo menos uma vez a cada poucas horas. Para isto, o campo de visão nunca deveria ser bloqueado, em nenhuma época do ano. Para evitar o Sol, o campo de visão escolhido estava fora do plano da eclíptica e tinha o maior número possível de estrelas. A missão Kepler observou aproximadamente 4,5 milhões de estrelas, das quais cerca de 150.000 foram monitoradas continuamente.

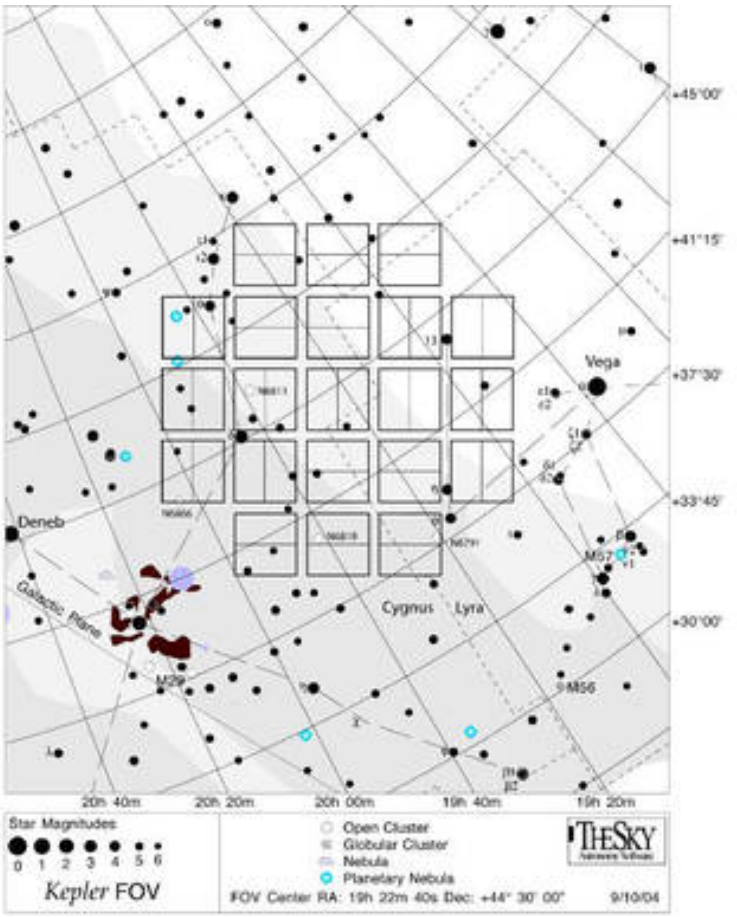


Figura 1.6: Campo de visão do Kepler na região próxima às constelações Cisne e Lira. O eixo horizontal representa ascensão reta (α) e o vertical representa declinação (δ).[6]

[6]Todas as informações desta seção podem ser consultadas na página da missão: https://www.nasa.gov/mission_pages/kepler.

O telescópio contava, em seu lançamento, com um sistema de quatro giroscópios onde, para manter o campo de visão fixo, eram necessários pelo menos três. Entretanto, em maio de 2013, o Kepler perdeu um segundo giroscópio. Isso encerrou, de certa forma, a missão principal, já que não era possível sustentar a estabilidade necessária para manter a observação do campo fixo com a precisão original pretendida pela missão. Nos anos finais (entre 2014 e 2018), o telescópio Kepler observou, então, diferentes campos distribuídos ao redor do plano da eclíptica. Isto permitia minimizar o torque exercido na espaçonave pela pressão do vento solar e, consequentemente, a deriva (desvio de apontamento) até o ponto em que o campo pode ser efetivamente controlado pelos propulsores e os dois giroscópios restantes. Tal procedimento mantinha estável a observação de um dado campo por aproximadamente 80 dias (duração de uma campanha). Este período de observação ficou conhecido como missão Kepler 2 ou, simplesmente, K2. O site Nasa Exoplanet Archive[7] mostra que a missão Kepler identificou, até então, 2711 exoplanetas e possui 2056 candidatos aguardando confirmação. Já a K2 identificou 537 exoplanetas e possui ainda 969 candidatos.

A missão Kepler foi pensada com o objetivo da busca por outros mundos, se interessando principalmente por aqueles semelhantes à Terra, localizados dentro ou próximo das zonas habitáveis de suas estrelas hospedeiras. Zona habitável é a região ao redor de uma estrela onde é possível manter água em estado líquido na superfície de um planeta. Essa suposição pressupõe que a vida extraterrestre é semelhante à terrestre e tem necessidades químicas, físicas e biológicas parecidas (Kasting, 1997). Entre os objetivos específicos da missão também estavam determinar: a porcentagem de planetas semelhantes a Terra, e maiores, dentro ou próximos das zonas habitáveis de estrelas de diferentes tipos espectrais; a distribuição de tamanhos e formas das órbitas desses planetas; porcentagem de planetas em sistemas estelares múltiplos; tamanhos, massas e densidades de planetas gigantes de curto período orbital; propriedades das estrelas hospedeiras, etc.

Até o momento, a missão Kepler foi a mais bem sucedida na busca por exoplanetas. O alto número de detecções que ela proporcionou, contribuiu para caracterizar a amostra de exoplanetas, principalmente quanto ao tamanho, possibilitando separá-los por classes. Segundo Petigura et al. (2018), entre as várias divisões, há evidências agora de um número substancial de três principais tipos de exoplanetas: gigantes gasosos (tipo-Júpiter), gigantes de gelo (subdivididos em Sub-Netunos e Sub-Saturnos) e super-Terras (principalmente as de tipo quente, que possuem órbitas de curto período). A Tabela 1.1, baseada nos dados de Petigura et al. (2018), apresentam os intervalos de raio que dividem estas classes exoplanetárias. Podemos também subclassificá-las de acordo com o período orbital (P): quente (p = 0-10 dias), morno (p = 10-100 dias) e frio (p = 100-350 dias). A tabela inclui também a proposta de Kopparapu et al. (2018) para planetas tipo-Terra.

[7]Em 12 de agosto de 2022 em: https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

Tabela 1.1: Classificação de exoplanetas por tamanho, segundo Petigura et al. (2018) e Kopparapu et al. (2018)

Classe	**Raio($R_\oplus$)**
Tipo-Terra	0,5-1,0
Super-Terra	1,0-1,7
Sub-Netuno	1,7-4,0
Sub-Saturno	4,0-8,0
tipo-Júpiter	8,0-24,0

Como vimos no início desta seção, apenas uma pequena fração das estrelas da missão Kepler foram observadas de forma contínua. A maior parte foi observada apenas por uma série de observações chamadas *Full Frame Images* (FFIs), como veremos a seguir.

1.2.1 Full-Frame Images

Devido à capacidade de armazenamento e transmissão de dados, a missão observou continuamente "apenas" cerca de 150.000 estrelas (Batalha et al., 2010), usando observações de cadência curta (58,85 segundos) e longa (29,4 minutos). Essas estrelas possuíam, predominantemente, magnitude K_p = 14-16. As demais estrelas (cerca de 4,35 milhões ou 97% da amostra total) eram mais fracas que $K_p = 16$ e foram observadas apenas através de *Full-Frame Images* (FFIs)[8]. Uma FFI é uma imagem de quadro cheio. *Full-frame* é um termo usado em cinematografia para o ato de capturar imagens com equipamento que possua um sensor do tamanho de sua largura e altura máximas. A maior vantagem de uma câmera full-frame é que ela não tem fator de corte. A presença de fator de corte significa que a imagem precisou ser cortada porque o sensor era muito pequeno para realizar a captura da imagem inteira.

As primeiras 8 imagens de tipo FFI foram observadas durante o período de comissionamento do telescópio ($\approx$ 36 horas) e constituíram as "golden FFIs", cada uma com tempo de exposição de $\approx 29,4$ minutos. Outras 45 FFIs "comuns" foram produzidas em intervalos de $\approx$ 1 mês, até o fim da missão, totalizando 53. No Kepler, uma FFI era composta por 21 módulos, onde cada módulo continha 2 CCDs e cada CCD continha 2 canais de saída. Cada FFI produziu, então, 84 imagens (vide Figura 1.7). Uma falha em um módulo ocasionava a perda de 2 CCDs (4 imagens). As FFIs foram classificadas quanto às orientações (ou *season*, divididas de 0 a 3) e trimestres (ou *quarter*, divididos de 0 a 16), de acordo com o período em que foram produzidas. Cada vez que o telescópio rotacionava 90° em relação ao eixo fixo de observação, ele mudava sua orientação. Como isso ocorria a cada $\approx$ 90 dias, eram observadas 4 orientações em um ano, e então voltava-se à orientação inicial, reiniciando a contagem.

[8] https://archive.stsci.edu/kepler/ffi_display.php

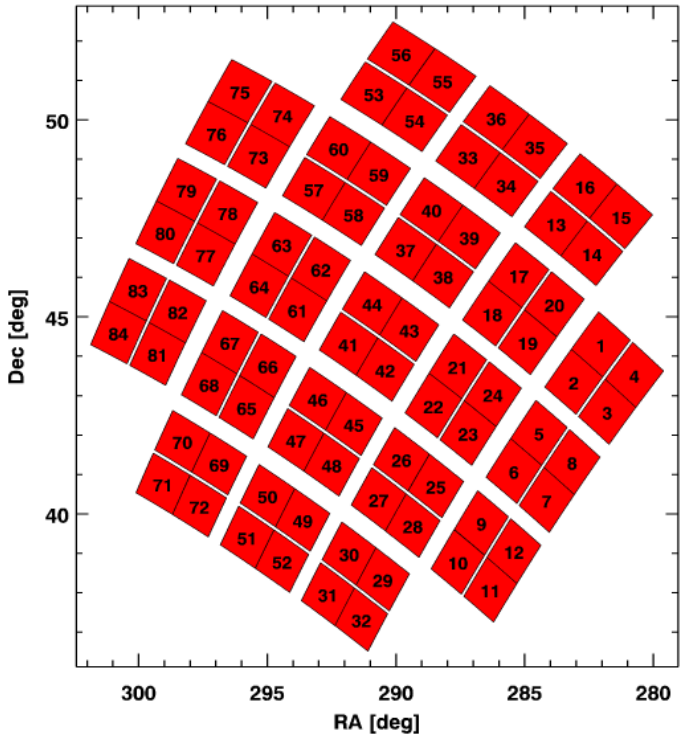


Figura 1.7: Representação de uma FFI para o campo de visão do Kepler. No Kepler, uma FFI era composta por 21 módulos (quadrados vermelhos de tamanho 2x2), onde cada módulo continha 2 CCDs (retângulos vermelhos de tamanho 2x1 dentro de um módulo) e cada CCD continha 2 canais de saída (quadrados vermelhos de tamanho 1x1; cada módulo continha 4 canais de saída). Cada FFI produziu, então, 84 imagens - a falha em um módulo ocasionava a perda de 4 delas.

1.3 Dependência entre parâmetros planetários e estelares

O estudo de exoplanetas depende de medições precisas para os parâmetros dos corpos avaliados. Os planetas, por sua vez, para serem bem caracterizados, dependem de parâmetros precisos para suas estrelas hospedeiras. Aqui apresentamos algumas destas dependências entre parâmetros estelares e planetários.

Vimos que, através da curva de luz da estrela, podemos detectar objetos em trânsito no disco estelar. O ponto mais baixo nesta curva, ou seja, aquele onde há a maior queda de fluxo, representa a profundidade máxima do trânsito (δ). Em geral, essa profundidade máxima corresponde ao momento do trânsito em que o objeto passa mais próximo do centro do disco estelar, que corresponde à região mais brilhante.

Também vimos que a maior parte das estrelas da missão Kepler não foi observada de forma contínua. Observações não contínuas de trânsitos não fornecem registros de todos os momentos dos mesmos. Neste caso, ao detectarmos um trânsito na curva de

luz de determinada estrela, não há garantia de que a profundidade máxima observada é a profundidade máxima real do trânsito. Na verdade, sem uma observação contínua é muito difícil que o trânsito tenha sido observado no momento ideal, ou seja, quando se alcança a profundidade máxima real. Desta forma, o que podemos inferir é a profundidade máxima observada do trânsito (Q; vide Figura 1.8).

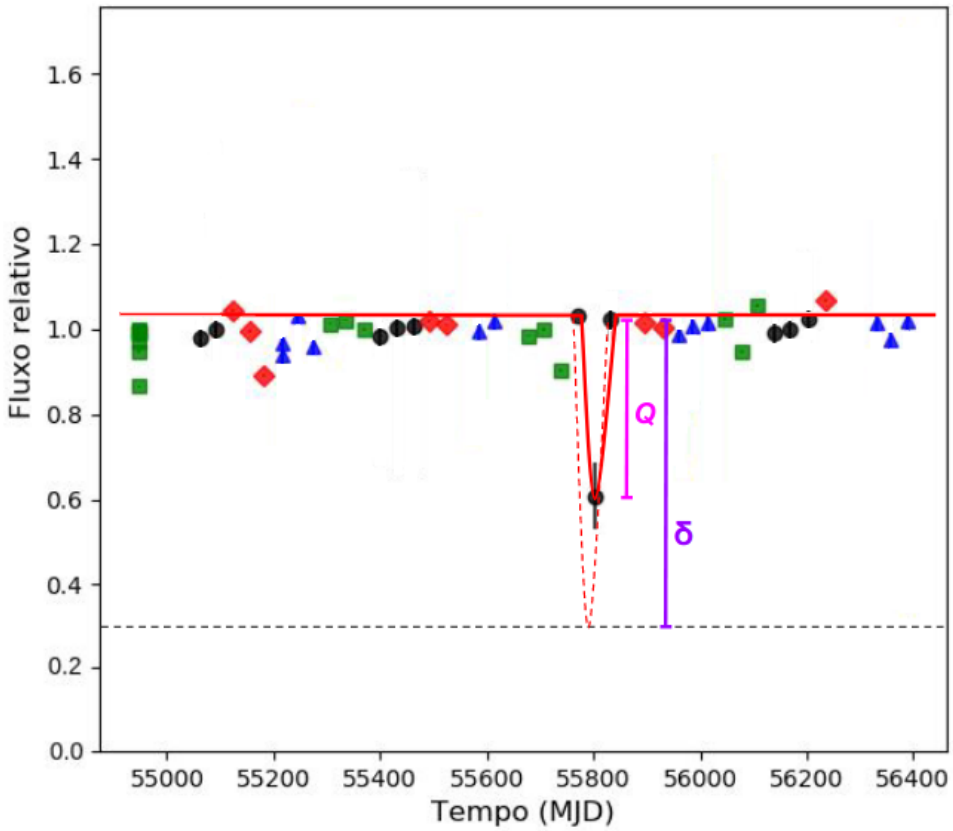


Figura 1.8: Curva de luz da estrela KIC 8561192, calculada por Nogueira (2020), exemplificando uma situação onde a profundidade máxima do trânsito (δ) é diferente da profundidade observada. Os quadrados verdes, triângulos azuis, círculos pretos e losangos vermelhos correspondem às 4 diferentes orientações do telescópio Kepler, comentadas na Subseção 1.2.1. A curva sólida em vermelho representa a aproximação da curva de luz calculada por Nogueira (2020) e a curva projetada (pontilhada em vermelho) se refere a curva de luz contínua desta estrela, segundo o Catálogo de Binárias Eclipsantes do Kepler (Kirk et al., 2016). Chamamos a profundidade observada de profundidade máxima observada do trânsito (Q), em alusão ao que ela representa, visto que $Q \leqslant \delta$. Como a profundidade real do trânsito não pode ser menor que Q, esta variável representará o valor mínimo da profundidade.

A profundidade máxima observada Q é a queda do fluxo F medida em uma curva de luz não contínua. Caso o trânsito seja observado no momento ideal, ela é igual à profundidade máxima real δ. Em alguns casos, podemos detectar mais de um trânsito na curva de luz. Cada trânsito tem sua própria Q. Podemos calcular Q através da Equação 1.1, desde que conheçamos os fluxos medidos logo antes (F_A) e logo depois (F_D)

da detecção, ou seja, usando o valor do fluxo nos dois pontos da curva de luz vizinhos ao ponto que corresponde ao trânsito (F_T).

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

É importante conhecer Q quando se quer conhecer o raio do objeto (R_p) causador do trânsito, por exemplo, um exoplaneta. Tão importante quanto, é conhecer o raio da estrela ($R_\star$) que ele orbita. Isto destaca a importância de obter um raio estelar com boa precisão. 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 podemos perceber que, para dois objetos que provocam uma mesma Q, quanto maior $R_\star$, maior R_p:

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

Podemos propagar as incertezas em R_p, σ_{R_P}, através de:

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

O raio estelar, por sua vez, é dependente de parâmetros físicos como temperatura efetiva ($T_{\rm ef}$) e luminosidade estelar (L), como nos mostra a Lei de Stefan-Boltzmann para a luminosidade (vide Equação 1.4; σ é a chamada constante de Stefan-Boltzmann, que tem o valor de 5,6697 x 10^{-5} erg cm^{-2} s^{-1} K^{-4}).

$$L = 4\pi R_\star^2 \sigma T_{\rm ef}^4 \tag{1.4}$$

Além de R_p, o método de trânsito também possibilita conhecer a periodicidade com a qual os trânsitos ocorrem, ao conhecer a duração do trânsito (d), que se relaciona ao período orbital (aqui chamado de T) através de:

$$d = \frac{T\alpha}{\pi} \rightarrow T = \frac{d\pi}{\alpha} \tag{1.5}$$

onde α é o ângulo entre a linha que liga o centro da estrela ao centro do planeta e o eixo x (direção do observador; vide Figura 1.9). A Figura 1.9 mostra quatro momentos do trânsito de um planeta (esfera azul) em frente ao disco da estrela hospedeira (esfera amarela).

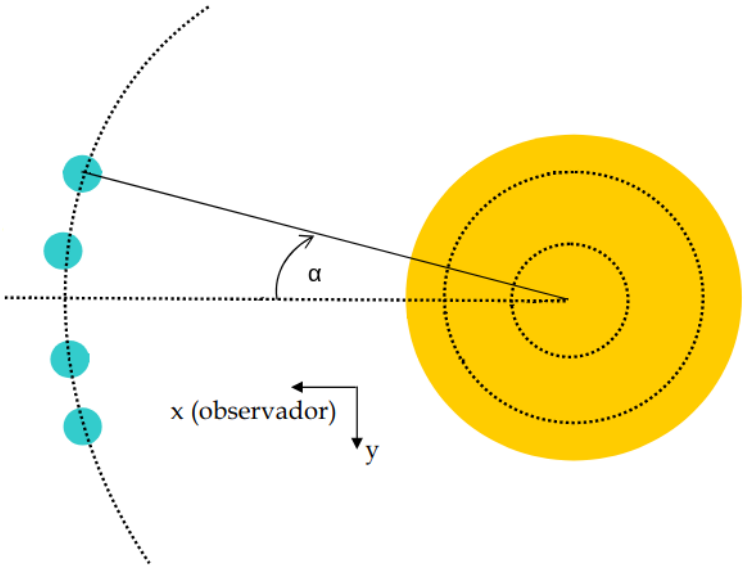


Figura 1.9: Quatro momentos do trânsito de um planeta (esfera azul) em frente ao disco da estrela hospedeira (esfera amarela). O ângulo α, usado na Equação 1.5, é formado pela linha que liga o centro da estrela ao centro do planeta e o eixo x, que representa a direção do observador (Aigrain, 2005).

Conhecendo T, podemos encontrar a distância orbital do objeto em trânsito (a, também conhecida como semieixo maior) através da 3^a lei de Kepler (vide Equação 1.6). A 3^a lei de Kepler enuncia que o quadrado do período orbital de um planeta é proporcional ao cubo do semieixo maior da órbita.

$$T^2 = \frac{4\pi^2}{GM}a^3 = \frac{5,92 \text{ x } 10^{11}}{M}a^3 = \frac{5,92 \text{ x } 10^{11}}{(m_1 + m_2)}a^3 \rightarrow a = \sqrt[3]{\frac{T^2(m_1 + m_2)}{5,92 \text{ x } 10^{11}}} \tag{1.6}$$

onde $G = 6,674184$ Nm^2/kg^2 é a constante de gravitação universal de Newton e $M = m_1 + m_2$ corresponde a massa do sistema (estrela + planeta). Uma das formas de estimar a massa de uma estrela é conhecendo sua luminosidade (a luminosidade pode ser obtida ao conhecer a magnitude aparente e a distância estrela-Terra, por sua vez, determinada pelo método de paralaxe), através da relação massa-luminosidade, que descreve uma relação linear entre o logaritmo da luminosidade (log L) e o logaritmo da massa (log M). Esse método é apenas uma aproximação e deve ser usado com cautela. Para a massa do planeta, são necessárias observações adicionais usando a técnica de velocidade radial. Mais informações sobre o uso da velocidade radial no estudo de exoplanetas podem ser consultadas em Wright (2018). Para a equação acima, podemos inclusive ignorar a massa do planeta, já que ela é desprezível em comparação com a massa da estrela, fazendo $M = m_1$ apenas.

Conhecendo a massa e o raio do planeta, podemos calcular sua densidade (vide Equa-

ção 1.7). Uma estimativa da densidade média do planeta é uma simples divisão da massa do planeta pelo seu volume (V_p), considerando aqui uma distribuição esférica:

$$\rho = \frac{m_2}{V_p} = \frac{m_2}{\frac{4\pi R_p^3}{3}} = \frac{3m_2}{4\pi R_p^3} \tag{1.7}$$

Algumas considerações importantes podem ser feitas sobre os parâmetros planetários definidos até aqui. Primeiro, a distância orbital do planeta, com relação a sua estrela hospedeira, é um parâmetro fundamental para determinar se tal planeta pode ou não suportar vida. Caso o planeta se encontre muito próximo da estrela hospedeira, ele será insuportavelmente quente e a maior parte da química molecular que ocorre na Terra será impossível. Certamente, a alta temperatura teria evaporado qualquer água líquida na superfície deste planeta e sua atmosfera teria sido varrida pelo vento estelar. Em contrapartida, se o planeta estiver muito longe da estrela, ele será muito frio para manter a água em estado líquido. Os limites para os valores da distância orbital que permitem uma temperatura superficial capaz de manter a água em estado líquido (0-100° C) definem a zona habitável desta estrela.

Em segundo lugar, a massa do planeta pode determinar a capacidade deste em manter uma atmosfera, uma vez que massas muito baixas proporcionam uma gravidade muito fraca. Em um planeta sem atmosfera, é improvável que a vida complexa possa evoluir na superfície. Além disso, sem atmosfera, a pressão é muito baixa e a água também não vai existir em estado líquido nesta superfície. Por outro lado, planetas muito massivos possuem gravidade tão alta que poderia impossibilitar a existência de vida na forma como conhecemos. A alta gravidade pode, por exemplo, sobrecarregar estes organismos ao submeter seus órgãos vitais a um excesso de pressão. Por último, a densidade pode nos ajudar a conhecer melhor a composição do planeta e se ele pode conter uma atmosfera significativa: planetas que não são muito densos têm maior probabilidade de terem uma atmosfera mais espessa, possivelmente até prejudicial para a vida, e suas composições são frequentemente de gelo e gás. Planetas densos são mais propensos a serem rochosos. Em contrapartida, estudos como o de Parkinson et al. (2007) sugerem que há evidências de que existam condições para a vida, abaixo da superfície, nos satétiles naturais Encélado e Europa, que não possuem atmosfera.

Além de $R_\star$, outros parâmetros estelares também podem contribuir para o entendimento das características dos planetas no sistema, como é o caso da correlação entre o raio planetário e a metalicidade da estrela hospedeira sugerida por Ghezzi et al. (2021). Eles compararam a distribuição de metalicidade para o disco fino da Galáxia (na vizinhança solar), e para estrelas hospedeiras segregadas, para entender a transição entre sub-Netunos e sub-Saturnos. Eles notaram que estas distribuições são cada vez mais diferentes à medida que o raio do maior planeta nos sistemas aumenta, principalmente para

planetas com $R_p > 2{,}7\ R_\oplus$. Além disso, eles também confirmaram que as distribuições gerais de metalicidade das estrelas hospedeiras variam entre sistemas planetários quentes e mornos de todos os tipos, sugerindo também uma correlação entre a metalicidade estelar e o período orbital. Outros estudos (e.g. Dong et al., 2018; Mulders et al., 2016; Petigura et al., 2018; Wilson et al., 2018, 2022) sustentam esta ideia, ao revelar a existência de correlação entre a metalicidade da estrela hospedeira e a presença de planetas quentes com períodos orbitais inferiores a 8-10 dias.

A melhora na precisão dos raios planetários revelou uma lacuna (*radius gap*) entre 1,5 e 2,0 $R_\oplus$, notada através dos dados do levantamento California-Kepler (Petigura et al., 2017), que revelaram uma distribuição bimodal para os raios de pequenos planetas (entre 0,7 e 3,5 $R_\oplus$; vide Figura 1.10; Fulton et al., 2017). O estudo de Fulton et al. (2017) sugere que esta lacuna pode ser explicada devido à perda de massa atmosférica impulsionada pela fotoevaporação por radiação estelar de alta energia, como a ultravioleta. A fotoevaporação é um processo em que a radiação energética ioniza o gás, fazendo com que ele se disperse para longe da fonte ionizante. Por exemplo, os fótons da radiação ultravioleta de uma estrela podem erodir a atmosfera planetária ao incidir sobre ela.

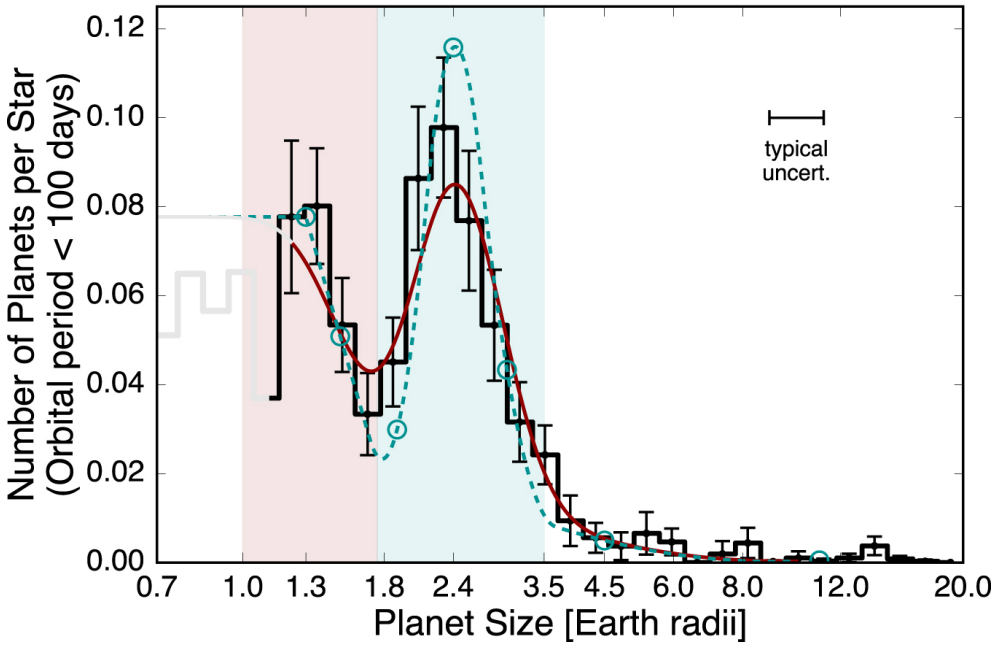


Figura 1.10: Histograma dos raios dos planetas com períodos orbitais menores que 100 dias. A região até 1,14 $R_\oplus$ sofre de baixa completeza. A linha sólida cinza é uma aproximação para os valores desta região. O modelo de melhor ajuste é representado pela linha sólida vermelha e não inclui a região branca à esquerda. A região sombreada corresponde a Super-Terras (rosa) e Sub-Netunos (azul). A linha pontilhada azul é um modelo proposto para a distribuição após considerar as incertezas nas medições dos raios. Alguns círculos azuis são marcados nesta linha e correspondem as posições dos nós. A lacuna (radius gap) está situada entre 1,5 e 2,0 $R_\oplus$. Fonte: Fulton et al. (2017)

O pico inferior da distribuição da Figura 1.10 ($\approx 1,3\ R_\oplus$) corresponde a planetas com núcleos rochosos que têm seu envoltório de H/He completamente removido por fotoevaporação ou que nunca tiveram uma atmosfera espessa. Como a radiação de alta energia diminui após os primeiros milhões de anos (Myr) da vida da estrela (para estrelas de tipo solar, isso ocorre após 100 Myr; Ribas et al., 2005), planetas maiores com envoltório de H/He mais espesso podem manter uma fração das suas atmosferas no momento em que a radiação de alta energia cessar, empurrando-os para o pico mais alto da distribuição de raios ($\approx 2,4\ R_\oplus$).

Nesta seção pudemos ver várias situações em que as análises se beneficiam de medidas mais precisas. A seguir apresentamos a motivação científica deste trabalho e nossas propostas para realizar medições mais precisas para parâmetros estelares.

1.4 Motivação científica

Vimos que, apesar do notável destaque da missão Kepler para a pesquisa de exoplanetas, os objetos detectados foram, em maioria, identificados através da observação contínua de "apenas" 150.000 estrelas presentes no seu campo de visão (cerca de 3%). Os 97% restantes correspondem às estrelas observadas pelas FFIs (estrelas fracas). As estrelas fracas da missão Kepler são, então, exemplos de objetos que não foram observados de forma contínua. Nogueira (2020) estudou as curvas de luz de estrelas fracas da missão Kepler e estimou o raio de objetos em trânsito, com base nos dados do catálogo de entrada do Kepler (Kepler Input Catalog - KIC; Latham et al., 2005). O KIC é um banco de dados público com informações para cerca de 13,2 milhões de objetos, entre os quais estão os $\approx 4,5$ milhões de alvos da missão Kepler (mais informações estão disponíveis na Seção 2.1).

Os dados de parâmetros físicos disponíveis no KIC não permitem uma caracterização precisa de objetos em trânsito, eventualmente detectados ao redor de suas estrelas. Isso ocorre por dois motivos. Primeiro, o KIC é uma compilação de diversos bancos de dados menores, disponibilizados por levantamentos astronômicos ao longo do tempo, cada um com seus respectivos objetos, métodos de cálculo e incertezas, não havendo homogeneidade nos dados. Brown et al. (2011) destacaram incertezas para os principais parâmetros físicos do KIC. Para a temperatura, elas podem chegar a $\pm$ 200 K para estrelas com $T_{\rm ef} < 7000$ K. Estrelas mais quentes podem alcançar incertezas tão grandes quanto 4000 K. Para log g, as incertezas podem chegar a 1,5 dex em estrelas gigantes. Para [Fe/H], essas incertezas são de $\pm$ 0,4 dex. O segundo motivo é que o KIC não fornece informações sobre os parâmetros físicos (temperatura efetiva, gravidade superficial, metalicidade e raio) para mais de 74% das estrelas da missão Kepler (3.299.079 objetos).

Como vimos na Seção 1.3, alguns parâmetros estelares são cruciais para a obtenção de certos parâmetros planetários. Diante disso, é necessário, antes de uma análise das

curvas de luz, que os parâmetros estelares sejam inferidos com um grau de confiança aceitável, com incertezas muito menores que as apresentadas pelo KIC, para o maior número de objetos possível. Breiman (2001) afirma que as técnicas de Aprendizagem de Máquina são capazes de realizar previsões para parâmetros físicos, desde que haja uma boa amostra de treinamento. De uma maneira geral, uma amostra de treinamento é a amostra oferecida a um algoritmo, baseado em Aprendizagem de Máquina, para que ele possa reconhecer padrões nela. O reconhecimento de padrões permite ao algoritmo identificar os valores esperados para objetos com as mesmas características dos objetos da amostra de treinamento. Para este trabalho, o algoritmo será capaz de reconhecer padrões de magnitude em estrelas com determinados parâmetros físicos. Essas magnitudes foram medidas pelo sistema de 12 filtros ópticos do *Javalambre Photometric Local Universe Survey*[9] (J-PLUS; Cenarro et al., 2019) e de 4 filtros do *Wide-field Infrared Survey Explorer* (WISE; Wright et al., 2010).

O uso de Aprendizagem de Máquina possibilitará tanto um ajuste nos parâmetros para os objetos da missão Kepler que já possuem dados disponíveis no KIC, como permitirá inferí-los, pela primeira vez, para as estrelas que ainda não os possuem. Isto permite uma melhor caracterização de objetos eventualmente encontrados ao redor delas. Além disso, será possível propôr um ajuste para o raio de alguns objetos em trânsito (R_p), caracterizados pela literatura.

1.5 Objetivos

Este trabalho tem o objetivo geral de caracterizar, com erro menor que o da literatura, as estrelas fracas da missão Kepler que foram observadas pelos levantamentos fotométricos de dados J-PLUS e WISE. Para cumprir este objetivo geral, os seguintes objetivos específicos devem ser atendidos:

- Identificar a melhor amostra de treinamento para o algoritmo;
- Calcular as cores para esta amostra, usando o sistema de filtros ópticos do J-PLUS e do WISE;
- Identificar e otimizar possíveis fatores de impacto, presentes no algoritmo de Aprendizagem de Máquina;
- Aplicar o algoritmo de Aprendizagem de Máquina nas estrelas fracas comuns entre J-PLUS/WISE e Kepler, a fim de mensurar parâmetros físicos tais como $T_{\rm ef}$, log g e [Fe/H];

[9] Mais informações sobre o J-PLUS estão disponíveis na Seção 2.2.

- Com os parâmetros físicos ($T_{\rm ef}$, log g e [Fe/H]) previstos pelo algoritmo, mensurar também luminosidades, raios e massas estelares.

Na continuação deste trabalho, no Capítulo 2, apresentamos a amostra de dados utilizada e como ela foi definida, bem como justificamos a escolha destes dados. No Capítulo 3 expomos a metodologia utilizada, detalhando o processo de Aprendizagem de Máquina. Nos Capítulos 4 e 5, respectivamente, discutimos os resultados alcançados e as conclusões obtidas a partir deles. Em seguida, estão as referências bibliográficas utilizadas neste trabalho. Por fim, apresentamos o Apêndice A, com um recorte da tabela de resultados finais, onde também fornecemos um endereço eletrônico de onde poderá ser acessada a sua versão completa.

Capítulo 2

Amostra de dados

Os dados utilizados nesta pesquisa provêm de diferentes levantamentos, sendo eles: um levantamento principal, de onde foram coletados os dados de magnitude em 12 filtros, o J-PLUS; um levantamento secundário (que chamaremos de levantamento auxiliar), de onde se seleciona dados de parâmetros físicos como temperatura efetiva ($T_{\rm ef}$), gravidade superficial (log g) e metalicidade ([Fe/H]); e o catálogo KIC, onde serão identificadas as estrelas observadas na missão Kepler e, nestas, aplicados os modelos de previsão calculados. Mais detalhes são apresentados nas seções seguintes. O objetivo de trabalhar com todos estes dados é criar uma modelagem para cada parâmetro, que permita calculá-los com maior precisão que aquela apresentada pelo KIC.

2.1 Kepler Input Catalog (KIC)

O catálogo de entrada do Kepler (Kepler Input Catalog - KIC) é um banco de dados público com informações para cerca de 13,2 milhões de objetos, e inclui dados como magnitude, temperatura efetiva, gravidade superficial, metalicidade, raio, massa, identificadores (nomes para o objeto), etc. O KIC foi criado porque nenhum catálogo reunia informações suficientes para a seleção de alvos do campo de visão da missão Kepler naquele momento. Nem todas as estrelas do KIC foram observadas pela missão Kepler, mas é possível, com auxílio da coluna "kct_num_season_onCCD", identificar os $\approx 4,5$ milhões de alvos da missão. Ela delimita as estrelas presentes em pelo menos uma orientação (*season*) do telescópio. Se isso não for satisfeito, significa que esta estrela não foi observada em nenhuma época, não fazendo parte do campo de visão do Kepler.

Existem aquelas estrelas que são designadas por KOI (*Kepler Object of Interest*). Todos os objetos classificados como KOIs são estrelas com sinais de trânsito, mas nem todas as estrelas do KIC - mesmo com trânsitos confirmados - possuem esta designação. Uma das razões para isto é que nem todas estas detecções foram feitas por membros da equipe Kepler, então não são incluídas oficialmente de forma imediata. Estas e outras informações estão disponíveis em Brown et al. (2011).

Os dados para parâmetros estelares, fornecidos pelo KIC, advém da literatura, onde não foram calculados de forma homogênea. Quando aplicados como informação de entrada para um algoritmo de Aprendizagem de Máquina, quase nenhum padrão de distribuição consistente é identificado. Geralmente isso ocorre devido a erros consideravelmente altos nas medidas (vide Figura 2.1).

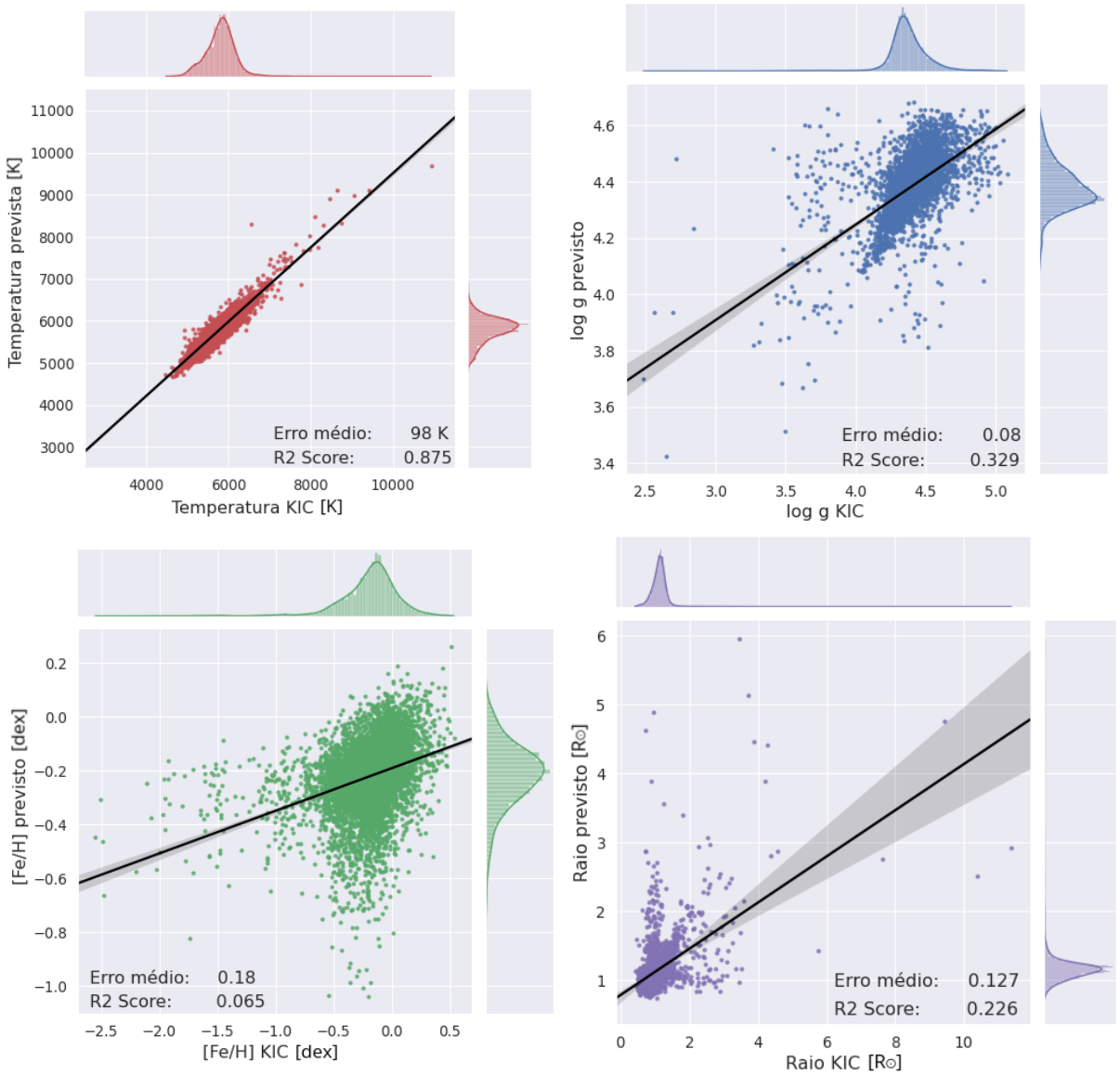


Figura 2.1: Simulação em Aprendizagem de Máquina (*Random Forest*) baseada nos filtros do J-PLUS, para estrelas do Kepler com parâmetros disponíveis no KIC. Desta forma, o algoritmo é alimentado com os dados do próprio KIC. Esta técnica se baseia no aprendizado com subamostras, contendo informações de frações aleatórias da amostra total, conhecidas como árvores. Aqui, utilizamos 100 árvores. Os baixos valores no R^2 *score* mostram que os dados do KIC não são adequados para este tipo de simulação. Isto geralmente se deve a erros consideravelmente altos nas medidas (apresentamos os erros do KIC na Subseção 2.1.2). Acima e a direita dos painéis, podemos ver a distribuição dos objetos da amostra de treinamento com relação ao valor do parâmetro.

Na Figura 2.1, podemos confirmar a falta de padrão de distribuição com o baixo valor de R^2 *score*. O R^2 *score* é um ajuste de um modelo estatístico linear generalizado, como a regressão linear, que representa o quadrado da correlação entre os valores observados de uma variável aleatória e os valores previstos (teóricos). Ele varia entre 0 e 1, onde 1 representa 100% e corresponde a uma correlação perfeita. Na figura, o R^2 *score* de log g (painel superior à direita), [Fe/H] (painel inferior à esquerda) e raio (painel inferior à direita) são bastante baixos, o que sugere que há um alto resíduo, facilmente notado na dispersão dos pontos nos painéis destes parâmetros. Este resíduo é a diferença entre o valor observado e o valor estimado. Para $T_{\rm ef}$ o resíduo é muito menor, como podemos ver pela baixa dispersão dos pontos, mas pode alcançar diferenças de até ≈ 1000 K.

Para entender como o cálculo do R^2 *score* é feito, precisamos calcular primeiro a soma total dos quadrados ($SQ_{\rm tot}$) das diferenças entre cada valor observado (y_i) e a média ($\bar{y}$) em um número n de observações:

$$SQ_{\rm tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2.1}$$

Também precisamos conhecer a soma dos quadrados dos resíduos ($SQ_{\rm res}$), que calcula a parte não explicada pelo modelo, com a soma dos quadrados das diferenças entre cada valor observado (y_i) e o valor estimado ($\widehat{y_i}$).

$$SQ_{\rm res} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{2.2}$$

Por fim, o R^2 *score* pode ser calculado através de:

$$R^2_{score} = 1 - \frac{SQ_{\rm res}}{SQ_{\rm tot}} \tag{2.3}$$

2.1.1 Sistema de magnitudes do KIC (K_p)

O KIC utiliza um sistema de magnitudes, denominado K_p, que foi obtido, principalmente, das magnitudes g, r e i do Sloan Digital Sky Survey (SDSS; Fukugita et al., 1996; Smith et al., 2002). Na prática, o processo de conversão de magnitudes entre sistemas fotométricos usa relações pré-determinadas para realizar a conversão, mas o processo total consiste em utilizar os valores de magnitudes do objeto celeste de interesse, ajustar um espectro teórico baseado neles e integrar este espectro sobre a curva de transmissão do instrumento para o qual deseja-se fazer a conversão. A Figura 2.2 mostra a curva de transmissão dos filtros do SDSS à esquerda e do Kepler à direita.

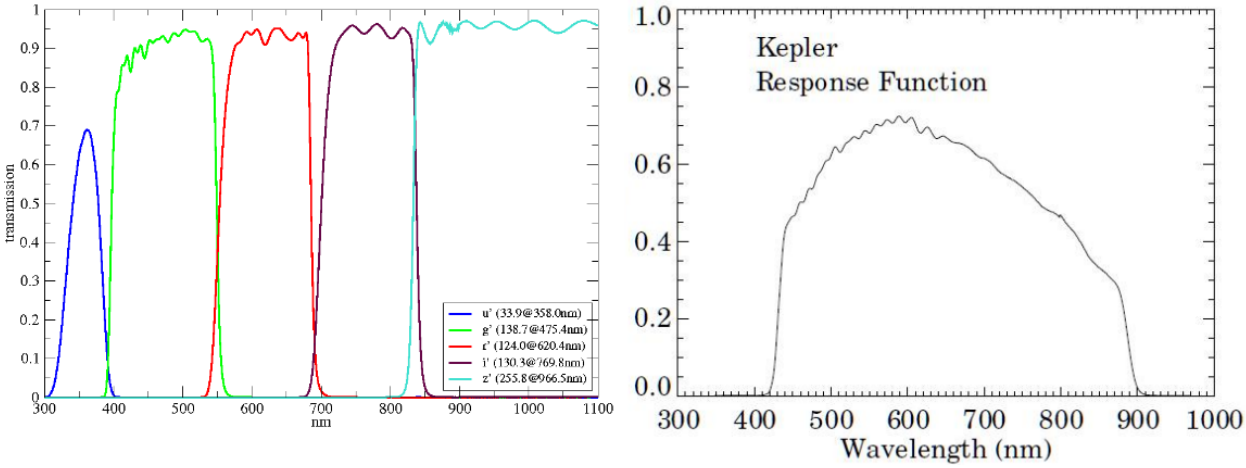


Figura 2.2: Curvas de transmissão dos filtros u (em azul), g (em verde), r (em vermelho), i (em roxo) e z (em ciano) do SDSS[10] à esquerda e do Kepler[11] (em cinza) à direita.

Brown et al. (2011) apresentam as relações matemáticas de conversão entre as magnitudes e discute incertezas de até 0,2 mag para K_p em estrelas de temperatura efetiva $T_{\rm ef} > 3500$ K com as 3 magnitudes do SDDS presentes - incertezas maiores são previstas por eles quando não existe valor para alguma das 3 magnitudes. O KIC não reúne estas incertezas. Quando há apenas uma das 3 magnitudes do SDSS citadas, ela é utilizada diretamente, sem qualquer conversão, por exemplo, caso haja apenas magnitude na banda r, $K_p = r$.

A magnitude K_p deve ser utilizada apenas para fins estatísticos. Ela é importante para estimar a precisão necessária do instrumento para se detectar trânsitos, já que quanto mais fracas as estrelas forem, mais difícil é detectar pequenas variações de brilho causadas por eles. Isso justifica o critério de seleção utilizado para as 150.000 estrelas de acompanhamento contínuo: apenas selecionou-se estrelas com $K_p < 16$, ou seja, estrelas razoavelmente brilhantes. Estrelas fracas são difíceis de acompanhar, já que sua fotometria é dependente de uma razão sinal-ruído particularmente alta.

2.1.2 Incertezas do KIC para os parâmetros de interesse

Brown et al. (2011) destacam os valores de incerteza na $T_{\rm ef}$ e log g do KIC, com base em uma comparação feita entre os seus valores e os obtidos espectroscopicamente, tomando o trabalho de Molenda-Zakowicz (2010) como referência. Para [Fe/H], a análise de Brown et al. (2011) considerou os dados obtidos por Fisher & Valenti (2005) com o pacote *Spectroscopy Made Easy* (SME; Piskunov & Valenti, 1996), uma ferramenta de análise para espectros estelares, escrita em 1996.

[10] Disponível em: https://old.aip.de/en/research/facilities/stella/instruments/data/sloanugriz-filter-curves

[11] Disponível em: https://www.nasa.gov/mission_pages/kepler

- **Temperatura efetiva:**

Um corpo negro é um objeto hipotético capaz de absorver toda a radiação eletromagnética incidente, não permitindo que nenhuma luz o atravesse ou seja refletida (corpo negro ideal). Idealmente, corpos negros emitem radiação na mesma proporção que absorvem. Ao gráfico da intensidade de emissão de radiação em função do comprimento de onda, damos o nome de curva de emissividade. Em uma aproximação, as estrelas podem ser consideradas como corpos negros. A temperatura efetiva de uma estrela é uma aproximação da temperatura na fotosfera estelar e pode ser estimada como a temperatura de um corpo negro que irradiaria a mesma quantidade total de fluxo que ela. Ao conhecermos a curva de emissividade de um corpo negro, é possível inferir sua temperatura efetiva, visto que, independentemente da sua composição, todos os corpos negros de mesma temperatura T emitem radiação térmica com mesmo espectro (vide Figura 2.3). Brown et al. (2011) apresentam as incertezas de temperatura efetiva no KIC. Para $T_{\rm ef} < 7000$ K, os valores do KIC e de Molenda-Zakowicz (2010) diferem em $\pm$ 200 K. Para temperaturas mais altas, os desvios são ainda maiores. Para a faixa de 9000 a 13500 K, as diferenças no KIC poder ser tão grandes quanto 4000 K.

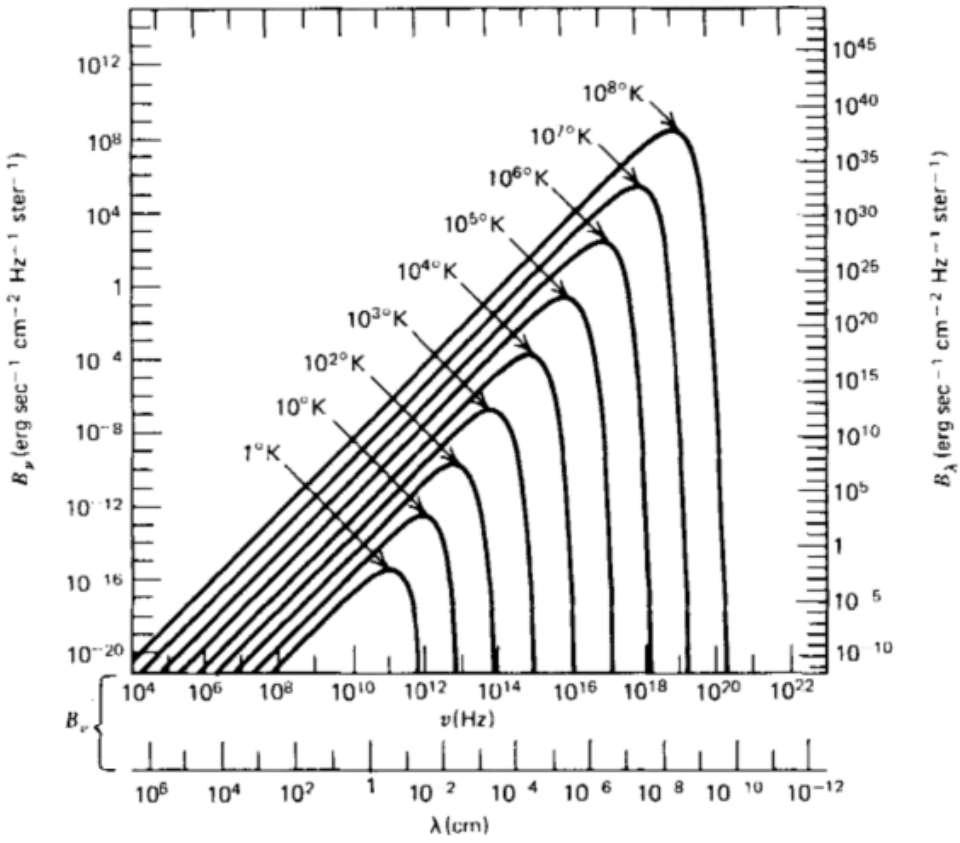


Figura 2.3: Espectro de radiação de corpo negro para várias temperaturas. Corpos negros de mesma temperatura emitem o mesmo espectro. Fonte: Rybicki & Lightman (2004).

- **Gravidade superficial:**

A gravidade superficial de um objeto astronômico (por exemplo, uma estrela) é a aceleração gravitacional experimentada em sua superfície no equador, considerando sua massa. Ela é medida em unidades de aceleração (no sistema CGS: cm/s^2), mas pode ser expressa como um múltiplo da gravidade superficial padrão da Terra ($g_{\text{Terra}} = 980{,}665$ cm/s^2) ou, mais frequentemente, como o logaritmo de base 10 da gravidade superficial da estrela (log g). Brown et al. (2011) afirmam que as estimativas de log g no KIC são satisfatórias para estrelas anãs, com precisão de $\approx 0,5$ dex. Para gigantes (incluindo aquelas com log g tão baixo quanto 1,5), as incertezas do catálogo podem chegar a 1,5 dex.

- **Metalicidade:**

As estrelas se formam da matéria presente em nuvens de gás interestelar (nuvens moleculares), que são compostas basicamente de hidrogênio, hélio e, também, poeira. Durante a vida da estrela, a atividade nuclear permite que elementos químicos mais pesados sejam formados a partir da fusão de elementos mais leves (nucleossíntese). A metalicidade é a proporção da matéria constituída destes elementos mais pesados, ou seja, dos elementos diferentes do hidrogênio e hélio. Em geral, ela também pode ser um indicativo da idade do objeto, onde quanto menor é esta proporção, mais velha é a estrela.

As primeiras estrelas do universo foram formadas do material pós-Big Bang e ficaram conhecidas como estrelas de População III. Por enquanto, elas são hipotéticas, pois ainda não conhecemos nenhuma estrela com metalicidade zero. Teoricamente, este material era livre de metais. Acredita-se que estas estrelas possuíam centenas de massas solares (Nakamura & Umemura, 2001) e, então, evoluíram muito rápido. Ao término da vida de uma estrela, ela libera parte de seu material (agora enriquecido pela nucleossíntese) de volta para o meio interestelar (enriquecendo o meio). Este material enriquecido vai, mais tarde, formar uma nova geração de estrelas. O processo se repete entre as gerações estelares. As estrelas mais jovens são mais ricas em metais e conhecidas como estrelas de População I.

Brown et al. (2011) também informam que, para [Fe/H], a comparação foi feita entre o KIC e os valores obtidos por Fisher & Valenti (2005) com o pacote *Spectroscopy Made Easy* (SME; Piskunov & Valenti, 1996), uma ferramenta de análise para espectros estelares, escrita em 1996. Os valores obtidos com o SME se basearam nos dados do espectrógrafo HIRES[12]. A cobertura de valores para [Fe/H], estimado pelo HIRES/SME, foi de cerca de 0,7 dex (-0,4 $<$ [Fe/H] $<$ 0,3 dex), onde a maioria das estrelas analisadas estavam agrupadas dentro de um intervalo ainda menor (-0,10 $<$ [Fe/H] $<$ 0,25 dex). A incerteza entre o KIC e o HIRES/SME foi de $\pm$ 0,4 dex.

[12]Para mais informações, consulte a página do HIRES: https://elt.eso.org/instrument/HIRES/

2.2 Levantamento de dados J-PLUS

O *Javalambre-Photometric Local Universe Survey* (J-PLUS; Cenarro et al., 2019), com início em 2018, consiste em um levantamento fotométrico de 8500 graus quadrados do halo da Galáxia, visto do hemisfério norte, e está utilizando um telescópio de 80 cm, instalado no Observatório Astrofísico de Javalambre (OAJ), em Teruel, Espanha. O OAJ conta com dois telescópios de campo de visão (*field of view*, FOV) amplo: o *Javalambre Survey Telescope* (JST/T250) e o *Javalambre Auxiliary Survey Telescope* (JAST/T80).

O JST/T250 é um telescópio de 250 cm de diâmetro com FOV de 3 graus de diâmetro, particularmente definido para grandes levantamentos como o *Javalambre Physics of the Accelerating Universe Astrophysical Survey* (J-PAS)[13], que fará observações com um conjunto óptico de 56 filtros. O JAST/T80, criado para realizar as calibrações necessárias para o J-PAS, foi posteriormente direcionado para um levantamento de dados paralelo, o J-PLUS (que utiliza um conjunto óptico de 12 filtros). O JAST/T80 possui um FOV de 2 graus de diâmetro e uma câmera de campo amplo equipada com um CCD de alta eficiência de 9200x9200 pixels de 10μm, instalada no seu foco Cassegrain (a T80Cam).

O J-PLUS, em sua segunda liberação de dados, DR2[14] (lançada em julho de 2020, com dados coletados de novembro de 2015 a fevereiro de 2020), cobriu uma área de 2176 graus quadrados, distribuída em 1088 campos espalhados pelo céu do hemisfério norte (vide Figura 2.4). O DR3 foi liberado em julho de 2022, mas não foi usado neste trabalho, por falta de tempo hábil. O J-PLUS está obtendo uma poderosa visão 3D do universo próximo, observando e caracterizando dezenas de milhões de galáxias e estrelas do halo da Via Láctea. Estas e outras informações estão disponíveis nas páginas da missão[15] e do próprio OAJ[16].

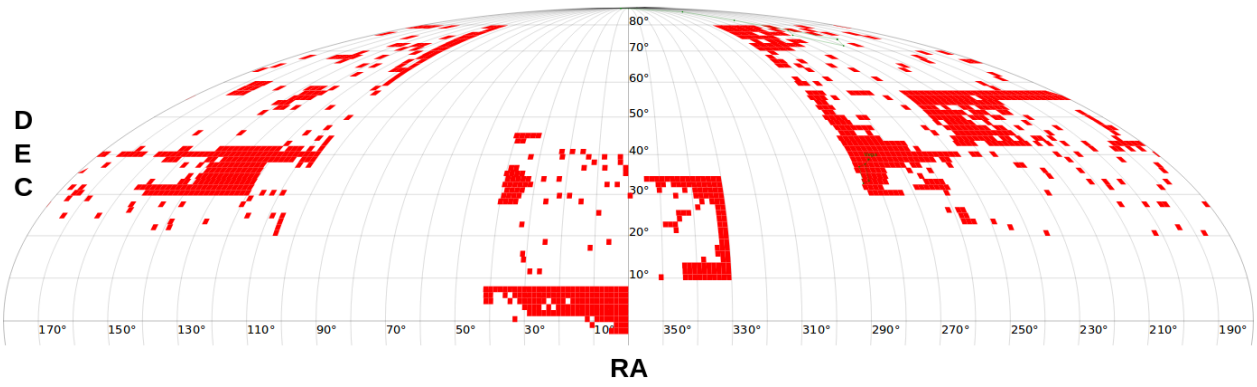


Figura 2.4: Distribuição da área observada pelo J-PLUS no DR2, em 1088 campos (quadrados/retângulos vermelhos) no céu do Hemisfério Norte. O eixo horizontal representa α (ascensão reta) e o vertical representa δ (declinação).[14]

[13] Mais informações em: http://www.j-pas.org/
[14] https://www.j-plus.es/datareleases/data_release_dr2
[15] https://www.j-plus.es/
[16] http://oajweb.cefca.es/

O sistema de 12 filtros ópticos (8 estreitos e 4 largos) possui uma cobertura espectral de 3500 Å a 10000 Å: uJAVA, J0378, J0395, J0410, J0430, gSDSS, J0515, rSDSS, J0660, iSDSS, J0861 e zSDSS. Alguns filtros são comuns a outros levantamentos: três são comuns ao J-PAS (uJAVA, J0378 e J0660) e quatro ao SDSS (e que tem o seu acrônimo). Os demais filtros analisam assinaturas espectrais específicas, como linhas de O II, Ca H e K, banda G, tripletos de cálcio e magnésio e linhas de H_α e H_δ. A magnitude limite para os diferentes filtros (assumindo uma razão sinal/ruído ≈ 3, uma abertura de 3" e uma PSF de 1,2" em uma noite não ideal) varia de 20,35 a 21,77 (vide Figura 2.5). Devido à turbulência atmosférica, geralmente a imagem produzida por um telescópio não é ideal, ou seja, não é um disco de difração de Airy (disco até o primeiro mínimo de difração), degradando a resolução. A Função de Espalhamento Pontual (*Point Spread Function*, PSF) é a função que descreve a distribuição de luz produzida por uma imagem pontual, no plano da imagem, devido à atmosfera, e sua largura mede a resolução real da imagem.

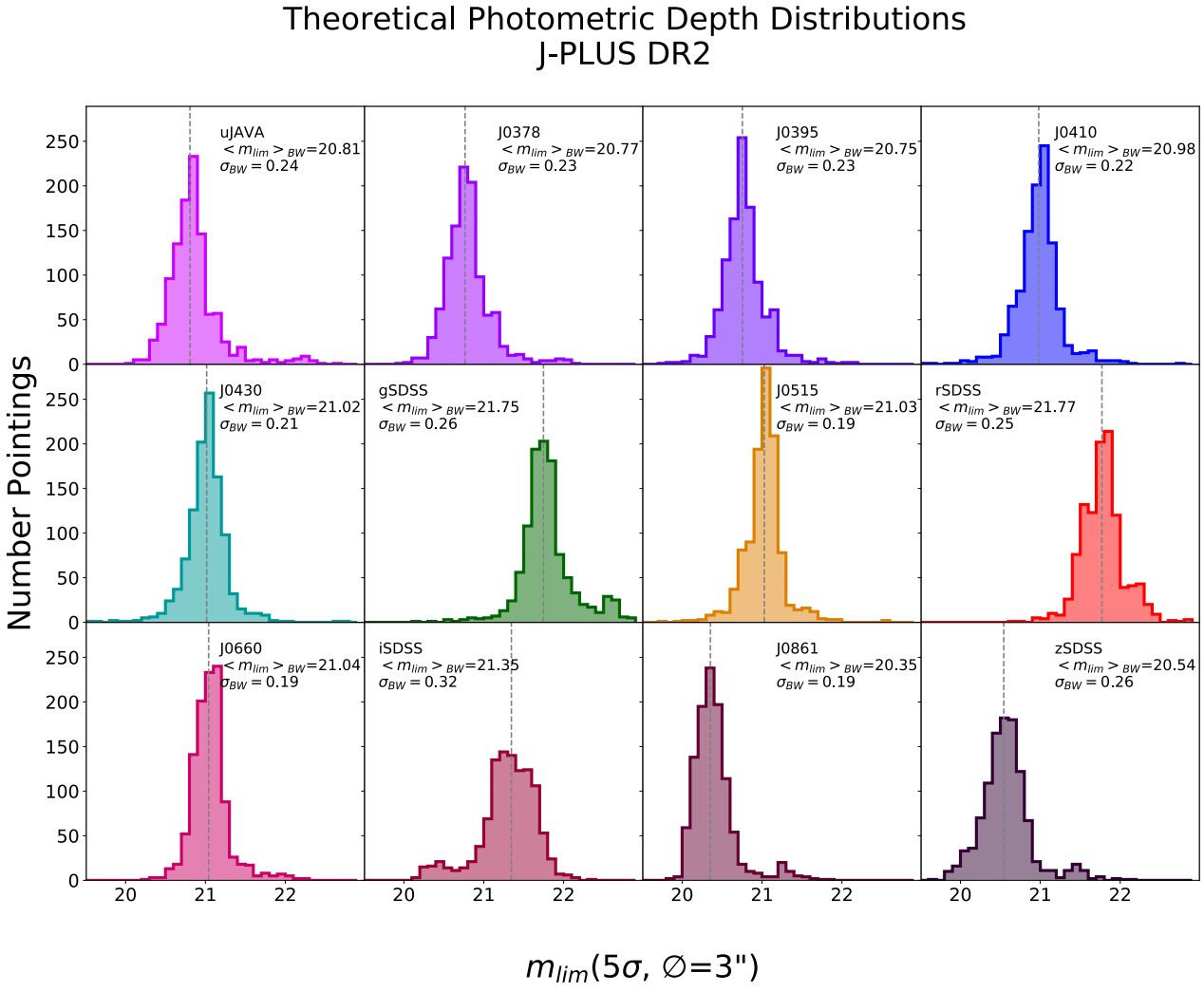


Figura 2.5: Distribuição da magnitude limite (m_{lim}) teórica de cada filtro associado, calculada a partir do desvio padrão do ruído de fundo (σ_{bg}) dentro de uma abertura de 3" e considerando um limite de fluxo de $5\sigma_{bg}$.[14]

A Figura 2.5 mostra a magnitude limite teórica (m_{lim}) para cada um dos filtros, calculada a partir do desvio padrão do ruído de fundo (σ_{bg}) dentro de uma abertura de 3" e considerando um limite de fluxo de $5\sigma_{bg}$. Cada subgráfico possui uma anotação no canto superior direito que informa o nome do filtro e o valor exato da m_{lim} correspondente a ele. A Figura 2.6 mostra as curvas de transmissão deles.

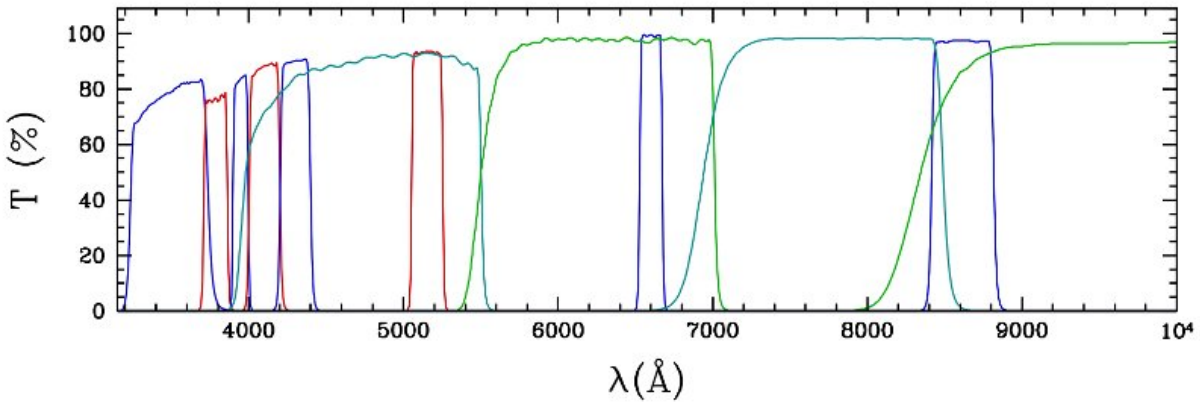


Figura 2.6: Curvas de transmissão para os filtros J-PLUS, na sequência em ordem crescente de comprimento de onda: uJAVA, J0378, J0395, J0410, J0430, gSDSS, J0515, rSDSS, J0660, iSDSS, J0861 e zSDSS.[15]

Nem todos os objetos da missão foram observados nos 12 filtros. Para este trabalho, é importante ter uma amostra sem lacunas, portanto com as doze medições. A seleção de dados, feita no ambiente interno do usuário, permite incluir esta e outras restrições que serão apresentadas pela Subseção 2.2.1. É possível treinar um modelo inteligente, baseado em algoritmo de Aprendizagem de Máquina, para estrelas com 11 medições. Tal algoritmo é programado para prever a magnitude do objeto no filtro faltante (ver Subseção 3.2.2) e então analisar se a amostra resultante, com adição da magnitude sintética, produz relevância para a amostra de treinamento. Caso possua, as estrelas com medições de magnitude nos 11 filtros + magnitude sintética podem ser adicionadas à amostra de interesse.

2.2.1 Seleção de dados

Nesta etapa, refinaremos o que a amostra de interesse do J-PLUS conterá. Alguns requisitos mínimos devem ser atendidos na seleção de objetos, de forma a passar informações suficientes para o algoritmo de previsão. Atender estes requisitos impacta diretamente na acurácia do modelo de previsão de parâmetros. Os objetos da amostra de interesse do J-PLUS mais restrita deverão:

- Ter sido observados com abertura circular fixa de 6";
- Ter sido observados por todos os doze (12) filtros do levantamento;

- Possuir erro na magnitude menor que 0,1 (e_mag $<$ 0,1) em todos os 12 filtros do J-PLUS;
- Possuir mais de 90% de probabilidade de ser uma estrela (prob_star $>$ 0,9);
- Possuir valor para correção de extinção interestelar calculado.

Note que o limite de erro de magnitude será aplicado para todos os doze filtros do J-PLUS, ou seja, o erro de magnitude em nenhum dos doze filtros deve ser maior que este limite. A fim de reduzir outras possíveis contaminações, a prob_star foi confirmada com os dados de paralaxe do Gaia, o que permitiu avaliar a distância em que o objeto observado se encontrava e que a sua classificação como estrela tinha fundamento. Todos os objetos da amostra passaram por esta avaliação e foram confirmados como estrelas (todos a uma distância $<$ 5 Kpc).

Também serão testados outros requisitos menos restritivos (amostra J-PLUS menos restrita) como:

- Ter sido observados com abertura circular fixa de 3";
- Ter sido observados por onze (11) filtros do levantamento;
- Possuir erro na magnitude menor que 0,2 (e_mag $<$ 0,2) em todos os 12 filtros J-PLUS.

A aplicação da segunda listagem de requisitos permite que objetos com 0,1 $<$ e_mag $<$ 0,2 sejam acrescentados a amostra. Ela será incluída apenas se comprovarmos que este erro adicional não prejudica significativamente as acurácias obtidas pelo algoritmo. Note que a segunda listagem inclui o teste para utilização de estrelas observadas por onze filtros, através da geração de um 12° filtro sintético.

Reduzir o erro na magnitude para menor que 0,1 melhora a confiança das previsões, porém reduz a amostra de treinamento (o que pode reduzir esta confiança), então seus prós e contras também serão avaliados. Para fins de análise, também podem ser testados objetos com erro de magnitude menor que 0,3, embora seja provável que estes objetos adicionem um ruído insatisfatório à amostra. Ressalta-se que, em função de todos os parâmetros listados serem modificáveis, sempre serão analisadas qualidade e quantidade, visto que o objetivo deste trabalho é o de calcular parâmetros estelares com o maior grau de confiança possível, e amostras de treinamento pequenas podem trazer instabilidade para a modelagem.

Entre os dados retornados nesta subseção, estão as colunas de correção de extinção, rotuladas no formato "Ax_" + nome do filtro. É possível usar estas colunas para corrigir os valores fornecidos pelas colunas de magnitudes para cada filtro (medida de magnitude do filtro + correção). Assim, cada medida de magnitude, em cada filtro, para cada objeto, estará corrigida dos efeitos de extinção.

2.3 Levantamentos de dados auxiliares

Como o J-PLUS não estima parâmetros físicos como temperatura efetiva, gravidade superficial, metalicidade, raio, luminosidade e massa, é necessário que se faça uso do banco de dados de outro telescópio, que tenha estes dados calculados com bom grau de confiança - a estes, daremos o nome de levantamentos de dados auxiliares. Vários levantamentos disponibilizam seus dados de forma pública, o que permite testá-los e avaliar a acurácia de previsão que eles proporcionam.

Para eleger um levantamento de dados auxiliar, é preciso primeiramente identificar possíveis candidatos que tenham observado regiões do céu em comum com o J-PLUS (de onde se utilizará os filtros). Isso permite que eles tenham objetos observados em comum, o que é fundamental para a modelagem. Quanto maior o número de objetos, maior é a quantidade de informação fornecida ao modelo. Pensando nesta informação fornecida, resolvemos, posteriormente, adicionar também as magnitudes medidas por quatro filtros do WISE (W1, W2, W3 e W4), para os objetos pré-selecionados do J-PLUS. Os objetos em comum entre J-PLUS/WISE e o levantamento de dados auxiliar podem ser encontrados a partir da correlação cruzada dos catálogos, usando os parâmetros astrométricos de ascensão reta (α) e declinação (δ).

A lista de objetos em comum entre o levantamento auxiliar e o J-PLUS/WISE forma uma amostra de treinamento. Após isso, um *script* é escrito para testar cada amostra de treinamento. O algoritmo permite dividir esta amostra em treino e teste, na taxa que o usuário preferir. Isto possibilita que sejam realizados testes por um número ilimitado de vezes, direcionando diferentes porcentagens da amostra para treino/teste e facilitando definir quais percentuais retornam os melhores resultados. Mais detalhes sobre como todo este processo ocorre são dados a seguir.

Capítulo 3

Metodologia

Após a seleção de dados (descrita na Subseção 2.2.1) e posterior correção de extinção destes dados, dividiu-se a metodologia deste trabalho em fases, agrupadas entre: 1) otimização de hiperparâmetros (apresentada ao final da Seção 3.1); 2) testagem do algoritmo (descrita na Seção 3.2) e 3) seleção das estrelas alvo, onde as modelagens da fase 2 serão aplicadas (vide Seção 3.3). Os cálculos de outros parâmetros físicos estelares (luminosidade, raio e massa) serão realizados com base nos resultados da fase 3 (nas Seções 3.4 e 3.5). Abaixo, descrevemos como cada fase funciona e o que esperamos alcançar em cada uma delas.

3.1 Aprendizagem de Máquina (*Machine Learning*)

A Aprendizagem de Máquina (*Machine Learning* ou AM) é uma subárea da inteligência artificial e da evolução de algoritmos computacionais projetados para emular a inteligência humana, onde a máquina aprende com o ambiente circundante. As técnicas de AM são altamente aplicáveis para qualquer projeto que utilize reconhecimento de padrões. O grau de complexidade desse processo pode variar e envolver estágios de tomadas de decisão, onde algoritmos baseados neste tipo de inteligência auxiliam para otimizar e automatizar este processo (Yao & Liu, 2013).

Estes algoritmos possuem a capacidade de aprender a partir de um contexto e generalizar para situações semelhantes. As tomadas de decisão na AM são baseadas nas chamadas árvores de decisão. As árvores de decisão utilizadas para um modelo são divididas em dois tipos: árvores de classificação e de regressão (definiremos cada tipo a seguir, na Subseção 3.1.1). Seu uso depende do objetivo da análise. A forma como a máquina aprenderá com os dados é baseada no tipo de árvore que o usuário definiu. Cada árvore contribui fornecendo um argumento sobre determinada amostra (Yao & Liu, 2013).

3.1.1 *Features* e árvores de decisão

Os algoritmos de AM são divididos em 3 tipos principais de aprendizagem: supervisionada, semisupervisionada e não supervisionada. O tipo supervisionado é usado principalmente para treinamentos guiados, onde o usuário auxilia o algoritmo, fornecendo informações relevantes sobre uma amostra que pode possuir características confusas. Essas informações recebem o nome de *features*[17]. Quanto melhor for o conjunto de *features* fornecido a um treinamento, melhor é a previsão de um modelo. A Figura 3.1 ilustra um exemplo clássico. Como a amostra é pura (ou seja, só contém um tipo de dado), o modelo retorna sua conclusão de previsão sem nenhuma dificuldade.

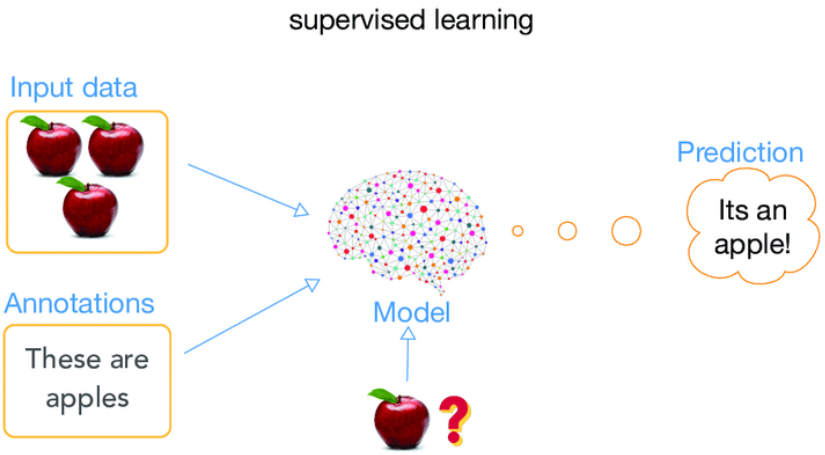


Figura 3.1: Exemplo de classificação com árvores de decisão em aprendizagem supervisionada. Nele, vemos os dados de entrada (*input data*) e a *feature* "*these are apples*"[18]. Os dados de entrada formam uma amostra pura (ou seja, que só contém um tipo de dado), então o algoritmo não enfrenta dificuldades no processo de previsão. Fonte: Yan et al. (2018)

Suponha que a *feature* empregada neste exemplo fosse o tamanho. Para isso, poderíamos definir um intervalo númerico que descrevesse o tamanho médio de uma maçã. Caso uma nova fruta de tamanho semelhante seja inserida no exemplo da Figura 3.1, sem informar nenhuma nova *feature* ao modelo, ele continuará fazendo previsões com base no tamanho, o que fará com que seja previsto que a nova fruta é uma maçã. Caso seja adicionada uma *feature* de cor ao modelo e a nova fruta tiver uma cor diferente de vermelho, as árvores poderão realizar a previsão corretamente baseadas na nova informação. Ainda assim, dificilmente teremos tantas informações que nos permitam construir um algoritmo 100% eficaz, principalmente quando desejamos fazer previsões em um conjunto de objetos com características muito semelhantes. O erro do algoritmo, que é influenciado pela

[17]Termo utilizado em AM que define uma propriedade individual mensurável ou característica de um objeto. Por este motivo, manteremos sua grafia original, em inglês.

[18]Geralmente *features* são númericas, mas strings também são aceitas em casos com árvores de classificação.

qualidade das informações fornecidas pelo conjunto de *features*, é o que conhecemos como função de perda e caracteriza a perda esperada para a previsão.

Como mencionado, as árvores de decisão são divididas em árvores de classificação e regressão. A principal diferença entre elas é que árvores de classificação são baseadas na moda de uma amostra e árvores de regressão em sua média. Podemos analisar isto em um exemplo. Considere uma estrela para a qual foram realizadas 3 medições distintas da temperatura efetiva: 5100, 5100 e 5400 K. Para este caso, uma árvore de classificação reconhece a moda de 5100 K e retorna este valor como previsão da temperatura da estrela e pode aplicá-lo para estrelas com características físicas semelhantes. Porém, o uso da moda desconsidera o valor de 5400 K, o que pode gerar uma incerteza máxima de 300 K. Uma árvore de regressão não desconsidera nenhum valor. Ao invés disso, ela realiza a média ponderada dos valores de temperatura para a sua previsão, encontrando o valor de 5200 K que, embora não esteja entre as medições desta estrela, oferece uma incerteza máxima menor (200 K). Árvores de decisão estão presentes em algumas ferramentas da AM, como a técnica *Random Forest*, utilizada neste trabalho.

3.1.2 *Random Forest*

Random Forest (Floresta Aleatória) é uma ferramenta de algoritmos de Aprendizagem de Máquina baseada em preditores de árvores de decisão. A amostra é inicialmente dividida entre treino e teste pela variável de percentual de teste, com valor definido pelo usuário. Um percentual de teste de 30% fará com que uma amostra aleatória de 70% dos dados de entrada seja destinada para o treinamento do algoritmo (percentual de treino).

Cada árvore da floresta recebe as informações desse percentual de treino. O *Random Forest* generaliza os padrões encontrados na amostra de treino, realiza previsões para os 30% que foram destinados para teste e compara os valores reais e previstos. O erro da previsão final dependerá do peso dos valores das árvores individuais e da correlação entre elas. As estimativas internas monitoram o erro, o peso e a correlação e são usadas para mostrar a resposta ao aumento do número de recursos usados - quanto mais informação fornecida, melhores serão as estimativas. O número de árvores de decisão é definido dentro do código (Breiman, 2001).

Uma árvore de decisão é formada por nós internos (na Figura 3.2, representados pelos círculos cinza), nós terminais (retângulos azuis) e arestas (linhas cinza) que conectam todos os nós de forma hierárquica. Cada nó interno tem uma borda de entrada e duas de saída e armazena um teste. Após o teste, os dados são enviados ao longo da borda de saída que corresponde ao seu resultado. Os nós terminais (ou folhas) são aqueles que armazenam o preditor que relaciona os dados de entrada à resposta final.

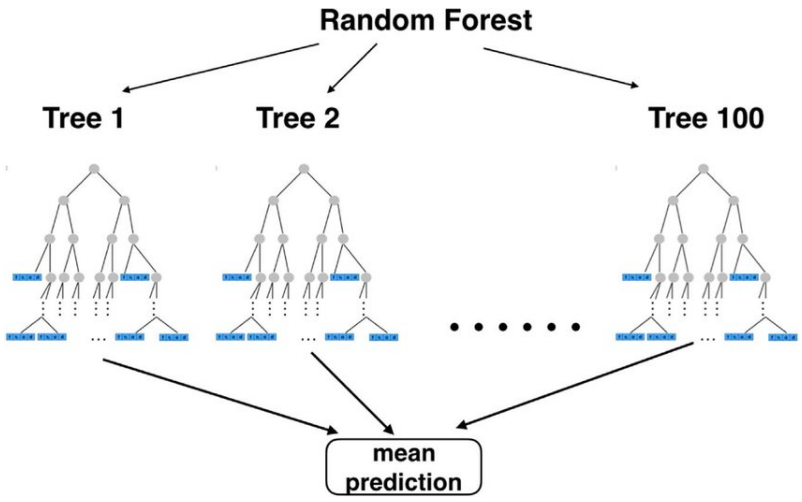


Figura 3.2: *Random Forest* formado por uma coleção de árvores de decisão de regressão. Cada árvore é formada por nós internos (círculos cinza), nós terminais (retângulos azuis) e arestas (linhas cinza) que conectam os nós de forma hierárquica. Cada nó interno tem uma borda de entrada e duas de saída e armazena um teste que, após a execução, define ao longo de qual borda de saída os dados serão enviados. O resultado final é a média ponderada da estimativa de cada árvore de decisão individual. Fonte: Nedjati-Gilani et al. (2017)

3.1.3 Amostra de treinamento

Uma amostra de treinamento é aquela amostra oferecida à maquina para que ela possa aprender e aplicá-la a um algoritmo de previsões. É importante que a amostra de treinamento apresente colunas de identificação do objeto, possíveis *features* e dados amostrais bem calculados dos parâmetros que se quer treinar a máquina para prever, bem como quaisquer informações que o usuário considerar relevantes para a previsão. Para este trabalho (e similares, na área de astronomia), é fundamental garantir que parâmetros como ascensão reta e declinação estejam dentro da pré-amostra de treinamento, já que os identificadores (nome do objeto observado) geralmente são diferentes de missão para missão, embora haja objetos em comum entre elas.

3.1.4 Hiperparâmetros

Vimos que, na AM, os valores dos parâmetros são estimados por meio do treinamento, com o uso de amostras de treinamento. O objetivo final de um algoritmo de aprendizado típico é encontrar uma função f que minimize a perda esperada - isto é, que diminua a dispersão entre o valor observado e o valor previsto, já que o algoritmo reconhece que é quase impossível que a amostra de treinamento contenha todas as informações necessárias para realizar uma previsão idêntica aos valores observados. Este algoritmo de aprendizagem mapeia um conjunto de dados (X_train) para esta função. Cada ferramenta de AM (no nosso caso, *Random Forest*) possui um conjunto de parâmetros padrão (θ), onde os

parâmetros deste conjunto possuem valores pré-determinados. Um exemplo de parâmetro do conjunto padrão é o número de iterações, que define a quantidade de repetições que o algoritmo deve realizar antes de entregar a previsão final (seu valor padrão é 100). Frequentemente, f é produzida por meio da otimização dos critérios de treinamento com relação a θ.

Também é comum que algoritmos de aprendizagem possuam um conjunto de hiperparâmetros (λ). O algoritmo de aprendizagem real é aquele obtido após a escolha de λ (Bergstra & Bengio, 2012). Um hiperparâmetro é um parâmetro cujo valor é usado para controlar o processo de aprendizagem. Diante deste controle, os hiperparâmetros devem então ser definidos antes da aplicação do algoritmo de aprendizagem real, a fim de evitar um viés que conceda uma configuração aparentemente otimizada, mas que só atende à amostra de treinamento. Algoritmos de treinamento com objetivos diferentes exigem configurações de hiperparâmetros diferentes (Hutter et al., 2009). Um exemplo de hiperparâmetro do conjunto λ é o número de árvores de decisão (seu valor padrão é 20).

Hutter et al. (2009) classificam os hiperparâmetros em dois tipos: hiperparâmetros do modelo e hiperparâmetros do algoritmo. No primeiro caso, referem-se a parâmetros que não devem ser inferidos durante o ajuste da máquina à amostra de treinamento, ou seja, devem ser definidos antes do treinamento (por exemplo, o número de árvores). No segundo, possuem influência direta na velocidade de aprendizagem. Por exemplo, temos o min. samples leaf (msl), que representa a amostra mínima permitida por folha e tem valor padrão igual a 1. O processo de otimização seleciona a configuração que produz um modelo "ideal", que minimiza a função de perda (*loss function*). Este processo precisa ser reprodutível em um grande número de sementes aleatórias (*random seed*).

Uma semente aleatória é um parâmetro da AM que permite a seleção de uma subamostra de treinamento aleatória. Ela é usada para garantir que os resultados sejam reprodutíveis, ou seja, que qualquer pessoa que execute novamente seu código obterá as mesmas saídas. Por exemplo, um percentual de treino de 70% em uma amostra de 1000 objetos, seleciona uma subamostra de treinamento aleatória de 700 objetos. A aleatoriedade dos dados se deve à presença do parâmetro de semente aleatória. Caso contrário, cada vez que o código fosse reproduzido, ele selecionaria os mesmos 700 objetos.

O número de hiperparâmetros é normalmente pequeno ($\leqslant 5$), mas pode variar até centenas para algoritmos de aprendizagem complexos. Bergstra & Bengio (2012) demonstraram que, em muitos casos, apenas alguns hiperparâmetros afetam significativamente a aprendizagem. Van Rijn & Hutter (2018) estudaram vários destes hiperparâmetros para três técnicas de aprendizagem de máquina: *Support Vector Machines* (SVM), *Random Forest* e *Adaboost*. Geralmente, estes hiperparâmetros estão embutidos dentro do código. A variação dos seus valores padrão, com o objetivo de melhorar a aprendizagem de um algoritmo, é chamada de otimização de hiperparâmetros. A Tabela 3.1 mostra os intervalos de valores testados pelos autores, para cada hiperparâmetro, para o *Random Forest*.

Tabela 3.1: Intervalo de valores testados na otimização de hiperparâmetros do Random Forest para a amostra de Van Rijn & Hutter (2018). As melhores configurações foram usadas na Figura 3.3.

Hiperparâmetro	**Valores**
bootstrap	[true, false]
max. features	[0.1, 0.9]
min. samples leaf	[1, 20]
min. samples split	[2, 20]
imputation	[mean, median, mode]
split criterion	[entropy, gini]

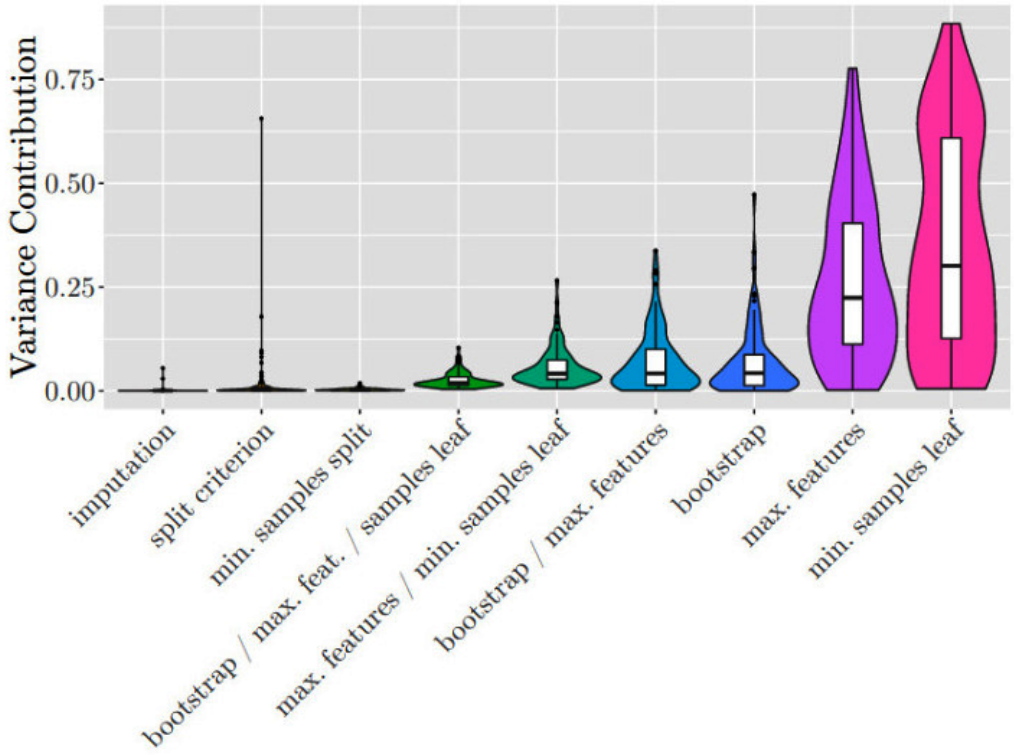


Figura 3.3: Hiperparâmetros otimizados de maior impacto em *Random Forest*, nos conjuntos de dados de Van Rijn & Hutter (2018). A análise mostra destaque para os hiperparâmetros *min. samples leaf* e *max. features*. *Imputation* se refere a estratégia usada para atribuir valores a variáveis numéricas ausentes (moda, média ou mediana da amostra). *Split criterion* é a função usada para determinar a qualidade da divisão. *Min. samples split* é a amostra mínima em um nó interno. *Bootstrap* permite que o algoritmo treine com múltiplos subconjuntos de dados com substituição (isso significa que podemos selecionar o mesmo valor várias vezes). *Max. features* é o número máximo de *features* em cada divisão. *Min. samples leaf* (msl) é a amostra mínima por folha.

A Figura 3.3 mostra os resultados para nove hiperparâmetros (ou combinações de hiperparâmetros) na técnica de *Random Forest*, também utilizada no presente trabalho, testados por Van Rijn & Hutter (2018). Nesta figura, os resultados mostram que a maior parte da variância na aprendizagem é atribuída a um par de hiperparâmetros: *min. samples leaf* (msl; amostra mínima por folha) e *max. features* (número máximo de *features* em cada divisão).

Vimos acima que algoritmos de treinamento com objetivos diferentes requerem configurações diferentes para os hiperparâmetros. Os testes foram refeitos por Cordeiro (2022, in prep.)[19], para o mesmo objetivo do algoritmo deste trabalho, a fim de encontrar tal configuração ideal para cada um dos parâmetros que se espera estimar com a técnica de *Random Forest*. O processo de otimização foi realizado nos dois hiperparâmetros mais relevantes, apresentados por Van Rijn & Hutter (2018) na Figura 3.3. Cordeiro (2022, in prep.) também analisou o impacto do número de *features* (*n_features*) e de árvores de decisão (*n_trees*) na aprendizagem do algoritmo, concluindo que um maior número de árvores revela mais informações para o algoritmo, porém a partir de um determinado valor, nenhuma informação relevante é acrescentada, entregando apenas mais das mesmas informações que o algoritmo já aprendeu. Usar todas as *features* se revelou também desnecessário, já que algumas delas podem fornecer apenas informações irrelevantes. O autor realizou a otimização por meio de 5 validações cruzadas (cada uma repetida 3 vezes), em 75% da amostra inicial. No total, foram testadas 100 combinações diferentes de hiperparâmetros. As combinações usaram os dois tipos de hiperparâmetros (do modelo e do algoritmo).

A Figura 3.4 mostra a variação do R^2 *score* para temperatura efetiva com a otimização. No painel superior, vemos que o melhor modelo é o que possui 45 *features*, *max_features* = 0,25 e 100 árvores de decisão - a diferença entre ele o segundo melhor modelo está na quarta casa decimal. Podemos escrever essa configuração como (45, 0,25, 100). O valor de *max_features* corresponde a fração das *features* que será passada à cada uma das árvores de decisão. Esta fração é baseada em uma amostra aleatória, o que resulta em árvores de decisão diferentes. Apesar de parecer contraditório, o *Random Forest* aprende melhor com este tipo de árvore. Árvores distintas passam informações distintas, fazendo com que o algoritmo capte detalhes que poderiam ser omitidos por um conjunto de árvores idênticas. No painel superior da Figura 3.4, não foi explicitado no código um valor para o hiperparâmetro msl, então o algoritmo utiliza o valor padrão (msl = 1). No painel inferior, comparamos o rendimento do modelo com um valor diferente para este hiperparâmetro (msl = 10). Esta alteração produz uma queda no R^2 *score* do modelo, melhor notado para log g e [Fe/H] (vide painel inferior das Figuras 3.6 e 3.8). O círculo amarelo destaca o modelo com maior R^2 *score* e, portanto, com a melhor combinação de hiperparâmetros para a nossa amostra.

[19] https://github.com/cordeirossauro/JPLUS-STEPEs

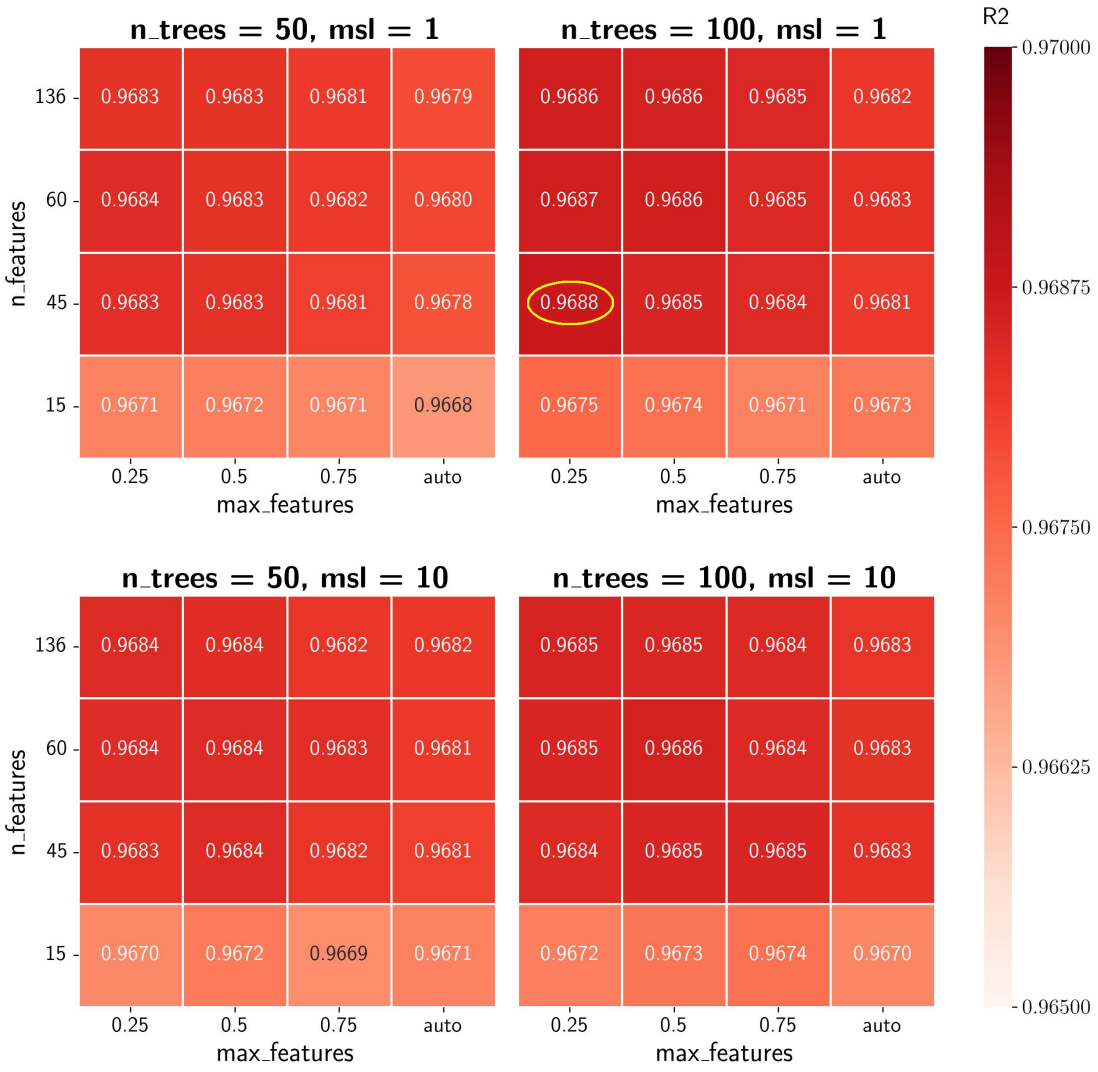


Figura 3.4: Variação do R^2 *score* para $T_{\rm ef}$, quando se varia *n_features*, *max_features*, *n_trees* e msl. No painel superior, mantemos o valor padrão de msl (msl = 1) e, no painel inferior, usamos msl = 10. Msl = 10, em geral, provoca uma queda no rendimento do modelo (melhor notada para log g e [Fe/H] nos painéis inferiores das Figuras 3.6 e 3.8), então manteremos msl = 1. O melhor modelo, por uma diferença na quarta casa decimal, possui 45 *features*, *max_features* = 0,25 e 100 árvores de decisão - podemos escrever essa configuração como (45, 0,25, 100). Fonte: Cordeiro (2022, in prep.)

Apesar de msl = 10 impactar positivamente no tempo de processamento, tornando o algoritmo mais rápido (vide Figura 3.5), ele causa uma queda na acurácia do modelo e isto não é interessante. Esta queda ocorre porque o algoritmo é limitado a usar objetos que tenham um número mínimo de pontos de dados em cada folha (no caso, 10). Qualquer objeto que não atenda isto é descartado e a amostra fica menor (amostras menores são mais rápidas de processar), porém, menos dados são fornecidos para a geração do modelo. Optou-se por manter o valor padrão de msl. A Figura 3.5 mostra também que um *max_features* menor (mais próximo de 0 que de 1) e um menor número de *features* também causa

este efeito. Como visto, *max_features* menor gera árvores diferentes, com informações relevantes. O menor número de *features* permite descartar aquelas irrelevantes (como visto na Figura 3.4, após 45 *features* quase não há variação no R^2 *score*). Mais árvores de decisão produzem melhor (ou igual) R^2 *score*, mas não existe, para a temperatura efetiva, variação significativa entre 50 e 100 árvores. Além disso, vemos na Figura 3.5 que um maior número de árvores eleva o tempo de processamento do algoritmo.

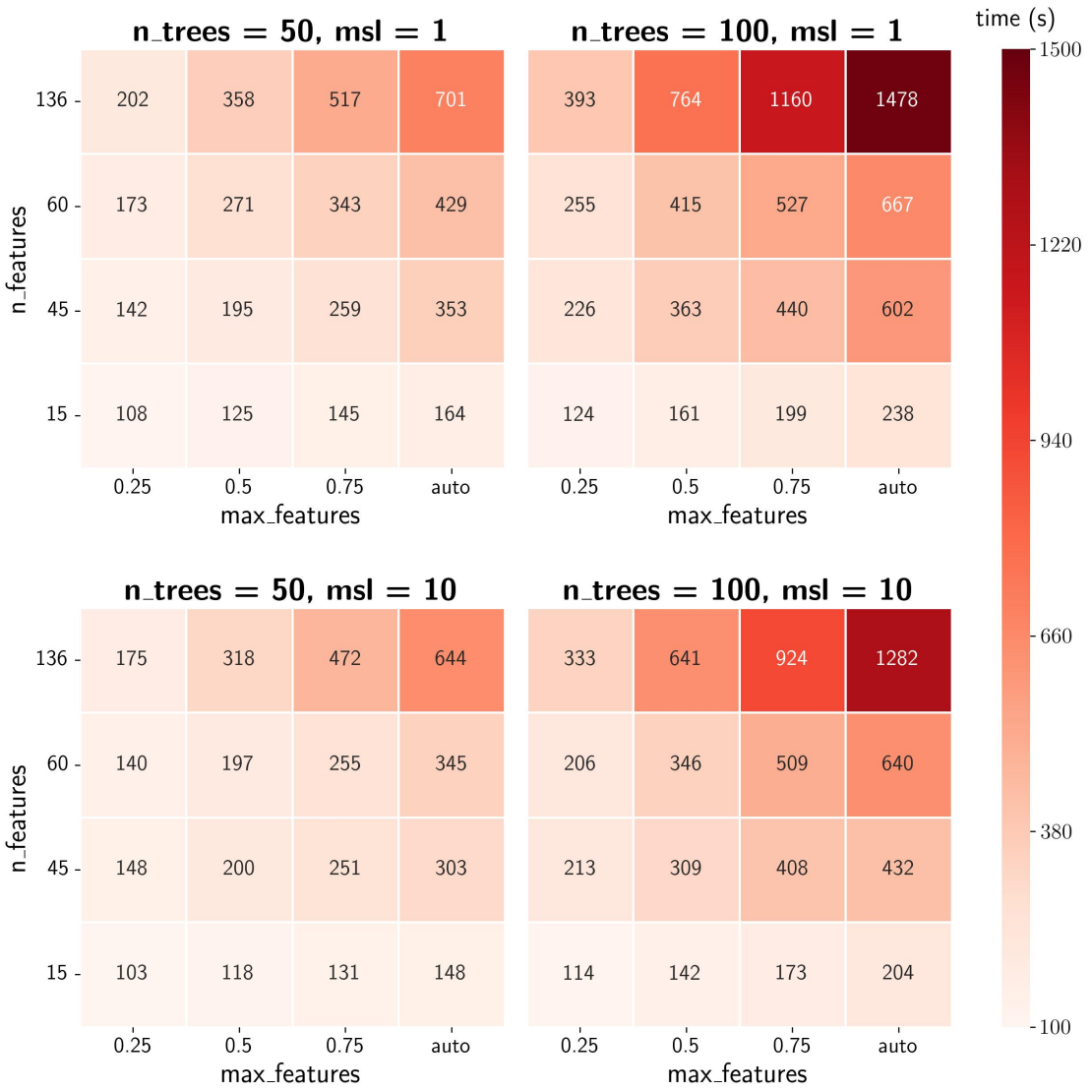


Figura 3.5: Variação do tempo de processamento para T_{ef}, quando se varia *n_features*, *max_features*, *n_trees* e msl. Note que msl = 10, em geral, resulta em tempos de processamento menores. Fonte: Cordeiro (2022, in prep.)

A Figura 3.6 mostra a variação do R^2 *score* para a gravidade superficial (log g) com a otimização. Nela, vemos que o melhor modelo de log g é o que possui 60 *features*, *max_features* de 0,25 e 100 árvores de decisão (configuração 60, 0,25, 100). O segundo melhor modelo (45, 0,25, 100) é inferior por uma diferença na quarta casa decimal. Embora o desempenho dos preditores para log g seja consideravelmente inferior ao de T_{ef}, seus

resultados ainda são excelentes: R^2 *score* = 0,8294 (correlação de 91,1% com os valores reais). No painel inferior, como já citado, vemos que o aumento de msl não mostra relevância para a acurácia do modelo.

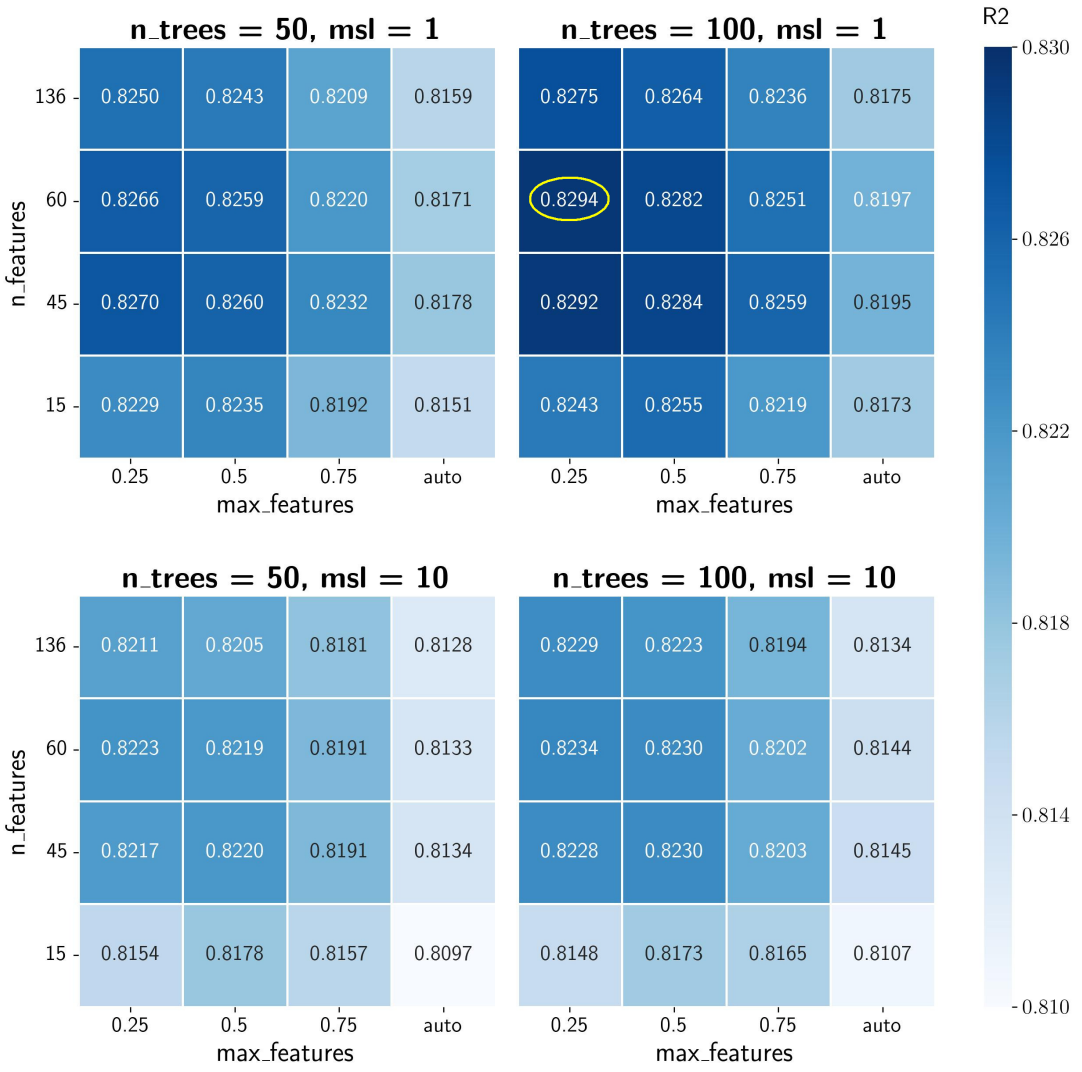


Figura 3.6: Variação do R^2 *score* para log g, quando se varia *n_features, max_features, n_trees* e msl. O painel superior usa msl = 1 (valor padrão) e o painel inferior usa msl = 10. Note que msl = 10 provoca uma queda indesejada no rendimento do modelo, que pode ser evitada facilmente usando msl = 1. O melhor modelo, por uma diferença na quarta casa decimal, possui 60 *features*, *max_features* = 0,25 e 100 árvores de decisão - configuração (60, 0,25, 100). Fonte: Cordeiro (2022, in prep.)

A Figura 3.7 é semelhante à Figura 3.5 e mostra o tempo de processamento do algoritmo para log g, considerando as mesmas variações de *max_features*, número de *features* e número de árvores de decisão. Todos os gráficos apresentados nesta subseção foram submetidos às 100 combinações testadas por Cordeiro (2022, in prep.). Mais detalhes estão incluídos no repositório do autor[20]. Para os três parâmetros, *max_features* = 0,25

[20] https://github.com/cordeirossauro/JPLUS-STEPEs

é ideal, tanto para o ganho de R^2 *score* (já que proporciona um treinamento com árvores com distintas informações), como na redução do tempo de processamento.

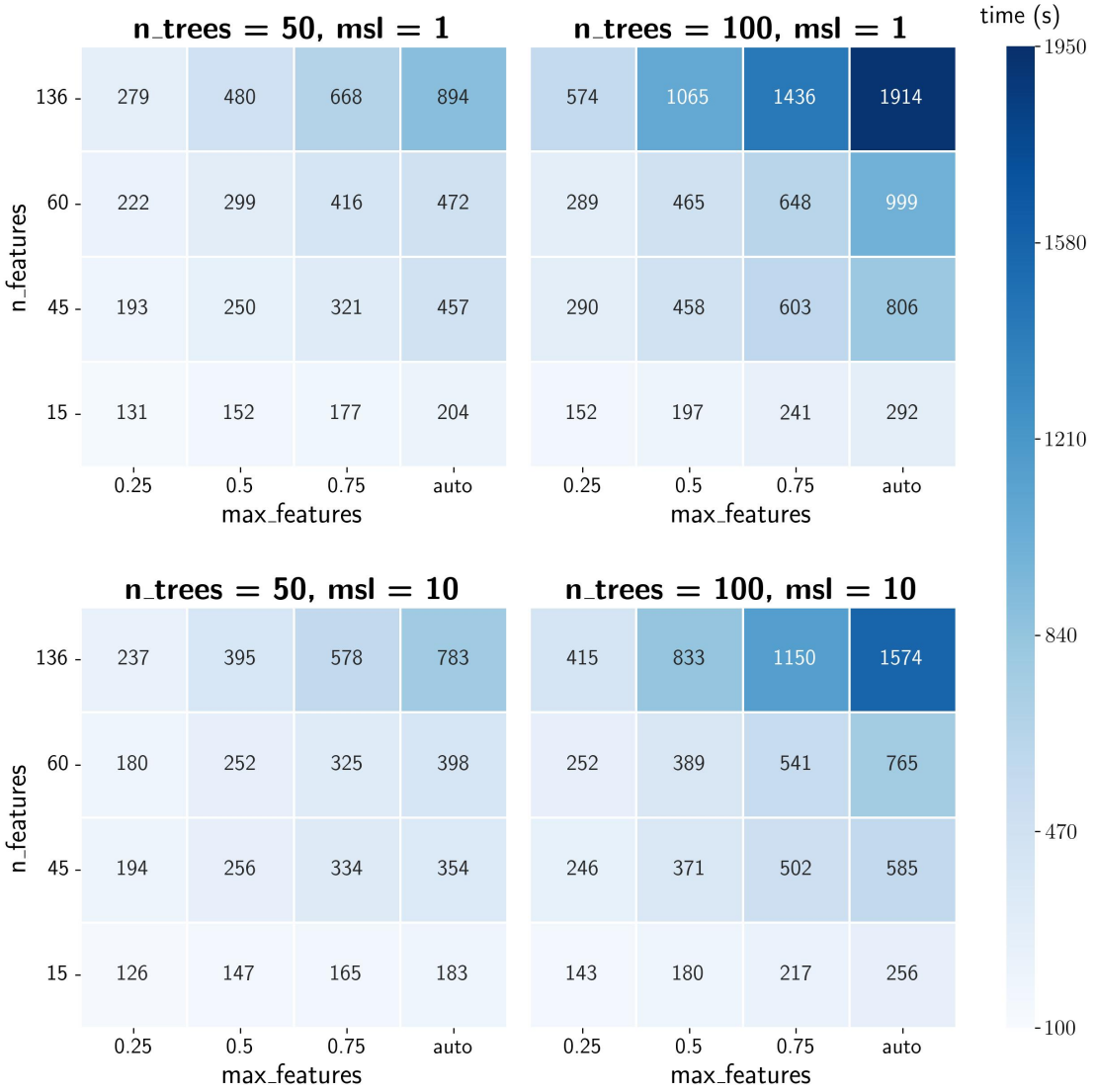


Figura 3.7: Variação do tempo de processamento para log g, quando se varia *n_features, max_features, n_trees* e msl. Note que msl = 10, em geral, resulta em tempos de processamento menores. Fonte: Cordeiro (2022, in prep.)

Por fim, a Figura 3.8 mostra a variação do R^2 *score* para [Fe/H], com melhor modelo na configuração (60, 0,25, 100), representando um rendimento de 0,8592 (o que equivale a uma correlação de 92,7% com os valores reais). Um resumo das configurações que resultaram nas melhores acurácias para cada parâmetro e serão utilizadas na construção do algoritmo deste trabalho são apresentadas na Tabela 3.6. A Figura 3.9 é semelhante às Figuras 3.5 e 3.7 e mostra o tempo de processamento do algoritmo para [Fe/H], considerando as mesmas variações de *max_features*, número de *features* e número de árvores de decisão. As conclusões para o tempo de processamento na amostra de [Fe/H] são as mesmas dos casos de $T_{\rm ef}$ e log g.

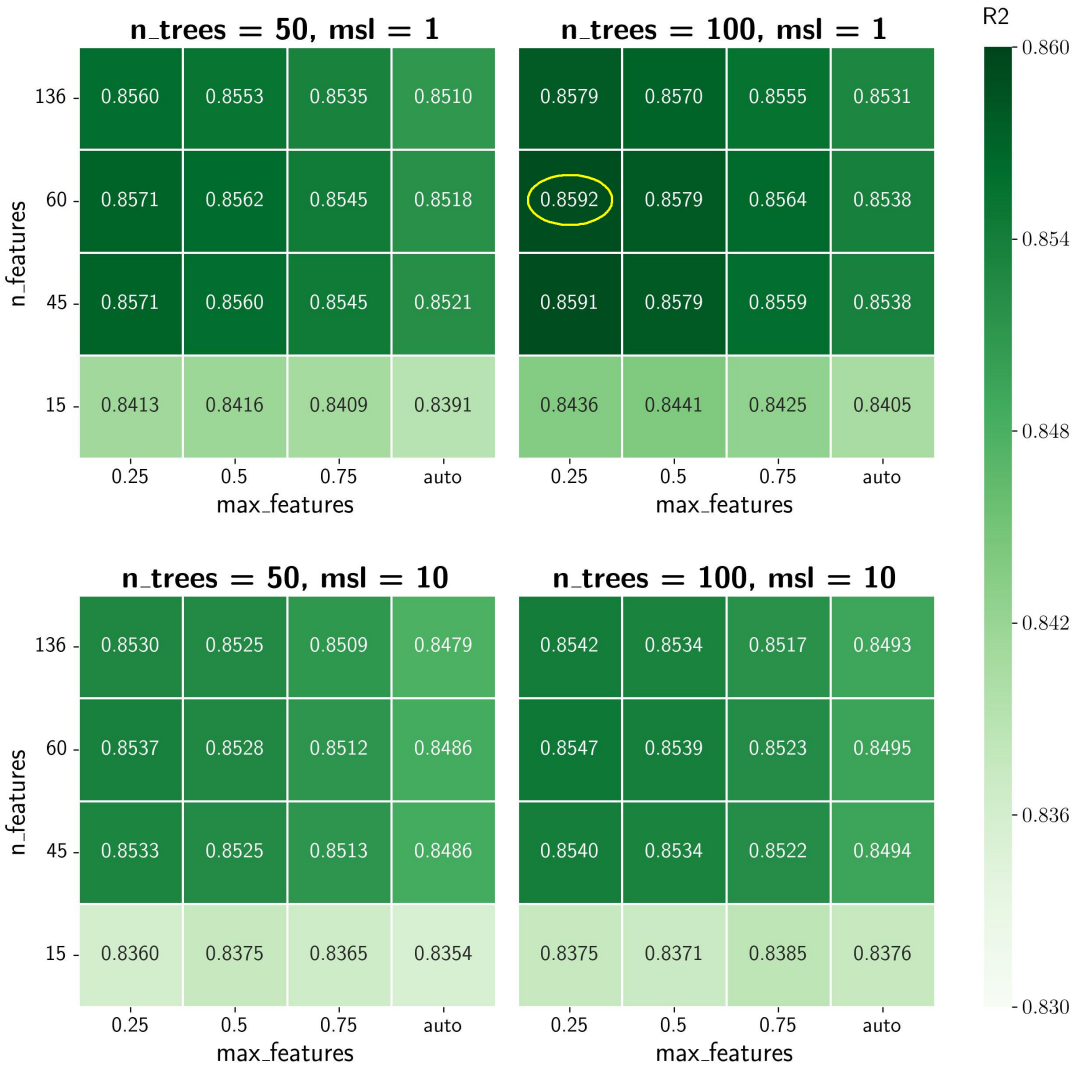


Figura 3.8: Variação do R^2 *score* para [Fe/H], quando se varia *n_features, max_features, n_trees* e msl. O painel superior usa msl = 1 (valor padrão) e o painel inferior usa msl = 10, que provoca novamente uma perda indesejada no rendimento do modelo. O melhor modelo, por uma diferença na quarta casa decimal, possui 60 *features*, *max_features* = 0,25 e 100 árvores de decisão - configuração (60, 0,25, 100). Fonte: Cordeiro (2022, in prep.)

Tabela 3.2: Configurações dos modelos de melhor rendimento, após otimização, usando msl = 1.

Parâmetro	Configuração do melhor modelo
$T_{\rm ef}$	(45, 0,25, 100)
log g	(60, 0,25, 100)
[Fe/H]	(60, 0,25, 100)

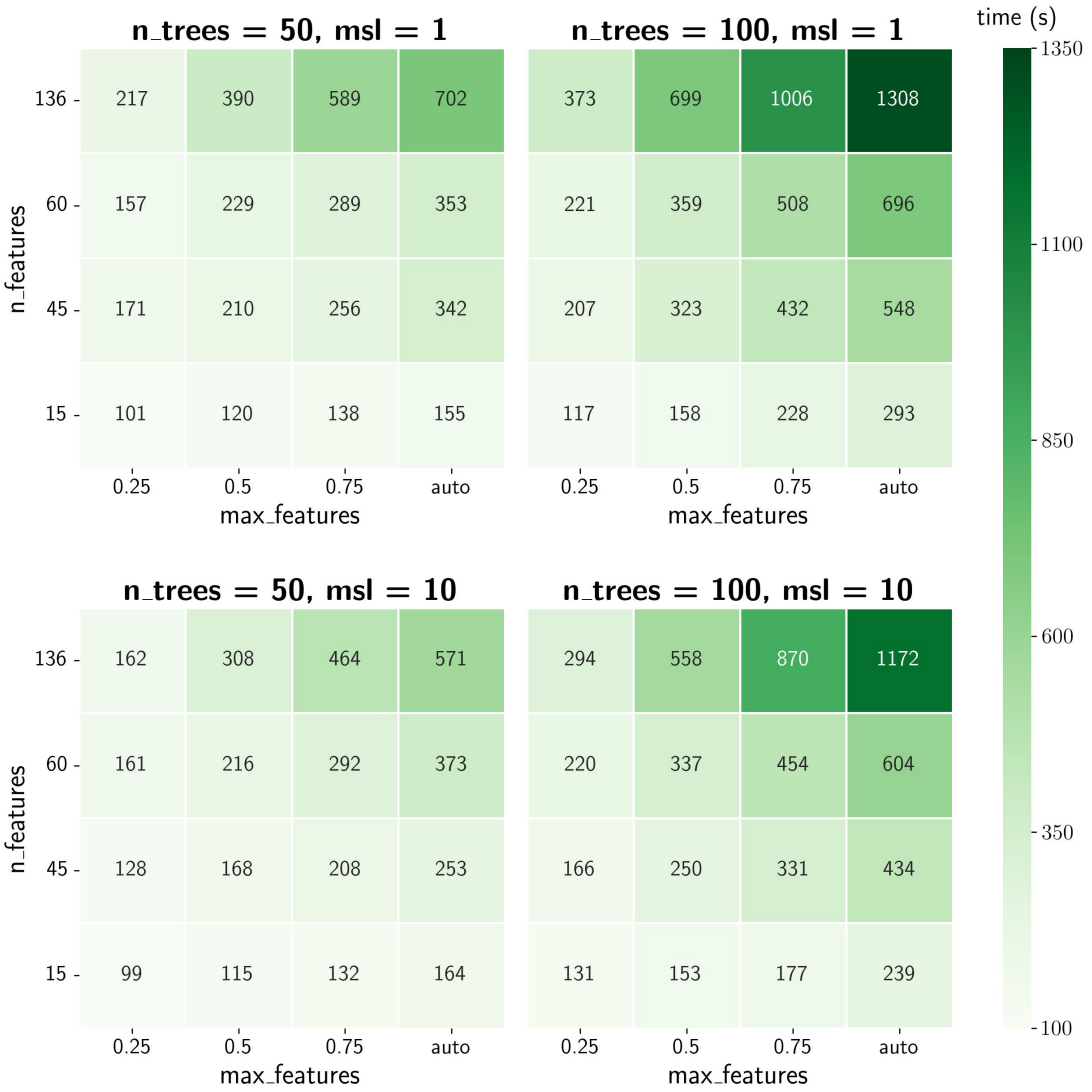


Figura 3.9: Variação do tempo de processamento para [Fe/H], quando se varia *n_features, max_features, n_trees* e msl. Note que msl = 10, em geral, resulta em tempos de processamento menores. Fonte: Cordeiro (2022, in prep.)

3.2 Testagem do algoritmo

Apesar da análise ser baseada nas magnitudes do catálogo de objetos do J-PLUS, previamente selecionadas na Subseção 2.2.1, vimos que apenas estes dados não sustentam o algoritmo, já que o J-PLUS não calculou os parâmetros físicos ($T_{\rm ef}$, log g, [Fe/H]) destes objetos. Um código baseado em Aprendizagem de Máquina precisa, antes de tudo, aprender sobre os parâmetros para então poder prevê-los para outros objetos (reconhecimento e aplicação de padrões). Como visto na Seção 2.3, para suprir esta necessidade de treinamento do algoritmo, usamos um levantamento de dados auxiliar que fornece estes valores. Esta seção reserva-se a expor todos os testes relevantes que foram realizados, antes de se decidir pelo levantamento auxiliar mais adequado.

Além das magnitudes dos filtros do J-PLUS, também utilizamos suas combinações em pares, que resultaram em 66 cores. Ao adicionarmos as magnitudes calculadas pelos 4 filtros do *Wide-field Infrared Survey Explore* (WISE; W1, W2, W3 e W4; Wright et al., 2010), expandimos nossas combinações para 120 cores (combinações usando mais de duas magnitudes não apresentaram relevância para o rendimento do algoritmo). Essas 120 cores, junto às 12 magnitudes fornecidas pelo J-PLUS e as 4 magnitudes fornecidas pelo WISE, são o que, anteriormente, definimos como *features*. O uso das 136 *features* no modelo demanda muito tempo computacional.

Quanto maior é a amostra de treinamento, maior é o tempo computacional por *feature*, já que o algoritmo precisa analisar o peso (quantidade de correlação) que uma *feature* possui, para um maior número de alvos - a Aprendizagem de Máquina dá maior importância a uma *feature*, se ela possui alta correlação com o parâmetro que está sendo estimado, ou seja, as informações de *features* de baixa correlação (próxima de 0) são praticamente descartadas. Seleciona-se, então, o maior conjunto de *features* viável para a máquina ou o conjunto mais relevante. Usamos as configurações sugeridas por Cordeiro (2022, in prep.), para o nosso tipo de análise, na Subseção 3.1.4. Um conjunto de *features* não é necessariamente igual para todos os parâmetros alvo, já que a seleção deste conjunto depende da correlação obtida entre cor-parâmetro e os parâmetros variam de modelo para modelo.

A amostra de treinamento do algoritmo, composta por estrelas observadas tanto pelo J-PLUS/WISE como pelo levantamento auxiliar (obtida através da correlação cruzada a partir da ascensão reta e declinação dos objetos), é o arquivo de entrada do algoritmo. Nele está contida toda a informação das 136 *features* e dos parâmetros físicos de interesse. Para alguns levantamentos auxiliares, pode ser que não haja boa correlação cor-parâmetro para a maioria das *features* do modelo, o que reduz drasticamente a quantidade de *features* utilizadas. Uma baixa correlação implica que não existe um padrão claro nos dados (por exemplo, estrelas de determinado valor de magnitude estão apresentando valores de log g muito diferentes). Isso faz com que poucas informações relevantes sejam passadas ao algoritmo, o que prejudica muito a previsão dos parâmetros. É difícil para o algoritmo fazer uma previsão se ele conclui que não há um padrão nos dados fornecidos pela amostra de treinamento ou se estes são insuficientes. Esta conclusão do algoritmo afirma que, para a Aprendizagem de Máquina, aqueles dados não fazem sentido. Assinala-se, então, os dois principais motivos de um modelo com baixa acurácia: os dados são insuficientes ou a correlação cor-parâmetro é baixa para a maioria das *features*.

A seguir, definimos as amostras de treinamento usadas, a quantidade de objetos que as compõem e os parâmetros que fornecem, sejam do levantamento auxiliar ou da amostra J-PLUS/WISE de interesse. Nesta última, são aplicadas a listagem de requisitos referidos na Subseção 2.2.1. Iremos dividir a análise em duas partes, onde a primeira analisa os resultados obtidos pela amostra mais restrita (estrelas observadas pela primeira lista de requisitos) e a segunda, pela amostra menos restrita (segunda lista de requisitos).

3.2.1 Amostra J-PLUS mais restrita

Vimos, na Subseção 2.2.1, que a amostra mais restrita é uma amostra de treinamento que contém apenas objetos com mais de 90% de probabilidade de serem estrelas e que foram observadas na abertura 6", em todos os doze filtros do equipamento. Além disso, estes objetos possuem correção de extinção para seus filtros e a incerteza na medida da magnitude é menor que 0,1 (e_mag $<$ 0,1). Atendendo a todos estes requisitos, obteve-se uma amostra de 1.365.454 objetos.

Esta amostra de treinamento é uma tabela de dados de 156 colunas, sendo elas: 2 de identificação do objeto, 2 de parâmetros astrométricos (α e δ), as 12 magnitudes observadas pelo J-PLUS, as 12 colunas de correção de extinção delas, as 4 magnitudes observadas pelo WISE, as 4 colunas de correção das delas e as 120 cores calculadas a partir das 16 magnitudes corrigidas por extinção. Muitos outros parâmetros são fornecidos pelo catálogo J-PLUS, mas são desnecessários para o objetivo deste trabalho. Esta tabela (bem como seus ajustes) é a base de qualquer amostra de treinamento oferecida ao algoritmo de Aprendizagem de Máquina, escrito para este trabalho, para treinamento de seus modelos.

Primeiramente, serão feitas as modelagens para os parâmetros físicos de temperatura efetiva ($T_{\rm ef}$), gravidade superficial (log g) e metalicidade ([Fe/H]). Inicia-se por eles, já que estes parâmetros previstos serão utilizados no cálculo de parâmetros como luminosidade, raio e massa estelares. Para $T_{\rm ef}$, log g e [Fe/H] serão realizados testes, principalmente, com 4 levantamentos auxiliares: TESS, SEGUE, GALAH e LAMOST. A escolha destes é justificada em suas subseções.

3.2.1.1 TESS

O *Transiting Exoplanet Survey Satellite* (TESS; Ricker et al., 2014) é uma missão espacial semelhante à missão Kepler, que também objetiva a busca por exoplanetas, incluindo aqueles que poderiam sustentar vida, através do método de trânsito. O TESS foi lançado em abril de 2018 e busca trânsitos nas 200.000 estrelas mais brilhantes próximas ao Sol - 30 a 100 vezes mais brilhantes do que as da missão Kepler e em uma área 400 vezes maior. Espera-se a detecção de pelo menos 300 exoplanetas do tipo-Terra e Super-Terra (classificados através do raio do exoplaneta, obtido pelo método de trânsito). Destes, pelo menos 50 devem ter suas massas estimadas com auxílio de dados adicionais obtidos com o método de variação de velocidade radial. Além das 200.000 estrelas monitoradas para produção de curvas de luz, o TESS capturará dados de milhares de objetos presentes no campo da missão, através de FFIs com tempo de exposição de 30 minutos. Mais detalhes estão disponíveis na página *web* da missão[21].

Pela semelhança com os objetivos da missão Kepler, o TESS foi o primeiro levantamento auxiliar testado. Os parâmetros físicos dos objetos no campo de visão do TESS

[21] https://www.nasa.gov/content/about-tess e https://tess.mit.edu/

também foram retirados de um compilado de catálogos menores, chamado catálogo de entrada do TESS (TIC, *TESS Input Catalog*; Stassun et al., 2018). O TIC foi utilizado na seleção de alvos para a missão e também apoia os cálculos das propriedades físicas e observacionais de candidatos a planetas. Sua versão atual é a TIC-8, que usa o catálogo Gaia DR2 como base e combina ainda um grande número de catálogos fotométricos, incluindo 2MASS, UCAC4, APASS, SDSS e WISE. Ele pode ser acessado e baixado diretamente através do *Mikulski Archive for Space Telescopes* (MAST)[22].

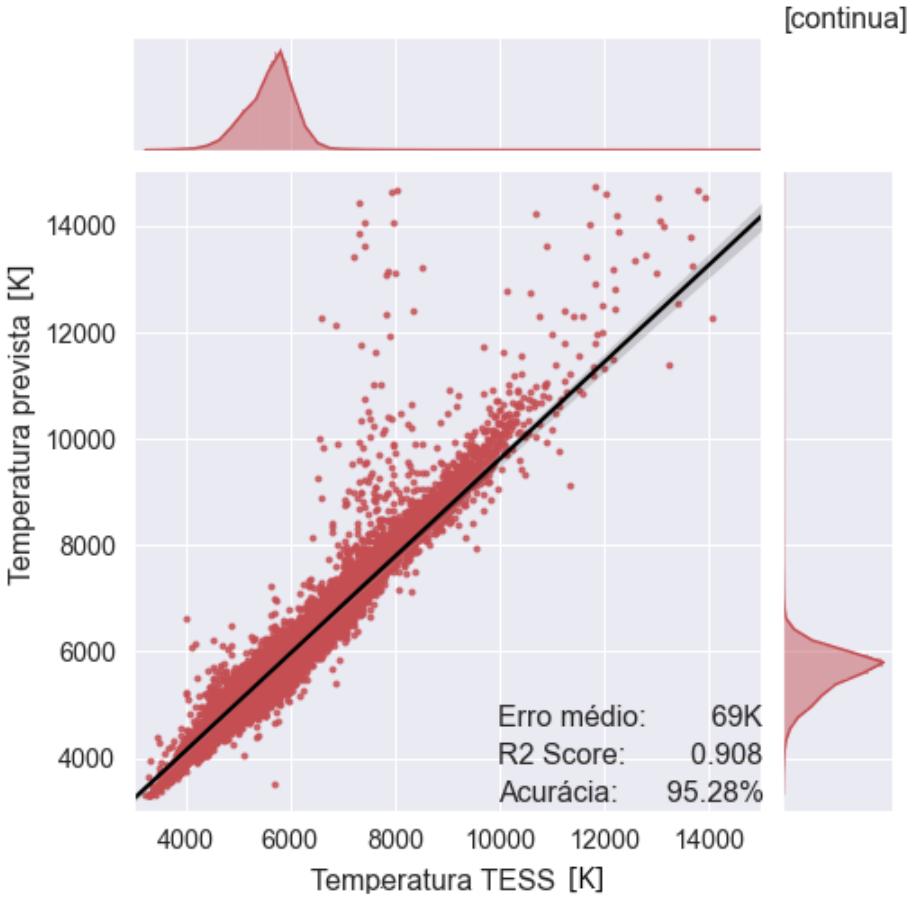
Segundo Stassun et al. (2019), as temperaturas efetivas e as metalicidades do TIC, quando disponíveis, advêm de catálogos como SPOCS, PASTEL, Gaia-ESO DR3, TESS-HERMES DR1, GALAH DR2, APOGEE-2 DR14, LAMOST DR4 e RAVE DR5. Já o log g é calculado com base na massa e no raio relatados pelos catálogos. Um problema notável é que existe um deficit nos dados para metalicidade, principalmente em comparação à $T_{\rm ef}$ e log g. O erro para $T_{\rm ef}$ no TIC é de até 150 K. Para log g, os erros podem chegar a 0,6 dex e para [Fe/H] a 0,5 dex, porém, os valores destes dois últimos não estão disponíveis para a maioria das estrelas do TIC. Dos erros declarados no catálogo: 1.243.419 estrelas possuem erro em $T_{\rm ef} < 150$ K, das quais apenas 67.009 possuem erro < 100 K; 37.971 estrelas possuem erro em log g $< 0{,}2$ dex; 142.688 estrelas possuem erro em [Fe/H] $< 0{,}3$ dex.

A Figura 3.10 mostra a acurácia do algoritmo, treinado com objetos do J-PLUS/WISE que foram observados pelo TESS e possuíam dados no TIC (amostra de treinamento com dados de objetos em comum da amostra mais restrita do J-PLUS, selecionados na Subseção 3.2.1, e do WISE e TESS/TIC). Nela, comparamos os valores previstos por ele com os disponíveis na amostra de teste (25% da amostra de treinamento). Os valores de erro médio absoluto, acurácia e R^2 *score* obtidos estão em seus painéis. À direita e acima dos painéis, temos histogramas representando a distribuição dos objetos (em quantidade) com respeito ao parâmetro. Estes histogramas estarão em todas as figuras que mostrem as simulações dos modelos.

Ainda na Figura 3.10 é possível perceber que não há muita consistência nos dados de log g. Mesmo após uma análise adicional, onde separamos estrelas em diferentes fases evolutivas (anãs e gigantes), o problema do log g não pôde ser solucionado. Isso é justificado internamente no modelo, já que existe uma correlação muito baixa entre as *features* e este parâmetro: apenas 6 cores apresentam correlação $> 0{,}19$. Apesar da metalicidade apresentar uma acurácia satisfatória, a quantidade de dados é muito inferior às utilizadas para os demais parâmetros. A Tabela 3.3 apresenta a amostra de treinamento desta modelagem, bem como a quantidade destes objetos com parâmetros fornecidos pelo TIC. Um número muito menor de objetos com [Fe/H] calculado é um problema direto, já que adquirimos um desequilíbrio na modelagem e isso pode resultar em um viés pela falta de uniformidade dos dados.

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

[continua]


Temperatura prevista [K]
14000
12000
10000
8000
6000
4000
Erro médio: 69K
R2 Score: 0.908
Acurácia: 95.28%
4000
6000
8000
10000
12000
14000
Temperatura TESS [K]

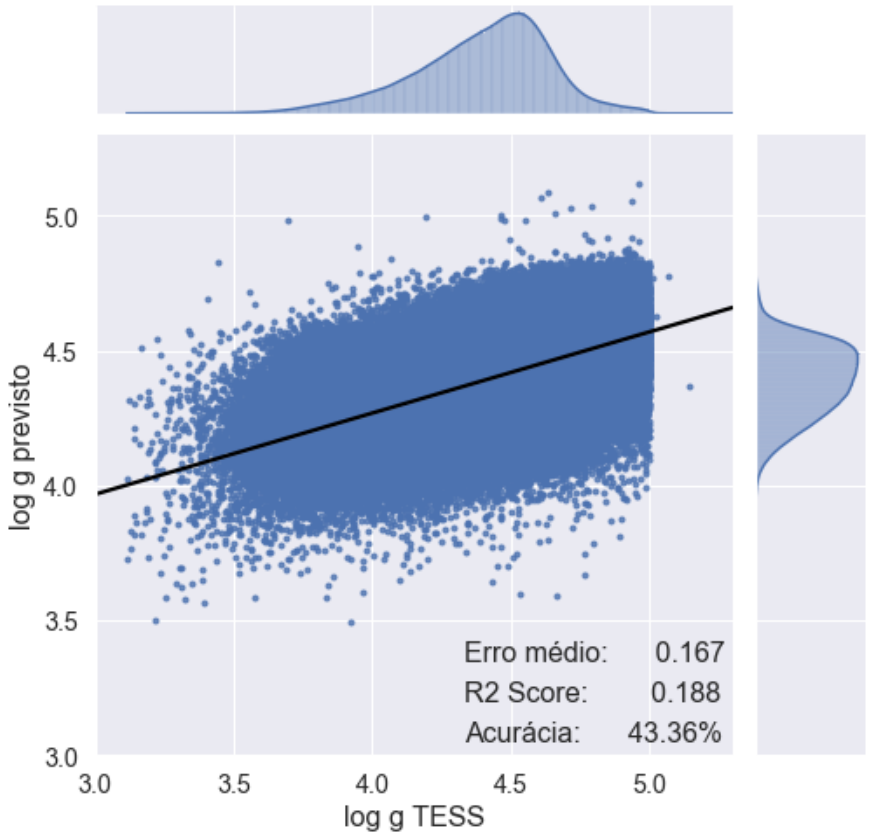

log g previsto
5.0
4.5
4.0
3.5
3.0
Erro médio: 0.167
R2 Score: 0.188
Acurácia: 43.36%
3.0
3.5
4.0
4.5
5.0
log g TESS

Figura 3.10: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas do TESS com parâmetros disponíveis no TIC. O painel superior (pontos vermelhos) corresponde à simulação de $T_{\rm ef}$ e compara os valores previstos pelo algoritmo com os valores disponíveis no TIC. O painel central (pontos azuis) representa o mesmo para log g e o painel inferior (pontos verdes) para [Fe/H]. A linha sólida preta em cada painel corresponde a linha de tendência dos dados. Cada painel possui informações sobre o erro médio, R^2 *score* e acurácia de seus respectivos modelos. Essa simulação mostra alta dispersão em $T_{\rm ef}$, principalmente para valores maiores que 6000 K, e log g. O rendimento para [Fe/H] não pode ser comparado, já que seu modelo utilizou uma amostra muito menor - esse desbalanceamento pode adicionar um viés ao modelo.

Tabela 3.3: Amostra de treinamento: Amostra J-PLUS mais restrita + WISE + TESS

Amostra de treinamento	Estrelas com $T_{\rm ef}$	Estrelas com log g	Estrelas com [Fe/H]
1.335.865	1.291.942	1.173.496	142.694

3.2.1.2 SEGUE

O *Sloan Extension for Galactic Understanding and Exploration* (SEGUE; Rockosi, 2005), um dos levantamentos que compõem a colaboração do grande levantamento de dados SDSS (*Sloan Digital Sky Survey*), obteve o espectro de cerca de 240.000 estrelas com velocidades radiais típicas de 10 km/s para estudar a estrutura da Via Lactea e investigar a formação de seus componentes.

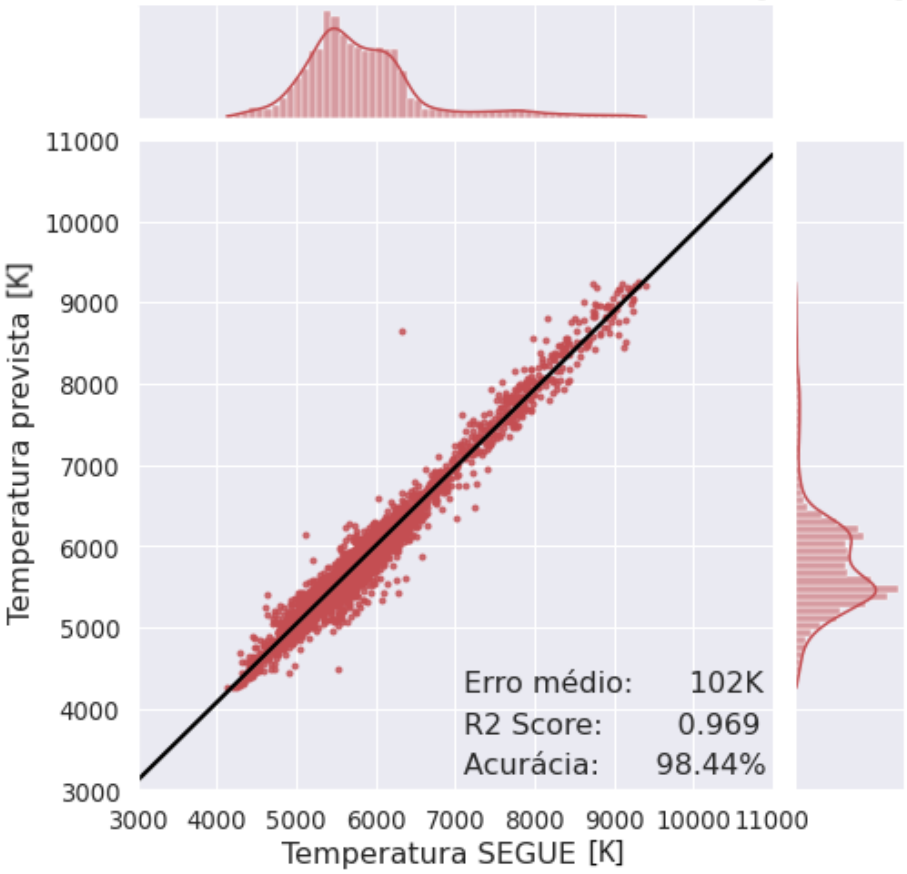
Para a modelagem com SEGUE utilizamos o *data release* 8 (DR8), onde os parâmetros físicos de interesse ainda eram calculados usando o Preditor de Parâmetros Estelares do SEGUE (*SEGUE Stellar Parameter Pipeline*, SSPP; Lee et al., 2008), com dados de espectroscopia de baixa resolução e incertezas médias de 117 K para $T_{\rm ef}$, 0,26 dex para log g e 0,22 dex para [Fe/H]. Essas incertezas dependem do tipo espectral e da razão sinal-ruído. Os valores listados são para estrelas com 4500 K $\leqslant T_{\rm ef} \leqslant$ 7500 K. Eles são relativamente altos e impactam no rendimento do algoritmo. O DR8 pode ser acessado em duas fontes principais: o *Science Archive Server* (SAS)[23] e o *Catalog Archive Server* (CAS, também conhecido como *SkyServer*)[24]. Em geral, o SAS permite acesso aos espectros e imagens FITS. O CAS dá acesso aos resultados em forma de catálogo por meio de uma interface de banco de dados SQL.

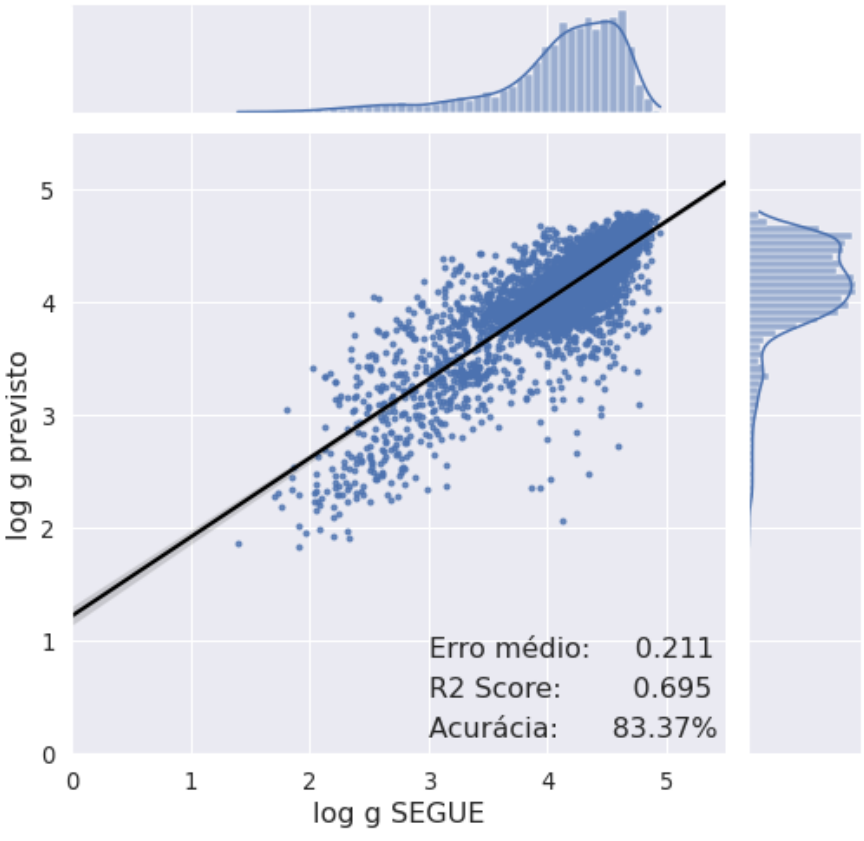
A Figura 3.11 mostra o desempenho do algoritmo para os dados do SEGUE, comparando os valores previstos por ele com os disponíveis na amostra de teste. Todas as amostras de teste deste trabalho correspondem a 25% de sua respectiva amostra de treinamento. Da figura, podemos perceber que o rendimento do log g é claramente superior ao apresentado pela modelagem com TESS, na Figura 3.10. Mesmo assim, o rendimento para log g está longe do ideal e ainda apresenta muita dispersão. Na mesma figura, notamos uma melhora bastante relevante na acurácia do modelo para $T_{\rm ef}$ (ganho $> 3\%$), principalmente na redução de dispersão e incerteza máxima - podemos perceber isso ao observarmos os pontos distantes da linha de tendência. O mesmo ocorre para o modelo de [Fe/H] (ganho $> 2\%$), onde o SEGUE também oferece uma maior cobertura de valores.

A amostra de treinamento do SEGUE é de 14.831 objetos, todos eles com os três parâmetros calculados. Esta amostra é muito menor que a utilizada pelo TESS. Em outro cenário, decidiríamos por escolher, neste ponto, as modelagens do TESS para $T_{\rm ef}$ e [Fe/H] e do SEGUE para log g. Manteríamos os modelos do TESS pela baixa variância na acurácia, já que ele oferece uma amostra maior. Porém, como visto, o TESS coleta as informações de seus parâmetros da literatura, fazendo do SEGUE (que possui parâmetros calculados de forma homogênea por espectroscopia) um melhor candidato para os três parâmetros - mesmo que apenas provisoriamente.

[23] https://data.sdss.org/sas/
[24] http://skyserver.sdss.org/dr8/en/

[continua]


Temperatura prevista [K]
Temperatura SEGUE [K]
Erro médio: 102K
R2 Score: 0.969
Acurácia: 98.44%


log g previsto
log g SEGUE
Erro médio: 0.211
R2 Score: 0.695
Acurácia: 83.37%

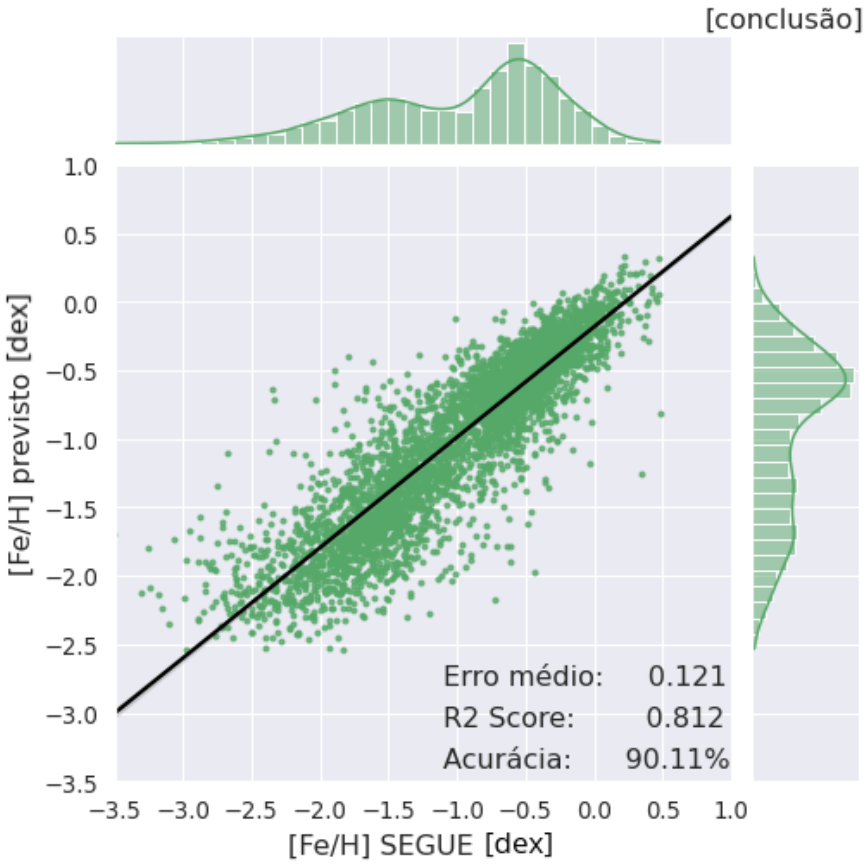
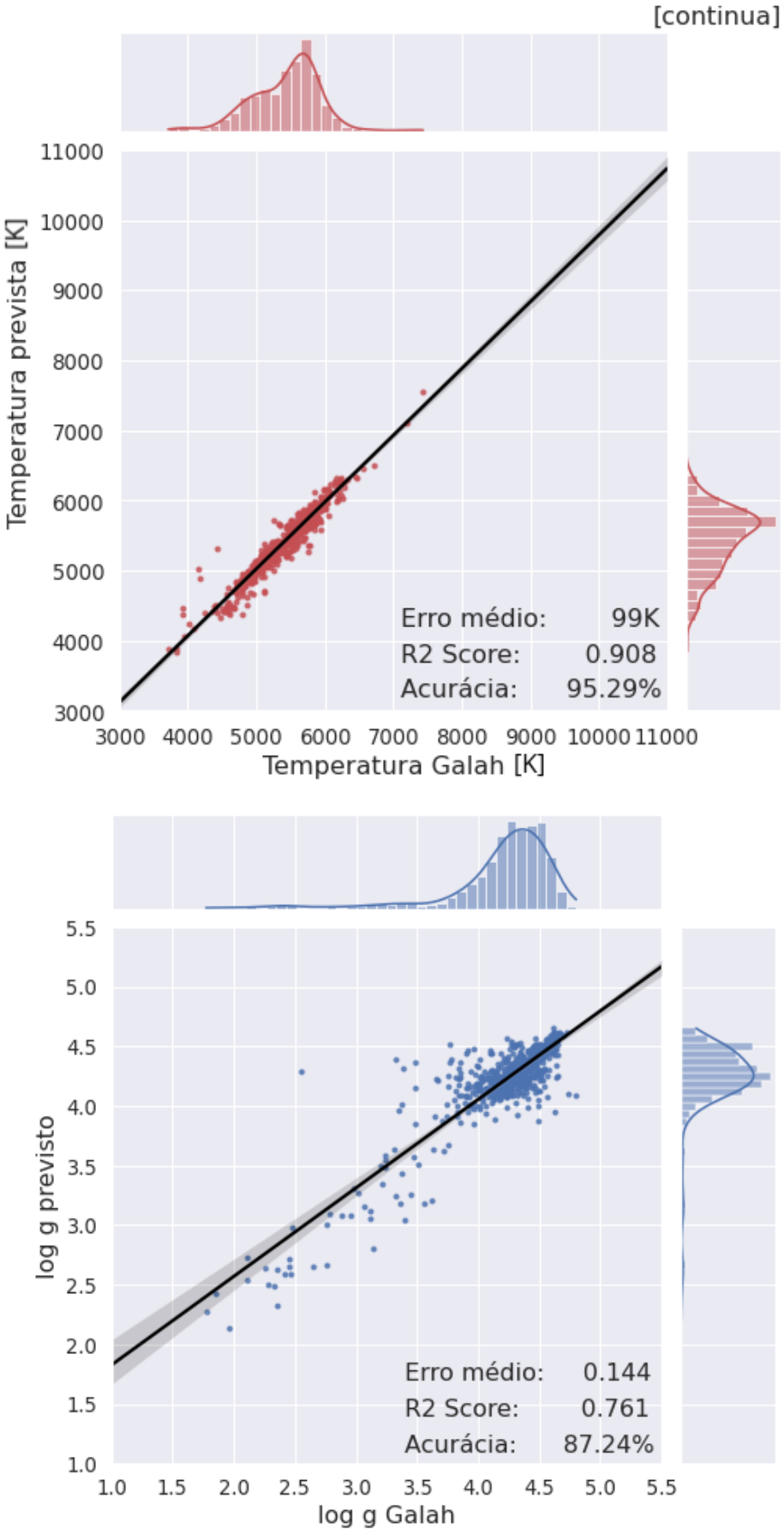


Figura 3.11: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas J-PLUS/WISE+SEGUE, com parâmetros disponíveis no SEGUE. Os painéis são semelhantes aos da Figura 3.10, e comparam os valores previstos pelo algoritmo com os valores disponíveis no SEGUE: superior (pontos vermelhos) para $T_{\rm ef}$; central (pontos azuis) para log g; inferior (pontos verdes) para [Fe/H]. Essa simulação mostra redução na dispersão dos 3 parâmetros, com relação à simulação da Figura 3.10.

3.2.1.3 GALAH

O *Galactic Archaeology with HERMES* (GALAH; De Silva et al., 2015) é um programa de observação que utiliza o espectrógrafo HERMES, com o objetivo de entender a formação e evolução da Via Láctea. O HERMES (*High Efficiency and Resolution Multi Element Spectrograph*; Sheinis et al., 2015) é alimentado pelo sistema de posicionamento de fibra óptica 2dF (*Two Degree Field*) que permite medições detalhadas de 400 estrelas por vez, auxiliando na obtenção de conjuntos de dados multidimensionais de alta resolução para mais de um milhão de estrelas de todas as idades e locais na Via Láctea. Aqui usamos o *data release* 3 (DR3)[25], descrito em Buder et al. (2021). Os parâmetros do catálogo são estimados com o BSTEP (*Bayesian Stellar Parameter Estimation code*; Sharma et al., 2018) - e levam a etiqueta "_bstep" em seus nomes. O BSTEP fornece uma estimativa bayesiana a partir de parâmetros observados, fazendo uso de isócronas estelares.

[25] https://www.galah-survey.org/dr3/the_catalogues/


[continua]
Temperatura prevista [K]
Temperatura Galah [K]
Erro médio: 99K
R2 Score: 0.908
Acurácia: 95.29%
log g previsto
log g Galah
Erro médio: 0.144
R2 Score: 0.761
Acurácia: 87.24%

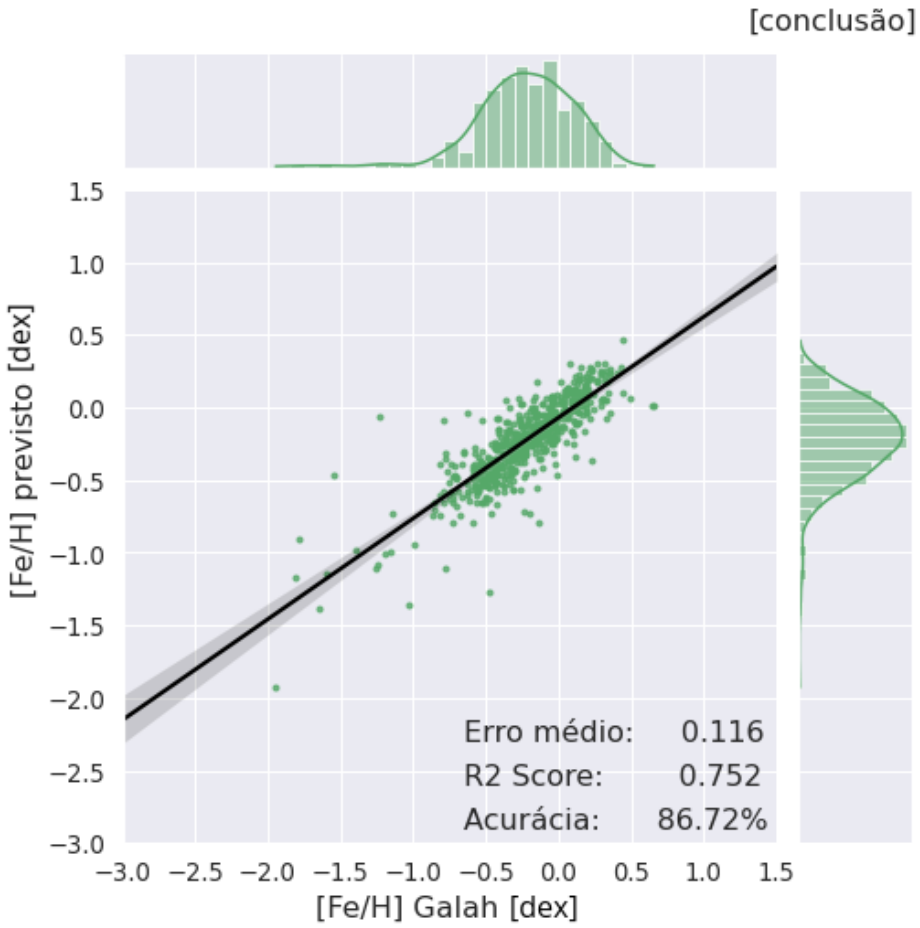


Figura 3.12: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas J-PLUS/WISE+GALAH, com parâmetros disponíveis no GALAH. Os painéis são semelhantes aos da Figura 3.10, e comparam os valores previstos pelo algoritmo com os valores disponíveis no GALAH: superior (pontos vermelhos) para $T_{\rm ef}$; central (pontos azuis) para log g; inferior (pontos verdes) para [Fe/H]. Estas simulações se mostraram muito instáveis (observe a região cinza que representa a margem de erro da linha de tendência), com alta variação no R^2 *score*, quando se realizava uma nova execução do código. Isto fere diretamente o princípio de reprodutibilidade[26].

Do catálogo, 65% são anãs, 34% são gigantes e 1% são estrelas não classificadas. A Tabela 3.4 apresenta a amostra de treinamento que combina J-PLUS/WISE e GALAH. Na Figura 3.12, vemos a modelagem do algoritmo para estrelas comuns entre a amostra J-PLUS/WISE e GALAH. Note que esta amostra é muito pequena (cerca de 2 mil estrelas) e não é suficiente para manter a estabilidade do modelo. Notamos isto ao executar o algoritmo mais de uma vez. Quando a amostra é suficiente, o rendimento do algoritmo é estável e o R^2 *score* não varia significativamente (mais que 0,05) entre uma execução e outra. Sendo assim, apesar do R^2 *score* das previsões ser alto, não podemos considerar a amostra do GALAH como uma boa amostra de treinamento.

[26]Importante princípio do método científico que implica que um resultado obtido por um experimento ou estudo observacional deve manter-se com alto grau de concordância quando replicado, com a mesma metodologia, por terceiros.

Tabela 3.4: Amostra de treinamento: Amostra J-PLUS mais restrita + WISE + GALAH

Amostra de treinamento	Estrelas com $T_{\rm ef}$	Estrelas com log g	Estrelas com [Fe/H]
2.018	2.018	2.018	1.998

3.2.1.4 LAMOST

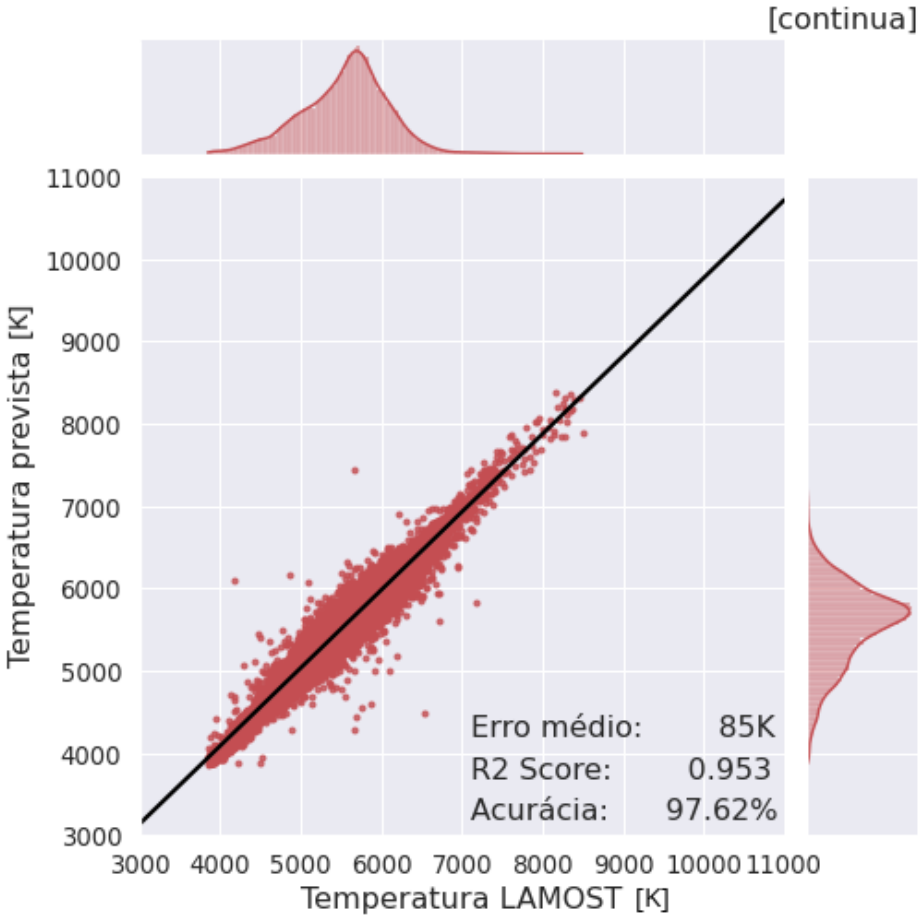
O *Large Sky Area Multi-Object Fiber Spectroscopic Telescope* (LAMOST; Gang et al., 2012)[27], também conhecido como *Shoujing Guo Telescope*, é um levantamento espectroscópico de mais de 8000 e 3500 graus quadrados de área, nos polos norte e sul galáctico, respectivamente. Foi construído para conduzir uma pesquisa espectroscópica de 10 milhões de estrelas da Via Láctea e outros objetos extragalácticos. O telescópio possui 4 metros de diâmetro e plano focal coberto com 4000 fibras, transmitindo luz para dezesseis espectrógrafos. Cada espectrógrafo possui duas câmeras CCD com lado azul (370-590 nm) e vermelho (570-900 nm). Seu design permite que alcance uma magnitude limite tão fraca quanto r = 19 com resolução R = 1500. Essa resolução pode chegar a R = 10.000 para objetos mais brilhantes.

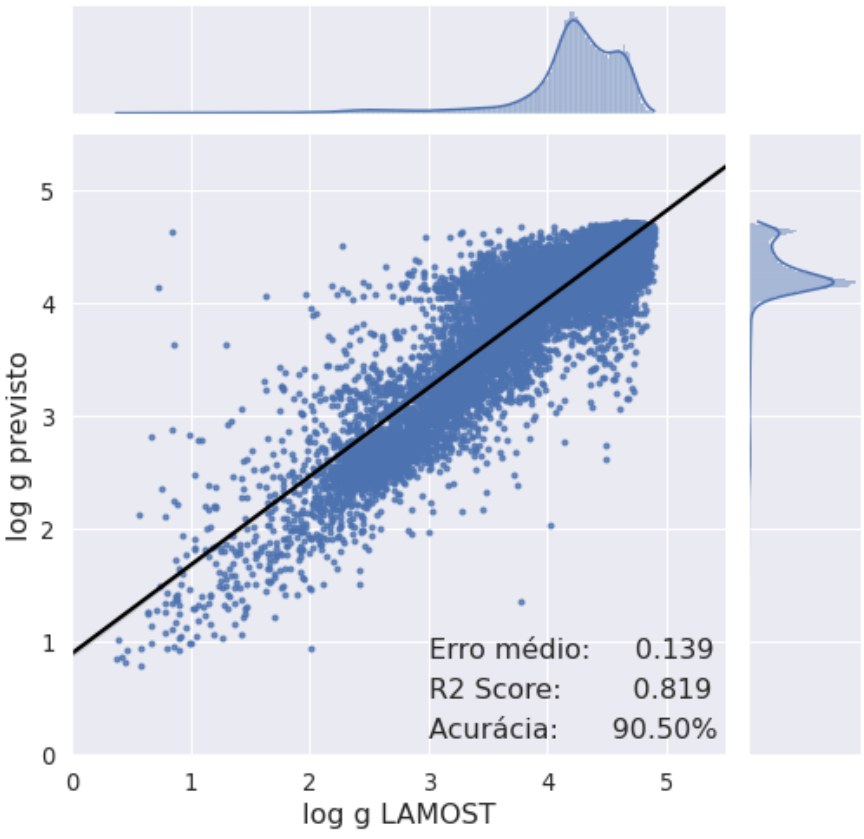
Foi dividido em dois projetos principais: o *Lamost ExtraGAlactic Survey* (LEGAS) e o *Lamost Experiment for Galactic Understanding and Exploration* (LEGUE). O LEGUE foi projetado para estudar a estrutura do halo Galáctico e os componentes do disco (incluindo regiões de formação estelar e aglomerados abertos). Pretende-se alcançar, a partir disto, melhor compreensão da formação estelar, da história da formação da Galáxia, da estrutura do potencial gravitacional, incluindo o buraco negro central e a (sub)estrutura da matéria escura. O LEGAS foi planejado com o objetivo de explorar o meio extragaláctico.

A amostra de treinamento resultante da correlação cruzada do DR8[28] do LAMOST com a amostra J-PLUS/WISE, referida na Subseção 3.2.1, reuniu 186.374 objetos, todos com $T_{\rm ef}$, log g e [Fe/H] calculados pelo LAMOST. Os parâmetros foram calculados pelo Preditor de Parâmetros Estelares do LAMOST (*LAMOST stellar parameter pipeline*, LASP; Xiang et al., 2015). A Figura 3.13 mostra o desempenho do algoritmo com esta amostra. Nela, temos uma amostra de treinamento que: a) atende aos três parâmetros físicos, ou seja, produz simulações com bons resultados; b) possui estabilidade no R^2 *score*; c) possui o melhor rendimento em log g e [Fe/H], e menos de 1% de variação em relação ao melhor rendimento de $T_{\rm ef}$ (dado pelo SEGUE). Como a quantidade de objetos do LAMOST oferece modelos mais estáveis, suas informações foram escolhidas para compor o conjunto de dados auxiliar.

[27] www.lamost.org/
[28] http://www.lamost.org/dr8/

[continua]


Temperatura prevista [K]
Temperatura LAMOST [K]
Erro médio: 85K
R2 Score: 0.953
Acurácia: 97.62%


log g previsto
log g LAMOST
Erro médio: 0.139
R2 Score: 0.819
Acurácia: 90.50%

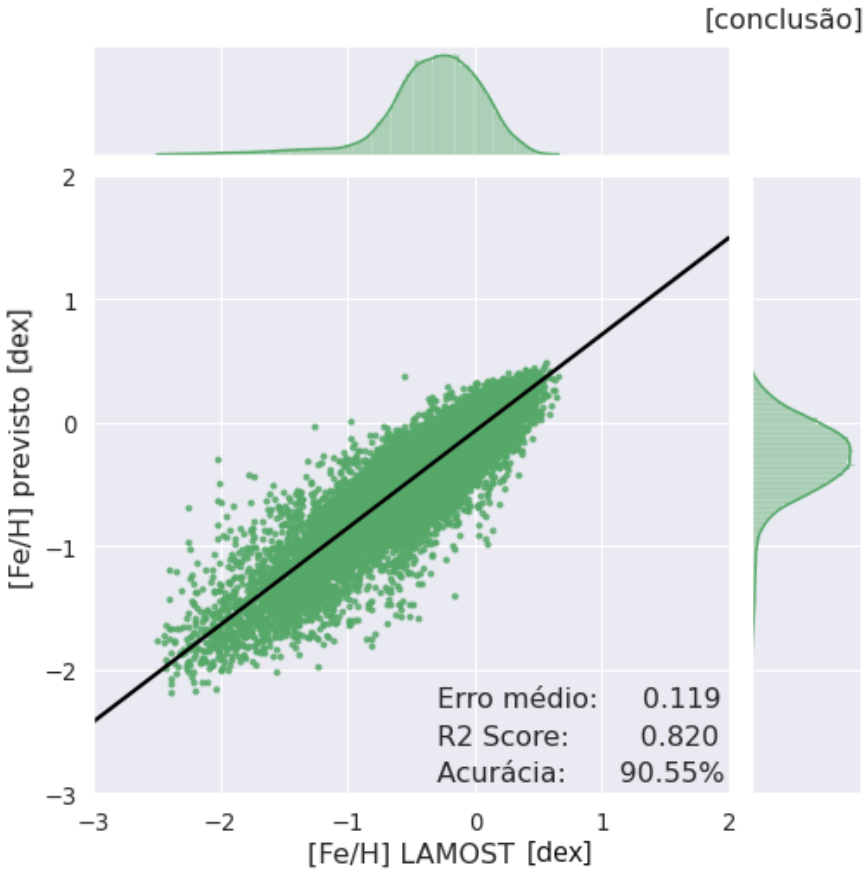


Figura 3.13: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas J-PLUS/WISE+LAMOST, com parâmetros disponíveis no LAMOST. Os painéis são semelhantes aos da Figura 3.10, e comparam os valores previstos pelo algoritmo com os valores disponíveis no LAMOST: superior (pontos vermelhos) para $T_{\rm ef}$; central (pontos azuis) para log g; inferior (pontos verdes) para [Fe/H]. Essa simulação apresenta os modelos de log g e [Fe/H] de melhor rendimento entre os 4 levantamentos auxiliares testados e o segundo melhor modelo de $T_{\rm ef}$, com um rendimento apenas 0,82% inferior ao apresentado pelo SEGUE na Figura 3.11.

A grande quantidade de objetos nos permite ser mais restritivos. A amostra LAMOST utilizada na simulação da Figura 3.13 contém estrelas com qualquer valor de incerteza. Podemos limitar estas incertezas a fim de refinar os modelos e elevar seus rendimentos. A Figura 3.14 mostra a distribuição das incertezas para a temperatura efetiva nesta amostra. Observa-se, nesta figura, que a grande maioria dos objetos (cerca de 170.000) possui uma incerteza na medida de temperatura abaixo de 300 K. Cerca de 138.000 objetos apresentam incerteza menor que 200 K. Podemos então adotar 200 K como corte. Não foi necessário ser mais restritivo que 200 K nas incertezas da temperatura pois o modelo de $T_{\rm ef}$ já apresenta, desde o início, excelente acurácia. Além disso, restringir ainda mais a incerteza proporcionou uma amostra de treinamento muito pequena.

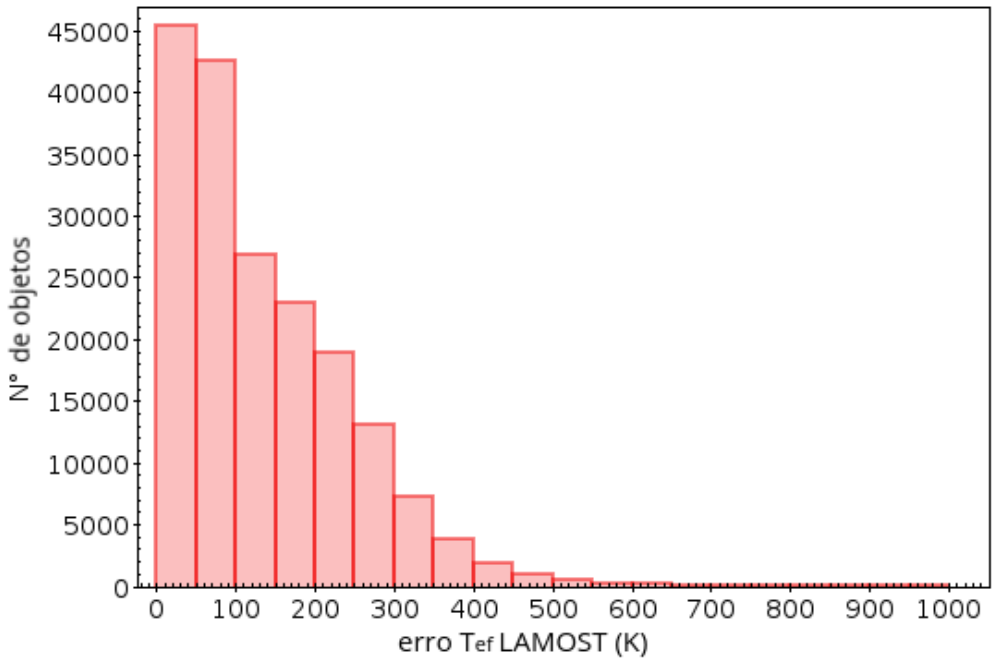


Figura 3.14: Distribuição das incertezas para a temperatura efetiva na amostra de 186.374 objetos J-PLUS/WISE+LAMOST. Destes, 169.937 objetos possuem incerteza $\leqslant$ 300 K e 137.865 possuem incerteza $\leqslant$ 200 K. Foram analisados objetos com incertezas ainda menores (até 100 K) mas isto produzia amostras de treinamento muito pequenas.

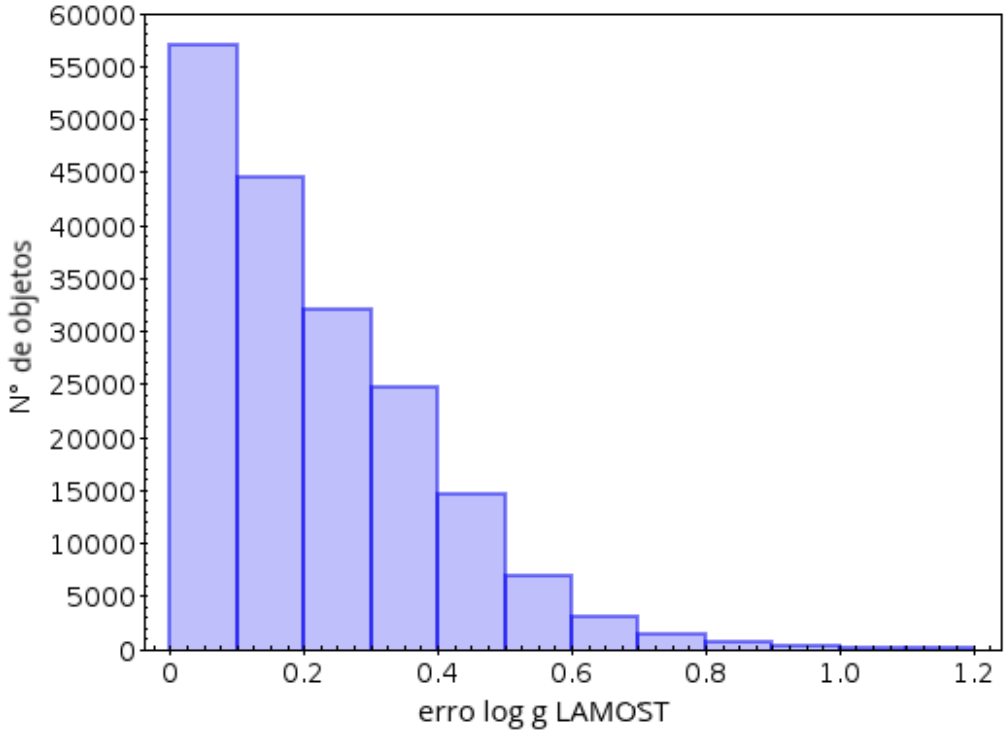


Figura 3.15: Distribuição das incertezas para o log g na amostra de 186.374 objetos J-PLUS/WISE+LAMOST. Destes, 158.267 objetos possuem incerteza $\leqslant$ 0,4 e 101.740 possuem incerteza $\leqslant$ 0,2.

A Figura 3.13 mostrou que os modelos de log g e [Fe/H] possuem um rendimento consideravelmente inferior ao de $T_{\rm ef}$. Podemos analisar também as incertezas nestes parâmetros. A presença de altos valores de incerteza obviamente prejudica estes modelos. Na Figura 3.15 vemos a distribuição das incertezas para o log g na amostra de treinamento. Nela, cerca de 160.000 objetos possuem incerteza $\leqslant 0{,}4$. O treinamento com estes 160.000 objetos resultou em um R^2 *score* de 0,851. Como a amostra com incerteza $\leqslant 0{,}2$ apresenta quase 102.000 objetos (o que ainda é uma grande amostra e que resulta em um modelo estável com R^2 *score* de 0,872), podemos ser ainda mais restritivos e adotar este valor como limite de incerteza.

Apresentamos as incertezas para [Fe/H] na Figura 3.16. Dois limites de incerteza foram analisados: incerteza $\leqslant 0{,}2$ e incerteza $\leqslant 0{,}3$, já que a maioria dos objetos apresentam incertezas inferiores a estes limites. Ambos os valores produziram um R^2 *score* muito semelhante ao já observado na Figura 3.13. Cerca de 144.000 estrelas possuem incerteza na [Fe/H] $\leqslant 0{,}2$. Para incerteza $\leqslant 0{,}3$, este número é de mais de 173.000 objetos. Adotaremos a incerteza $\leqslant 0{,}2$. A Tabela 3.5 apresenta um resumo das incertezas adotadas e suas respectivas amostras. Vale lembrar que elas se referem a amostras de objetos comuns entre LAMOST e J-PLUS/WISE. Estas amostras serão usadas, respectivamente, para um novo treinamento dos modelos.

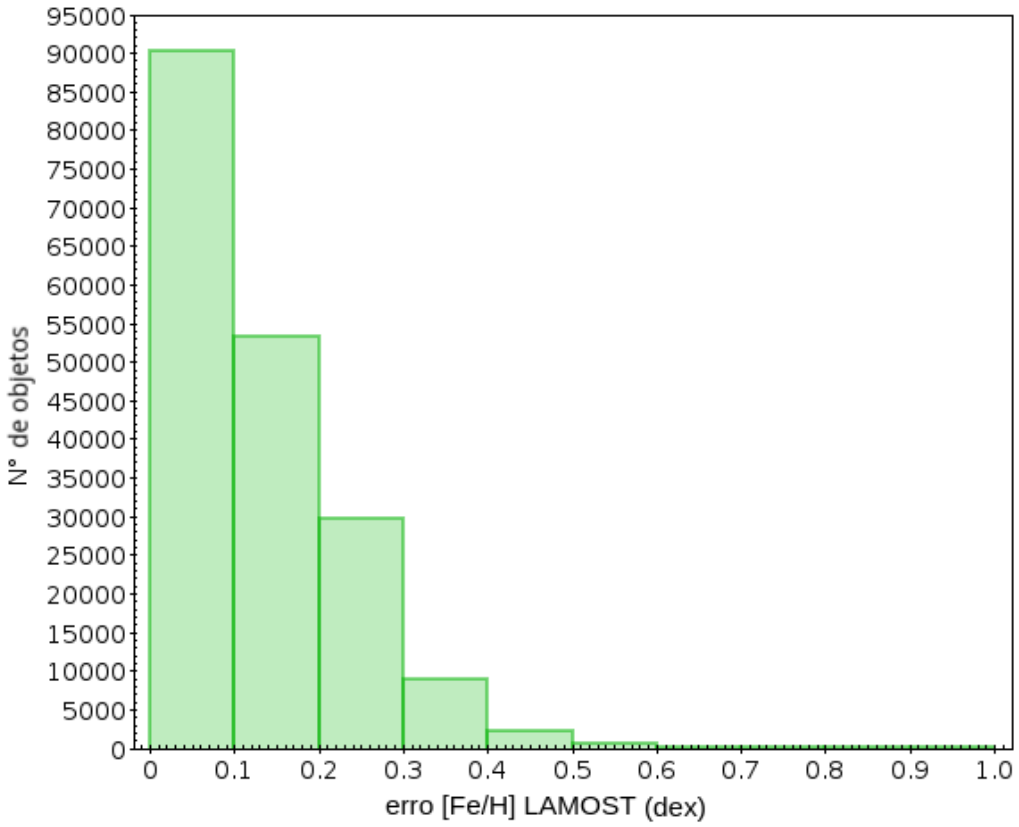


Figura 3.16: Distribuição das incertezas para a [Fe/H] na amostra de 186.374 objetos J-PLUS/WISE+LAMOST. Destes, 173.327 objetos possuem incerteza $\leqslant 0{,}3$ e 143.787 possuem incerteza $\leqslant 0{,}2$.

Tabela 3.5: Limites de incerteza adotados para T_{ef}, log g e [Fe/H] e definição das novas amostras de treinamento

-	Número de objetos
Amostra de treinamento	186.374
Estrelas com incerteza de $T_{ef} \leqslant 200$ K	137.865
Estrelas com incerteza de log g $\leqslant$ 0,2 dex	101.740
Estrelas com incerteza de [Fe/H] $\leqslant$ 0,2 dex	143.787

A otimização original foi apresentada na Subseção 3.1.4. Considerando os limites de incerteza estabelecidos para cada parâmetro, recalculamos esta otimização para as novas amostras (explicitadas na Tabela 3.5), para definir se os hiperparâmetros utilizados até aqui também se aplicam à amostra refinada pelas incertezas. O painel superior da Figura 3.17 mostra a otimização para T_{ef}. No painel central temos a otimização para log g e no painel inferior, para [Fe/H] - como fizemos até aqui, sempre que nos referirmos a um parâmetro, usaremos uma cor específica no gráfico: vermelho para T_{ef}, azul para log g e verde para [Fe/H]. Além disso, como pode ser visto nas barras de cor da Figura 3.17, quanto mais escura a região do gráfico, maior é o R^2 *score* da configuração.

Em ambos os casos, a melhor configuração de hiperpârametros é a mesma para [Fe/H]. Para log g, há uma mudança no número de *features* recomendadas (que agora são 45 *features* ao invés de 60) e, para T_{ef}, a fração de *features* por árvore passa a ser 0,5 (vide Tabela 3.6). Diferenças consideráveis surgem da otimização de log g, que mostra um ganho de 0,052 no R^2 *score* (ganho de $\approx 3\%$ na acurácia deste parâmetro). Na T_{ef}, esta variação é de 0,005 (+0,3% na acurácia) e na [Fe/H] é de 0,004 (+0,24% na acurácia). A Figura 3.18 apresenta a modelagem real, que considera os limites de incerteza, para amostra J-PLUS/WISE + LAMOST com erro de magnitude $\leqslant$ 0,1. Os cortes nas incertezas proporcionaram um ganho no R^2 *score* de 0,019 (0,97%) em T_{ef}, 0,053 (2,88%) em log g e 0,011 (0,61%) em [Fe/H].

Tabela 3.6: Configurações de hiperparâmetros dos modelos de melhor rendimento, na otimização original e na amostra com incerteza de $T_{ef} \leqslant 200$ K, incerteza de log g $\leqslant$ 0,2 e incerteza de [Fe/H] $\leqslant$ 0,2

Parâmetro	Config. original	Config. pós corte na incerteza
T_{ef}	(45, 0,25, 100)	(45, 0,5, 100)
log g	(60, 0,25, 100)	(45, 0,25, 100)
[Fe/H]	(60, 0,25, 100)	(60, 0,25, 100)

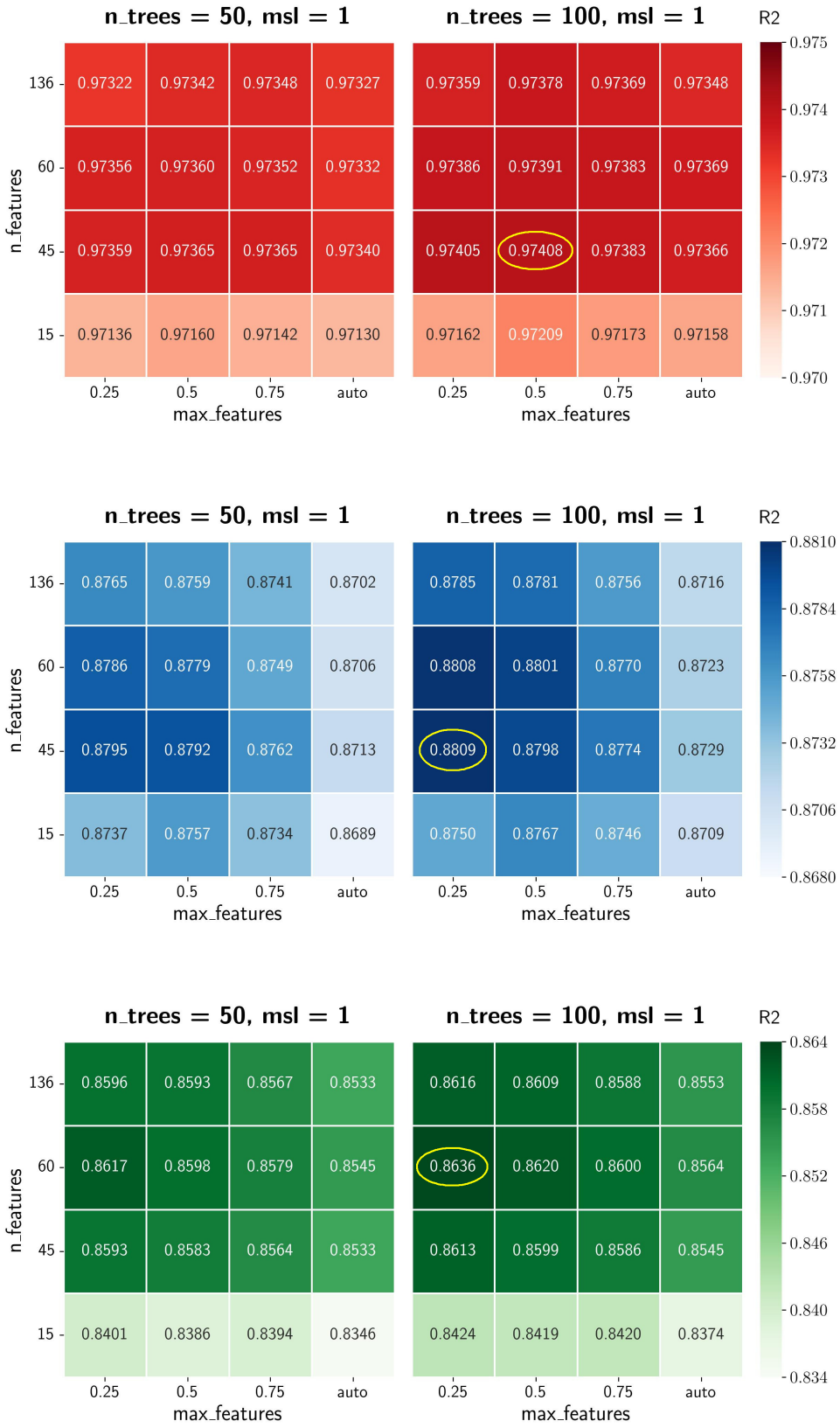
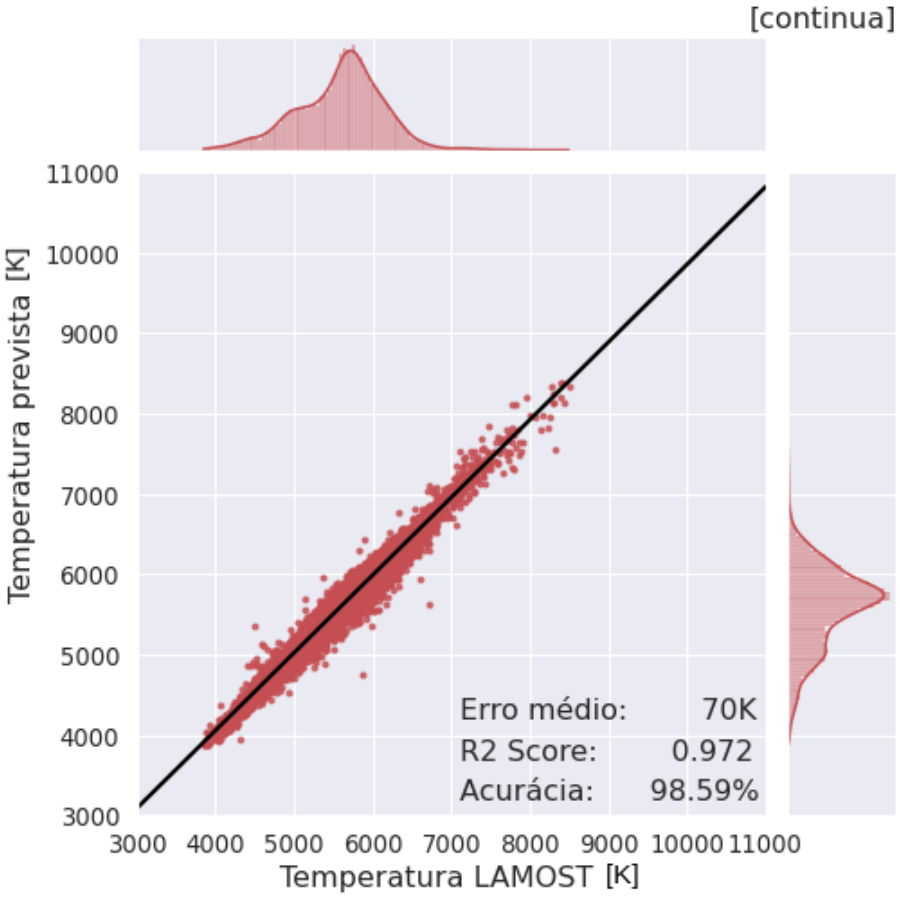
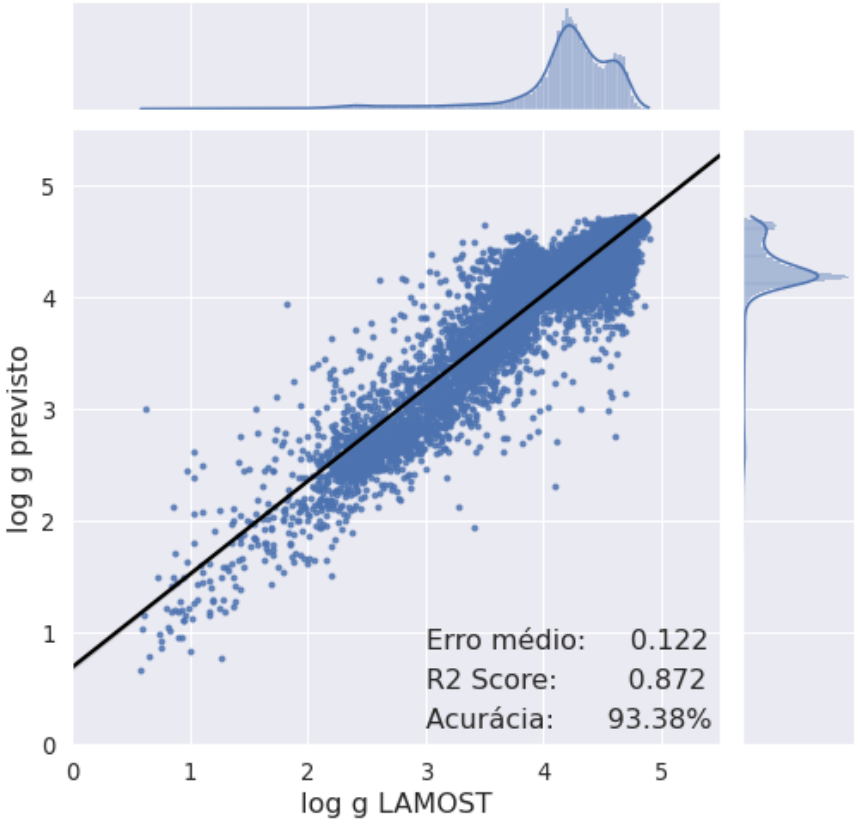


Figura 3.17: Otimização de hiperparâmetros para as amostras apresentadas na Tabela 3.5, que consideram uma limitação nas incertezas dos parâmetros da amostra de treinamento. Os painéis mostram os resultados da otimização: superior (em vermelho) para $T_{\rm ef}$; central (em azul) para log g; inferior (em verde) para [Fe/H].

[continua]


Temperatura prevista [K]
11000
10000
9000
8000
7000
6000
5000
4000
3000
3000 4000 5000 6000 7000 8000 9000 10000 11000
Temperatura LAMOST [K]
Erro médio: 70K
R2 Score: 0.972
Acurácia: 98.59%


log g previsto
5
4
3
2
1
0
0 1 2 3 4 5
log g LAMOST
Erro médio: 0.122
R2 Score: 0.872
Acurácia: 93.38%

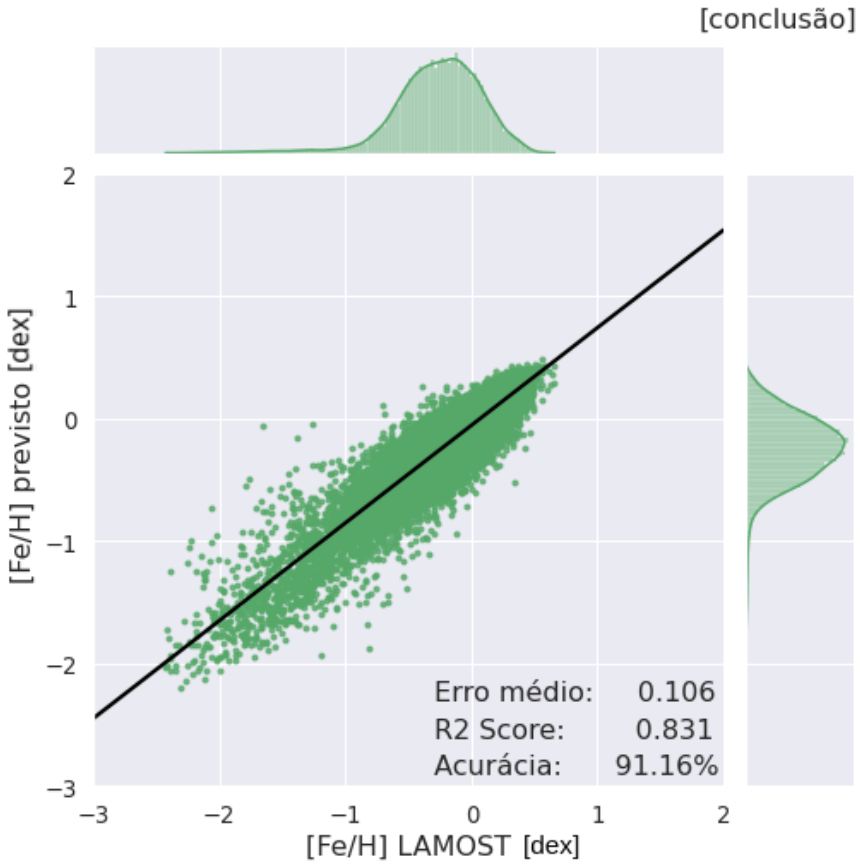


Figura 3.18: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas J-PLUS/WISE+LAMOST, com as seguintes limitações nas incertezas dos parâmetros: incerteza em $T_{\rm ef} \leqslant 200$ K, incerteza em log g $\leqslant 0{,}2$ e incerteza em [Fe/H] $\leqslant 0{,}2$. Os painéis são semelhantes aos da Figura 3.10 e comparam os valores previstos pelo algoritmo com os valores disponíveis no LAMOST. Todos os parâmetros tiveram ganho no R^2 *score*, em comparação com as simulações apresentadas na Figura 3.13.

Uma última otimização para a previsão de log g pode ser feita usando as distância das estrelas, baseada no valor de paralaxe do Gaia, para estimar suas magnitudes absolutas e incluí-las junto as 136 magnitudes utilizadas até então. Isso causa uma melhora significativa no rendimento do modelo de log g. Se selecionarmos apenas distâncias com incerteza menor que 30%, limitamos a amostra de 101.740 estrelas para 97.526 objetos. A inclusão desta *feature* eleva o R^2 *score* do log g em 0,073, que representa um ganho de acurácia de 3,83% com relação ao modelo da Figura 3.18 (vide painel superior da Figura 3.19).

Schlaufman & Casey (2014) apresentam a influência positiva dos dados das bandas J, H e K do 2MASS, no cálculo de metalicidade. Por este motivo, incluímos os dados de magnitude destas bandas e recalculamos as cores (agora totalizando 190 *features*). A inclusão desta *feature* eleva o R^2 *score* da [Fe/H] em 0,025, que representa um ganho de acurácia de 1,36% com relação ao modelo apresentado na Figura 3.18 (vide painel inferior da Figura 3.19). Na Figura 3.19, vemos o modelo final de log g e [Fe/H], para erro de magnitude $< 0{,}1$.

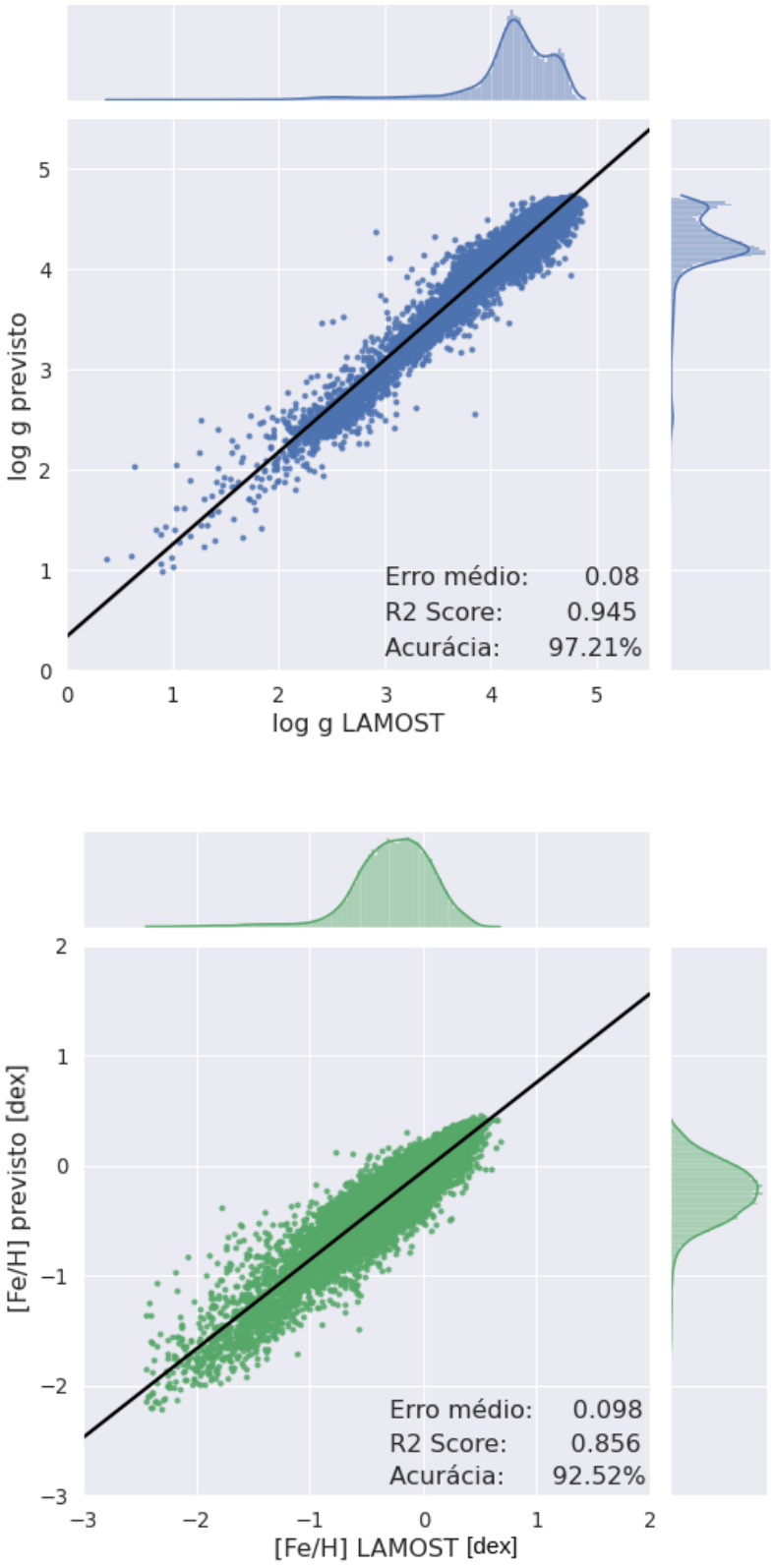


Figura 3.19: Otimização final dos modelos de log g e [Fe/H], com limitações na incerteza de log g ($\leqslant$ 0,2) e de [Fe/H] ($\leqslant$ 0,2). Os painéis são semelhantes aos da Figura 3.18. Com a inserção das magnitudes absolutas, o R^2 *score* do log g teve um ganho de 0,073 (+3,83%). No [Fe/H], a inclusão das bandas J, H e K provocou ganho de 0,025 no R^2 *score* (+1,36%).

A Tabela 3.7 apresenta, de forma resumida, os rendimentos alcançados pelos levantamentos auxiliares testados. No LAMOST, considera-se as amostras com limitação de incertezas apresentadas na Tabela 3.5.

Tabela 3.7: Rendimento dos modelos para os 4 levantamentos de dados testados

Levantamento	Amostra cruzada	R^2 *score*: T_{ef}	R^2 *score*: log g	R^2 *score*: [Fe/H]
TESS	1.291.942	0,908 (95,28%)	0,188 (43,36%)	0,775 (88,03%)
SEGUE	14.831	0,969 (98,44%)	0,695 (83,37%)	0,812 (90,11%)
GALAH	2.018	0,908 (95,29%)	0,761 (87,24%)	0,752 (86,72%)
LAMOST	Ver Tab. 3.5	0,972 (98,59%)	0,945 (97,21%)	0,856 (92,52%)

Já que os dados de log g do TESS não apresentam consistência, GALAH não possui uma amostra suficiente para treinamento e SEGUE é muito inferior a LAMOST, em quantidade de objetos, a segunda listagem de requisitos (definida na Subseção 2.2.1) será aplicada apenas ao levantamento auxiliar de melhor rendimento: o LAMOST. Na seção seguinte, vamos avaliar se estes requisitos extras contribuem, de alguma forma, para a melhoria do algoritmo - seja elevando o rendimento de suas previsões ou simplesmente atendendo a uma maior quantidade de estrelas do Kepler (isto será ponderado apenas no caso de que não se perca a qualidade do modelo).

3.2.2 Amostra J-PLUS menos restrita

Vimos, na Subseção 2.2.1, alguns ajustes que podem ser feitos na amostra J-PLUS de interesse, a fim de expandir a amostra de treinamento e permitir a caracterização de mais estrelas do Kepler - chamamos este ajuste de segunda listagem de requisitos. Essas novas condições permitem que usemos agora objetos observados: a) em uma abertura diferente (abertura 3"); b) em menos de 12 filtros; c) com erro de magnitude (e_mag) maior que 0,1 (e_mag $<$ 0,2). Manteremos observações de objetos com mais de 90% de probabilidade de serem estrelas. Chamaremos esta nova amostra J-PLUS de amostra menos restrita.

As amostras manterão as 156 colunas J-PLUS de dados usadas anteriormente (2 colunas de identificação, 2 referentes aos parâmetros astrométricos α e δ, as 12 magnitudes J-PLUS, 12 colunas de correção de extinção do J-PLUS, as 4 magnitudes WISE, 4 colunas de correção de extinção do WISE e as 120 cores calculadas a partir das 16 magnitudes) e adicionaremos os dados de distância Gaia para o modelo de log g e as bandas J, H e K para [Fe/H]. A amostra menos restrita será novamente cruzada com os dados do LAMOST e

oferecida ao algoritmo para treinamento de seus modelos.

A primeira mudança na pesquisa de objetos de interesse foi realizada no erro de magnitude aceito. Alteramos este valor para e_mag $<$ 0,2. Isto permitiu que as estrelas com 0,1 $<$ e_mag $<$ 0,2 fossem adicionadas à amostra de base. Isto alterou o número de estrelas na amostra de 1.365.454 para 1.935.674 objetos (vide Tabela 3.8). Neste momento, todos os demais requisitos foram mantidos. A maior vantagem dos modelos com erro de até 0,2 está no acréscimo apreciável do número de estrelas do Kepler caracterizadas - de 29.164 para 44.483 objetos (um aumento maior que 52%).

Tabela 3.8: Amostra de interesse, considerando também a abertura de 3" e e_mag $<$ 0,2

Abertura 6" e e_mag $<$ 0,1	Abertura 6" e e_mag $<$ 0,2	Abertura 3" e e_mag $<$ 0,1	Abertura 3" e e_mag $<$ 0,2
1.365.454	1.935.674	1.656.236	2.297.550

Também avaliamos como variava a amostra de base, vista na Subseção 3.2.1, com uma mudança na abertura. Ao invés de 6", permitimos a seleção de objetos da abertura 3". A abertura 3" permitiu a seleção de mais estrelas tanto com e_mag $<$ 0,1 quanto com e_mag $<$ 0,2 (vide Tabela 3.8). Isso ocorre porque estamos selecionando objetos de uma faixa de erro específica. Na verdade, se considerarmos os catálogos totais para as aberturas 3" e 6" (ou seja, a relação de objetos que contêm qualquer erro de magnitude e que foram observados em qualquer quantidade de filtros), ambas as aberturas contém um total de 7.242.301 estrelas. Não obtivemos uma resposta conclusiva sobre o motivo disto acontecer, mas é fato que existem mais objetos com erro menor que 0,2 na abertura 3".

Podemos ver, na Figura 3.20, a distribuição dos erros de magnitude por filtro nas duas aberturas analisadas. Para qualquer filtro, a abertura 3" sempre apresenta número superior de objetos para erro de magnitude muito baixo (e_mag $<$ 0,05). Para os filtros uJAVA, J0378, J0395, J0410, J0430, gSDSS e J0515, os erros na abertura 3" são sempre menores que na abertura 6", na faixa de erro adotada (e_mag $<$ 0,2). Para os demais filtros, a abertura 6" começa a predominar para a faixa de erro entre 0,1 e 0,2 mag, apresentando mais objetos nesta faixa.

Em uma última etapa, pretendíamos selecionar estrelas que possuíam observações em apenas 11 filtros, a fim de preencher esta lacuna de observação com inteligência artificial, realizando previsões para os valores no(s) filtro(s) faltante(s). Entretanto, a etapa revelou uma quantidade insignificante de boas estrelas. A grande maioria das estrelas vistas em menos de 12 filtros apresentaram um erro de magnitude muito alto. Mesmo que fosse viável aceitar estrelas com erro de magnitude até 1,0, isso só adicionaria cerca de 10.000 objetos à amostra. As estrelas desta etapa foram invalidadas e descartadas.

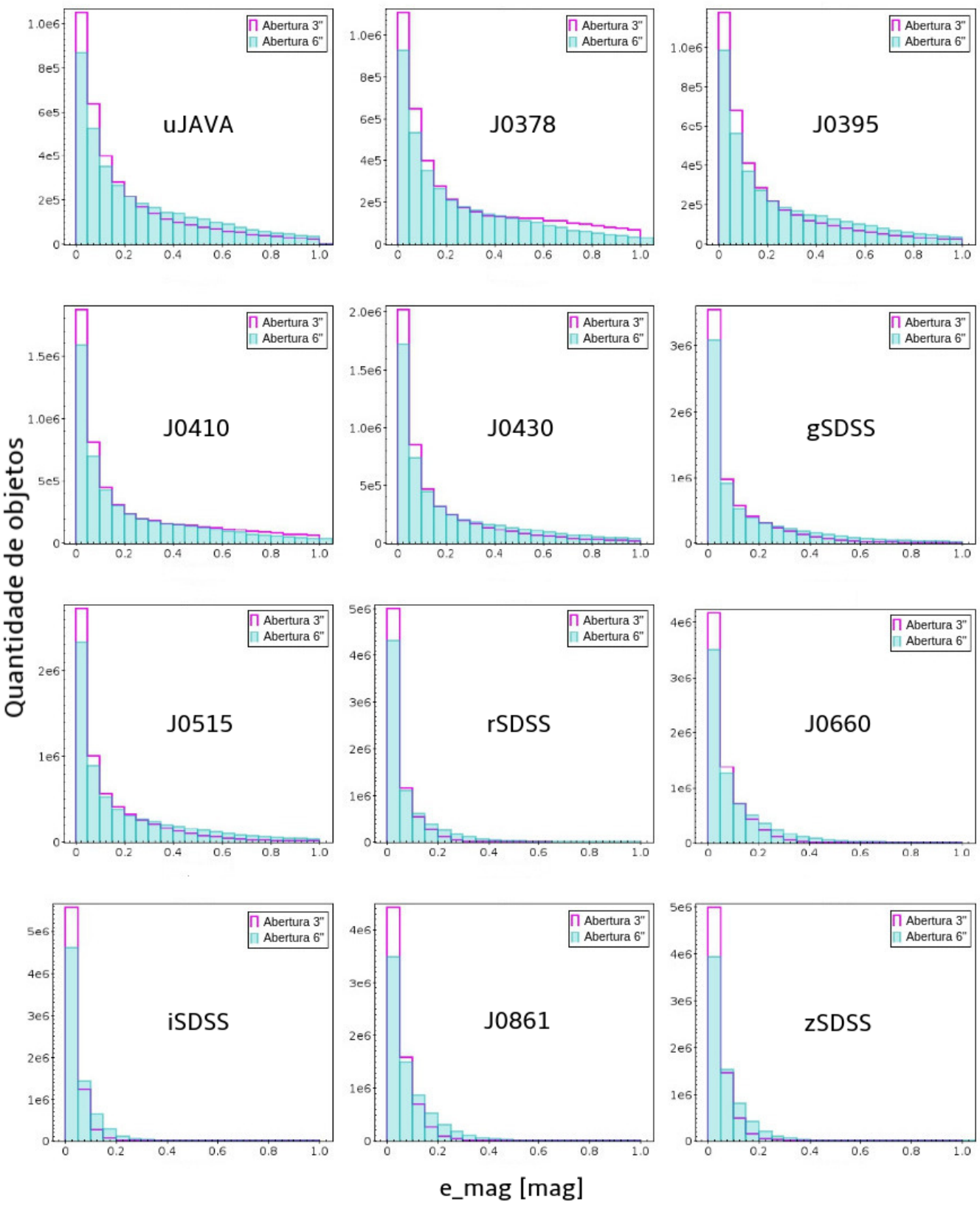


Figura 3.20: Distribuição de e_mag por filtro para as aberturas 3" e 6". A abertura 3" sempre apresenta mais objetos em e_mag $<$ 0,05. Para uJAVA, J0378, J0395, J0410, J0430, gSDSS e J0515, os erros em 3" são sempre menores que em 6" para e_mag $<$ 0,2.

3.2.2.1 Remodelagem para T_{ef}, log g e [Fe/H]

Vimos, na Tabela 3.8, quatro possíveis amostras base para o J-PLUS, 3 delas expandidas com a mudança nos requisitos da seleção. Na Subseção 3.2.1.4, vimos a testagem do algoritmo para o melhor levantamento auxiliar na primeira amostra base: com abertura 6" e e_mag $< 0{,}1$. A seguir vemos os testes com as outras 3 amostras da tabela, cruzadas com a do LAMOST. Como estas amostras são todas maiores que a primeira, elas possivelmente caracterizam mais estrelas da missão Kepler. A Figura 3.21 compara os campos de observação do Kepler e do J-PLUS.

A caracterização poderá ser feita nas estrelas da missão Kepler que foram observadas no DR2 do J-PLUS, isto é, as estrelas que aparecem sobrepostas na Figura 3.21. A aplicação do modelo *Random Forest* (e outras técnicas de AM) exige que os objetos da previsão tenham sido observados pelas mesmas *features* do modelo, ou seja, um modelo baseado nos filtros do J-PLUS/WISE consegue caracterizar objetos observados por estes mesmos filtros. O algoritmo foi escrito de modo a adaptar-se às futuras liberações de dados do J-PLUS. Assim, à medida que o J-PLUS observe mais estrelas do campo do Kepler, novos objetos podem ser caracterizados.

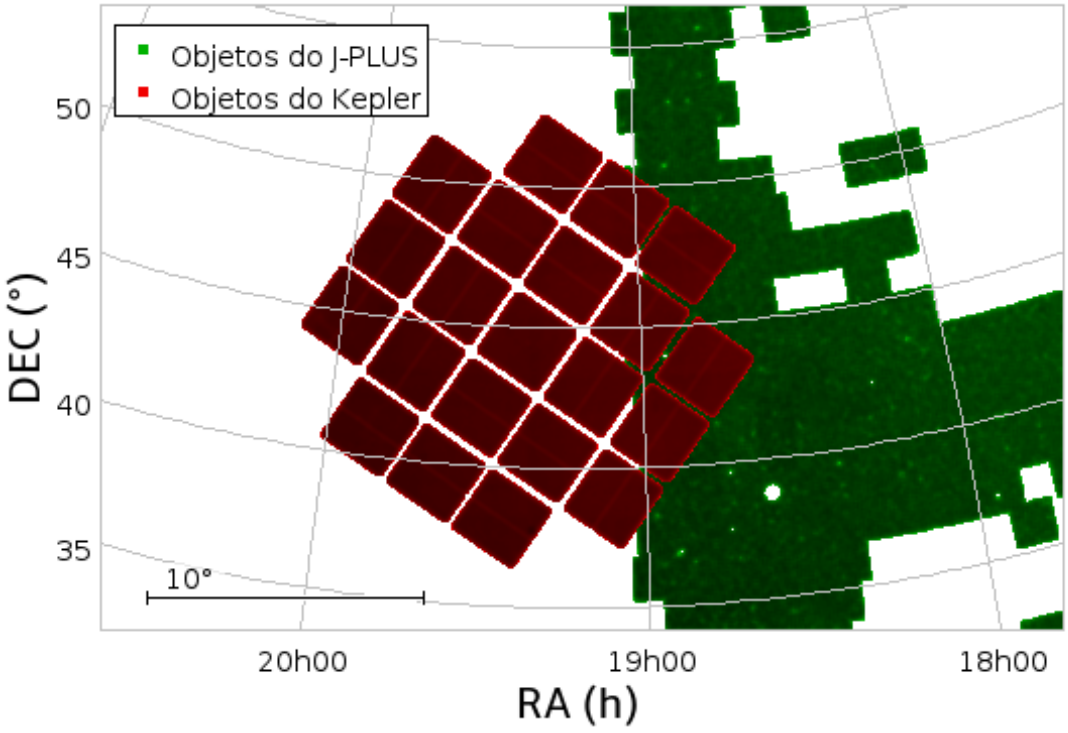
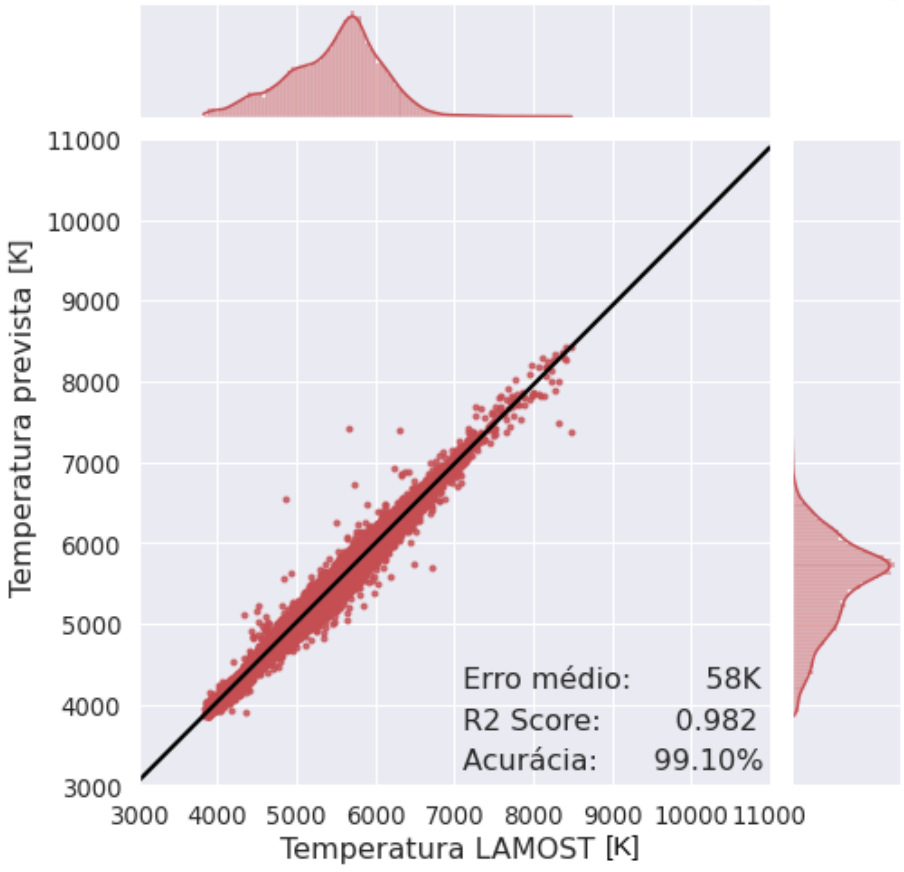
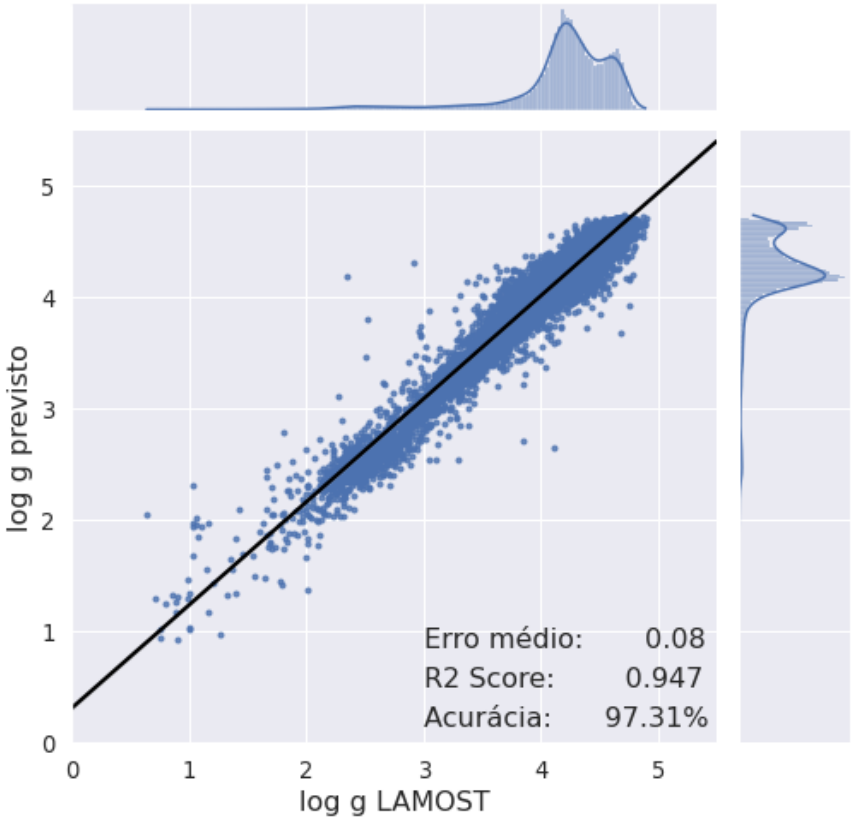


Figura 3.21: Cobertura do J-PLUS (em verde) na região observada pelo Kepler (em vermelho). A caracterização poderá ser feita nas estrelas do Kepler que foram observadas no DR2 do J-PLUS, ou seja, as estrelas que aparecem sobrepostas nesta figura. Isso permite que as estrelas caracterizadas tenham sido observadas pelas *features* utilizadas nos modelos.

[continua]


Temperatura prevista [K]
Temperatura LAMOST [K]
Erro médio: 58K
R2 Score: 0.982
Acurácia: 99.10%


log g previsto
log g LAMOST
Erro médio: 0.08
R2 Score: 0.947
Acurácia: 97.31%

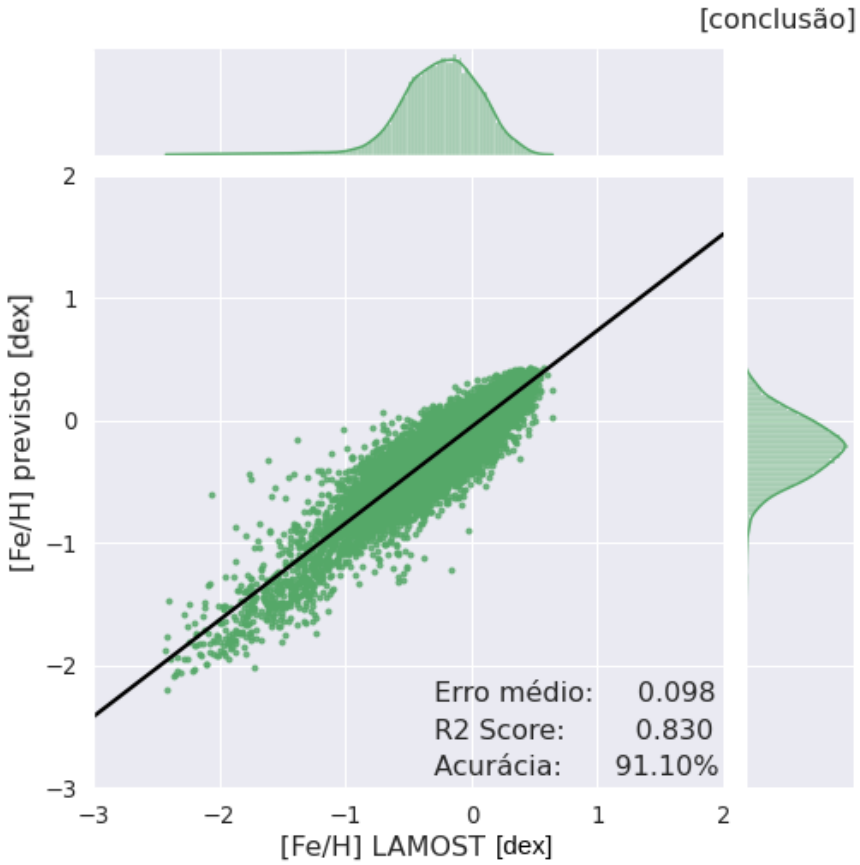


Figura 3.22: Simulação em modelagem *Random Forest* baseada nos filtros do J-PLUS/WISE, para estrelas J-PLUS/WISE+LAMOST observadas na abertura 6" e com e_mag $<$ 0,2. Para os parâmetros físicos, as incertezas são: $\leqslant$ 200 K em $T_{\rm ef}$, $\leqslant$ 0,2 em log g e $\leqslant$ 0,2 em [Fe/H]. Os painéis são semelhantes aos da Figura 3.10 e comparam os valores previstos pelo algoritmo com os disponíveis no LAMOST.

Na Figura 3.22, observamos a modelagem baseada na amostra de treinamento com erro de magnitude menor que 0,2. Nela, nota-se um crescimento sutil na acurácia de $T_{\rm ef}$, com relação ao visto na Figura 3.18, de 0,51%. A falta de oscilação negativa já é suficiente para validar a expansão da amostra para $T_{\rm ef}$ (vide Tabela 3.9). As variações no R^2 *score* de [Fe/H] e de log g (vistos na Figura 3.19) são de -0,002 e +0,002, respectivamente. Apesar da variação em [Fe/H] ser negativa, isso não invalida a nova amostra. É esperado que as acurácias possam cair, já que a amostra de treinamento destes modelos considerou um erro maior nas magnitudes dos filtros J-PLUS/WISE. Além disso, a variação representa uma oscilação de apenas 0,11% na acurácia deste parâmetro.

É importante observar os limites dos valores dos parâmetros usados na amostra de treinamento, porque isso mostra os limites em que o modelo foi treinado e, portanto, o tipo de estrela que ele aprendeu a reconhecer. Se o modelo é aplicado em estrelas que estão fora deles, não existe garantia que não seja criado um viés nas previsões. Os limites são os mesmos nas duas amostras de treinamento: 3790 K $<$ $T_{\rm ef}$ $<$ 8500 K; 0,11 $<$ log g $<$ 4,90; -2,5 $<$ [Fe/H] $<$ 0,72.

Tabela 3.9: Expansão da amostra de treinamento J-PLUS/WISE + LAMOST 6". Ambas as amostras consideram as seguintes incertezas: incerteza $\leqslant$ 200 K para $T_{\rm ef}$ e $\leqslant$ 0,2 para log g e [Fe/H].

Amostra de treinamento para 6" e e_mag $<$ 0,1	Amostra de treinamento para 6" e e_mag $<$ 0,2
$\leqslant$ 143.787 objetos (vide Tab. 3.5)	$\leqslant$ 152.275 objetos[29]

As amostras de treinamento produzidas pelas estrelas observadas na abertura 3" não mostraram rendimento que justifique seu uso. Estes rendimentos estão expressos na Tabela 3.10. Por este motivo, mantivemos a modelagem apresentada pelas amostras de abertura 6". Por fim, os parâmetros para as estrelas do Kepler serão calculados com dois modelos: um de maior precisão (baseado em estrelas com e_mag $<$ 0,1) e outro de precisão um pouco inferior, mas ainda assim excelente (baseado em estrelas com e_mag $<$ 0,2).

Tabela 3.10: Rendimento dos modelos para as 4 amostras de treinamento que usaram os filtros do J-PLUS/WISE + os parâmetros físicos do LAMOST

-	Amostra cruzada	R^2 *score*: $T_{\rm ef}$	R^2 *score*: log g	R^2 *score*: [Fe/H]
LAMOST + J-PLUS 6" + WISE (e_mag $<$ 0,1)	$\leqslant$ 173.327	0,972 (98,59%)	0,945 (97,21%)	0,856 (92,52%)
LAMOST + J-PLUS 6" + WISE (e_mag $<$ 0,2)	$\leqslant$ 189.146	0,982 (99,10%)	0,947 (97,31%)	0,830 (91,10%)
LAMOST + J-PLUS 3" + WISE (e_mag $<$ 0,1)	197.698	0,953 (97,62%)	0,685 (82,76%)	0,757 (87,00%)
LAMOST + J-PLUS 3" + WISE (e_mag $<$ 0,2)	214.499	0,953 (97,62%)	0,663 (81,42%)	0,743 (86,20%)

3.3 Aplicação para estrelas alvo

Depois que todos os levantamentos auxiliares forem testados e estiver decidido qual amostra de treinamento oferece as melhores previsões para um número suficiente de objetos, é o momento de salvar estes modelos e aplicá-los em um segundo algoritmo, que

[29] Existem, para e_mag $<$ 0,2: 144.774 objetos na amostra de $T_{\rm ef}$; 103.341 na de log g; 152.275 na de [Fe/H].

aplica a modelagem obtida pelo primeiro em qualquer estrela da missão Kepler que tiver sido observada por cada uma das duas amostras J-PLUS, definidas na Subseção 2.2.1. Para identificar estas estrelas comuns, novamente será aplicada a correlação cruzada por ascensão reta e declinação. Ainda é necessário que as estrelas da missão Kepler atendam a estes requisitos porque as modelagens foram baseadas em estrelas que os atendiam. Outros parâmetros como luminosidade e raio serão inferidos para estas estrelas com base na paralaxe fornecida pelo Gaia eDR3[30].

3.4 Cálculo de luminosidade e raio

Na etapa anterior, $T_{\rm ef}$, log g e [Fe/H] das estrelas da missão Kepler que também foram observadas pelo J-PLUS/WISE puderam ser previstos pelos modelos de Aprendizagem de Máquina deste trabalho. Para a determinação da luminosidade ($L_\star$) e do raio ($R_\star$), utilizamos uma abordagem diferente. O Gaia eDR3 fornece medidas de posição celeste e magnitude aparente na banda G para aproximadamente 1,8 bilhão de fontes. Destas, 1,5 bilhão contêm dados de paralaxe (bem como as distâncias dos objetos, calculadas com base nesta paralaxe), movimento próprio e cor ($G_{BP} - G_{RP}$) (Brown et al., 2021). Para as amostras do J-PLUS de interesse (1ª e 2ª coluna da Tabela 3.8), temos dados de paralaxe/distância para 1.335.796 das 1.365.454 estrelas com e_mag $<$ 0,1 ($\approx 97,8\%$). Para e_mag $<$ 0,2, 1.895.292 das 1.935.674 estrelas possuem este valor ($\approx 97,9\%$).

Apesar da Aprendizagem de Máquina não ser usada para prever estes parâmetros diretamente, ela não se torna desnecessária, já que usaremos o valor de $L_\star$ para estimar $R_\star$ e o cálculo de $L_\star$ depende da $T_{\rm ef}$ prevista pelo algoritmo. Com os valores de paralaxe/distância (d, em parsecs) e magnitude aparente na banda G (m_G) do Gaia eDR3, podemos estimar a magnitude absoluta nesta banda (M_G, Equação 3.2) para os objetos J-PLUS, isolando-a na Equação 3.1:

$$m_G = M_G + 5\log(d) - 5 \tag{3.1}$$

$$M_G = m_G - 5\log(d) + 5 \tag{3.2}$$

A incerteza em M_G, σ_{M_G}, é baseada nas incertezas em m_G e em d (σ_{m_G} e σ_d, respectivamente). Porém, para muitos objetos, não temos σ_{m_G}. Em outros, σ_d deve ser analisado com cuidado, pois pertencem a objetos muito distantes. Resolvemos não considerar estas incertezas neste trabalho, mas pretendemos incluí-las futuramente, seguindo a estratégia proposta no Capítulo 5. Por este motivo, os erros dos parâmetros diretamente dependentes destas incertezas poderão estar subestimados - isto não inclui $\sigma_{\log g}$, dependente de M_G, pois o modelo de log g já foi treinado apenas com objetos com $\sigma_d < 30\%$.

[30] https://cds.u-strasbg.fr/gaia#gedr3

Com M_G calculada, podemos encontrar a magnitude bolométrica (M_{bol}) das estrelas. Sua diferença para a M_G é que ela considera o fluxo total da estrela, em todos os comprimentos de onda. Para fins de cálculo, é necessário ainda definir a chamada correção bolométrica (BC): uma correção feita na M_G de um objeto, de modo a convertê-la em M_{bol}. Seu valor é alto para estrelas que irradiam grande parte de sua energia fora da faixa visível do espectro eletromagnético. A Tabela 3.11 mostra valores típicos de correção bolométrica para algumas classes estelares, de acordo com Kaler (1989).

Tabela 3.11: Correção bolométrica típica com relação a magnitude visual (M_v) para algumas classes espectrais, segundo Kaler (1989).

Classe	Seq. principal	Gigantes	Supergigantes
O3	-4.30	-4.20	-4.00
G0	-0.10	-0.13	-0.10
G5	-0.14	-0.34	-0.20
K0	-0.24	-0.42	-0.38
K5	-0.66	-1.19	-1.00
M0	-1.21	-1.28	-1.30

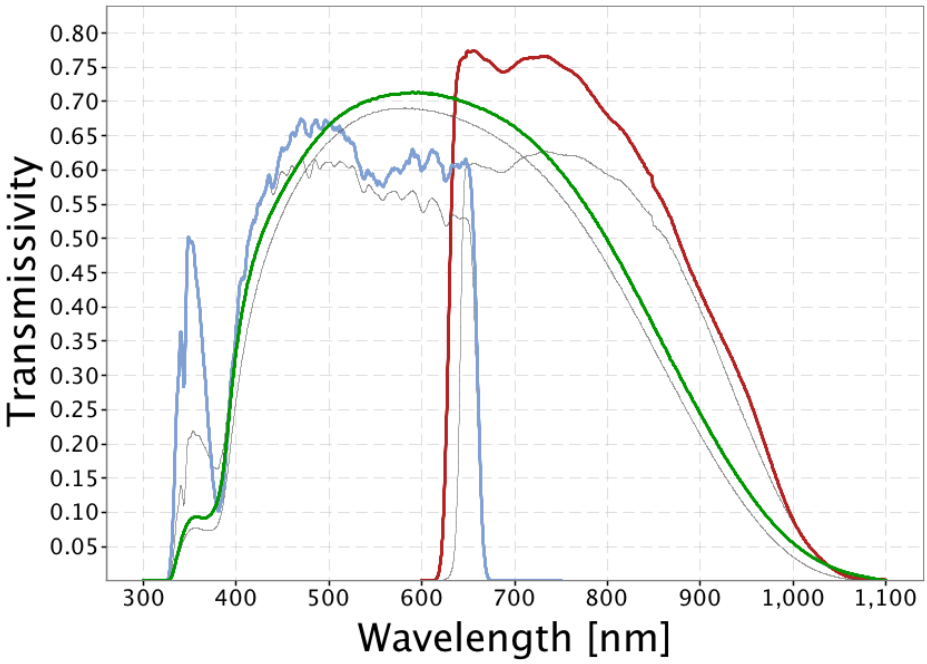


Figura 3.23: Curvas de transmissão do Gaia. As linhas coloridas mostram as bandas revisadas para G (verde), GBP (azul) e GRP (vermelha), definindo o sistema fotométrico do Gaia DR2. As linhas finas e cinzas mostram as bandas pré-lançamento publicadas em Jordi et al. (2010), usadas no Gaia DR1.

As curvas de transmissão do Gaia podem ser vistas na Figura 3.23. As linhas finas e cinzas mostram as bandas pré-lançamento publicadas em Jordi et al. (2010), usadas no Gaia DR1[31]. Houve uma revisão dos filtros entre o DR1 e DR2. As linhas coloridas na figura mostram as bandas revisadas para G (verde), GBP (azul) e GRP (vermelha), definindo o sistema fotométrico usado no Gaia DR2.

O trabalho de Jordi et al. (2010) calcula e fornece os dados de BC para uma amostra de estrelas Gaia com base em seus valores de $T_{\rm ef}$, log g e [Fe/H]. Podemos calcular a BC aproximada na banda G das estrelas para as quais realizamos a previsão de $T_{\rm ef}$, log g e [Fe/H]. Para isso, também usamos um algoritmo *Random Forest* de regressão linear, treinado com os dados fornecidos pelos autores. O algoritmo para BC mostrou acurácia de 99,89% na previsão geral. O gráfico do treinamento é apresentado na Figura 3.24. Na Tabela 3.12 vemos as incertezas da previsão de BC.

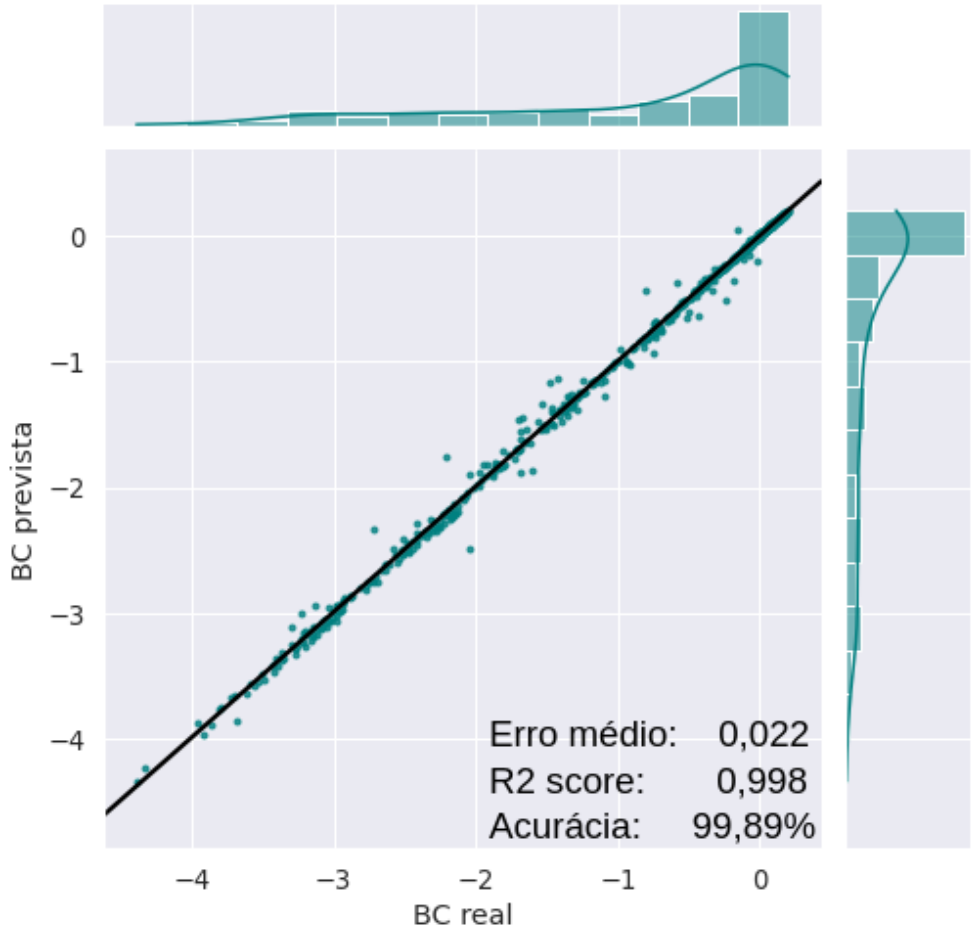


Figura 3.24: Simulação em modelagem *Random Forest* para correção bolométrica, usando regressão linear, baseada nos dados de Jordi et al. (2010). Nesta figura, podemos ver o comparativo entre BC de Jordi et al. (2010) e a BC prevista pelo algoritmo.

[31] Consulte https://www.cosmos.esa.int/web/gaia/iow_20180316

Tabela 3.12: Incertezas para o algoritmo de correção bolométrica, baseado em regressão linear.

	Erro
Erro absoluto médio	0,022
Erro absoluto mediano	0,007
Desvio padrão (σ)	0,052
Erro máximo	0,562
R^2 *score*	0,998

A Figura 3.25 mostra a distribuição das incertezas, ou seja, a diferença entre o valor real e o valor previsto pelo algoritmo. O desvio padrão (σ) nos dá um retorno sobre o valor aceitável para a incerteza. Vamos considerar aceitável até 3 desvios padrões (3σ = 0,156): 97,8% dos objetos possuem incertezas $\leqslant 3\sigma$; 7 objetos possuem incertezas $\geqslant 5\sigma$ (0,9%).

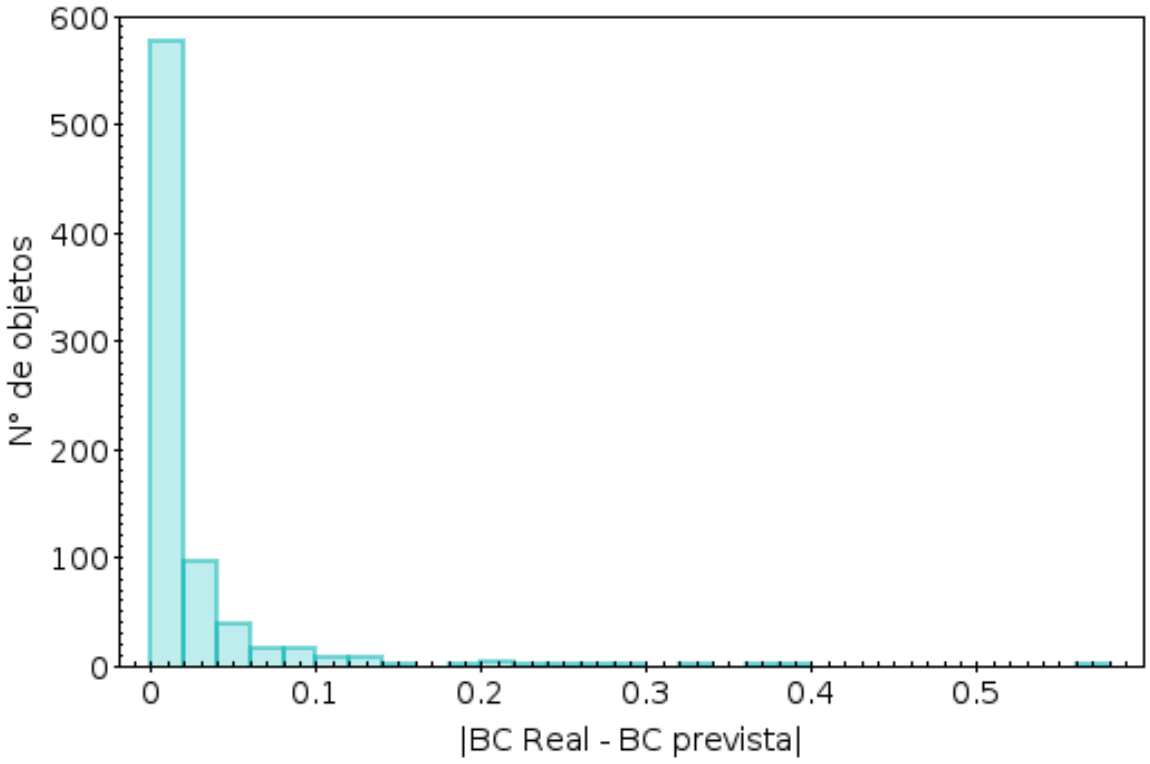


Figura 3.25: Distribuição de incertezas da simulação: |BC real - BC prevista|.

Na Figura 3.26 podemos analisar a dependência da incerteza (BC real - BC prevista) com a $T_{\rm ef}$ das estrelas. Com ela, somos capazes de perceber que incertezas maiores que 3σ se concentram abaixo de 3.500 K.

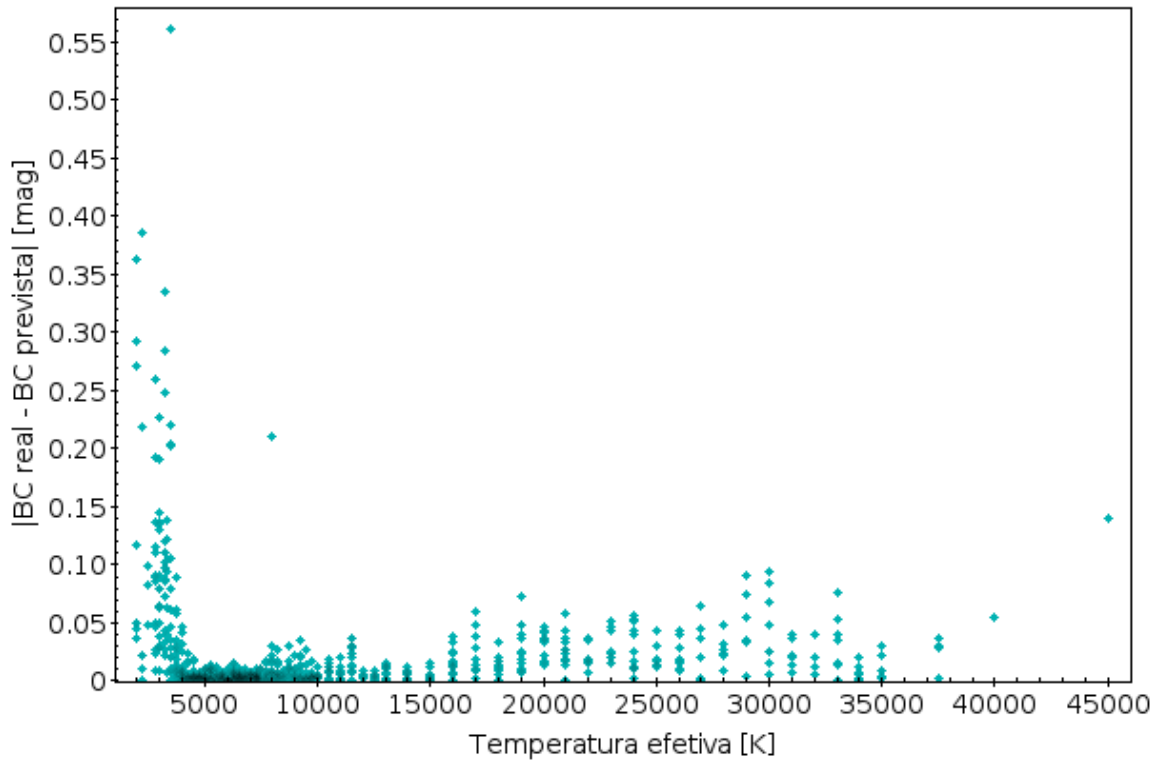


Figura 3.26: Dependência da incerteza na simulação de BC com a $T_{\rm ef}$ da estrela. Note que incertezas maiores que 3σ se concentram abaixo de 3.500 K.

Também de acordo com Jordi et al. (2010), conhecendo agora BC, podemos encontrar a M_{bol} através de:

$$BC_G = M_{bol} - M_G \rightarrow M_{bol} = M_G + BC_G \tag{3.3}$$

Como não obtivemos a incerteza em M_G, a incerteza em M_{bol} será tomada como a incerteza em BC_G, $\sigma_{M_{bol}} = \sigma_{BC_G}$. Sabendo o valor de M_{bol}, podemos finalmente calcular a luminosidade da estrela (L) com a relação apresentada por Carroll & Ostlie (2007):

$$M_{bol} = -2,5\log_{10}\left(\frac{L}{L_0}\right) \rightarrow L = 10^{-0,4M_{bol}}L_0 \tag{3.4}$$

onde L_0 á a luminosidade no ponto zero = 3,0128 x 10^{28} W. Podemos dividir L pelo valor da luminosidade solar (3,828 x 10^{26} W) para obter L em unidades solares ($L/L_\odot$). Para propagar as incertezas de L (σ_L), podemos usar a Equação 3.5.

$$\sigma_L^2 = [(-0,4)10^{-0,4M_{bol}}\ln(10)\sigma_{M_{bol}}L_0]^2 + (10^{-0,4M_{bol}}\sigma_{L_0})^2 \tag{3.5}$$

Como não conhecemos σ_{L_0}, a Equação 3.5 pode ser reduzida a:

$$\sigma_L = (-0,4)10^{-0,4M_{bol}}\ln(10)\sigma_{M_{bol}}L_0 \tag{3.6}$$

Com a luminosidade calculada pela Equação 3.4 e $T_{\rm ef}$ estimada pelo algoritmo, podemos estimar o raio (R) destas estrelas através da Equação 3.7:

$$L = 4\pi R^2\sigma T_{\rm ef}^4 \rightarrow R = \left(\frac{L}{4\pi\sigma T_{\rm ef}^4}\right)^{1/2} \tag{3.7}$$

onde σ é a chamada constante de Stefan-Boltzmann, que tem o valor de 5,6697 x 10^{-5} erg cm^{-2} s^{-1} K^{-4}. Também podemos usar a equação equivalente, que expressa seus parâmetros em unidades solares:

$$\frac{L}{L_\odot} = \left(\frac{R}{R_\odot}\right)^2\left(\frac{T_{\rm ef}}{T_\odot}\right)^4 \rightarrow \frac{R}{R_\odot} = \left[\frac{(L/L_\odot)}{(T_{\rm ef}/T_\odot)^4}\right]^{1/2} \tag{3.8}$$

Para propagarmos a incerteza em R ($\sigma_{R_\star}$), simplificaremos a Equação 3.8. Chamaremos $R/R_\odot = R_\star$ e $L/L_\odot = L_\star$. A unidade de medida da $T_{\rm ef}$ das estrelas caracterizadas será o Kelvin, então manteremos a notação $T_{\rm ef}/T_\odot$ nas equações a seguir e faremos a conversão para unidades solares somente quando necessário.

$$R_\star = L_\star^{1/2}(T_{\rm ef}/T_\odot)^{-2} \tag{3.9}$$

Como a União Astronômica Internacional (IAU, *International Astronomical Union*)[32] não define um valor para a incerteza da temperatura solar ($\sigma_{T_\odot}$), não podemos inserir $\sigma_{T_\odot}$ na propagação de incertezas de $R_\star$. Vamos considerar $T_\odot = 5772$ K, como sugerido pela IAU. A incerteza $\sigma_{R_\star}$ pode ser obtida através de:

$${\sigma_{R_\star}}^2 = \left(\frac{1}{2}L_\star^{-1/2}\sigma_{L_\star}(T_{\rm ef}/T_\odot)^{-2}\right)^2 + \left((-2)L_\star^{1/2}(T_{\rm ef}/T_\odot)^{-3}(\sigma_{T_{\rm ef}}/T_\odot)\right)^2 \tag{3.10}$$

$$\sigma_{R_\star} = \sqrt{\left(\frac{1}{2}L_\star^{-1/2}\sigma_{L_\star}(T_{\rm ef}/5772)^{-2}\right)^2 + \left((-2)L_\star^{1/2}(T_{\rm ef}/5772)^{-3}(\sigma_{T_{\rm ef}}/5772)\right)^2} \tag{3.11}$$

[32] A IAU, na resolução B1 de 2015, que pode ser consultada em: https://www.iau.org/static/resolutions/IAU2015_English.pdf, atualizou os valores dos parâmetros solares.

3.5 Cálculo de massa

O trabalho de Torres et al. (2010) calculou as massas de 190 estrelas em 95 sistemas binários separados. Estes são sistemas não interagentes (não existe troca de matéria) e, portanto, suas estrelas evoluíram como estrelas isoladas (do inglês, *single stars*). O trabalho também se aplica para a derivação de massa de estrelas isoladas de (pós-) sequência principal acima de 0.6 $M_{\odot}$. Para isto, os autores utilizaram uma função polinomial utilizando os valores de $T_{\rm ef}$, log g e [Fe/H], o que ocasionou em massas com erros de aproximadamente 6,4%. A função de massa também utiliza os coeficientes de calibração a_i e seus valores estão listados na Tabela 3.13.

Tabela 3.13: Coeficientes de calibração para a equação de Torres et al. (2010).

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

Nos interessa utilizar as calibrações de Torres et al. (2010) para obter as massas estelares ($M_{\star}$) dos objetos da missão Kepler que puderam ser caracterizados pelo algoritmo deste trabalho. Segundo Torres et al. (2010), o logaritmo de $M_{\star}$ pode ser obtido através de:

$$\log M_{\star} = a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_5 (\log g)^2 + a_6 (\log g)^3 + a_7 [\mathrm{Fe/H}] \tag{3.12}$$

onde X = log($T_{\rm ef}$) - 4,1. E como afirmam os autores:

$$\sigma_{M_{\star}} = \frac{6,4}{100} M_{\star} \tag{3.13}$$

Capítulo 4

Resultados

Como visto, o nosso algoritmo de Aprendizagem de Máquina baseado no Random Forest passou por uma série de testes, com variações dos requisitos para seleção das melhores amostras de treinamento. Nestes testes, selecionamos o modelo de melhor acurácia, que inclui em sua amostra de treinamento estrelas observadas pelo J-PLUS e pelo levantamento auxiliar LAMOST. Duas amostras de treinamento foram utilizadas, uma com erro de magnitude menor que 0,1 e outra com erro menor que 0,2. Este limite de erro foi aplicado para todas as 12 bandas do J-PLUS, ou seja, o erro na magnitude de nenhuma das bandas deve ser maior que este limite. Ambas as amostras contêm apenas objetos observados em uma abertura de 6", com probabilidade de ser estrela superior a 90%. Todos os objetos foram confirmados como estrelas através de dados adicionais do Gaia, onde foi possível inferir suas distâncias. Os rendimentos do algoritmo para estas amostras foram apresentados na Tabela 3.10.

Cada amostra de treinamento resultou em 3 modelos treinados: um para temperatura efetiva ($T_{\rm ef}$), um para gravidade superficial (log g) e um para metalicidade ([Fe/H]). Usando estes modelos, estimamos estes 3 parâmetros físicos para as estrelas observadas pela missão Kepler que foram também observadas pelo DR2 do J-PLUS. Serão apresentados, então, duas estimativas para cada parâmetro: uma com o modelo baseado em estrelas com erro de magnitude $< 0{,}1$ e outra com o modelo baseado em estrelas com erro de magnitude $< 0{,}2$.

4.1 Temperatura efetiva ($T_{\rm ef}$)

Na Figura 3.21, vimos as estrelas da missão Kepler que também pertencem ao campo de visão do DR2 do J-PLUS. Essas estrelas representam um total de 29.164 objetos com erro de magnitude menor que 0,1. Das 29.164 estrelas, 2.146 possuem seus parâmetros calculados por espectroscopia pelo LAMOST. A Figura 4.1 mostra a correlação existente entre os dados de temperatura destas 2.146 estrelas e as temperaturas previstas pelo modelo de Aprendizagem de Máquina deste trabalho, que considerou um erro de magnitude

menor que 0,1.

O modelo em questão mostrou uma eficiência de 98,59% durante o treinamento (vide Figura 3.18). Para a amostra de 29.164 estrelas, os erros obtidos durante o treinamento podem ser considerados: erro médio absoluto de 70 K e o erro mediano absoluto de 58 K. Destas estrelas, 3.235 não possuem nenhum dado de temperatura registrado no KIC. As mesmas estrelas também não apresentam dados para log g e/ou [Fe/H]. Na amostra de 2.146 estrelas, a aplicação do modelo demonstra uma correlação de aproximadamente 99,3% entre a temperatura prevista e a temperatura LAMOST, com erro médio absoluto de apenas 42 K. O erro mediano absoluto foi de aproximadamente 25 K.

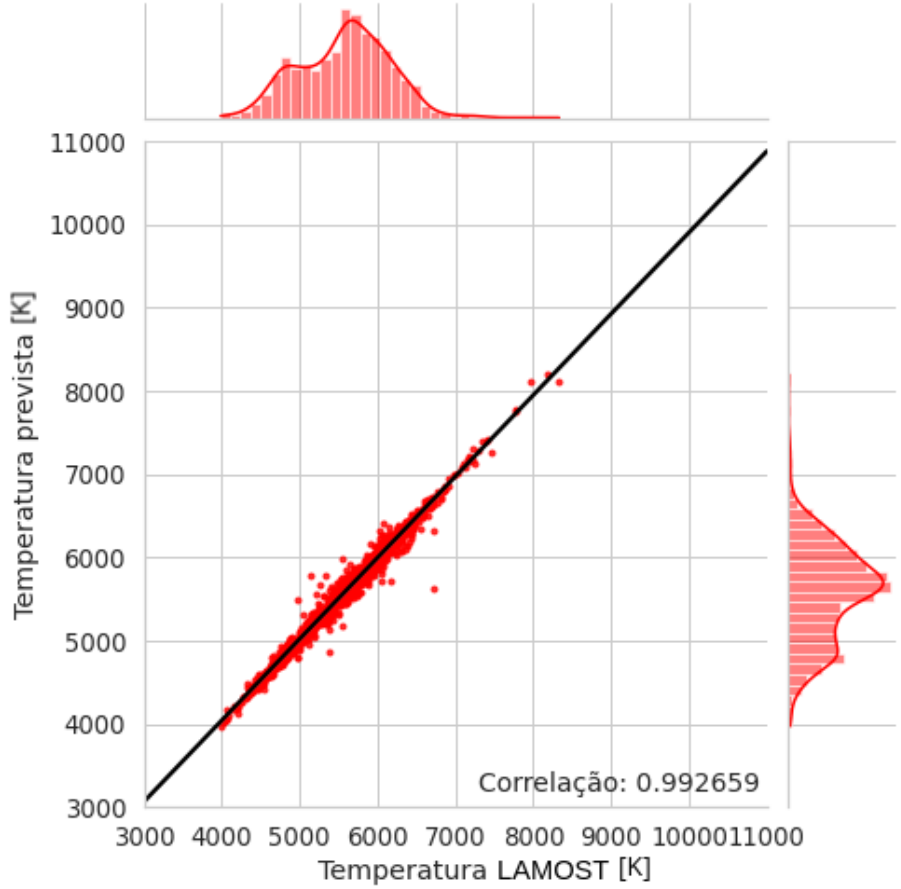


Figura 4.1: Correlação entre as temperaturas previstas pelo algoritmo e as temperaturas apresentadas pelo LAMOST para 2.146 das 29.164 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,1 nos filtros do J-PLUS.

Se considerarmos agora as estrelas da missão Kepler que também pertencem ao campo de visão do DR2 do J-PLUS, mas que possuem um erro de magnitude nos filtros J-PLUS menor que 0,2, obtemos uma amostra de 44.483 estrelas. Como também possuímos um modelo de temperatura efetiva baseado em uma amostra de treinamento com este limite de erro, podemos prever as temperaturas das estrelas Kepler contidas nesta nova amostra.

O segundo modelo treinado para temperatura, que considera estrelas com erro de magnitude menor que 0,2, apresentou rendimento de 99,10% durante o treinamento (vide Figura 3.22). Para a amostra de 44.483 estrelas, podemos considerar os erros obtidos durante o treinamento: erro médio absoluto de 58 K e erro mediano absoluto de 49 K. Destas estrelas, 5.115 não apresentam nenhum dado de temperatura, log g e/ou [Fe/H] no KIC. Das 44.483 estrelas, 2.197 possuem dados de temperatura no LAMOST. A Figura 4.2 mostra a correlação entre a temperatura prevista pelo algoritmo e aquela apresentada pelo LAMOST. Na amostra de 2.197 estrelas, a aplicação do modelo mostra uma correlação maior que 98% para estas temperaturas, com erro médio absoluto de 37 K e erro mediano absoluto de 23 K.

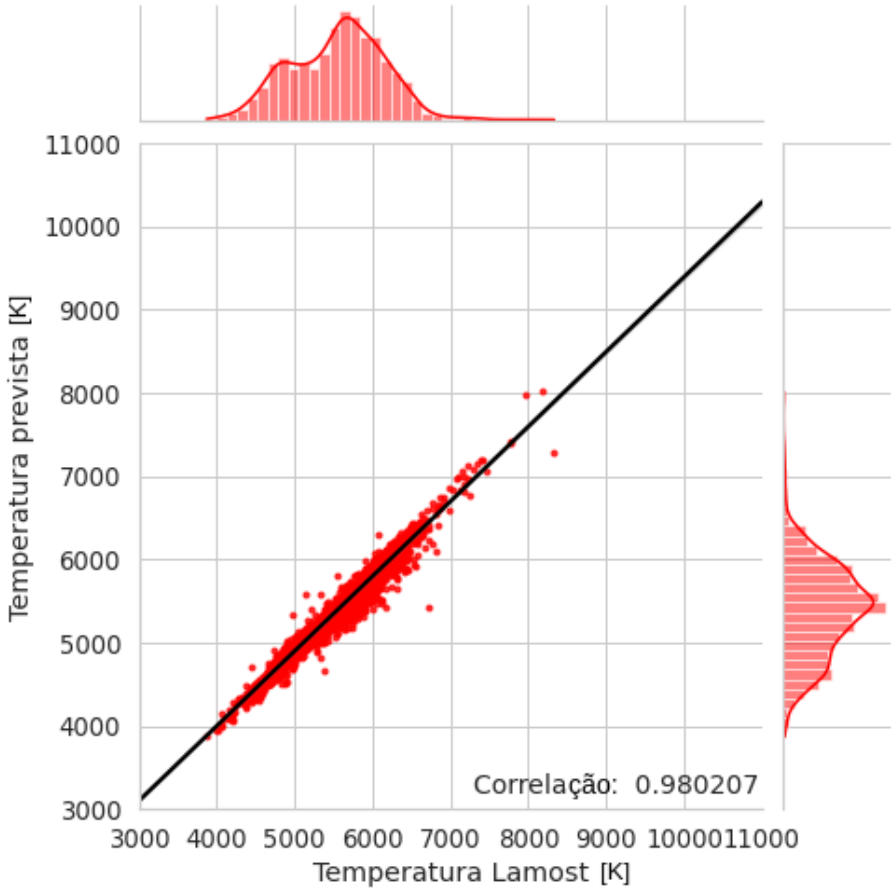


Figura 4.2: Correlação entre as temperaturas previstas pelo algoritmo e as temperaturas apresentadas pelo LAMOST para 2.197 das 44.483 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,2 nos filtros do J-PLUS.

4.2 Gravidade superficial (log g)

Também podemos analisar a correlação entre os valores de gravidade superficial previstos pelo modelo e presentes nos dados do LAMOST. O modelo treinado com uma amostra

de estrelas de erro de magnitude menor que 0,1 apresentou uma acurácia de 97,21% durante o treinamento (vide Figura 3.19). O modelo também foi aplicado nas 29.164 estrelas com este limite de erro, presentes no campo comum entre Kepler e J-PLUS. Para esta amostra, os erros do modelo devem ser considerados: erro médio absoluto de 0,08 e erro mediano absoluto de 0,06. Como visto na Seção 4.1, 2.146 destas estrelas possuem dados calculados pelo LAMOST. A Figura 4.3 mostra a correlação de 97,72% entre o log g previsto pelo modelo e o log g do LAMOST para estas 2.146 estrelas. O erro médio absoluto nesta amostra específica é de 0,07 e o erro mediano absoluto de 0,04.

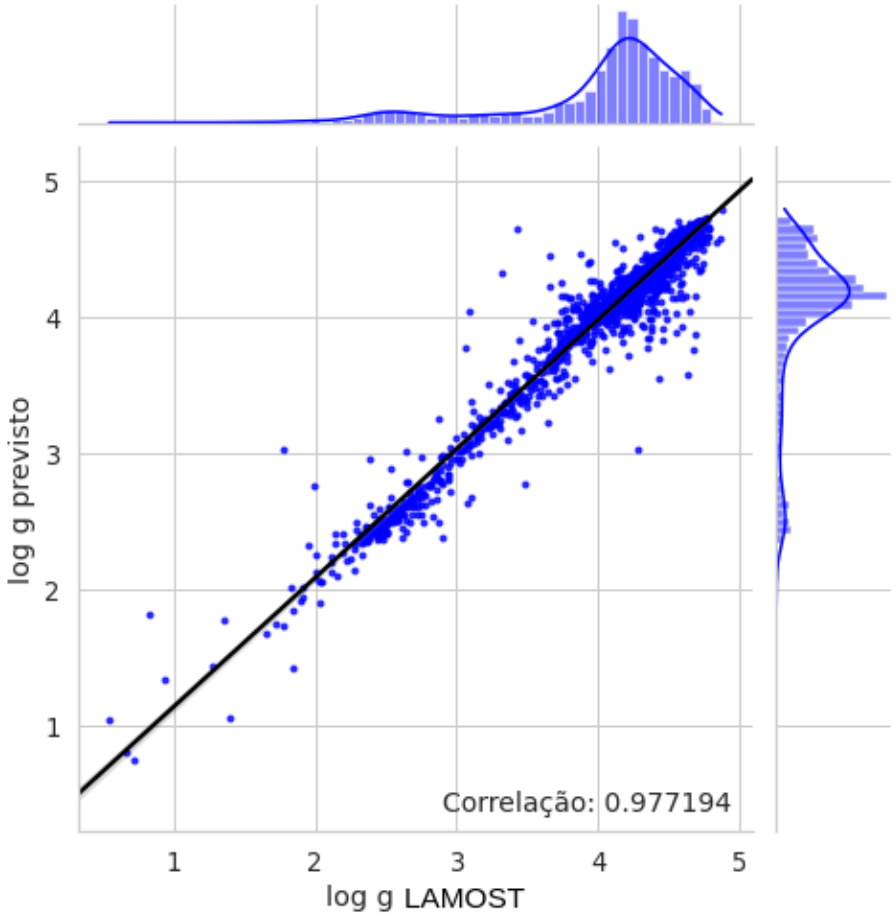


Figura 4.3: Correlação entre os valores de log g previstos pelo algoritmo e aqueles apresentados pelo LAMOST para 2.146 das 29.164 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,1 nos filtros do J-PLUS.

Considerando agora as 44.483 estrelas do campo comum entre Kepler e J-PLUS que possuem erro de magnitude de até 0,2, podemos aplicar o modelo de log g baseado em estrelas com este limite de erro. O modelo obteve 97,31% de acurácia durante o treinamento (vide Figura 3.22). Para a amostra de 44.483, consideramos os erros do modelo: erro médio absoluto de 0,08 e erro mediano absoluto de 0,06. A Figura 4.4 mostra a correlação de 96,83% entre o log g previsto por este modelo e o log g de 2.197 das 44.483 estrelas,

que possuem dados de log g no LAMOST. O erro médio absoluto para as previsões destas 2.197 estrelas é de 0,08 e o erro mediano absoluto de 0,05.

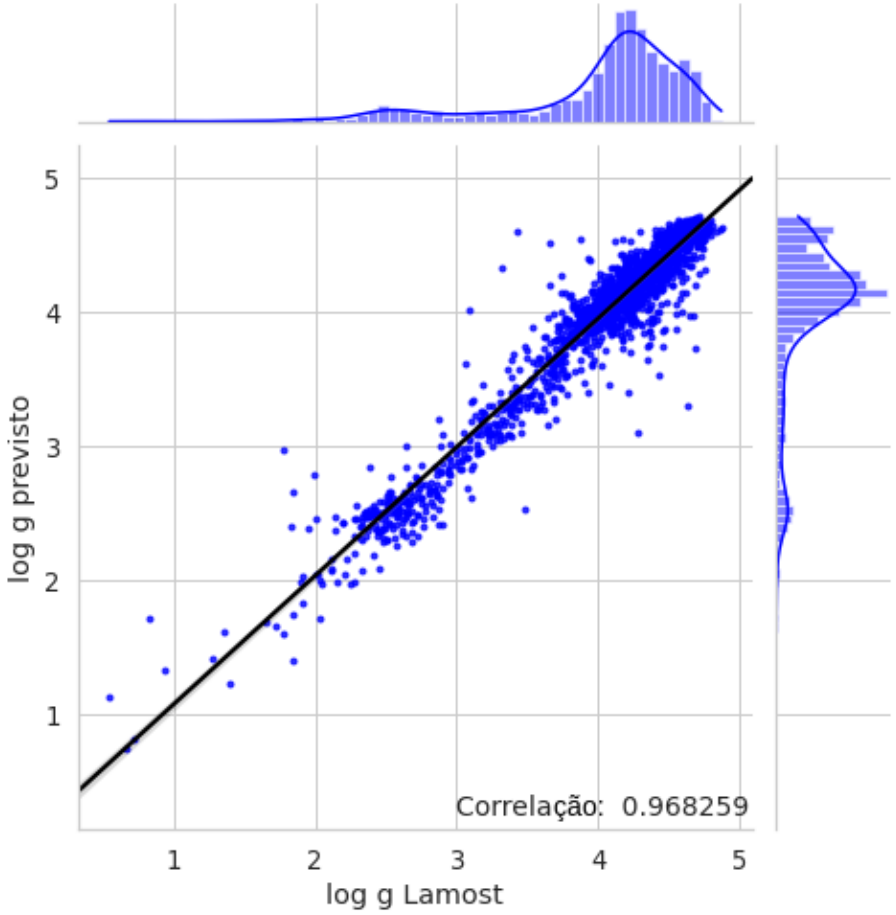


Figura 4.4: Correlação entre os valores de log g previstos pelo algoritmo e aqueles apresentados pelo LAMOST para 2.197 das 44.483 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,2 nos filtros do J-PLUS.

4.3 Metalicidade [Fe/H]

A Aprendizagem de Máquina também foi usada para desenvolver dois modelos treinados para previsão de [Fe/H]: um treinado com estrelas com erro de magnitude de até 0,1 (que será aplicado nas 29.164 estrelas do campo comum entre Kepler e J-PLUS que possuem erro menor que 0,1 nos filtros J-PLUS) e outro treinado com estrelas com erro de magnitude de até 0,2 (que pode ser aplicado nas 44.483 estrelas do campo Kepler e J-PLUS que estão dentro deste limite de erro).

O modelo que atende até 0,1 de erro mostrou eficiência de 92,52% durante o treinamento (vide Figura 3.19). Para a amostra de 29.164 estrelas, consideramos os erros do modelo: erro médio absoluto de 0,10 e erro mediano absoluto de 0,08. A Figura 4.5 apre-

senta a correlação entre a [Fe/H] prevista pelo modelo e a [Fe/H] da literatura, calculada pelo LAMOST, para 2.146 das 29.164 estrelas. A correlação obtida entre os valores foi de 95,63%. O erro médio absoluto na amostra de 2.146 estrelas foi de 0,06 e o erro mediano absoluto de 0,04.

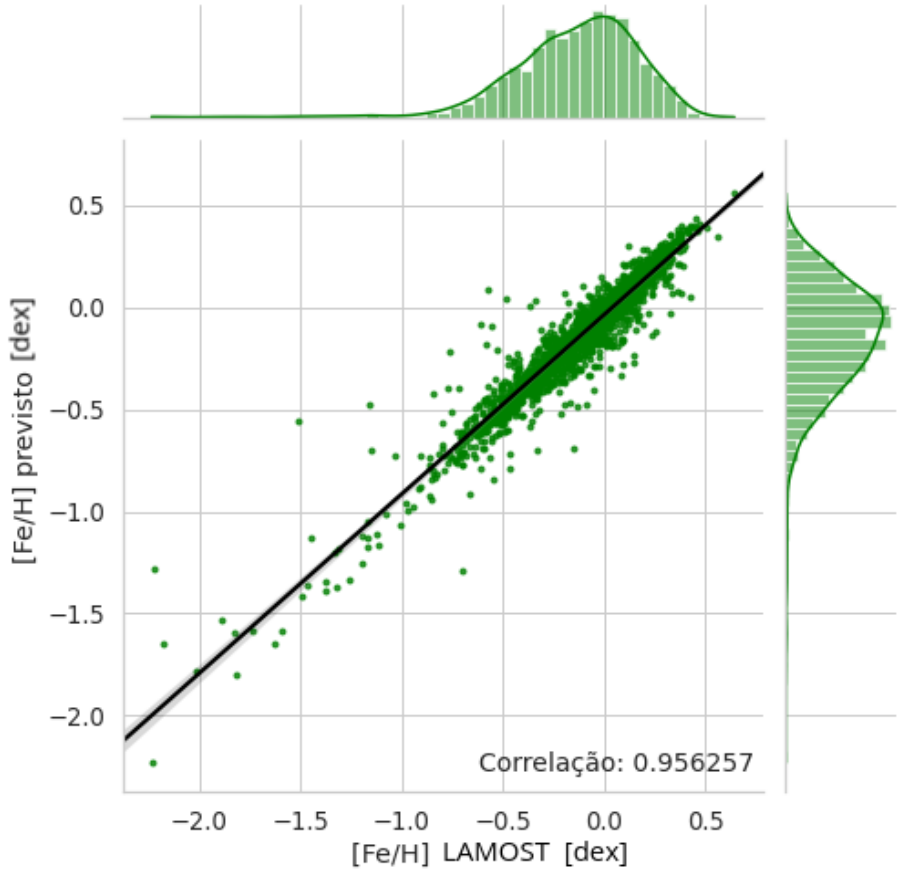


Figura 4.5: Correlação entre os valores de [Fe/H] previstos pelo algoritmo e aqueles apresentados pelo LAMOST para 2.146 das 29.164 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,1 nos filtros do J-PLUS.

Vamos considerar agora as 44.483 estrelas do campo comum entre Kepler e J-PLUS com erro de magnitude menor que 0,2 e seu modelo de previsão equivalente. O modelo apresentou rendimento de 91,10% durante o treinamento (vide Figura 3.22). Para a amostra de 44.483 estrelas, vamos considerar os erros do modelo: erro médio absoluto de 0,10 e erro mediano absoluto de 0,09. A Figura 4.6 apresenta a correlação de 88,83% entre a [Fe/H] prevista pelo algoritmo e a [Fe/H] de 2.197 destas estrelas, disponíveis na literatura. O erro médio absoluto das previsões destas 2.197 estrelas foi de 0,06 e o erro mediano absoluto de 0,04.

Vimos que, para cada parâmetro, há dois modelos de previsão: um treinado para atender estrelas com erro de magnitude menor que 0,1 (modelo mais restrito) e outro para erro menor que 0,2 (modelo menos restrito). A variação no rendimento do par de modelos, isto é, entre os modelos que prevêem um mesmo parâmetro, é sempre menor

que 1,5%. As flutuações para $T_{\rm ef}$, log g e [Fe/H] são, respectivamente, 0,51%, 0,10% e 1,42%. Como é esperado, o modelo menos restrito apresenta maior dispersão na previsão de alguns objetos. O erro máximo deste modelo é maior do que o erro máximo do modelo mais restrito.

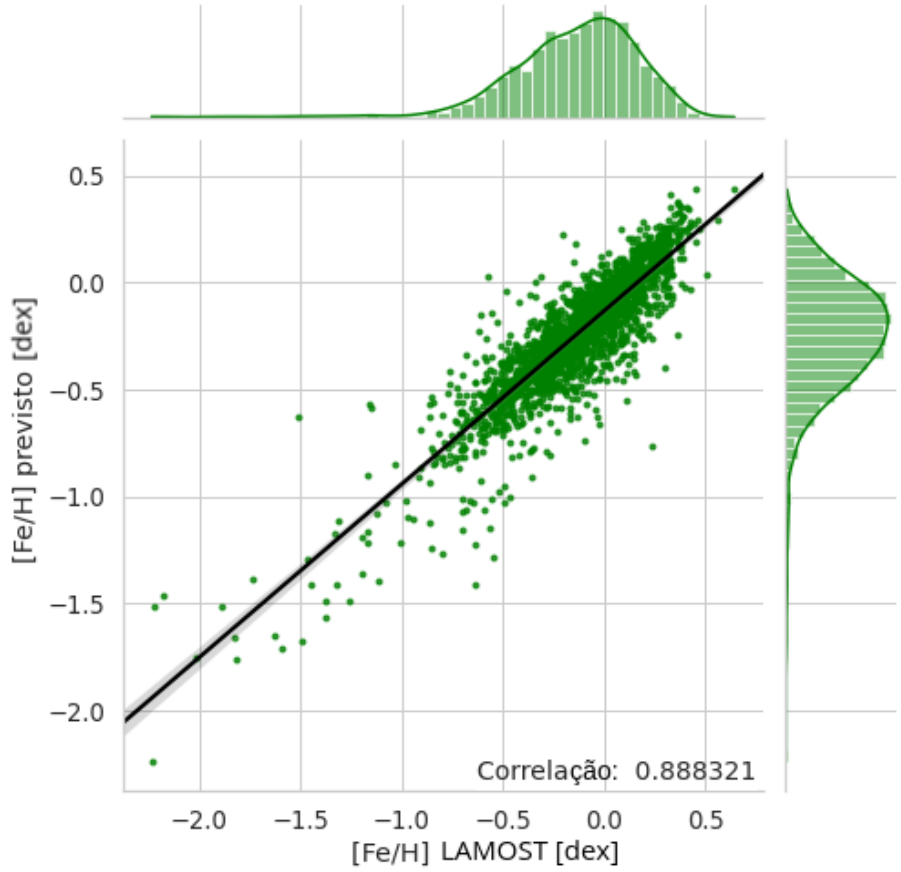


Figura 4.6: Correlação entre os valores de [Fe/H] previstos pelo algoritmo e aqueles apresentados pelo LAMOST para 2.197 das 44.483 estrelas presentes no campo comum Kepler e J-PLUS, que apresentaram e_mag $<$ 0,2 nos filtros do J-PLUS.

Como vimos anteriormente, a maior vantagem dos modelos com erro de até 0,2 é o acréscimo no número de estrelas Kepler caracterizadas - de 29.164 para 44.483 objetos (ganho $>$ 52%). Entretanto, sempre que possível, recomendamos realizar previsões com o modelo mais preciso, por exemplo, caso os dois modelos calculem a $T_{\rm ef}$ de uma estrela específica, deve ser priorizada aquela baseada no modelo mais preciso.

Com os valores dos parâmetros previstos, podemos gerar um diagrama de Kiel. Este diagrama é uma versão simplificada do diagrama Hertzsprung-Russell (diagrama H–R) para evolução estelar, exibindo as gravidades superficiais estelares contra as temperaturas efetivas correspondentes. Também é comum adicionar os dados de metalicidade ao diagrama de Kiel, a fim de observar sua distribuição na amostra. A Figura 4.7 apresenta este diagrama para as 29.164 estrelas Kepler observadas pelo J-PLUS que puderam ser

caracterizadas pelo modelo mais preciso.

No diagrama, podemos observar uma concentração horizontal de estrelas com log g próximo e acima de 4,0. Estas são estrelas de sequência principal. A concentração vertical de estrelas mostra objetos com log g abaixo de 4,0, principalmente no intervalo de $T_{\rm ef}$ da classe K (3700–5200K). Algumas estrelas de tipo K como a gigante Aldebaran podem exemplificar o log g típico deste tipo de estrela: 1,59. Para uma anã de sequência principal de tipo K como α Centauri B, o log g é de 4,54. Subgigantes de tipo K como η Cephei, possuem log g de 3,47. O ramo vertical mostra então as estrelas evoluídas de nossa amostra (4688 objetos possuem log g $<$ 4,0, onde 1793 tem log g $<$ 3,5 e 47 tem log g $<$ 2,0). Uma parte considerável (49%) das nossas estrelas possuem [Fe/H] entre -0,8 e -0,3. Cerca de 1% apresentam [Fe/H] $<$ -1,5. Apenas 10 objetos possuem [Fe/H] $<$ -2,0, sendo os objetos mais pobres em metais da nossa amostra.

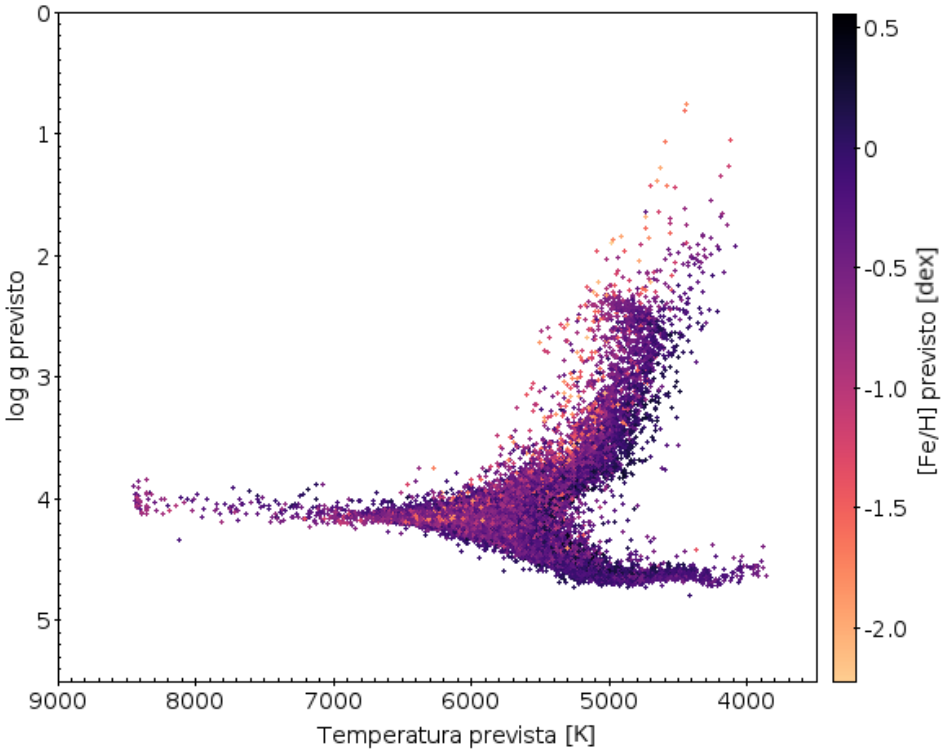


Figura 4.7: Diagrama de Kiel com as previsões do algoritmo para as 29.164 estrelas presentes no campo comum entre Kepler e J-PLUS, que apresentaram erro de magnitude menor que 0,1 nos filtros J-PLUS.

4.4 Luminosidade e Raio

Luminosidades e raios foram estimados seguindo a descrição da Seção 3.4. Podemos comparar os resultados obtidos a partir do modelo mais restrito (mais preciso) e a partir do menos restrito (menos preciso), para as 29.164 estrelas em comum, ou seja, as estrelas caracterizadas por ambos os modelos.

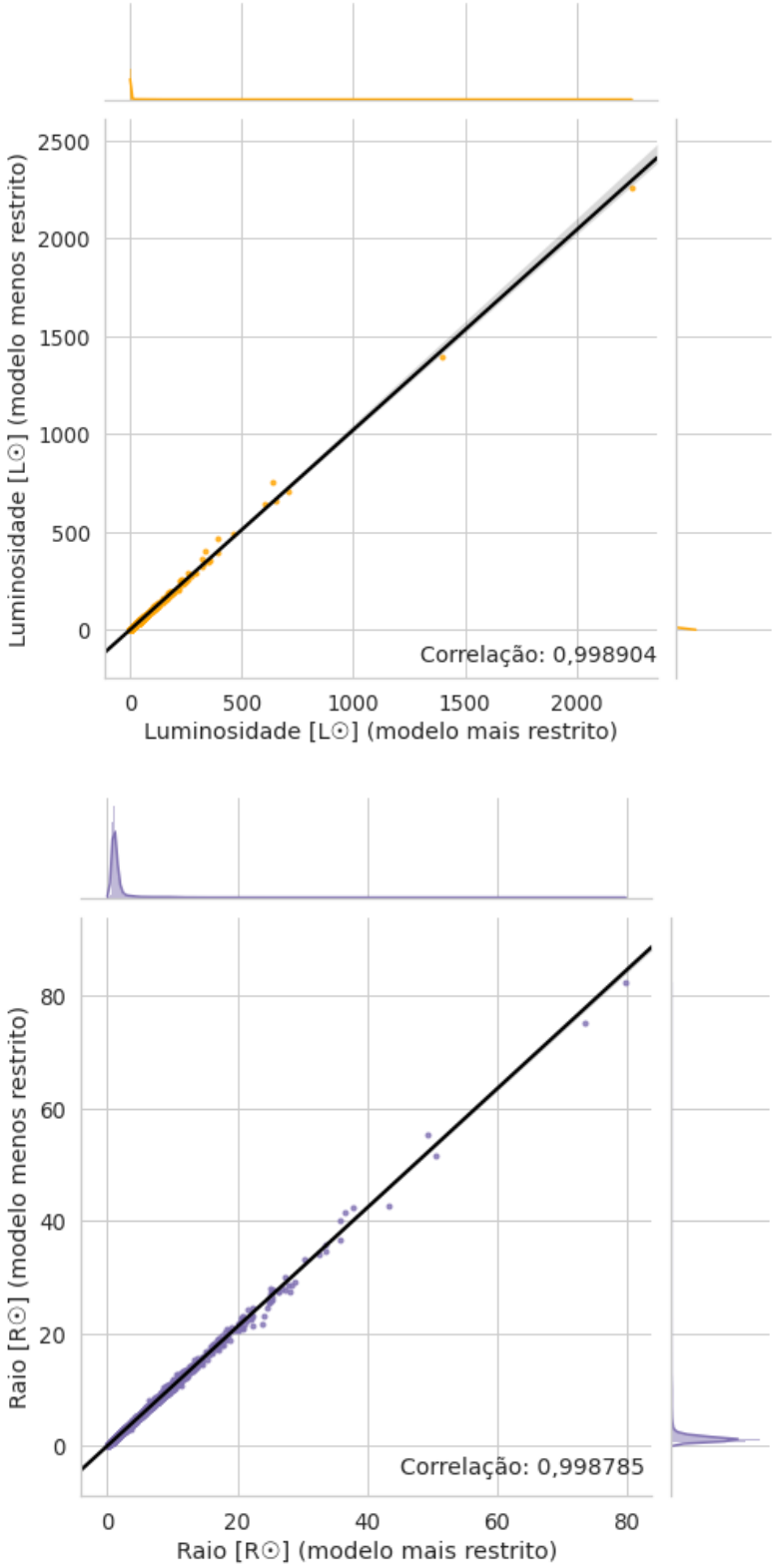


Figura 4.8: Comparativos para luminosidade (painel superior) e raio (painel inferior), calculados por cada modelo, para as 29.164 estrelas em comum, ou seja, as estrelas caracterizadas por ambos os modelos.

Como podemos notar na Figura 4.8, existe uma alta correlação entre as previsões dos dois modelos ($\approx 99,9\%$ em ambos os parâmetros). Isto nos dá segurança para usar as

previsões do modelo menos preciso, quando necessário, isto é, quando não hajam previsões do modelo mais preciso.

4.5 Comparação com os dados do KIC

Na Figura 4.9, comparamos os resultados obtidos por este trabalho para $T_{\rm ef}$, log g, [Fe/H] e $R_\star$ com os dados do KIC. Podemos comparar as incertezas que observamos com as incertezas apresentadas por Brown et al. (2011). No painel superior esquerdo da Figura 4.9 ($T_{\rm ef}$, pontos vermelhos), temos uma correlação de 95,49% entre os resultados do nosso modelo e os do KIC. Neste painel temos as maiores dispersões acima de 9000 K, região limite das previsões do modelo. Cinco objetos se destacam neste painel, formando o agrupamento vertical destacado no retângulo vermelho: KIC 5682485, KIC 6424652, KIC 10118750, KIC 10904353 e KIC 11953267. Este comportamento sugere que: 1) o KIC apresenta incertezas particularmente alta para estes objetos ou 2) estas estrelas são mais quentes do que o modelo é capaz de prever (o que é mais provável). Para Brown et al. (2011), as incertezas foram de $\pm$ 200 K para $T_{\rm ef} < 7000$ K e até 4000 K para $T_{\rm ef}$ de 9000 a 13500 K. Na análise dos nossos resultados, para $T_{\rm ef} < 7000$ K, observamos uma variação média de $\pm$ 124 K. Para $T_{\rm ef} > 7000$ K, nossos valores diferiam até 3414 K.

Ainda na Figura 4.9, no painel superior direito (log g, pontos azuis), apresentamos uma correlação de 76,62% entre os resultados do nosso modelo e os do KIC. Destacamos o objeto KIC 11953267 (retângulo azul). No KIC, ele apresenta log g = 6,16, o que excede o limite de previsão do modelo de log g deste trabalho. Para log g, Brown et al. (2011) observaram incertezas de 0,5 dex para anãs e 1,5 dex para gigantes. Nossa análise revelou incertezas médias de 0,22 dex para anãs e até 2,1 dex para gigantes (vide painel superior direito da Figura 4.9). Brown et al. (2011) observaram incertezas de $\pm$ 0,4 dex entre a [Fe/H] do KIC e a obtida por Fisher & Valenti (2005), em uma pequena amostra de estrelas. Nossos resultados, em maioria (cerca de 70%), tiveram uma variação menor que 0,5 dex para o [Fe/H] do KIC, mas mais de 4000 objetos apresentaram uma diferença tão grande quanto 2,0 dex (vide painel inferior esquerdo da Figura 4.9). Para [Fe/H], a correlação entre os nossos resultados e os do KIC foi de aproximadamente 38%.

Brown et al. (2011) não avaliaram a precisão de raio do KIC, mas nós obtivemos uma incerteza média de 0,34 $R_\odot$ para um raio KIC entre 0,23 e 1,5 $R_\odot$, o que representa 79% dos objetos. Para os 4% que possuem raios entre 1,5 e 2,5 $R_\odot$ no KIC, a incerteza sobe para 0,84 $R_\odot$. Cerca de 6% dos objetos possuem raios entre 2,5 e 45 $R_\odot$ no KIC e as incertezas podem ser tão grandes quanto 58 $R_\odot$ (vide painel inferior direito da Figura 4.9). Aproximadamente 11% não possuíam raio calculado no KIC. A correlação para log g, [Fe/H] e $R_\star$, entre os nossos resultados e os do KIC, é baixa e/ou os resultados apresentam alta dispersão quando comparados. Isto já era esperado, já que vimos que as medidas do KIC possuem incertezas consideravelmente altas. Para $R_\star$, a correlação entre

os nossos resultados e os do KIC foi de aproximadamente 82%.

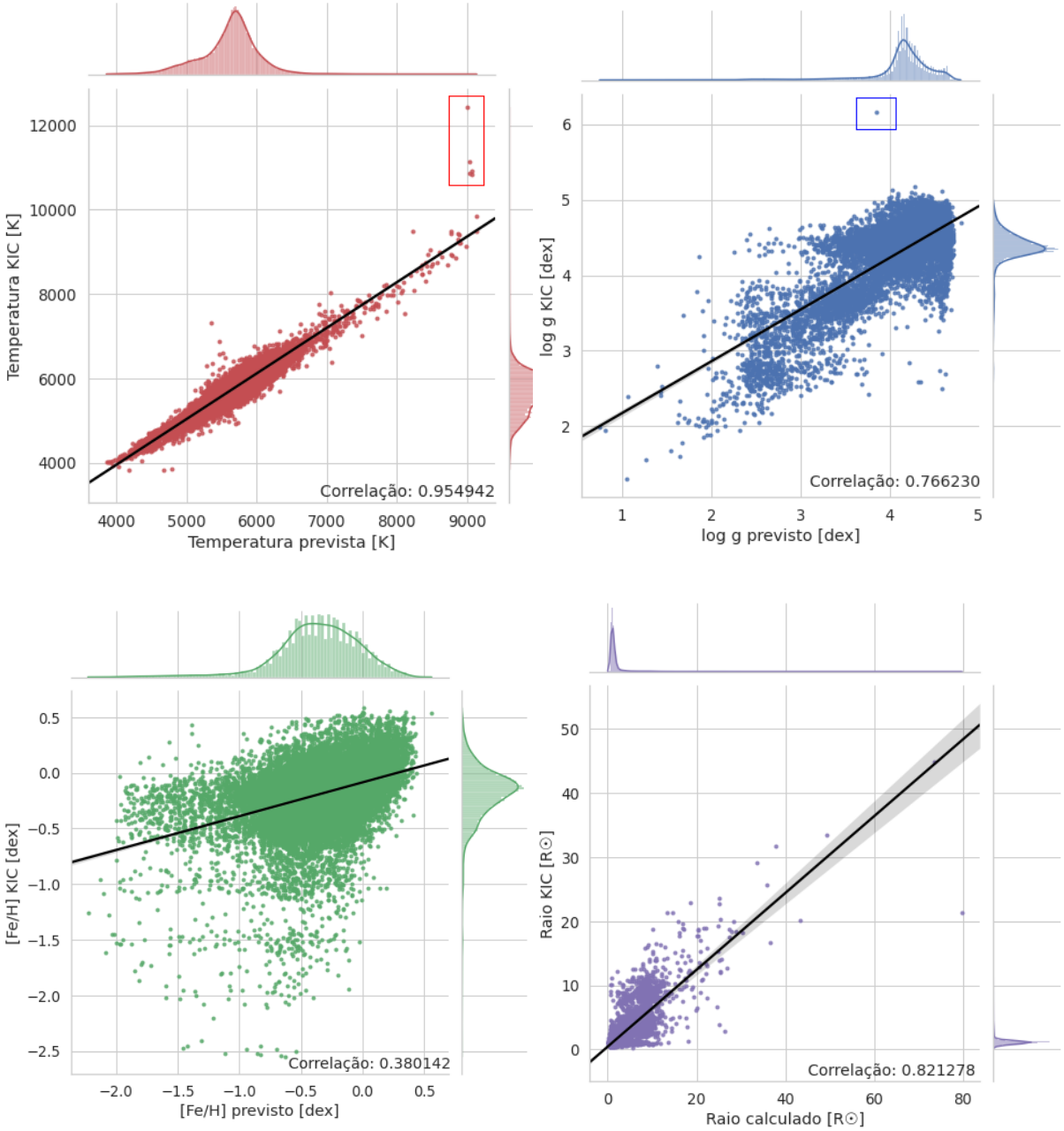


Figura 4.9: Comparativo entre os valores dos parâmetros previstos pelo algoritmo e os fornecidos no KIC, para $T_{\rm ef}$ (pontos vermelhos), log g (pontos azuis), [Fe/H] (pontos verdes) e $R_\star$ (pontos roxos). Na figura, podemos perceber como os dados mais precisos propostos por este trabalho diferem dos dados originais do KIC. Este resultado já era esperado, já que vimos que as medidas do KIC possuem incertezas consideravelmente altas.

4.6 Discussão de objetos interessantes

Nogueira (2020) analisou as estrelas fracas da missão Kepler que possuíam magnitude

$16 < K_p < 16{,}5$ a fim de identificar possíveis trânsitos. Inicialmente, ele obteve uma amostra de 453 estrelas com 468 possíveis trânsitos detectados - os números de estrelas e trânsitos diferem porque algumas estrelas apresentaram indícios de mais de um trânsito, na mesma curva de luz. Após um processo de validação que eliminou falsos trânsitos, causados por efeitos sistemáticos, restaram 230 estrelas com 236 possíveis trânsitos. Entre as 44.483 estrelas que puderam ser caracterizadas pelos modelos deste trabalho, encontramos 7 objetos de Nogueira (2020), que serão discutidos adiante. Seus parâmetros estão expressos na Tabela 4.1. Destes, 5 puderam ser caracterizados pelo modelo mais restrito (ou seja, aquele de maior precisão).

Tabela 4.1: Parâmetros físicos calculados pelos modelos deste trabalho para os objetos em comum com o trabalho de Nogueira (2020).

Estrela [KIC ID]	$T_{\rm ef}$ [K]	log g [dex]	[Fe/H] [dex]	L [$L_\odot$]	$R_\star$ [$R_\odot$]	$R_{\star KIC}$ [$R_\odot$]
Modelo mais restrito						
7661441	5580	4,34	-0,30	0,93±0,02	1,03±0,03	1,04
7799491	5414	4,28	-0,12	0,80±0,02	1,02±0,03	1,08
7801955	5365	4,48	0,18	0,49±0,01	0,82±0,02	0,71
8933590	5473	4,40	0,18	0,66±0,01	0,91±0,02	1,02
10515164	5701	4,16	-0,46	1,14±0,02	1,10±0,03	1,18
Modelo menos restrito						
7658948	5194	4,47	0,11	0,58±0,01	0,94±0,03	0,86
10708854	4810	3,67	-0,15	4,22±0,09	2,97±0,09	1,08

As colunas da Tabela 4.1 se dividem em:

- Estrela: Número identificador do objeto no catálogo KIC (KIC ID).
- $T_{\rm ef}$: Temperatura efetiva estelar, calculada pelo algoritmo deste trabalho (expressa em Kelvin, K).
- log g: Logaritmo de base 10 da gravidade superficial da estrela, calculado pelo algoritmo deste trabalho (expresso em expoente decimal, dex).
- [Fe/H]: Metalicidade estelar, calculada pelo algoritmo deste trabalho (expressa em dex).

- L: Luminosidade estelar, calculada com base nos parâmetros previstos pelo algoritmo (em unidades solares, $L_{\odot}$; vide Seção 3.4).

- $R_{\star}$: Raio estelar, calculado com base nos parâmetros previstos pelo algoritmo (em unidades solares, $R_{\odot}$; vide Seção 3.4).

- $R_{\star KIC}$: Raio estelar dado pelo catálogo KIC (em unidades solares, $R_{\odot}$).

Como visto, a Equação 1.2 da Seção 1.3 nos permite calcular os raios de objetos em órbita (R_p), desde que se conheça a profundidade máxima observada (Q) do trânsito e o raio da estrela hospedeira ($R_{\star}$). A Tabela 4.1 nos fornece $R_{\star}$ e Nogueira (2020) mediu os valores de Q. Organizamos estes dados na Tabela 4.2. Nela apresentamos também o R_p obtido por este trabalho através da Equação 1.2, e R_p obtido por Nogueira (2020) (R_p^{Nog}), que fez uso dos dados do KIC.

Tabela 4.2: Raio dos objetos em trânsito ao redor das estrelas da Tabela 4.1.

Estrela [KIC ID]	$R_{\star}$ [$R_{\odot}$]	Q [%]	R_p [R_{Jup}]	R_p^{Nog} [R_{Jup}]
Modelo mais restrito				
7661441	1,03±0,03	8,91±22,33	3,00±0,04	3,03±0,07
7799491	1,02±0,03	12,58±22,24	3,51±0,04	3,71±0,02
7801955	0,82±0,02	7,66±33,78	2,20±0,04	1,91±0,03
8933590	0,91±0,02	8,04±29,91	2,50±0,03	2,80±0,04
10515164	1,10±0,03	14,2±27,47	4,03±0,06	4,32±0,06
Modelo menos restrito				
7658948	0,94±0,03	13,14±32,61	3,31±0,05	3,03±0,02
10708854	2,97±0,09	3,91±15,30	5,71±0,08	2,08±0,07

As duas primeiras colunas da Tabela 4.2 são comuns às da Tabela 4.1. O restante se divide em:

- Q: Profundidade máxima observada do trânsito (em percentual). Neste caso, representa a queda percentual do fluxo recebido da estrela hospedeira, causada pela passagem do corpo pelo seu disco, registrada na curva de luz analisada. A incerteza de Q, σ_Q, equivale a um percentual de Q. Por exemplo, para KIC 7661441, σ_Q equivale a 22,33% do valor de Q desta estrela.

- R_p: Raio mínimo do objeto em trânsito (em raios de Júpiter, $R_{\rm Jup}$), calculado com base no $R_\star$ medido por este trabalho que, por sua vez, foi baseado nos parâmetros previstos pelo algoritmo.

- R_p^{Nog}: Raio mínimo do objeto em trânsito (em raios de Júpiter, $R_{\rm Jup}$), calculado por Nogueira (2020), com base nos parâmetros do KIC.

Como explicado na Seção 3.4, para calcularmos $R_\star$ utilizamos a Equação 3.8. Esta equação utiliza a $T_{\rm ef}$ prevista pelo algoritmo e a luminosidade, L, calculada a partir da magnitude bolométrica, M_{bol}, do objeto. Podemos notar o impacto do algoritmo ao comparar o $R_\star$ calculado com o auxílio dele e aquele fornecido pelo KIC. A seguir, vamos analisar tal impacto para os 7 objetos da Tabela 4.2. Essa variação atinge proporcionalmente os valores de R_p, já que este depende do raio da estrela. É importante destacar que R_p depende de Q, que é um valor observacional. Como mencionado anteriormente, a profundidade real do trânsito só pode ser inferida com segurança caso se observe todos os momentos da órbita. Sendo assim, chamamos Q de profundidade máxima observada pois a profundidade real do trânsito pode ser ainda maior do que foi possível observar com os dados da curva de luz não contínua da hospedeira. O raio R_p então é o raio mínimo do objeto em trânsito.

Podemos classificar os objetos em órbita, com base em R_p, em: binárias eclipsantes, anãs marrons e exoplanetas. Binárias eclipsantes são sistemas estelares nos quais duas estrelas, gravitacionalmente ligadas, orbitam um centro de massa comum, enquanto seu plano orbital é aproximadamente perpendicular a nossa linha de visada. Isto permite a detecção de eclipses na curva de luz, causados pela passagem periódica de uma estrela em frente à outra (Aitken, 1935). As anãs marrons são objetos que tem massa menor do que a massa mínima necessária para ocorrer a fusão de hidrogênio no núcleo ($M < \approx 0,075\ M_\odot$), porém maior do que a massa limite para a classificação como exoplaneta ($M > \approx 0,013\ M_\odot$; Smith, 2004). Estes objetos realizam fusão termonuclear de apenas de um isótopo do hidrogênio (deutério, ^{2}H, resultando em ^{3}He) ou, para anãs marrons mais massivas (M > 0,06 $M_\odot$), também de um isótopo do lítio (^{7}Li, resultando em ^{4}He). Isto faz com que anãs marrons possuam baixas luminosidades e temperaturas. Elas podem ser detectadas através do método de trânsito ou observações diretas no infravermelho (Chabrier & Baraffe, 2000).

Objetos ainda menos massivos ($M <$ 0,013 $M_\odot$) são classificados como exoplanetas. Para diferenciarmos anãs marrons e exoplanetas de binárias eclipsantes, utilizamos o critério de raio máximo para um exoplaneta de $24R_\oplus$ ($2{,}14R_{\rm Jup}$), descrito por Petigura et al. (2018). Não pudemos distinguir exoplanetas de anãs marrons pois, para isso, seria necessário obter também a massa destes objetos. Um trânsito que produza $R_p > 2{,}14R_{\rm Jup}$ é, provavelmente, causado por uma binária eclipsante.

Os cinco primeiros objetos citados na Tabela 4.2 foram caracterizados pelo modelo

mais restrito deste trabalho e classificados por ele como estrelas de tipo G. O primeiro objeto é a estrela KIC 7661441. Nogueira (2020) identificou 1 trânsito ao redor desta estrela que, segundo ele, deveria pertencer a uma componente binária eclipsante de raio $R_p^{Nog} = 3{,}03 \pm 0{,}07\ R_{\rm Jup}$. Entre as 7, esta estrela apresenta a menor variação (apenas 0,94%) entre o raio estelar dado pelo KIC ($R_{\star KIC} = 1{,}04\ R_\odot$) e o proposto por este trabalho ($R_\star = 1{,}03 \pm 0{,}03\ R_\odot$). Este trabalho propõe então que o trânsito pertence a um objeto de raio $R_p = 3{,}00 \pm 0{,}04\ R_{\rm Jup}$, o que está dentro da margem de erro de Nogueira (2020) e mantém a classificação de binária eclipsante.

O segundo objeto é a estrela KIC 7799491. Nogueira (2020) também identificou apenas 1 trânsito ao redor desta estrela, que deveria pertencer a uma componente binária eclipsante de raio $R_p^{Nog} = 3{,}71 \pm 0{,}02\ R_{\rm Jup}$. Para esta estrela, $R_{\star KIC} = 1{,}08\ R_\odot$ e $R_\star = 1{,}02 \pm 0{,}03\ R_\odot$. Neste caso, o raio proposto por este trabalho é 5,41% menor que o do KIC. Assim, com $R_\star$, o trânsito deve pertencer a um objeto de raio $R_p = 3{,}51 \pm 0{,}04\ R_{\rm Jup}$, ainda mantendo a classificação de binária eclipsante.

O terceiro objeto é a estrela KIC 7801955. Tanto para o KIC quanto para este trabalho, ela é a estrela de menor raio entre as 7 ($R_{\star KIC} = 0{,}71\ R_\odot$ e $R_\star = 0{,}82 \pm 0{,}02\ R_\odot$). O raio proposto por este trabalho é 15,21% maior que o do KIC. Ela também apresentou a menor luminosidade entre os 7 objetos ($L = 0{,}49 \pm 0{,}01\ L_\odot$). Nogueira (2020) também identificou apenas 1 trânsito ao redor desta estrela mas que, dessa vez, poderia pertencer a um exoplaneta ou anã marrom de raio $R_p^{Nog} = 1{,}91 \pm 0{,}03\ R_{\rm Jup}$. No entanto, considerando $R_\star$, o raio do objeto em trânsito é maior que o de Nogueira (2020), $R_p = 2{,}20 \pm 0{,}04\ R_{\rm Jup}$. Isto faz com que o objeto seja reclassificado como uma estrela secundária, compondo um sistema binário eclipsante, mas com um raio muito próximo do limite que divide exoplanetas e anãs marrons de estrelas binárias.

O quarto objeto é a estrela KIC 8933590. Nogueira (2020) identificou 1 trânsito que deveria pertencer a uma componente da binária eclipsante de raio $R_p^{Nog} = 2{,}80 \pm 0{,}04\ R_{\rm Jup}$. Para esta estrela, $R_{\star KIC} = 1{,}02\ R_\odot$ e $R_\star = 0{,}91 \pm 0{,}02\ R_\odot$. Neste caso, o raio proposto por este trabalho é 10,73% menor que o apresentado pelo KIC. Assim, com $R_\star$, o trânsito deve pertencer a um objeto de raio $R_p = 2{,}50 \pm 0{,}03\ R_{\rm Jup}$, mantendo a classificação de binária eclipsante.

O quinto objeto é a estrela KIC 10515164, que apresentou a maior $T_{\rm ef}$ entre os 7 objetos analisados. Nogueira (2020) identificou 3 possíveis trânsitos em sua curva de luz, que poderiam ser gerados por uma componente binária eclipsante de curto período, de raio $R_p^{Nog} = 4{,}32 \pm 0{,}06\ R_{\rm Jup}$. Para esta estrela, $R_{\star KIC} = 1{,}18\ R_\odot$ e $R_\star = 1{,}10 \pm 0{,}03\ R_\odot$. Neste caso, o raio proposto por este trabalho é 6,71% menor que o do KIC. Assim, com $R_\star$, o trânsito deve pertencer a um objeto de raio $R_p = 4{,}03 \pm 0{,}06\ R_{\rm Jup}$, mantendo a classificação de binária eclipsante.

O sexto e sétimo objeto foram caracterizados apenas pelo modelo menos restrito e foram ambos classificados como estrelas de tipo K. O sexto objeto é a estrela KIC 7658948.

Nogueira (2020) identificou 1 trânsito que deveria pertencer a uma componente da binária eclipsante de raio $R_p^{Nog} = 3{,}03 \pm 0{,}02\ R_{\rm Jup}$. Para esta estrela, $R_{\star KIC} = 0{,}86\ R_\odot$ e $R_\star = 0{,}94 \pm 0{,}03\ R_\odot$. Neste caso, o raio proposto por este trabalho é 9,36% maior que o do KIC. Assim, com $R_\star$, o trânsito deve pertencer a uma estrela secundária de raio $R_p = 3{,}31 \pm 0{,}05\ R_{\rm Jup}$, mantendo a classificação do sistema como binária eclipsante.

O sétimo objeto é a estrela KIC 10708854 e é o mais interessante. Para esta estrela, $R_{\star KIC} = 1{,}08\ R_\odot$. Ela possui a menor temperatura entre os 7 objetos ($T_{\rm ef} = 4810$ K), além da maior luminosidade ($L_\star = 4{,}22 \pm 0{,}09\ L_\odot$) e raio ($R_\star = 2{,}97 \pm 0{,}09\ R_\odot$) da subamostra. Este raio é 174,16% maior que o proposto pelo KIC. Por causa desta grande variação, decidimos coletar dados adicionais sobre ela e os outros 6 objetos desta seção: analisamos a magnitude absoluta na banda G (M_G), calculada com base nos dados do Gaia, para os 7 objetos. Para os seis primeiros, M_G variou entre +4,5 e +5,5 e log g entre 4,0 e 4,5, caracterizando estrelas de tipo G e K de sequência principal, mas KIC 10708854 apresentou $M_G = +3{,}2$, além de o algoritmo prever um log g = 3,67 para ela. Estes dados são indícios de uma estrela subgigante.

Nogueira (2020) analisou a curva de luz da estrela KIC 10708854 e identificou 1 trânsito que poderia pertencer a um exoplaneta ou anã marrom de raio $R_p^{Nog} = 2{,}08 \pm 0{,}07\ R_{\rm Jp}$. No entanto, como $R_\star$ é bem maior que $R_{\star KIC}$, o raio do objeto em trânsito proposto aqui também é bem maior que o de Nogueira (2020), $R_p = 5{,}71 \pm 0{,}08\ R_{\rm Jp}$. Isto permite que o objeto seja reclassificado como uma estrela secundária, caracterizando um sistema binário eclipsante. Dessa forma, os trânsitos em todos os 7 objetos apresentam indícios de serem provocados por estrelas companheiras.

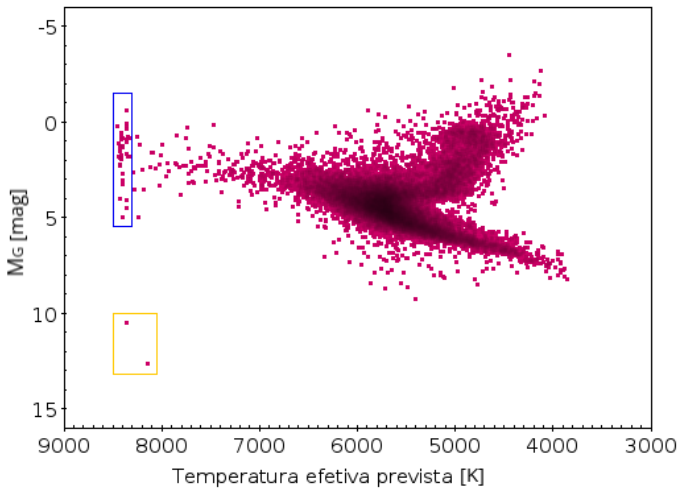


Figura 4.10: Diagrama HR com as previsões do algoritmo para as 29.164 estrelas presentes no campo comum entre Kepler e J-PLUS que apresentaram erro de magnitude menor que 0,1 nos filtros J-PLUS. Observe que algumas estrelas parecem agrupar-se verticalmente próximo ao limite máximo do modelo de temperatura (retângulo azul). Estas estrelas são, provavelmente, mais quentes mas o modelo as agrupa em torno do seu limite. Destacamos as candidatas a anãs brancas no retângulo amarelo.

Na Figura 4.10 apresentamos o Diagrama H–R com as previsões do modelo mais preciso, onde destacamos um grupo de estrelas que parecem ser mais quentes que o limite máximo do modelo de temperatura (retângulo azul). Concluímos isto pois estas estrelas agrupam-se verticalmente próximo a este limite, já que o modelo não foi treinado para ultrapassá-lo. Fornecemos, no entanto, uma possível solução para as previsões de $T_{\rm ef}$ das estrelas que parecem ter temperaturas maiores que os limites do algoritmo (vide Figura 4.11). Notamos que a barreira ocorre em $T_{\rm ef} > 8300$ K e que afeta as previsões de 40 objetos. Nossa proposta é utilizar, exclusivamente para estas estrelas, o modelo de $T_{\rm ef}$ apresentado na Subseção 3.2.1.2. Este modelo foi baseado nos dados do levantamento auxiliar SEGUE e mostrou-se aplicável a estrelas com $T_{\rm ef}$ de até 9450 K. A Figura 4.11 mostra a proposta de correção para as estrelas com $T_{\rm ef} > 8300$ K. Este modelo, apesar de também ter se mostrado eficiente, só deve ser aplicado em casos especiais, como este, já que apresentou acurácia inferior ao modelo baseado no LAMOST.

Dois objetos extras merecem ser destacados: KIC 7797992 e KIC 8804387 (retângulo amarelo na Figura 4.10). Ambas as estrelas foram caracterizadas pelo modelo mais preciso deste trabalho (vide Tabela 4.3), onde KIC 8804387 foi corrigida pelo modelo do SEGUE, já que apresentava $T_{\rm ef} > 8300$ K. As magnitudes absolutas, temperaturas e luminosidades destes objetos fornecem evidências de que se tratam de anãs brancas[33]. Seus raios também sugerem isto - Shipman (1979) afirma que anãs brancas tem raios típicos de $0{,}008 < R_\star < 0{,}020\ R_\odot$. Embora KIC 8804387 esteja fora deste intervalo ($R_\star = 0{,}0313 \pm 0{,}0007\ R_\odot$), os demais parâmetros sugerem que ela também seja uma anã branca.

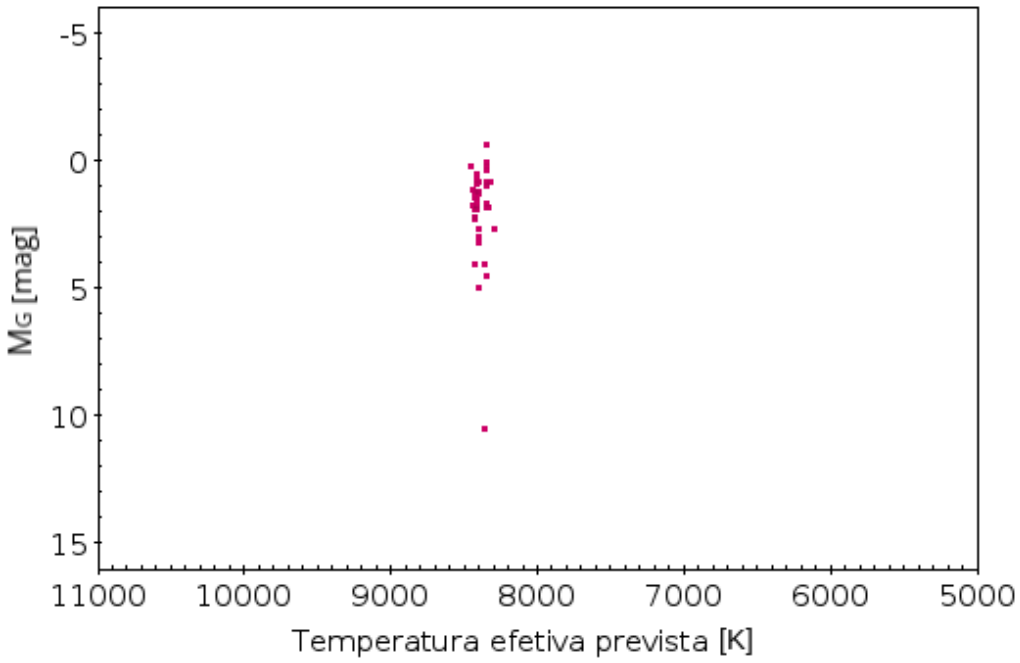
Tabela 4.3: Parâmetros físicos calculados pelos modelos deste trabalho para as candidatas a anãs brancas da amostra.

Estrela [KIC ID]	M_G [Mag]	$T_{\rm ef}$ [K]	log g [dex]	[Fe/H] [dex]	$L_\star$ [$L_\odot$]	$R_\star$ [$R_\odot$]
7797992	+12,7	8139	4,06	-0,96	0,0007 ±0,00001	0,0134 ±0,0005
8804387	+10,5	9016	4,12	-0,85	0,0058 ±0,0001	0,0313 ±0,0007

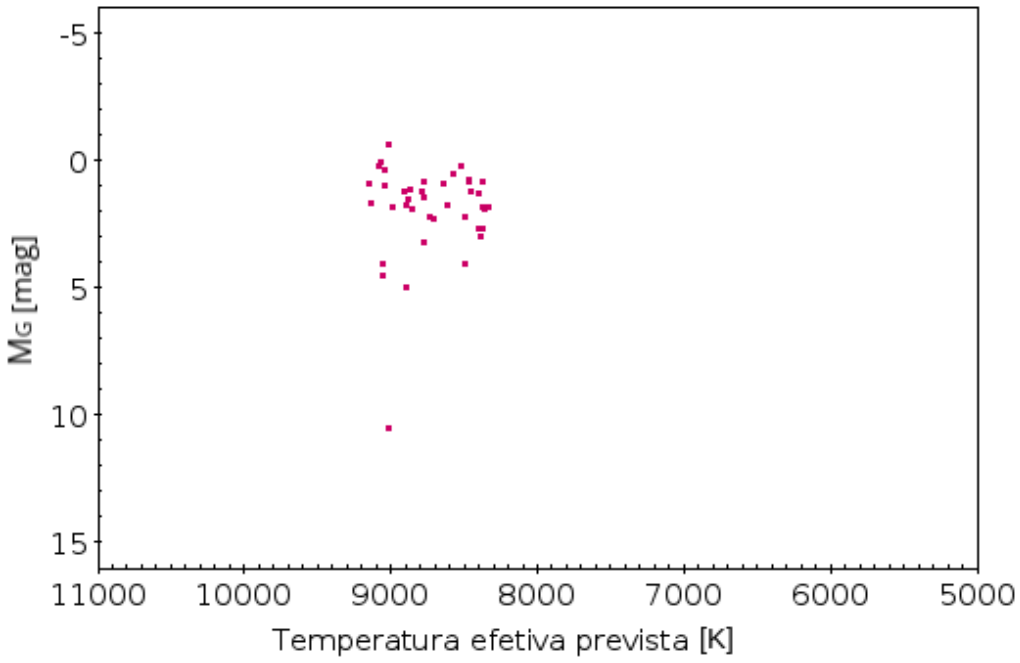
Chamamos atenção também para os valores de log g das candidatas a anãs brancas, apresentados na Tabela 4.3. Comumente, anãs brancas não binárias possuem log g $\approx 8,0$ (Dufour et al., 2008), o que supera o limite máximo do modelo de log g ($0{,}11 <$ log g $<$ 4,90). Em casos como este, este parâmetro não pode ser inferido com precisão pelo modelo atual. Para ambas as candidatas, o KIC não fornece nenhum dado de $T_{\rm ef}$, log g, [Fe/H] ou

[33]Consulte o Diagrama H–R (banda V), de acordo com o European Southern Observatory (ESO), em: https://www.eso.org/public/brazil/images/eso0728c/?lang. Para o Diagrama H–R do Gaia, consulte: https://sci.esa.int/web/gaia/-/60208-hipparcos-hertzsprung-russell-diagram.

$R_\star$. Também não encontramos qualquer indício de classificação destas candidatas como anãs brancas na literatura, o que torna provável que sejamos os primeiros a classificá-las assim.



(a) Previsão de $T_{\rm ef}$ dada pelo LAMOST



(b) Correção de $T_{\rm ef}$ baseada nos dados do SEGUE

Figura 4.11: Diagrama HR com a proposta de correção para as previsões de $T_{\rm ef}$ das 40 estrelas destacadas pelo retângulo azul da Figura 4.10. O painel superior apresenta as previsões do modelo do LAMOST, com comportamento que mostra que elas são provavelmente mais quentes que o limite do algoritmo. O painel inferior mostra as previsões depois da correção pelo modelo baseado nos dados do SEGUE.

Em contraste às anãs brancas, temos KIC 7658030. Ela é a maior ($R_\star \approx 80 \pm 2{,}63\ R_\odot$) e mais luminosa ($L \approx 2247 \pm 46\ L_\odot$) estrela da amostra. No KIC, seu raio é de apenas 21,4 $R_\odot$. Seus parâmetros (M_G = -3,48, $T_{\rm ef}$ = 4455 K e log g = 0,81) a caracterizam como uma gigante luminosa (Morgan et al., 1943). Baseados na classificação de Yerkes e nos valores de M_G, log g e $T_{\rm ef}$, concluímos que a maioria das estrelas da amostra são anãs de sequência principal, mas há alguns casos de gigantes normais, gigantes luminosas e subgigantes. Não há nenhuma supergigante na amostra (a maior massa estimada pertence a KIC 7658030, $M_\star = 4{,}84 \pm 0{,}31\ M_\odot$).

Também avaliamos as estrelas hospedeiras de exoplanetas confirmados e candidatos da Enciclopédia de Exoplanetas[34] e encontramos 13 objetos (11 candidatos e 2 confirmados) também caracterizados por este trabalho. Isso nos permitiu comparar nossos resultados com os do catálogo. A Tabela 4.4 apresenta os parâmetros destes objetos. A 1ª, 2ª e 4ª coluna são semelhantes às da Tabela 4.1. A 5ª e 7ª coluna equivalem às da Tabela 4.2. A 3ª, 6ª e 8ª coluna se dividem em:

- $R_\star^{\rm Enc}$: Raio estelar fornecido pela Enciclopédia de Exoplanetas (em unidades solares, $R_\odot$).
- $R_p^{\rm Enc}$: Raio mínimo do objeto em trânsito (em raios de Júpiter, $R_{\rm Jup}$), fornecido pela Enciclopédia de Exoplanetas.
- Situação: Categoria que divide as detecções em confirmadas (conf) ou candidatas (cand).

Já que não tivemos acesso às curvas de luz das estrelas da Tabela 4.4, para saber o valor de Q recorremos a Equação 1.2, onde podemos obter Q sempre que se conheça $R_\star^{\rm Enc}$ e $R_p^{\rm Enc}$. Entre os 13 objetos, apenas para KIC 9932197 não temos estes valores e, logo, Q não pode ser calculada, não sendo possível obter R_p. A seguir, vamos analisar os demais objetos da tabela. Todos eles foram detectados através de trânsito primário.

O primeiro objeto é o candidato a exoplaneta conhecido atualmente como K06601.01, detectado em órbita da estrela KIC 5597361 com um período de aproximadamente 18±0,0002125 dias, segundo a Enciclopédia de Exoplanetas[34]. Este catálogo reune dados de raio estelar e planetário, mas não fornece informações sobre os autores responsáveis por estes cálculos, tampouco que métodos foram utilizados para obter os valores. Para a estrela, $R_\star^{\rm Enc} = 1{,}09^{+0{,}39}_{-0{,}16}\ R_\odot$ e, com o algoritmo, encontramos um valor de $R_\star$ = 1,40±0,04 $R_\odot$, dentro do limite de incertezas da Enciclopédia. Para K06601.01, $R_p^{\rm Enc} = 1{,}57^{+0{,}57}_{-0{,}24}$ $R_\oplus$ e nós encontramos R_p = 2,02±0,06 $R_\oplus$ (28,66% maior, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de Sub-Netuno (vide Tabela 1.1).

[34] http://www.exoplanet.eu/catalog/

Tabela 4.4: Estrelas hospedeiras da Enciclopédia de Exoplanetas que também foram caracterizadas pelo algoritmo deste trabalho.

Estrela [KIC ID]	$R_{\star KIC}$ $[R_\odot]$	$R_\star^{\text{Enc}}$ $[R_\odot]$	$R_\star$ $[R_\odot]$	Q [%]	R_p^{Enc} $[R_\oplus]$	R_p $[R_\oplus]$	Situação
Modelo mais restrito							
5597361	1,14	$1,09^{+0,39}_{-0,16}$	1,40±0,04	0,0175	$1,57^{+0,57}_{-0,24}$	2,02±0,06	cand.
6345758	1,22	$1,16^{+0,42}_{-0,22}$	1,56±0,04	0,0067	$1,04^{+0,37}_{-0,19}$	1,39±0,03	cand.
6664842	1,00	$1,01^{+0,47}_{-0,09}$	1,05±0,03	0,0070	$0,92^{+0,43}_{-0,08}$	0,96±0,02	cand.
8013289	-	$1,47^{+0,68}_{-0,44}$	1,41±0,04	0,0111	$1,69^{+0,78}_{-0,50}$	1,62±0,04	cand.
9693803	0,90	$1,02^{+0,46}_{-0,10}$	1,06±0,03	0,0214	$1,63^{+0,73}_{-0,15}$	1,70±0,04	cand.
9932197	1,91	-	2,14±0,06	-	-	-	cand.
10053138	0,86	$0,81^{+0,36}_{-0,07}$	0,94±0,03	0,0071	$0,74^{+0,33}_{-0,06}$	0,87±0,02	cand.
10255705	1,83	$2,12^{+0,44}_{-0,52}$	2,86±0,08	0,0993	7,29±2,58	9,82±0,28	conf.
10646620	1,16	$1,13^{+0,60}_{-0,12}$	1,21±0,03	0,0017	$0,50^{+0,27}_{-0,05}$	0,54±0,01	cand.
11068630	0,92	$0,97^{+0,43}_{-0,08}$	1,09±0,03	0,0169	$1,37^{+0,62}_{-0,11}$	1,54±0,04	cand.
11650543	-	$1,69^{+0,72}_{-0,58}$	1,02±0,03	0,0052	$1,33^{+0,56}_{-0,46}$	0,81±0,02	cand.
Modelo menos restrito							
8283875	0,68	$0,59^{+0,05}_{-0,06}$	0,78±0,03	0,0149	$0,78^{+0,07}_{-0,08}$	1,03±0,04	cand.
10450504	0,86	$0,75^{+0,10}_{-0,07}$	0,77±0,02	0,0565	$1,95^{+0,26}_{-0,19}$	1,99±0,06	conf.

A Enciclopédia de Exoplanetas pode ser consultada em: http://www.exoplanet.eu/catalog/

O segundo objeto é o candidato a exoplaneta conhecido atualmente como K06689.01, detectado em órbita da estrela KIC 6345758 com um período de aproximadamente 3,4±0,000064 dias. Para esta estrela, $R_\star{}^{\rm Enc} = 1,16^{+0,42}_{-0,22}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 1{,}56\pm0{,}04\ R_\odot$, dentro do limite de incertezas do catálogo. Para K06689.01, $R_p^{\rm Enc} = 1,04^{+0,37}_{-0,19}\ R_\oplus$ e nós encontramos $R_p = 1{,}39\pm0{,}03\ R_\oplus$ (33,65% maior, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de Super-Terra.

O terceiro objeto é o candidato a exoplaneta conhecido atualmente como K06750.01, detectado em órbita da estrela KIC 6664842 com um período de aproximadamente 5,2±0,000081 dias. Para a estrela, $R_\star{}^{\rm Enc} = 1,01^{+0,47}_{-0,09}\ R_\odot$. Com o algoritmo, encontramos $R_\star = 1{,}05\pm0{,}03\ R_\odot$, que está dentro da incerteza de $R_\star{}^{\rm Enc}$. Para K06750.01, $R_p^{\rm Enc} = 0,92^{+0,43}_{-0,08}\ R_\oplus$ e nós encontramos $R_p = 0{,}96\pm0{,}02\ R_\oplus$ (4,34% maior, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de tipo-Terra.

O quarto objeto é o candidato a exoplaneta conhecido como K06952.01, detectado em órbita da estrela KIC 8013289 com um período de aproximadamente 10,6±0,00028 dias. Para a estrela, a Enciclopédia apresenta $R_\star{}^{\rm Enc} = 1,47^{+0,68}_{-0,44}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 1{,}41\pm0{,}04\ R_\odot$, dentro do limite de incerteza do catálogo. Para K06952.01, $R_p^{\rm Enc} = 1,69^{+0,78}_{-0,50}\ R_\oplus$ e nós encontramos $R_p = 1{,}62\pm0{,}04\ R_\oplus$ (4,14% menor, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de Super-Terra.

O quinto objeto é o candidato a exoplaneta conhecido atualmente como K07224.01, detectado em órbita da estrela KIC 9693803 com um período de aproximadamente 13,5±0,00026 dias. Para a estrela, $R_\star{}^{\rm Enc} = 1,02^{+0,46}_{-0,10}\ R_\odot$. Com o algoritmo, nós encontramos $R_\star = 1{,}06\pm0{,}03\ R_\odot$, dentro do limite de incerteza de $R_\star{}^{\rm Enc}$. Para K07224.01, $R_p^{\rm Enc} = 1,63^{+0,73}_{-0,15}\ R_\oplus$ e nós encontramos $R_p = 1{,}70\pm0{,}04\ R_\oplus$ (4,29% maior, mas ainda dentro da incerteza do catálogo), o que o coloca na categoria de Super-Terra.

O sétimo objeto é o candidato a exoplaneta conhecido como K07279.01, detectado em órbita da estrela KIC 10053138 com um período de aproximadamente 11,8±0,00016 dias. Para a estrela, a Enciclopédia apresenta $R_\star{}^{\rm Enc} = 0,81^{+0,36}_{-0,07}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 0{,}94\pm0{,}03\ R_\odot$, dentro da incerteza do catálogo. Para K07279.01, a Enciclopédia apresenta $R_p^{\rm Enc} = 0,74^{+0,33}_{-0,06}\ R_\oplus$ e nós encontramos $R_p = 0{,}87\pm0{,}02\ R_\oplus$ (17,57% maior, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de tipo-Terra.

O oitavo objeto é o exoplaneta confirmado conhecido como KIC 10255705b, detectado em órbita da estrela KIC 10255705 (a de maior raio da amostra) com um período de aproximadamente $707,38^{+0,004}_{-0,016}$ dias. Para a estrela, $R_\star{}^{\rm Enc} = 2,12^{+0,44}_{-0,52}\ R_\odot$. Com o algoritmo, encontramos $R_\star = 2{,}86\pm0{,}08\ R_\odot$, fora do limite de incerteza da Enciclopédia. Para KIC 10255705b, $R_p^{\rm Enc} = 7{,}29\pm2{,}58\ R_\oplus$ e nós encontramos $R_p = 9{,}82\pm0{,}28\ R_\oplus$ (34,7% maior, mas ainda dentro da incerteza de $R_p^{\rm Enc}$), o que o coloca na categoria de tipo-Júpiter. Com o valor da Enciclopédia, ele seria classificado como Sub-Saturno. Este exoplaneta

também é o maior da amostra de 13 objetos.

O nono objeto é o candidato a exoplaneta conhecido atualmente como K07352.01, detectado em órbita da estrela KIC 10646620 com um período de aproximadamente 2,41±0,00028 dias. Para a estrela, $R_\star{}^{\rm Enc} = 1,13^{+0,60}_{-0,12}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 1{,}21{\pm}0{,}03\ R_\odot$, dentro da incerteza do catálogo. Para K07352.01, a Enciclopédia apresenta $R_p{}^{\rm Enc} = 0,50^{+0,27}_{-0,05}\ R_\oplus$ e nós encontramos $R_p = 0{,}54{\pm}0{,}01\ R_\oplus$ (8,00% maior, mas ainda dentro da incerteza da Enciclopédia), o que o coloca na categoria de tipo-Terra.

O décimo objeto é o candidato a exoplaneta conhecido como K07404.01, detectado em órbita da estrela KIC 11068630 com um período de aproximadamente 1,82±0,000008 dias. Para a estrela, a Enciclopédia apresenta $R_\star{}^{\rm Enc} = 0,97^{+0,43}_{-0,08}\ R_\odot$. Com o algoritmo, encontramos $R_\star = 1{,}09{\pm}0{,}03\ R_\odot$, dentro do limite de incerteza do catálogo. Para K07404.01, $R_p{}^{\rm Enc} = 1,37^{+0,62}_{-0,11}\ R_\oplus$ e nós encontramos $R_p = 1{,}54{\pm}0{,}04\ R_\oplus$ (12,4% maior, mas ainda dentro da incerteza de $R_p{}^{\rm Enc}$), o que o coloca na categoria de Super-Terra.

O décimo primeiro objeto é o candidato a exoplaneta conhecido atualmente como K07467.01, detectado em órbita da estrela KIC 11650543 com um período de aproximadamente 7,21±0,00009 dias. Para a estrela, $R_\star{}^{\rm Enc} = 1,69^{+0,72}_{-0,58}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 1{,}02{\pm}0{,}03\ R_\odot$, fora da incerteza de $R_\star{}^{\rm Enc}$. Para K07467.01, $R_p{}^{\rm Enc} = 1,33^{+0,56}_{-0,46}\ R_\oplus$ e nós encontramos $R_p = 0{,}81{\pm}0{,}02\ R_\oplus$ (39,1% menor e fora da incerteza do catálogo), o que o coloca na categoria de tipo-Terra. No caso da Enciclopédia ele seria classificado como Super-Terra.

Os dois objetos seguintes foram caracterizados apenas pelo modelo menos restrito deste trabalho. O décimo segundo objeto é o candidato a exoplaneta conhecido como K07007.01, detectado em órbita da estrela KIC 8283875 com um período de aproximadamente 0,80±0,000002 dias. Para a estrela, $R_\star{}^{\rm Enc} = 0,59^{+0,05}_{-0,06}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 0{,}78{\pm}0{,}03\ R_\odot$, fora do limite de incerteza da Enciclopédia. Para K07007.01, $R_p{}^{\rm Enc} = 0,78^{+0,07}_{-0,08}\ R_\oplus$ e nós encontramos $R_p = 1{,}03{\pm}0{,}04\ R_\oplus$ (32,1% maior e fora da incerteza de $R_p{}^{\rm Enc}$), o que o coloca no limite entre as categorias tipo-Terra e Super-Terra - vamos considerá-lo um exoplaneta tipo-Terra.

O décimo terceiro objeto é o exoplaneta confirmado conhecido como KOI-7327b, detectado em órbita da estrela KIC 10450504 com um período de aproximadamente 18±0,0001 dias. Para a estrela, a Enciclopédia apresenta $R_\star{}^{\rm Enc} = 0,75^{+0,10}_{-0,07}\ R_\odot$. Com o algoritmo, encontramos um valor de $R_\star = 0{,}77{\pm}0{,}02\ R_\odot$, dentro do limite de incerteza de $R_\star{}^{\rm Enc}$. Para KOI-7327b, $R_p{}^{\rm Enc} = 1,95^{+0,26}_{-0,19}\ R_\oplus$ e nós encontramos $R_p = 1{,}99{\pm}0{,}06\ R_\oplus$ (2,1% maior, mas ainda dentro da incerteza de $R_p{}^{\rm Enc}$), o que o coloca na categoria de Sub-netuno.

Dessa forma, com base nos resultados obtidos por este trabalho, a amostra de 13 objetos parece conter 5 exoplanetas tipo-Terra, 4 super-Terras, 2 sub-Netunos, 1 tipo-Júpiter e 1 objeto que não pôde ser analisado.

4.7 Massa

Obtivemos as massas aproximadas ($M_\star$) das estrelas da missão Kepler que também foram observadas pelo J-PLUS/WISE através da calibração de Torres et al. (2010), descrita na Seção 3.5. Comparamos os resultados obtidos com os encontrados na Enciclopédia de Exoplanetas[35] ($M_\star{}^{\text{autor}}$), para as estrelas hospedeiras descritas na Tabela 4.4 (vide Tabela 4.5). Inserimos a coluna "intervalo de $M_\star{}^{\text{autor}}$" para melhor visualização do intervalo de massa proposto pelos autores para cada objeto. Isto permite que comparemos mais facilmente nossos resultados.

Tabela 4.5: Massa aproximada das estrelas hospedeiras da Tabela 4.4.

Estrela [KIC ID]	$M_\star$ $[M_\odot]$	$M_\star{}^{\text{autor}}$ $[M_\odot]$	Intervalo de $M_\star{}^{\text{autor}}$ $[M_\odot]$
Modelo mais restrito			
5597361	1,20 ± 0,08	$0,967^{+0,147}_{-0,106}$	0,861-1,114
6345758	1,16 ± 0,07	$0,958^{+0,137}_{-0,079}$	0,879-1,095
6664842	1,05 ± 0,06	$1,081^{+0,217}_{-0,137}$	0,944-1,298
8013289	1,05 ± 0,06	$0,960^{+0,188}_{-0,113}$	0,847-1,148
9693803	1,09 ± 0,07	$1,102^{+0,212}_{-0,145}$	0,957-1,314
9932197	1,24 ± 0,08	-	-
10053138	0,97 ± 0,06	$0,856^{+0,104}_{-0,079}$	0,777-0,960
10255705	1,23 ± 0,08	$1,100^{+0,290}_{-0,170}$	0,930-1,390
10646620	1,29 ± 0,08	$1,149^{+0,241}_{-0,150}$	0,999-1,390
11068630	1,11 ± 0,07	$1,083^{+0,198}_{-0,147}$	0,936-1,281
11650543	1,13 ± 0,07	$1,107^{+0,295}_{-0,176}$	0,931-1,402
Modelo menos restrito			
8283875	0,57 ± 0,04	$0,574^{+0,058}_{-0,045}$	0,529-0,632
10450504	0,71 ± 0,05	$0,727^{+0,112}_{-0,057}$	0,670-0,839

Em maioria, os resultados concordam com as estimativas da literatura. É importante notar que, se os resultados de $M_\star{}^{\text{autor}}$ tiverem sido determinados através de trajetórias evolutivas, eles são mais precisos que as calibrações de Torres et al. (2010). O polinômio de Torres et al. (2010) depende de boas estimativas de T_{ef}, log g e [Fe/H] e, apesar de ser efetivo para estrelas de (pós-) sequência principal, ele não realizou boas previsões nas candidatas a anãs brancas. Claramente isto poderia ser ocasionado pela má medição do algoritmo para o log g destes objetos (vide Tabela 4.3), então adotamos o valor de log g

[35] http://www.exoplanet.eu/catalog/

sugerido por Dufour et al. (2008) para anãs brancas não binárias ($\log g \approx 8,0$) e refizemos o cálculo de $M_\star$, usando ainda a calibração de Torres et al. (2010), para KIC 7797992 e KIC 8804387. Entretanto, o problema persistiu e não identificamos a sua causa e por isso descartamos as massas de Torres et al. (2010) para estes objetos. Para os demais casos, as massas obtidas podem ser utilizadas e estarão disponíveis na tabela final de dados.

Para o problema das massas das candidatas a anãs brancas, o trabalho de Smalley (2005) sugere outra relação para $M_\star$, relacionando-a a log g e $R_\star$:

$$\log g - \log g_\odot = \log M_\star - 2\log R_\star \tag{4.1}$$

Podemos isolar $\log M_\star$:

$$\log M_\star = \log g - \log g_\odot + 2\log R_\star \tag{4.2}$$

Considerando o log g sugerido por Dufour et al. (2008), a relação de Smalley (2005) permitiu encontrar para KIC 7797992 uma $M_\star = 0{,}66\ M_\odot$, condizente com o esperado para uma anã branca. Para KIC 8804387 o problema persistiu e encontramos $M_\star = 3{,}58\ M_\odot$. Visto isso, descartamos a massa dessa candidata a anã branca. Serão necessários dados adicionais para a caracterização deste objeto.

Capítulo 5

Conclusões

Este trabalho propôs a criação de modelos de previsão de parâmetros físicos estelares, tais como $T_{\rm ef}$, log g e [Fe/H]. Estes modelos foram baseados em algoritmos de Aprendizagem de Máquina e possuíam o objetivo de caracterizar, com menor erro que o da literatura, as estrelas da missão Kepler, principalmente aquelas com $K_p > 16$ (estrelas fracas). Para chegarmos aos resultados finais deste trabalho, seguimos algumas etapas, resumidas brevemente a seguir.

Inicialmente, foi necessário selecionar os dados de magnitude provenientes do sistema de 12 filtros do levantamento de dados J-PLUS, os quais elegemos apenas aqueles que atendiam os seguintes requisitos: objetos com mais de 90% de probabilidade de serem estrelas, observados na abertura circular fixa de 6", em todos os 12 filtros do levantamento. Além disso, as magnitudes obtidas deveriam possuir um erro de medição menor que 0,1 (e_mag $<$ 0,1 mag) e valores para correção de extinção (vide Subseção 2.2.1). Também foram acrescentadas as magnitudes medidas para os objetos pelo sistema de filtros no infravermelho do levantamento de dados WISE. Assim, cada objeto possuía 16 medições de magnitude, sendo 12 provenientes do J-PLUS e 4 do WISE. Combinamos, em pares, as medições dos 16 filtros, o que retornou 120 cores (vide Seção 3.2). Desta forma, no total, cada objeto reunia 136 informações de magnitude. Estes dados foram a base do modelo de previsão de parâmetros mais preciso deste trabalho, o qual chamamos simplesmente de modelo mais restrito.

Os dados acima foram cruzados com os bancos de dados de 4 levantamentos (TESS, SEGUE, GALAH e LAMOST; vide Subseção 3.2.1), a fim de coletarmos as medidas de $T_{\rm ef}$, log g e [Fe/H] para estes objetos. Chamamos estes de levantamentos auxiliares. Cada um destes cruzamentos gerou uma amostra de treinamento, que são tabelas de dados usadas para alimentar algoritmos baseados em Aprendizagem de Máquina. A amostra de treinamento baseada no TESS possuia 1.291.942 objetos e mostrou uma acurácia razoável para $T_{\rm ef}$ (95,28%), com erro médio de $\pm$ 69 K, porém com alta dispersão entre as previsões do algoritmo e os dados do TESS. Para log g e [Fe/H] as dispersões foram ainda maiores e os rendimentos mais baixos (43,36% com erro médio de $\pm$ 0,167 dex e 88,03% com

erro médio de ± 0,124 dex, respectivamente). Em particular, nos chamou atenção o desbalanceamento dos dados de [Fe/H] na amostra, onde apenas 142.694 das estrelas possuíam medida para este parâmetro.

A amostra de treinamento baseada no SEGUE apresentou 14.831 objetos, todos eles com medições nos três parâmetros. Sua acurácia foi bastante superior ao TESS: 98,44% para $T_{\rm ef}$ com erro médio de ± 102 K, 83,37% para log g com erro médio de ± 0,211 dex e 90,11% para [Fe/H] com erro médio de ± 0,121 dex. A dispersão diminuiu nos três parâmetros. A amostra de treinamento baseada no GALAH precisou ser descartada, pois a baixa quantidade de objetos (apenas 2018 estrelas) não permitia a reprodutibilidade do modelo.

A amostra de treinamento baseada no LAMOST apresentou 186.374 objetos. Esta quantidade de objetos permitiu realizar restrições no erro das medidas destes parâmetros. Para $T_{\rm ef}$, selecionamos apenas objetos com erro $<$ 200 K, o que reduziu a amostra para 137.865 estrelas. Para log g, o corte foi feito em 0,2 dex, o que selecionou 101.740 das 186.374 estrelas iniciais. Fizemos uma restrição adicional para log g, selecionando apenas objetos que possuíam medidas de distâncias disponibilizadas pelo Gaia eDR3, com erro menor que 30%. Isso gerou uma amostra final de treinamento para log g de 97.526 objetos. Para [Fe/H], o corte também foi feito em 0,2 dex. Também incluímos os dados de magnitude das bandas J, H e K do 2MASS sugeridos por Schlaufman & Casey (2014), que demonstraram a influência positiva destes no cálculo de metalicidade. A amostra final de treinamento para [Fe/H] foi de 143.787 objetos. As acurácias dos modelos foram de: 98,59% para $T_{\rm ef}$ com erro médio de ± 70 K, 97,21% para log g com erro médio de ± 0,08 dex e 92,52% para [Fe/H] com erro médio de ± 0,10 dex.

O LAMOST foi, então, o melhor levantamento auxiliar, pois apresentou as melhores acurácias, além de uma quantidade elevada de objetos. A partir da amostra de treinamento deste levantamento auxiliar, desenvolvemos o modelo final de previsão de parâmetros para as 29.164 estrelas do Kepler também observadas pelo J-PLUS/WISE. Conhecendo a magnitude aparente na banda G (m_G) e a distância dos objetos, fornecidas no Gaia eDR3, calculamos a magnitude absoluta na mesma banda (M_G) para os objetos. Com $T_{\rm ef}$, log g e [Fe/H] destas estrelas inferidas pelo modelo, desenvolvemos um modelo extra de estrutura semelhante, também baseado em Aprendizagem de Máquina, para a previsão de valores de correção bolométrica (BC). O modelo apresentou acurácia de 99,89% e erro médio de 0,022, onde 97,8% das previsões obtiveram erro $\leqslant 3\sigma$. Erros $> 3\sigma$ ocorreram para objetos com $T_{\rm ef} <$ 3.500 K e $T_{\rm ef} >$ 45.000 K, o que está fora da cobertura típica do Kepler (vide Seção 3.4).

Com os valores de BC previstos, pudemos calcular a magnitude bolométrica ($M_{\rm bol}$). Com ela, calculamos a luminosidade ($L_\star$) dos objetos através da relação fornecida em Carroll & Ostlie (2007) e, a partir de $L_\star$, calculamos o raio dos objetos ($R_\star$). Por último, usamos as calibrações de Torres et al. (2010) para obter a massa aproximada ($M_\star$) das

estrelas (vide Seção 3.5). Comparamos nossos resultados para objetos em comum caracterizados pela literatura. Nossas medidas de $R_\star$ possibilitaram, inclusive, propôr correções em estimativas de raios de objetos em trânsito (R_p) conhecidos (vide Seção 4.6).

Também desenvolvemos um segundo modelo, o qual chamamos de modelo menos restrito, que permitiu incluir objetos com um valor maior de erro na magnitude (e_mag $<$ 0,2 mag; vide Subseção 3.2.2). As amostras de treinamento foram de 144.774 objetos para $T_{\rm ef}$, 103.341 objetos para log g e 152.275 objetos para [Fe/H]. Este modelo, apesar de incluir um ruído extra nas magnitudes, apresentou excelentes resultados: 99,10% de acurácia para $T_{\rm ef}$, com erro médio de $\pm$ 58 K, 97,31% para log g, com erro médio de $\pm$ 0,08 dex e 91,10% para [Fe/H], com erro médio de $\pm$ 0,10 dex. Esta amostra, permitiu elevar o número de estrelas do Kepler caracterizadas para 44.483 objetos (ganho $>$ 52%). Assim como no modelo mais restrito, além de $T_{\rm ef}$, log g e [Fe/H], calculamos M_G, BC, $M_{\rm bol}$, $L_\star$, $R_\star$ e $M_\star$ dos objetos.

Vale ressaltar que, já que o algoritmo realiza previsões através do reconhecimento de padrões presentes na amostra de treinamento, ele é limitado a prever objetos semelhantes aos quais foi treinado. Os modelos finais cobrem objetos dentro dos seguintes valores: 3790 K $< T_{\rm ef} <$ 8500 K, 0,11 $<$ log g $<$ 4,90, e -2,5 $<$ [Fe/H] $<$ 0,72. Para casos particulares de objetos com $T_{\rm ef} >$ 8300 K é sugerido o uso do modelo de $T_{\rm ef}$ baseado no SEGUE, que possui cobertura até 9450 K.

Sobre as 29.164 estrelas da missão Kepler, caracterizadas pelo modelo mais restrito, pudemos observar que:

- 4688 objetos apresentaram log g $<$ 4,0, dos quais 1793 tiveram log g $<$ 3,5 e 47 tiveram log g $<$ 2,0.

- Uma parte considerável (49%) da amostra apresentou [Fe/H] entre -0,8 e -0,3. Cerca de 1% estão abaixo de [Fe/H] = -1,5. Apenas 10 objetos apresentaram [Fe/H] $<$ -2,0.

- A classificação de Yerkes e os valores de M_G, log g, $T_{\rm ef}$ e $M_\star$ das estrelas sugerem que a maioria é composta por anãs de sequência principal, mas há alguns casos de gigantes normais, gigantes luminosas e subgigantes.

- A M_G, $T_{\rm ef}$, $L_\star$ e $R_\star$ de KIC 7797992 e KIC 8804387 sugerem que estes objetos são anãs brancas.

- Destas estrelas, 3235 (cerca de 11%) não possuem nenhum dado de $T_{\rm ef}$, log g, [Fe/H] ou $R_\star$ registrado no KIC.

Para a amostra de 44.483 estrelas da missão Kepler, caracterizadas pelo modelo menos restrito, observamos que:

- Destas estrelas, 5115 (cerca de 17,5%) não possuem nenhum dado de $T_{\rm ef}$, log g, [Fe/H] ou $R_\star$ registrado no KIC.

- As previsões de $L_\star$ e $R_\star$ deste modelo apresentaram correlação de 99,89% e 99,88%, respectivamente, com as previsões do modelo mais restrito, o que nos dá segurança para utilizá-las.

Com isso, concluímos que este trabalho permitiu inferir parâmetros mais precisos para dezenas de milhares de estrelas. Apresentamos a seguir nossas perspectivas futuras:

- Primeiramente, pretendemos obter σ_{m_G} para os objetos que não possuem este dado no Gaia. Para isso, planejamos realizar uma aproximação destes valores, baseando-nos na incerteza média de estrelas com características semelhantes (tipo de estrela, distância, m_G, etc);

- Com σ_{m_G} calculada, poderemos propagar σ_{M_G}, avaliando objetos muito distantes, que costumam possuir altos valores de σ_d;

- O ponto anterior nos permitirá inferir incertezas mais precisas para M_{bol}, $L_\star$ e $R_\star$;

- Parte dos dados usados neste trabalho (DR2 do J-PLUS e eDR3 do Gaia) foi recentemente atualizada por novas liberações de dados: DR3 para o J-PLUS e DR3 para o Gaia. Estes novos dados podem conter estrelas da missão Kepler ainda não caracterizadas pelo algoritmo deste trabalho. Além disso, liberações de dados futuras possuem o mesmo potencial;

- Pretendemos usar os dados futuros do *Javalambre-Physics of the Accelerated Universe Astrophysical Survey* (J-PAS; Dupke et al., 2019) para treinamento do algoritmo deste trabalho. O conjunto de 16 filtros (J-PLUS + WISE) usado neste trabalho gerou amostras de treinamento com 120 cores. O J-PAS contará com um conjunto óptico de 56 filtros, o que permitirá gerar amostras de treinamento com 1540 cores.

- Com os passos anteriores, seremos capazes de realizar uma melhor análise dos parâmetros planetários;

- Um artigo, discutindo os resultados apresentados neste trabalho e os resultados das etapas mencionadas acima, está em preparação.

Referências Bibliográficas

AIGRAIN, S. (2005) "Planetary transits and stellar variability". Tese de Doutorado em Astronomia. *Institute of Astronomy & Corpus Christi College*, http://www.astro.ex.ac.uk/people/suz/docs/thesis/thesis.pdf

AITKEN, R. G. (1935) "Binary Stars". *Nature* 136, 590-591, https://doi.org/10.1038/136590a0

BATALHA, N. M.; BORUCKI, W. J.; KOCH, D. G.; BRYSON, S. T.; HAAS, M. R. et al. (2010) "Selection, prioritization, and characteristics of kepler target stars". *The Astrophysical Journal Letters* 713, L109, https://doi.org/10.1088/2041-8205/713/2/L109

BERGSTRA, J.; BENGIO, Y. (2012) "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research* 13, 281-305, https://doi.org/10.5555/2188385.2188395

BORUCKI, W. J.; KOCH, D. G.; BASRI, G. B.; BATALHA, N.; BROWN, T. et al. (2010) "Kepler Planet-Detection Mission: Introduction and First Results". *Science* 327, 977-980, https://doi.org/10.1126/science.1185402

BREIMAN, L. (2001) "Random Forests. Machine Learning". *Kluwer Academic Publishers* 45, 5-32, https://doi.org/10.1023/A:1010933404324

BROWN, A. G. A.; VALLENARI, A.; PRUSTI, T.; DE BRUIJNE, J. H. J.; BABUSIAUX, C. et al., (2021) "Gaia Early Data Release 3: Summary of the contents and survey properties". *Astronomy & Astrophysics*, 650, 1432-0746, http://dx.doi.org/10.1051/0004-6361/202039657e

BROWN, T. M.; LATHAM, D. W.; EVERETT, M. E.; ESQUERDO, G. A. et al. (2011) "Kepler Input Catalog: Photometric Calibration and Stellar Classification". *The Astronomical Journal* 142, 112, https://doi.org/10.1088/0004-6256/142/4/112

BUDER, S.; SHARMA, S.; KOS, J.; AMARSI, A. M.; NORDLANDER, T. et al. (2021) "The GALAH+ survey: Third data release". *Monthly Notices of the Royal Astronomical Society* 506, 150-201, https://doi.org/10.1093/mnras/stab1242

CARROLL, B. W.; OSTLIE, D. A. (2007) "An Introduction to Modern Astrophysics". *Pearson*, San Francisco, CA, ISBN 978-0-321-44284-0.

CENARRO, A. J.; MOLES, M.; CRISTÓBAL-HORNILLOS, D.; MARÍN-FRANCH, A.; EDEROCLITE, A. et al. (2019) "J-PLUS: The javalambre photometric local universe survey". *Astronomy & Astrophysics* 622, A176, https://doi.org/10.1051/0004-6361/201833036

CHABRIER, G.; BARAFFE, I. (2000) "Theory of Low-Mass Stars and Substellar Objects". *Astronomy & Astrophysics* 38, 337-377, https://doi.org/10.1146/annurev.astro.38.1.337

CORDEIRO, V. (2022, in prep.) "STEllar Parameter Predictors (STEPPs)". *Zenodo*, https://doi.org/10.5281/zenodo.5628938

CROSTA, M.; VECCHIATO, A. (2010) "Gaia relativistic astrometric models I. Proper stellar direction and aberration". *Astronomy & Astrophysics* 509, A37, https://doi.org/10.1051/0004-6361/200912691

DAMASIO, F. (2011) "O início da revolução científica: questões acerca de Copérnico e os epiciclos, Kepler e as órbitas elípticas". *Revista Brasileira de Ensino de Física* 33, 1-6, https://doi.org/10.1590/S1806-11172011000300020

DANIELSON, D. R. (2001) "The great Copernican cliché". *American Journal of Physics* 69, 1029-1035, https://doi.org/10.1119/1.1379734

DE SILVA, G. M.; FREEMAN, K. C.; BLAND-HAWTHORN, J.; MARTELL, S.; DE BOER, E. et al. (2015) "The GALAH survey: scientific motivation". *Monthly Notices of the Royal Astronomical Society* 449, 2604-2617, https://doi.org/10.1093/mnras/stv327

DÍAZ, R. F.; DAMIANI, C.; DELEUIL, M.; ALMENARA, J. M.; MOUTOU, C. et al., (2013) "SOPHIE velocimetry of Kepler transit candidates. VIII. KOI-205b: a brown-dwarf companion to a K-type dwarf". *Astronomy & Astrophysics* 551, L9, https://doi.org/10.1051/0004-6361/201321124

DONG, S.; XIE, J.-W.; ZHOU, J.-L.; ZHENG, Z.; LUO, A. (2018) "LAMOST telescope reveals that Neptunian cousins of hot Jupiters are mostly single offspring of stars that are rich in heavy elements". *PNAS* 115, 266, https://doi.org/10.1073/pnas.1711406115

DUFOUR, P.; FONTAINE, G.; LIEBERT, J.; SCHMIDT, G. D.; BEHARA, N. (2008) "Hot DQ White Dwarfs: Something Different". *The Astrophysical Journal* 683, 978, https://doi.org/10.1086/589855

DUPKE, R. A.; IRWIN, J.; BONOLI, S.; CENARRO, J.; ABRAMO, R. et al. (2019) "J-PAS: The Javalambre-Physics of the Accelerating Universe Astrophysical Survey". *American Astronomical Society*, 233, https://ui.adsabs.harvard.edu/abs/2019AAS...23338301D

FISHER, D. A.; VALENTI, J. (2005) "The Planet-Metallicity Correlation". *The Astrophysical Journal* 622, 1102, https://doi.org/10.1086/428383

FUKUGITA, M.; ICHIKAWA, T.; GUNN, J. E.; DOI, M.; SHIMASAKU, K. et al. (1996) "The Sloan Digital Sky Survey Photometric System". *The Astronomical Journal* 111, 1748, https://doi.org/10.1086/117915

FULTON, B. J.; PETIGURA, E. A.; HOWARD, A. W.; ISAACSON, H.; MARCY, G. W. et al. (2017) "The California-Kepler Survey. III. A Gap in the Radius Distribution of Small Planets". *The Astronomical Journal* 154, 109, https://doi.org/10.3847/1538-3881/aa80eb

GANG, Z.; YONG-HENG, Z.; YAO-QUAN, C.; YI-PENG, J.; LI-CAI, D. (2012) "LAMOST spectral survey - An overview". *Research in Astronomy and Astrophysics* 12, 723, https://doi.org/10.1088/1674-4527/12/7/002

GATTI, H. (2010) "Essays on Giordano Bruno". *Princeton University Press* 1ed, 376, ISBN 978-0-691-14839-7

GHEZZI, L.; MARTINEZ, C. F.; WILSON, R. F.; CUNHA, K.; SMITH, V. V.; MAJEWSKI, S. R. (2021) "A Spectroscopic Analysis of the California-Kepler Survey Sample. II. Correlations of Stellar Metallicities with Planetary Architectures". *The Astrophysical Journal* 920, 19, https://doi.org/10.3847/1538-4357/ac14c3

HUTTER, F.; HOOS, H. H.; LEYTON-BROWN, K; STUTZLE, T. (2009) "ParamILS: an automatic algorithm configuration framework". *Journal of Artificial Intelligence Research* 36, 267-306, https://doi.org/10.1613/jair.2861

JOHNSON, J. A.; APPS, K.; GAZAK, J. Z.; CREPP, J. R.; CROSSFIELD I. J. et al. (2011) "A transiting field brown dwarf discovered by the Kepler mission". *The Astrophysical Journal* 730, 79, https://doi.org/10.1088/0004-637X/730/2/79

JORDI, C.; GEBRAN, M.; CARRASCO, J. M.; DE BRUIJNE, J.; VOSS, H. et al., (2010) "Gaia broad band photometry". *Astronomy & Astrophysics*, 523, A48, https://doi.org/10.1051/0004-6361/201015441

KALER, James B. (1989) "Stars and their spectra: an introduction to the spectral sequence." *Cambridge University Press*, 2ed, 263, ISBN 978-0-521-89954-3.

KASTING, J. F. (1997) "Habitable zones around low mass stars and the search for extra-terrestrial life. Planetary and interstellar processes relevant to the origins of life". *Springer* 1, 291-307, https://doi.org/10.1007/978-94-015-8907-9_15

KIRK, B.; CONROY, K.; PRSA, A.; ABDUL-MASHI A.; KOCHOSKA, A. et al., (2016) "Kepler eclipsing binary stars. Vii. The catalog of eclipsing binaries found in the entire kepler data set". *The Astronomical Journal* 151, 68, https://doi.org/10.3847/0004-6256/151/3/68

KOCH, D. G.; BORUCKI, W. J.; BASRI, G. B.; BATALHA, N. M.; BROWN, T. M. et al. (2010) "Kepler Mission Design, Realized Photometric Performance, and Early Science". *Astrophysics Journal* 713, L79-L86, https://www.doi.org/10.1088/2041-8205/713/2/L79

KOPPARAPU, R. K.; HÉBRARD, E.; BELIKOV, R.; BATALHA, N. M.; MULDERS, G. D. et al. (2018) "Exoplanet Classification and Yield Estimates for Direct Imaging Missions". *The Astrophysical Journal* 856, 122, https://doi.org/10.3847/1538-4357/aab205

LATHAM, D. W.; BROWN, T. M.; MONET, D. G.; EVERETT, M.; ESQUERDO, G. A. et al. (2005) "The Kepler input catalog". *American Astronomical Society Meeting Abstracts* 207, 110, https://ui.adsabs.harvard.edu/abs/2005AAS..20711013L/exportcitation

LEE, Y. S.; BEERS, T. C.; SIVARANI, T.; PRIETO, C. A.; KOESTERKE, L. et al. (2008) "The SEGUE Stellar Parameter Pipeline. I. Description and Comparison of Individual Methods". *The Astronomical Journal* 136, 2022-2049 https://doi.org/doi:10.1088/0004-6256/136/5/2022

MARTINS, A.; LIMA, T. O.; BOLZAN, M. J. A.; SOUSA, P. A.; FILHO, V. B. L. et al., (2020) "Cálculo do valor da unidade astronômica: como o trânsito de Mercúrio nos indica a nossa distância ao Sol". *Revista Latino-Americana de Educação em Astronomia (RELEA)* 30, 51-64, https://doi.org/10.37156/RELEA/2020.30.051

MAYOR, M.; QUELOZ, D. (1995) "A Jupiter-mass companion to a solar-type star". *Nature* 378, 355-359, https://doi.org/10.1038/378355a0

MOLENDA-ZAKOWICZ, J.; JERZYKIEWICZ, M.; FRASCA, A.; CATANZARO, G.; KOPACKI, G.; LATHAM, D. W. (2010) "Characteristics of 100+ Kepler Aste-

roseismic Targets from Ground-Based Observations". *Astron. Nachr.* 999, 789-792, https://arxiv.org/pdf/1005.0985.pdf

MORGAN, W. W.; KEENAN, P. C.; KELLMAN, E. (1943) "An atlas of stellar spectra, with an outline of spectral classification". The University of Chicago press. Chicago, IL.

MOUTOU, C.; BONOMO, A. S.; BRUNO, G.; MONTAGNIER, G.; BOUCHY, F. et al., (2013) "SOPHIE velocimetry of Kepler transit candidates. IX. KOI-415b: a long-period, eccentric transiting brown dwarf to an evolved Sun". *Astronomy & Astrophysics* 558, L6, https://doi.org/10.1051/0004-6361/201322201

MULDERS, G. D.; PASCUCCI, I.; APAI, D.; FRASCA, A.; MOLENDA-ZAKOWICZ, J. (2016) "A Super-Solar metallicity for stars with hot rocky exoplanets". *The Astrophysical Journal* 152, 187, https://doi.org/10.3847/0004-6256/152/6/187

NAKAMURA, F.; UMEMURA, M. (2001) "On the Initial Mass Function of Population III Stars". *The Astrophysical Journal* 548, 19, https://doi.org/10.1086/318663

NEDJATI-GILANI, G. L.; SCHNEIDER, T.; HALL, M. G.; CAWLEY, N.; HILL, I. et al., (2017) "Machine learning based compartment models with permeability for white matter microstructure imaging". *NeuroImage* 150, 119-135, https://doi.org/10.1016/j.neuroimage.2017.02.013

NOGUEIRA, P. H. S. S. P. (2020) "Detecção de exoplanetas ao redor de estrelas fracas observadas pela missão Kepler". Dissertação de Mestrado em Astronomia. *Observatório Nacional*, https://antigo.on.br/conteudo/dppg_e_iniciacao/dppg/ferramenta_teses/teses/ASTRONOMIA/[472_32-26_C]dissertacao-pedronogueira.pdf

PARKINSON, C. D.; LIANG, M.-C.; HARTMAN, H.; HANSEN, C. J.; TINETTI, G. et al. (2007) "Enceladus: Cassini observations and implications for the search for life". *Astronomy & Astrophysics* 463, 353-357, https://doi.org/10.1051/0004-6361:20065773

PETIGURA, E. A.; HOWARD, A. W.; MARCY, G. W.; JOHNSON, J. A.; ISAACSON, H. et al. (2017) "The California-Kepler Survey. I. High-resolution Spectroscopy of 1305 Stars Hosting Kepler Transiting Planets". *The Astronomical Journal* 154, 107, https://doi.org/10.3847/1538-3881/aa80de

PETIGURA, E. A.; MARCY, G. W.; WINN, J. N.; WEISS, L. M.; FULTON, B. J. et al. (2018) "The California-Kepler Survey. IV. Metal-rich Stars Host a Greater Diversity of Planets". *The Astronomical Journal* 155, 89, https://doi.org/10.3847/1538-3881/aaa54c

PISKUNOV, N.; VALENTI, J. A. (1996) "Spectroscopy Made Easy: A new tool for fitting observations with synthetic spectra". *Astronomy & Astrophysics* 118, 595-603, https://doi.org/10.1051/aas:1996222

RIBAS, I.; GUINAN, E. F.; GUDEL, M.; AUDARD, M. (2005) "Evolution of the Solar Activity over Time and Effects on Planetary Atmospheres. I. High-Energy Irradiances (1-1700 Å)". *The Astrophysical Journal* 622, 680, https://doi.org/10.1086/427977

RICKER, G. R.; VANDERSPEK, R. K.; LATHAM, D. W.; WINN, J. N. (2014) "The Transiting Exoplanet Survey Satellite Mission". *American Astronomical Society Meeting Abstracts* 224, 113, https://ui.adsabs.harvard.edu/abs/2014AAS...22411302R/abstract

ROCKOSI, C. M. (2005) "The Sloan Extension for Galactic Understanding and Exploration". *American Astronomical Society Meeting Abstracts* 207, 147, https://ui.adsabs.harvard.edu/abs/2005AAS...20714701R/abstract

RYBICKI, G. B.; LIGHTMAN, A. P. (2004) "Radiative processes in astrophysics". WILEY-VCH Verlag GmbH & Co., Weinheim, Alemanha, ISBN 0-471-82759-2

SANTIAGO, B. (2005) "Aberração". https://www.if.ufrgs.br/oei/santiago/fis2005/textos/varcrds.htm

SANTOS, W. C.; AMORIM, R. G. G. (2017) "Descobertas de exoplanetas pelo método do trânsito". *Rev. Bras. Ensino Fís.* 39, e2308, https://doi.org/10.1590/1806-9126-RBEF-2016-0217

SCHLAUFMAN, K. C.; CASEY, A. R. (2014) "The Best and Brightest Metal-poor Stars". *The Astrophysical Journal* 797, 13, https://doi.org/10.1088/0004-637X/797/1/13

SHARMA, S.; STELLO, D.; BUDER, S.; KOS, J.; BLAND-HAWTHORN, J. et al. (2018) "The TESS-HERMES survey data release 1: high-resolution spectroscopy of the TESS southern continuous viewing zone". *Monthly Notices of the Royal Astronomical Society* 473, 2004-2019, https://doi.org/10.1093/mnras/stx2582

SHEINIS, A. I.; JIMENEZ, B. A.; ASPLUND, M.; BACIGALUPO, C.; BARDEN, S. C. et al. (2015) "First light results from the High Efficiency and Resolution Multi-Element Spectrograph at the Anglo-Australian Telescope". *Journal of Astronomical Telescopes, Instruments, and Systems* 1, 1-18, https://doi.org/10.1117/1.JATIS.1.3.035002

SHIPMAN, H. L. (1979) "Masses and radii of white-dwarf stars. III. Results for 110 hydrogen-rich and 28 helium-rich stars". *The Astrophysical Journal* 228, 240-256, https://doi.org/10.1086/156841

SMALLEY, B. (2005) "Teff and log g determinations". *Mem. S.A.* 75, 1, https://doi.org/10.48550/arXiv.astro-ph/0509535

SMITH, J. A.; TUCKER, D. L.; KENT, S.; RICHMOND, M. W.; FUKUGITA, M. et al. (2002) "The u'g'r'i'z' Standard-Star System". *The Astronomical Journal* 123, 2121-2144, https://doi.org/10.1086/339311

SMITH, J. C.; STUMPE, M. C.; CLEVE, J. E.; JENKINS, J. M.; BARCLAY, T. S. et al. (2012) "Kepler Presearch Data Conditioning II - A Bayesian Approach to Systematic Error Correction". *Astronomical Society of the Pacific* 124, 1000, https://doi.org/10.1086/667697

SMITH, M. D. (2004). "The Origin of Stars". Imperial College Press. pp. 57-68. ISBN 978-1-86094-501-4.

STASSUN, K. G.; OELKERS, R. J.; PAEGERT, M.; TORRES, G.; PEPPER, J. et al. (2019) "The Revised TESS Input Catalog and Candidate Target List". *The Astronomical Journal* 158, 138, https://doi.org/10.3847/1538-3881/ab3467

STASSUN, K. G.; OELKERS, R. J.; PEPPER, J.; PAEGERT, M.; DE LEE, N. et al. (2018) "The TESS Input Catalog and Candidate Target List". *The Astronomical Journal* 156, 102, https://doi.org/10.3847/1538-3881/aad050

TORRES, G.; ANDERSEN, J.; GIMÉNEZ, A. (2010) "Accurate masses and radii of normal stars: modern results and applications". *The Astronomy and Astrophysics Review* 18, 67-126, https://doi.org/10.1007/s00159-009-0025-1

VAN RIJN, J. N.; HUTTER, F. (2018) "Hyperparameter Importance Across Datasets". *Association for Computing Machinery*, 2367-2376, https://doi.org/10.1145/3219819.3220058

WARREN, J. (2004) "Ancient atomists on the plurality of worlds". *The Classical Quarterly* 54, 354-365, https://doi.org/10.1093/clquaj/bmh044

WILSON, R. F.; TESKE, J.; MAJEWSKI, S. R.; CUNHA, K.; SMITH, V. 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* 155, 68, https://doi.org/10.3847/1538-3881/aa9f27

WILSON, T. G.; GOFFO, E.; ALIBERT, Y.; GANDOLFI, D.; BONFANTI, A. 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* 511, 1043-1071, https://doi.org/10.1093/mnras/stab3799

WINN, J. N. (2010) "Exoplanet Transits and Occultations". *arXiv* 1, 2, https://doi.org/10.48550/arXiv.1001.2010

WRIGHT, E. L.; EISENHARDT, P. R. M.; MAINZER, A. K.; RESSLER, M. E.; CUTRI, R. M. et al. (2010) "The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance". *The Astronomical Journal* 140, 1868, https://doi.org/10.1088/0004-6256/140/6/1868

WRIGHT, J. T. (2018). "Radial Velocities as an Exoplanet Discovery Method". *Springer* 1, 619-631, https://doi.org/10.1007/978-3-319-55333-7_4

XIANG, M. S.; LIU, X. W.; YUAN, H. B.; HUANG, Y.; HUO, Z. Y. et al. (2015) "The LAMOST stellar parameter pipeline at Peking University-LSP3". *Monthly Notices of the Royal Astronomical Society* 448, 822-854, https://doi.org/10.1093/mnras/stu2692

YAN, M.; LIU, K.; GUAN, Z.; XINKAI, X.; QIAN, X. et al., (2018) "Background Augmentation Generative Adversarial Networks (BAGANs): Effective Data Generation Based on GAN-Augmented 3D Synthesizing". *Simmetry* 10, 734, https://doi.org/10.3390/sym10120734

YAO, X.; LIU, Y. (2013) "Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques: Machine Learning". *Springer Science + Business Media, Springer, Boston, MA,* ISBN 978-1-4614-6940-7, https://doi.org/10.1007/978-1-4614-6940-7_17

Apêndice A

Tabela de resultados finais

Os resultados finais deste estudo estão parcialmente apresentados na Tabela A.1. Apresentamos a divisão das colunas:

- Estrela: Número identificador do objeto no catálogo KIC (KIC ID).
- $T_{\rm ef}$: Temperatura efetiva estelar, calculada pelo modelo mais restrito deste trabalho (em Kelvin, K), cujo erro médio é $\pm$ 70 K.
- log g: Logaritmo de base 10 da gravidade superficial estelar, calculada pelo modelo mais restrito deste trabalho (em expoente decimal, dex), cujo erro médio é $\pm$ 0,08.
- [Fe/H]: Metalicidade estelar, calculada pelo modelo mais restrito deste trabalho (em dex), cujo erro médio é $\pm$ 0,10 dex.
- M_G: Magnitude absoluta na banda G, calculada com base nos dados de magnitude aparente na banda G e distância, ambas fornecidas pela missão Gaia eDR3 (em unidades de magnitude, mag).
- M_{bol}: Magnitude bolométrica calculada por este trabalho (em mag).
- BC: Correção bolométrica estimada por um algoritmo desenvolvido por este trabalho (em mag), cujo erro médio é $\pm$ 0,022 mag.
- $L_\star$: Luminosidade estelar, calculada com base nos parâmetros previstos pelo modelo mais restrito deste trabalho (em unidades solares, $L_\odot$).
- $R_\star$: Raio estelar, calculado com base nos parâmetros previstos pelo modelo mais restrito deste trabalho (em unidades solares, $R_\odot$).
- $M_\star$: Massa estelar, estimada com base nas calibrações de Torres et al. (2010) (em unidades solares, $M_\odot$).

Tabela A.1: Estrelas da missão Kepler caracterizadas pelo modelo mais restrito deste trabalho.

Estrela [KIC ID]	$T_{\rm ef}$ [K]	log g [dex]	[Fe/H] [dex]	M_G [mag]	$M_{\rm bol}$ [mag]	BC [mag]	$L_\star$ [$L_\odot$]	$R_\star$ [$R_\odot$]	$M_\star$ [$M_\odot$]
3215507	5680	4,23	-1,03	4,81	4,84	0,023	0,92±0,02	0,99±0,03	0,86±0,06
3215583	5788	3,95	-0,19	3,09	3,17	0,075	4,26±0,09	2,06±0,05	1,20±0,08
3215587	5852	4,21	-0,37	4,30	4,36	0,051	1,42±0,03	1,16±0,03	1,05±0,07
3322673	5983	4,13	-0,29	3,26	3,32	0,054	3,70±0,08	1,80±0,05	1,14±0,07
3322684	5605	4,33	-0,63	5,32	5,36	0,034	0,57±0,01	0,80±0,02	0,89±0,06
3322688	5642	4,13	-0,40	4,72	4,77	0,049	0,97±0,02	1,03±0,03	1,02±0,07
3322715	5447	4,41	-0,42	5,39	5,43	0,034	0,53±0,01	0,82±0,02	0,88±0,06
3322737	5337	4,38	-0,24	5,24	5,27	0,024	0,62±0,01	0,92±0,03	0,89±0,06
3322764	5829	4,19	-0,31	4,34	4,40	0,051	1,37±0,03	1,15±0,03	1,07±0,07
3322775	6012	4,07	-0,46	3,58	3,64	0,054	2,76±0,06	1,54±0,04	1,13±0,07
3322797	5763	4,23	-0,41	4,51	4,57	0,051	1,17±0,02	1,09±0,03	1,01±0,06
3322805	5545	4,46	-0,64	5,64	5,68	0,034	0,42±0,01	0,71±0,02	0,85±0,05
3322837	6355	4,15	-0,97	3,95	4,00	0,047	1,97±0,04	1,16±0,03	1,05±0,07
3322860	5357	4,52	-0,16	5,57	5,59	0,023	0,46±0,01	0,79±0,02	0,89±0,06
3322872	6169	4,12	-0,18	3,65	3,75	0,099	2,48±0,05	1,38±0,03	1,23±0,08
3322889	5731	4,29	-0,28	4,61	4,66	0,046	1,08±0,02	1,06±0,03	1,01±0,06
3322953	5035	3,17	-0,97	1,14	1,12	-0,015	27,93±0,57	6,96±0,21	1,30±0,08
					...				

Continua em: https://github.com/lethyciacarvalho/Kepler-stars

O restante da Tabela A.1 pode ser acessada online no endereço eletrônico disponível no rodapé da mesma tabela. Nele, também está a tabela com a caracterização feita pelo modelo menos restrito deste trabalho. Em ambos os casos, algumas estrelas só apresentam $T_{\rm ef}$ e [Fe/H], além do ID da estrela (125 objetos para o modelo mais restrito e 177 objetos para o modelo menos restrito). Os demais dados não foram possíveis de obter por causa da ausência de dados de distância Gaia. Todos os parâmetros posteriores se relacionam à ela direta ou indiretamente. Conhecer a distância do objeto é de suma importância para a previsão do seu log g, bem como para calcular M_G. O log g, por sua vez, é necessário para obter o valor de BC. Consequentemente, sem BC, não podemos calcular a M_{bol} e sem M_{bol} não obtemos $L_\star$, que por sua vez é importante para o cálculo de $R_\star$. Sem log g também não podemos obter $M_\star$.