Observatório Nacional

DISSERTAÇÃO DE MESTRADO

IDENTIFICAÇÃO DE SUBANÃS QUENTES EM LEVANTAMENTOS ASTRONÔMICOS

MARCOS VINICIUS EMANUEL CORDEIRO DA SILVA

RIO DE JANEIRO
2023

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

Dissertação de Mestrado

IDENTIFICAÇÃO DE SUBANÃS QUENTES EM LEVANTAMENTOS ASTRONÔMICOS

por

Marcos Vinicius Emanuel Cordeiro da Silva

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 Fernantes

Rio de Janeiro, RJ – Brasil
Maio de 2023

C794 Cordeiro da Silva, Marcos Vinicius Emanuel
Identificação de Subanãs Quentes em Levantamentos Astronômicos [Rio de Janeiro] 2023.
xx, 113 p. 29,7 cm: graf. il. tab.

Dissertação (mestrado) - Observatório Nacional - Rio de Janeiro, 2023.

1. Florestas Aleatórias. 2. Subanãs Quentes. 3. Previsão de Parâmetros Estelares. 4. J-PLUS. 5. S-PLUS. I. Observatório Nacional. II. Título.

CDU 000.000.000

"IDENTIFICAÇÃO DE SUBANÃS QUENTES EM LEVANTAMENTOS ASTRONÔMICOS"

MARCOS VINICIUS EMANUEL CORDEIRO DA SILVA

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 Fernantes – Observatório Nacional (Orientador)

Dr. Francisco Ferreira Souza Maia – UFRJ

Dr. Marcos Perez Diaz – IAG/USP

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2023

Agradecimentos

Agradeço primeiramente ao meu orientador, Dr. Marcelo Borges Fernandes, pelo apoio e compreensão que ao longo desse mestrado foram essenciais para o desenvolvimento e finalização desse trabalho. Agradeço também pela sua disponibilidade em responder às minhas dúvidas e por compartilhar conhecimentos valiosos.

Agradeço à minha noiva Mayra, que sempre esteve ao meu lado em todos os momentos difíceis desse processo, me dando força para continuar. Agradeço também por me incentivar e celebrar comigo todas as conquistas que vieram durante o desenvolvimento desse trabalho. Sem você nada disso seria possível.

Agradeço aos meus pais e minha família, por todo o incentivo e por possibilitarem que eu seguisse o caminho que me trouxe até aqui.

Agradeço aos amigos que fiz no ON, especialmente à Lethycia Carvalho, que foi minha companheira de trabalho desde o início do desenvolvimento dessa dissertação.

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

Marcos Vinicius Emanuel Cordeiro da Silva

IDENTIFICAÇÃO DE SUBANÃS QUENTES EM LEVANTAMENTOS ASTRONÔMICOS

RESUMO

Apesar de terem sido descritas pela primeira vez no final da década de 70 as estrelas subanãs quentes ainda não são completamente compreendidas, principalmente no que diz respeito às condições necessárias e aos mecanismos responsáveis por sua formação. Um dos maiores impedimentos desse campo está relacionado com a pouca quantidade de objetos catalogados e confirmados, o que dificulta o desenvolvimento e confirmação de possíveis hipóteses em relação a dados observados. Apesar disso, as subanãs quentes são importantes para diversos campos da astronomia e impactam estudos que vão desde a evolução estelar, até as propriedades físicas da nossa e de outras galáxias. Assim, esse trabalho tem como objetivo principal o desenvolvimento de soluções voltadas à identificar novas candidatas à subanãs quentes e com isso expandir os catálogos atuais dessa classe de objetos. Para esse fim são utilizados algoritmos de Aprendizado de Máquina (*Machine Learning*, ou ML), capazes de processar uma grande quantidade de dados de maneira eficiente. Os modelos criados neste trabalho são baseados em algoritmos de *Random Forest* e os dados utilizados nesse processo são as 12 magnitudes fotométricas fornecidas pelo J-PLUS e o S-PLUS. Após cruzar esses dois levantamentos com um catálogo de subanãs quentes confirmadas, foram geradas amostras com cerca de 17 mil objetos (entre estrelas gerais e subanãs quentes) para cada um deles. Com uma cuidadosa otimização de seus hiperparâmetros, os modelos criados obtiveram ótimos resultados nas amostras de teste (Score F1 de 0,88 no J-PLUS e de 0,94 no S-PLUS), e a partir deles foi possível identificar 2896 novas candidatas à subanãs quentes dentro dos dois levantamentos considerados. Além disso, como forma de validação dos algoritmos utilizados, foram desenvolvidos também modelos de previsão de parâmetros estelares (T_{eff}, log(g) e Fe/H) baseados em *Random Forest* para objetos do J-PLUS e do S-PLUS. As amostras de desenvolvimento desses modelos foram criadas a partir do cruzamento dos dois levantamentos com dados de parâmetros estelares do LAMOST, resultando em cerca de 211 mil objetos para o J-PLUS e cerca de 66 mil para o S-PLUS. Para essa previsão, testou-se o uso das magnitudes absolutas como variáveis de entrada dos modelos, o que trouxe uma melhora de $\sim 30\%$ nas previsões de log(g) dos objetos. Após também realizar uma otimização dos hiperparâmetros, os modelos obtidos neste trabalho demonstraram uma boa performance nas amostras de teste, com erros de $\sim 50K$ para T_{eff}, $\sim 0,07$ dex para log(g) e $\sim 0,08$ dex para [Fe/H]. Por fim, a partir desses estimadores foram criados catálogos de parâmetros estelares para o J-PLUS e o S-PLUS com 3 milhões e 5 milhões de objetos respectivamente.

Marcos Vinicius Emanuel Cordeiro da Silva

HOT SUBDWARF IDENTIFICATION IN ASTRONOMICAL SURVEYS

ABSTRACT

Despite being first described in the late 1970s, hot subdwarf stars are still not fully understood, especially when it comes to the necessary conditions and mechanisms responsible for their formation. One of the biggest problems in this field is related to the small number of catalogued and confirmed objects, which hinders the development and confirmation of possible hypotheses in relation to observed data. Despite this, hot subdwarf are important for several fields of astronomy and they impact studies ranging from stellar evolution to the physical properties of our own and other galaxies. Thus, the main objective of this study is the development of solutions aimed at identifying new candidates for hot subdwarfs and thus expand the current catalogues. To this end, Machine Learning (ML) algorithms are used, which are capable of processing a large amount of data in an efficient way. The models created in this work are based on Random Forest algorithms and the data used are the 12 photometric magnitudes provided by J-PLUS and S-PLUS. After crossmatching these two surveys with a catalogue of confirmed hot subdwarfs, samples with about 17,000 objects (between general stars and hot subdwarfs) were generated for each of them. With a careful optimization of their hyperparameters, the models created obtained excellent results in the test samples (F1 Score of 0.88 in J-PLUS and 0.94 in S-PLUS), and they were used to identify 2896 new candidates for hot subdwarfs within the two surveys considered. Furthermore, as a way of validating the algorithms used, stellar parameter (T_{eff}, log(g) and Fe/H) prediction models based on Random Forest were also developed for J-PLUS and S-PLUS objects. The development samples for these models were created from crossmatching the two surveys with LAMOST stellar parameter data, resulting in about 211 thousand objects for J-PLUS and about 66 thousand for S-PLUS. For these predictions, the use of absolute magnitudes as input variables for the models was tested, which brought an improvement of about $\sim 30\%$ in the log(g) predictions. After also performing an optimization of the hyperparameters, the models obtained in this work showed a good performance in the test samples, with errors of $\sim 50K$ for T_{eff}, ~ 0.07 dex for log(g) and ~ 0.08 dex for [Fe/H]. Finally, the models were used to create stellar parameter catalogues for J-PLUS and S-PLUS with 3 million and 5 million objects respectively.

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

Neste trabalho de mestrado pretendemos estudar uma classe de estrelas ainda pouco conhecidas e com poucos exemplos conhecidos e catalogados. Esses objetos são denominados de subanãs quentes, e um melhor entendimento de sua estrutura e mecanismos de formação pode trazer avanços para diversos campos da astronomia.

1.1 Subanãs Quentes

As estrelas subanãs quentes foram descritas pela primeira vez no final da década de 60 por SARGENT e SEARLE (1968) após realizarem observações espectrais de uma lista de estrelas azuis em altas latitudes galáticas compilada por FEIGE (1958). Inicialmente, as características que definiam esses objetos (que ocupam regiões do diagrama HR correspondentes aos tipos espectrais O e B) eram relacionadas às linhas de Balmer em seus espectros. No caso das subanãs quentes de tipo espectral B (sdBs), as linhas de Balmer são mais largas do que as linhas de estrelas da Sequência Principal (SP) de mesmo tipo espectral, enquanto que no caso das subanãs quentes do tipo O (sdOs), essas linhas são mais intensas do que as linhas de uma estrela na SP (SARGENT e SEARLE, 1968).

1.1.1 Estrutura Interna

Essa peculiaridade no espectro das sdOs e sdBs está diretamente correlacionada com sua estrutura, que consiste em um núcleo onde ocorre a queima de hélio rodeado por uma camada extremamente fina de hidrogênio ($< 0,01 M_{\odot}$), geralmente representando apenas uma fração da massa total da estrela ($\sim 0,5 M_{\odot}$) (HEBER, 1986). Para que essa estrutura se forme, é necessário que algum processo de perda de massa ocorra durante a evolução do objeto (HAN *et al.*, 2002).

Devido à perda de massa durante sua evolução as camadas mais internas (e mais quentes) desses objetos são expostas, de modo que eles têm temperaturas efetivas mais altas ($25000K < T_{eff} < 65000K$) do que as estrelas da sequência principal de mesma

classe espectral (HEBER, 1986). Além disso, a pouca massa de hidrogênio ao redor do núcleo impossibilita que ocorra queima nessa casca, de modo que as subanãs quentes são objetos mais compactos ($4,5 < log(g) < 6,5$) que as estrelas que seguem o caminho evolutivo comum (onde ocorre queima de hidrogênio na casca) (HEBER, 2016).

Em relação à sua posição no diagrama Hertzsprung-Russell (HR), as estrelas subanãs quentes se encontram na região entre a sequência principal e as anãs brancas, como se pode observar na Figura 1.1, onde estão indicadas as zonas que concentram as sdOs e as sdBs.

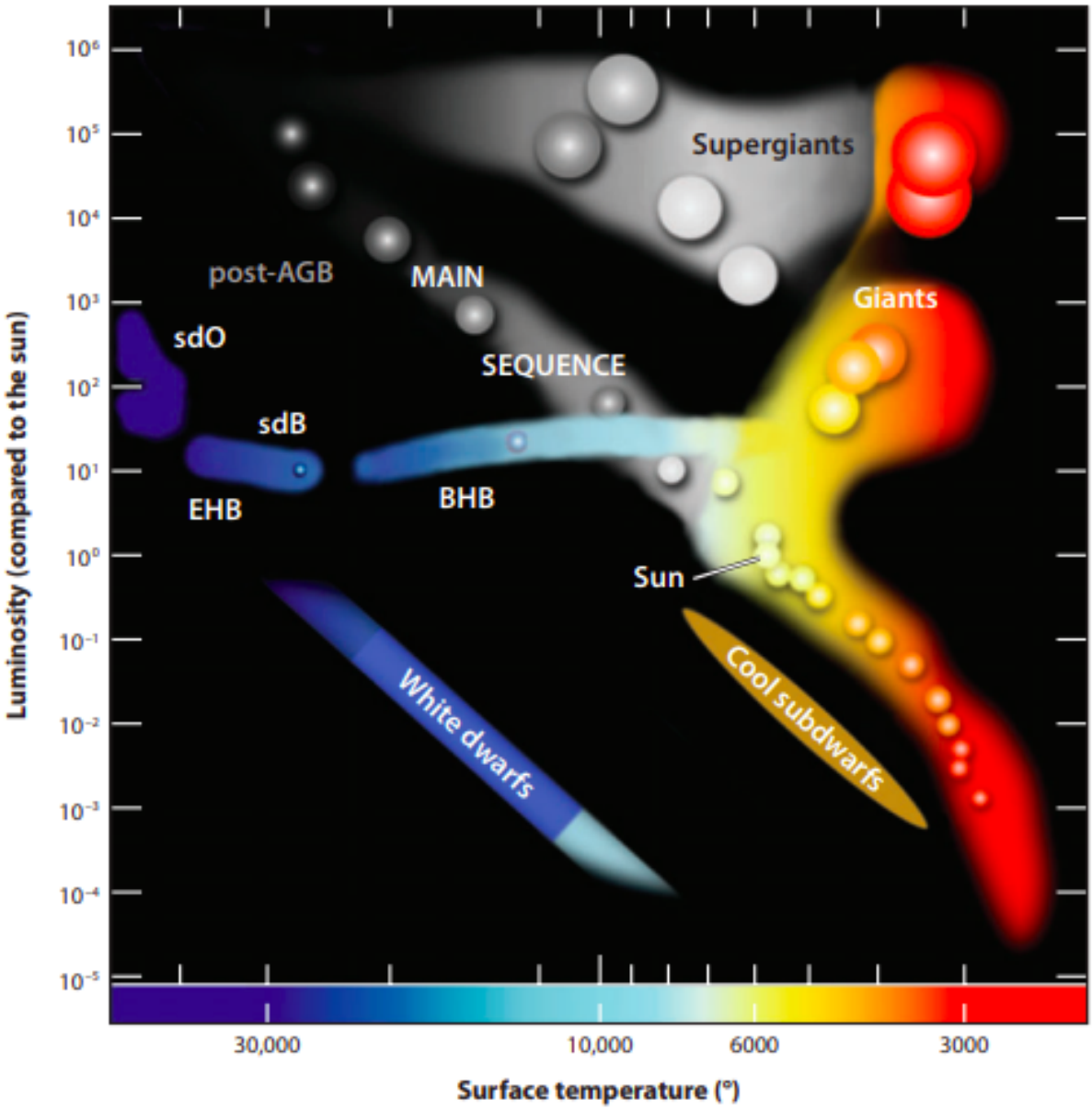

Figura 1.1: Diagrama Hertzsprung-Russell (HR) com destaque para as áreas de concentração das estrelas subanãs quentes sdOs e sdBs. Fonte: HEBER (2009)

Na Figura 1.1 também fica clara a diferença entre as subanãs quentes e as estrelas que seguem o caminho evolutivo canônico. Geralmente, ao passar pela fase de gigante

vermelha, o caminho evolutivo de uma estrela comum a leva para a parte azul do Ramo Horizontal (BHB). Enquanto isso, devido à sua temperatura efetiva mais alta, as subanãs quentes estão localizadas mais à esquerda no diagrama HR (o que significa que elas são ainda mais azuis).

1.1.2 Mecanismos de Formação

Por não serem fruto do processo canônico de evolução estelar, as sdOs e sdBs são consideradas objetos exóticos e os mecanismos de formação por trás delas são foco de uma grande quantidade de estudos (HAN *et al.* 2002; HAN *et al.* 2003; MILLER BERTOLAMI *et al.* 2008; JUSTHAM *et al.* 2011; ZHANG e JEFFERY 2012; PELISOLI *et al.* 2020).

1.1.2.1 Transferência de Massa

Um dos possíveis caminhos de formação das subanãs quentes se baseia em uma estrela que segue a rota evolutiva canônica até o topo do ramo das gigantes vermelhas, mas que durante essa evolução perde uma grande quantidade de massa de suas camadas mais externas (HAN *et al.*, 2002).

Após chegar no topo do ramo das gigantes vermelhas, as estrelas sofrem o flash do hélio e se movem para baixo e para a esquerda no diagrama HR, até chegarem no ramo horizontal (CATELAN, 2007). A posição inicial do objeto nesse ramo está diretamente relacionada com a massa da sua camada convectiva de hidrogênio no topo do RGB, com camadas menos massivas implicando em posições mais para a parte azul do Ramo Horizontal (CATELAN, 2007). No caso das subanãs, a perda de massa é tão extrema que elas ficam ainda mais para o azul do que o ramo horizontal canônico, numa região conhecida como ramo horizontal extremo, como pode se observar na Figura 1.1.

Além disso, essa perda de massa impossibilita que haja a queima de hidrogênio na camada ao redor do núcleo, o que por sua vez impede a estrela de continuar na rota evolutiva em direção ao Ramo Assintótico das Gigantes (AGB), de modo que as subanãs evoluem diretamente para a fase de anã branca (HEBER, 2016).

Apesar de existir esse consenso sobre a perda de massa durante a evolução pelo RGB ser responsável pelo nascimento de uma subanã quente, ao longo dos anos foram propostos diversos mecanismos possíveis para explicar como essa perda de massa pode ocorrer (HAN *et al.* 2002; HAN *et al.* 2003; HEBER 2016). A maioria desses mecanismos envolve interações binárias da estrela gigante vermelha com uma companheira, e isso implicaria diretamente no fato de uma fração significativa das subanãs quentes conhecidas estarem em sistemas binários (KUPFER *et al.* 2015; PELISOLI *et al.* 2020).

Além disso, alguns outros mecanismos de evolução a partir de estrelas simples já foram propostos, envolvendo processos internos como a perda de massa através de ventos estelares extremos (D'CRUZ *et al.*, 1996). Contudo, através da análise de um conjunto

de sistemas binários de longa duração com subanãs quentes conhecidas, PELISOLI *et al.* (2020) propuseram que subanãs quentes são fruto exclusivamente de interações binárias.

Essa hipótese baseou-se no fato de que, considerando que fosse possível uma estrela simples evoluir para uma subanã quente sem influência externa, ao menos uma fração das subanãs quentes em sistemas binários de longa duração deveria ter evoluído dessa maneira, e essas não teriam nenhum indício de interação com sua companheira no que diz respeito à sua rotação. No entanto, mais de 60% das subanãs na amostra de 123 sistemas considerados demonstraram evidências de interações em seu passado, e quando se leva em conta apenas os sistemas longe de zonas estelares densas (que podem dificultar a detecção dessas interações), essa fração chega a 78% para os 26 com magnitude G <= 13,5 e 100% para os 6 com magnitude G <= 12,0. Isso pode ser considerado um indício observacional a favor da necessidade de interações binárias para a formação de subanãs quentes (PELISOLI *et al.*, 2020).

Em relação à sua classificação, essas interações binárias progenitoras de subanãs quentes podem ser divididas em dois grandes grupos, onde cada um deles se baseia num mecanismo diferente de transferência de massa (HAN *et al.*, 2002):

- **Envelope Comum (CE):** Nesse caminho evolutivo, o objeto mais massivo do sistema binário inicial evolui ao longo do ramo das gigantes vermelhas e durante esse processo preenche totalmente seu lóbulo de Roche, resultando em uma transferência de massa instável para o outro objeto e a formação de uma envoltória comum englobando tanto a estrela gigante quanto sua companheira. Após essa etapa, as duas componentes no interior da envoltória passam a transferir energia orbital para ela, de modo que o período do sistema diminui. Além disso, a perturbação que as duas componentes causam na envoltória pode fazer com ela eventualmente seja ejetada, deixando para trás um sistema que consiste no núcleo degenerado da gigante vermelha (que posteriormente dá origem à subanã quente) e uma companheira (HAN *et al.*, 2002);

- **Preenchimento Estável do Lóbulo de Roche (*Roche Lobe Overflow* - RLOF):** Similar ao primeiro caminho, a estrela de maior massa nesse caso também evolui pelo ramo das gigantes vermelhas e preenche seu lóbulo. No entanto, aqui a transferência de massa para a companheira é estável, de modo que as camadas externas da gigantes são removidas mais lentamente, e uma envoltória comum nunca é formado. Nesse caso não haverá encolhimento das órbitas, e o sistema resultante tem períodos mais longos (HAN *et al.*, 2002).

De acordo com a Figura 1.2, que ilustra os dois mecanismos citados acima, ambos resultam em sistemas binários, onde a subanã formada é acompanhada por um outro objeto. No entanto, cerca de um terço das estrelas subanãs de campo observadas não

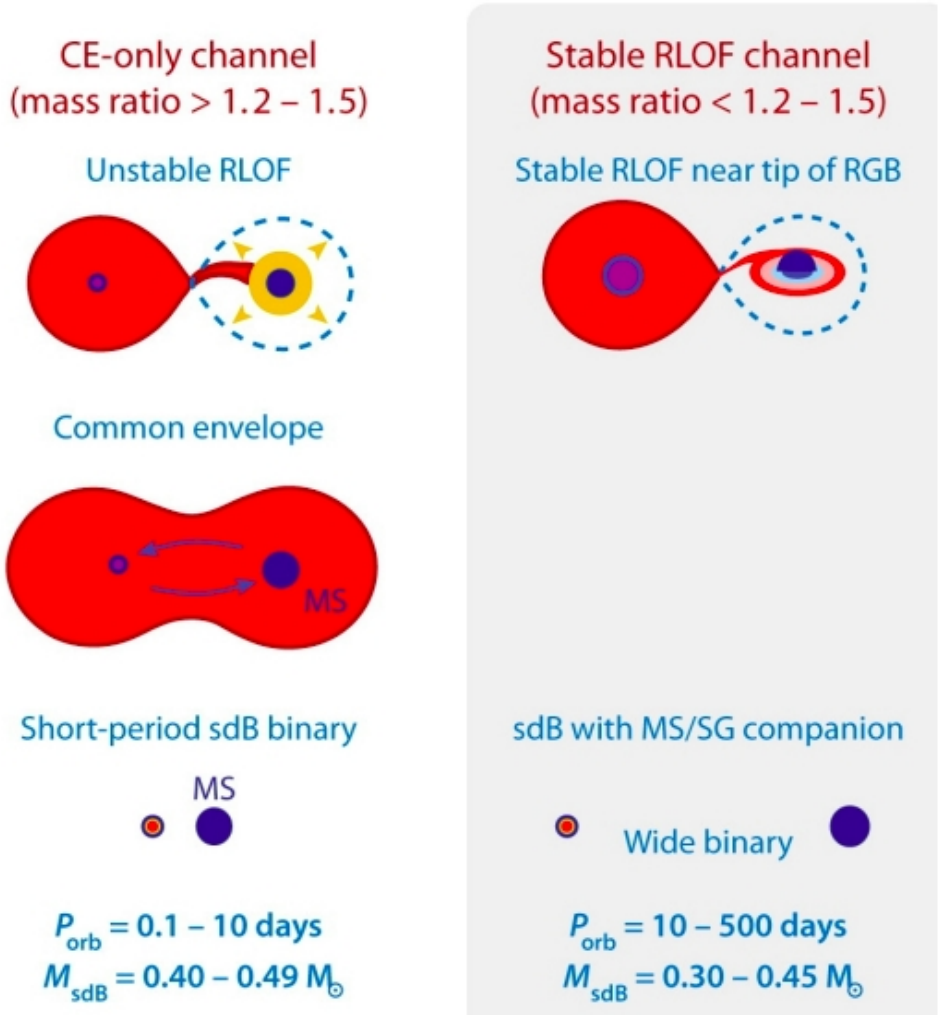

Figura 1.2: Mecanismos de formação de subanãs quentes por transferência de massa em um sistema binário. Fonte: Adaptado de HEBER (2009).

mostram nenhum sinal de binaridade, enquanto que em aglomerados essa fração pode ser ainda maior (LATOUR *et al.*, 2018). Para explicar a formação desses sistemas sem companheiros, um outro tipo de mecanismo é necessário.

1.1.2.2 Por Fusão

A proposição de uma fusão entre duas anãs brancas de hélio (HeWD) ser capaz de gerar uma subanã quente foi feita inicialmente por WEBBINK (1984) em um trabalho que teorizava a possibilidade de existência desse tipo de sistema binário e explorava os possíveis produtos gerados por eles.

A partir disso, e com a evolução dos algoritmos computacionais de evolução estelar, nos últimos anos diversos modelos vêm sendo simulados para explicar melhor a fusão num sistema binário de anãs brancas e esclarecer a possibilidade de esse fenômeno ser o progenitor da população de subanãs quentes sem companheiros (ZHANG e JEFFERY 2012; SCHWAB 2018).

No início dos anos 2010, ZHANG e JEFFERY (2012) modelaram três mecanismos diferentes de fusão e compararam seus resultados com dados observacionais de uma amostra de subanãs quentes com temperaturas efetivas, log(g) e abundâncias de nitrogênio e carbono para concluir qual deles seria o melhor candidato.

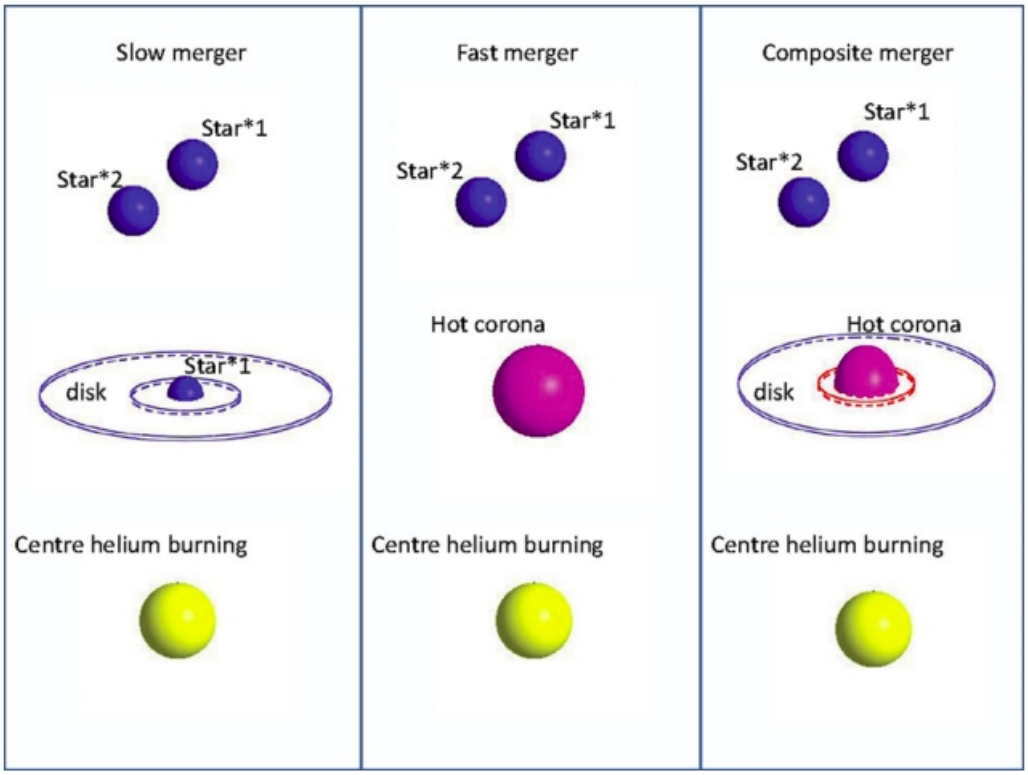

Figura 1.3: Mecanismos de formação de subanãs quentes por fusão de anãs brancas. Fonte: ZHANG e JEFFERY (2012)

Como se pode observar na Figura 1.3, todos os três mecanismos envolvem uma anã branca mais massiva (Star *1, ou primária) e uma uma menos massiva (Star *2, ou secundária), que iniciam o processo em uma órbita próxima e ao longo do tempo se aproximam devido à perda de energia na forma de radiação gravitacional (ZHANG e JEFFERY, 2012). Após um certo ponto de aproximação começa a ocorrer a transferência de massa entre as duas anãs brancas, o que faz com que o objeto secundário seja consumido pelo primário. Em todos esses casos a formação de uma subanã quente depende da massa resultante do objeto final, pois se ele ultrapassar o limite de Chandrasekhar $(1,4M_{\odot})$, o resultado da fusão possivelmente será uma Supernova Ia (TOONEN *et al.* 2012; MOLL *et al.* 2014).

No primeiro mecanismo proposto a massa transferida pela estrela secundária forma um disco ao redor da primária, que é então acrescido ao longo de milhões de anos. Pela escala de tempo necessária para a acreção esse mecanismo leva o nome de fusão lenta (ZHANG e JEFFERY, 2012).

No segundo mecanismo, ou fusão rápida, a massa transferida é depositada diretamente

na superfície da estrela primária e aquecida a temperaturas da ordem de 10^8 K. Essas altas temperaturas fazem com que o material da estrela secundária se expanda e forme uma coroa quente ao redor da primária. Como esse mecanismo não envolve um processo de acreção de longa duração como o primeiro, ele recebe o nome de fusão rápida.

Por fim, o terceiro mecanismo proposto é denominado de fusão composta e consiste da ocorrência dos dois mecanismos anteriores simultaneamente. Nesse caso, uma fração ($\sim 50\%$) da massa transferida forma um disco de acreção ao redor da estrela primária, enquanto que o restante da massa é depositado em sua superfície e forma a coroa de alta temperatura.

Utilizando os três mecanismos de fusão propostos, foi possível simular a evolução dos objetos resultantes pós-fusão para diferentes massas iniciais de anãs brancas ($0,25M_{\odot}, 0,30M_{\odot}, 0,35M_{\odot} e 0,40M_{\odot}$) e chegar nas seguintes conclusões em relação às propriedades desses objetos (ZHANG e JEFFERY, 2012):

- **T_{eff} e log(g):** Para todos os três casos, as estrelas resultantes evoluem para regiões do diagrama de T_{eff} - log(g) dominadas por subanãs quentes ricas em hélio.
- **Abundâncias Superficiais:** Enquanto que na fusão lenta as estrelas geradas foram todas ricas em nitrogênio, no caso da fusão rápida as estrelas resultantes foram todas ricas em carbono. Esse comportamento está relacionado com o fato dos objetos formados pela fusão lenta preservarem a composição superficial rica em nitrogênio das anãs originais, enquanto que os objetos formados pela fusão rápida possuem uma camada convectiva mais profunda e capaz de transportar carbono para sua superfície. Por fim, considerando que a fusão composta é uma junção dos dois métodos, ela foi capaz de gerar tanto estrelas ricas em N quanto em C.

Além disso, ao analisar mais profundamente os resultados das fusões compostas para as diversas massas iniciais consideradas, ZHANG e JEFFERY (2012) foram capazes de identificar uma correlação direta entre a massa final do sistema e as abundâncias superficiais: geralmente, sistemas com massas finais menores do que $0,7M_{\odot}$ produzem subanãs quentes ricas em nitrogênio, enquanto que sistemas com massas finais maiores do que $0,7M_{\odot}$ produzem subanãs quentes ricas em carbono (mas com uma abundância considerável de nitrogênio).

Como existem evidências observacionais para a existência desses dois tipos de subanãs quentes (ricas em C e ricas em N), os mecanismos simples de fusão não são capazes de explicar toda a população desses objetos. Com isso, o mecanismo da fusão composta se mostra o mais capaz de reproduzir as distribuições de parâmetros observados nas subanãs quentes (YU *et al.*, 2021).

Alguns anos depois, em meados dos anos 2010, HALL e JEFFERY (2016) expandiram os modelos desenvolvidos até então e passaram a considerar explicitamente a presença de

hidrogênio durante suas simulações. Nos modelos anteriores o H era ignorado baseando-se na suposição de que as altas temperaturas presentes no processo de fusão eram suficientes para transformar boa parte do H presente em He, resultando em subanãs com superfícies ricas em hélio (HALL e JEFFERY, 2016).

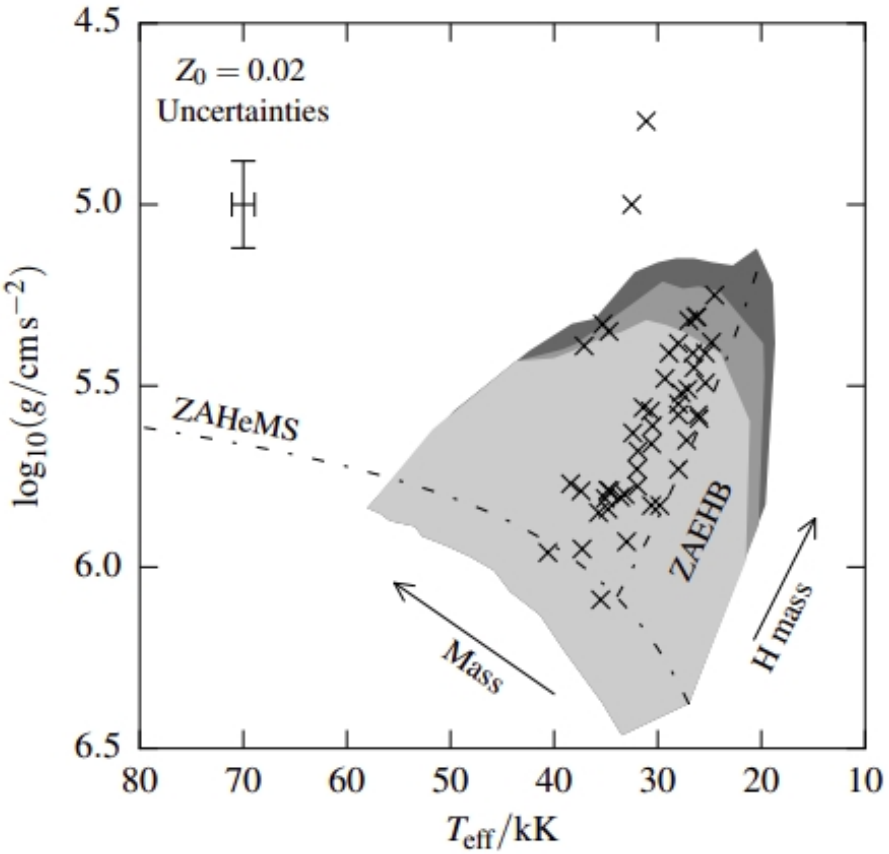

Figura 1.4: Zona do diagrama T_{eff} x log(g) onde as estrelas simuladas por HALL e JEFFERY (2016) se encontram. As três regiões em cinza indicam as áreas onde ocorre a queima de hélio no núcleo das estrelas resultantes de fusões de anãs brancas recém chegadas na curva de resfriamento (jovens, todas as três áreas cinza), 1 giga-ano depois de sua chegada na curva de resfriamento (idade média, área cinza médio + cinza claro) e 12 giga-anos depois de sua chegada na curva de resfriamento (velhas, área cinza claro). Também representadas no diagrama estão as 48 subanãs quentes com parâmetros atmosféricos conhecidos, indicadas acima com os símbolos X. Também indicadas no gráfico estão as linhas da *Zero Age Helium Main Sequence* (ZAHeMS, área no diagrama análoga à SP, mas para estrelas formadas por Hélio), e a *Zero Age Extreme Horizontal Branch* (ZAEHB, área no diagrama ocupada por estrelas imediatamente após sua chegada no ramo horizontal). Fonte: HALL e JEFFERY (2016)

Ao levar em consideração a massa de hidrogênio presente nas duas anãs brancas antes de fusão e obter estimativas para a massa de hidrogênio restante depois desse processo, HALL e JEFFERY (2016) foram capazes de propor um modelo de formação para as subanãs com superfícies ricas em Hidrogênio. Para validar esse modelo, as distribuições de T_{eff} e log(g) teóricas foram comparadas com as distribuições observadas em uma amostra

de 48 sdBs aparentemente isoladas cujos parâmetros atmosféricos foram calculados por HALL e JEFFERY (2016).

Como se pode observar na Figura 1.4, apenas duas subanãs da amostra conhecida não estão dentro das áreas onde teoricamente um objeto resultante do tipo de fusão proposto se posicionaria durante sua fase de queima de hélio no núcleo.

Nesses dois casos a gravidade superficial baixa das subanãs faz com que elas estejam numa área não coberta pelos modelos, o que pode indicar que elas foram formadas por um outro tipo de mecanismo, ou que as duas possuem companheiras ainda não detectados ou que elas ainda não estão na fase de queima de hélio em seu núcleo (HALL e JEFFERY, 2016).

Esse resultado mostrou que o mecanismo de fusão também é capaz, em teoria, de explicar a população de subanãs quentes com superfícies ricas em hidrogênio. No entanto, é importante ter em mente as diversas suposições à respeito da massa de hidrogênio nas anãs brancas progenitoras antes da fusão e de sua distribuição durante as fases iniciais, ainda não totalmente compreendidas, da fusão. Esses dois pontos trazem um grau de incerteza para a análise, e podem alterar consideravelmente as áreas obtidas no diagrama T_{eff} x log(g) (HALL e JEFFERY, 2016).

Além disso, o número pequeno de subanãs com parâmetros atmosféricos conhecidos diminui o impacto de qualquer análise comparativa entre distribuições teóricas e observadas. Para resolver essa questão, a observação e caracterização de mais objetos dessa classe é um passo extremamente necessário.

1.1.3 Importância

No contexto de sua importância, as estrelas subanãs quentes se destacam em diversos campos de estudo, e seu melhor entendimento pode trazer desenvolvimentos em diferentes áreas da astronomia. Como exemplos podemos citar:

- **Excesso de luz ultravioleta em galáxias elípticas**: A quantidade de luz UV emitida por galáxias elípticas, normalmente relacionada à estrelas azuis recém formadas, vai de encontro com o fato dessas galáxias serem formadas majoritariamente por estrelas vermelhas antigas, e não possuírem formação estelar suficiente para gerar a quantidade de luz ultravioleta observada. No entanto, esse excesso pode ser explicado pela presença de uma população de sdOs e sdBs, que devido à sua alta temperatura emitem consideravelmente no UV, e que por serem estágios avançados na evolução de estrelas de baixa massa, não implicam em formação estelar recente (PODSIADLOWSKI *et al.* 2008; HAN *et al.* 2010).

- **Cálculo da massa do halo de matéria escura**: No estudo das propriedades do halo de matéria escura da Via Láctea, uma das maneiras de se calcular sua massa

é através da observação e da análise de estrelas em hipervelocidade. Como algumas subanãs quentes podem alcançar altíssimas velocidades através de interações binárias em seus sistemas originais (GEIER *et al.*, 2015), elas são ótimas candidatas para aplicação nesse campo de estudo (TILLICH *et al.*, 2011).

- **Estudo da estrutura de núcleos de hélio**: Através da análise das séries temporais de subanãs quentes pulsantes, a astrosismologia permite o estudo do interior de seu núcleo de hélio bem como sua composição química detalhada (VAN GROOTEL *et al.* 2010; CHARPINET *et al.* 2011). Como a maioria das estrelas passa por uma fase de queima de hélio em seu núcleo durante a evolução canônica, esse tipo de entendimento pode expandir o estudo da estrutura e dos processos evolutivos de uma grande quantidade de estrelas.

- **Supernovas Tipo Ia**: Apesar de serem considerados ótimos indicadores de distância e poderem ser utilizadas como calibradores de cálculos desse tipo, os mecanismos por trás da origem das supernovas tipo Ia ainda não são completamente conhecidos. O consenso atual na astronomia é de que esses eventos ocorrem devido à explosão termonuclear de uma anã branca quando ela ultrapassa uma massa crítica de $1,4 M_{\odot}$ (limite de Chandrasekhar), mas para que isso ocorra ela precisa acretar massa ou passar por uma fusão com um outro objeto. Nesse sentido, PELISOLI *et al.* (2021) propuseram um possível progenitor de uma supernova tipo Ia através da determinação da massa do sistema binário HD 265435 formado por uma anã branca e uma subanã quente, que irão se fundir em cerca de 70 milhões de anos e resultar num objeto com massa superior à massa crítica.

- **Binárias de Verificação para Levantamentos**: Planejado como um dos equipamentos mais promissores da nova geração de observatórios astronômicos, o *The Laser Interferometer Space Antenna* (LISA) visa operar no intervalo das ondas gravitacionais de baixa frequência ($10^{-4} - 1$ Hz) (AMARO-SEOANE *et al.*, 2017). Nessa faixa de frequência, as observações são dominadas por objetos ultracompactos em órbitas binárias curtas (da ordem de horas) que dificilmente seriam observados por equipamentos em operação atualmente. Com isso, existem esforços sendo realizados para detectar sistemas binários capazes de serem observados pelo LISA (denominados 'binárias de verificação'), para que seu funcionamento e operação sejam validades e calibrados no início da operação. Assim, devido à sua estrutura compacta e processos de formação que dependem de interações binárias, espera-se que sistemas com subanãs quentes sejam possíveis candidatos à binários de verificação. Pelo menos um sistema binário formado por uma subanã quente e uma anã branca já foi confirmado para essa aplicação (KUPFER *et al.*, 2018).

Tendo em mente as aplicações aqui citadas, além de diversas outras ainda em desen-

volvimento, fica claro o potencial de ganho oriundo do estudo das propriedades e dos mecanismos de evolução das subanãs quentes, que são dois campos ainda não totalmente compreendidos (HEBER, 2016). Portanto, aumentar a amostra de subanãs conhecidas e determinar os seus parâmetros é fundamental.

Considerando o contexto descrito acima, esse trabalho tem como objetivo a utilização de algoritmos de Aprendizado de Máquina (ou *Machine Learning*, ML) para criar modelos capazes de identificar objetos candidatos à serem subanãs quentes dentro de grandes levantamentos (mais especificamente o J-PLUS e o S-PLUS).

Em relação à estrutura desse trabalho, no capítulo 2 estão explanados todos os métodos de ML aplicados durante o desenvolvimentos desses modelos, bem como a estrutura proposta para gerá-los. No capítulo 3 são descritos os levantamentos astronômicos e catálogos utilizados para criar as amostras de desenvolvimento dos modelos, e o capítulo 4 traz tanto os resultados obtidos com relação aos modelos, quanto uma discussão detalhada a respeito deles. Por fim, no capítulo 5 são apresentadas as conclusões do trabalho e as perspectivas futuras com relação à expansão e aprofundamento dos resultados obtidos.

Capítulo 2

Metodologia

Nesse capítulo serão descritos os métodos de Aprendizado de Máquina utilizados durante o desenvolvimento dos modelos de classificação de subanãs quentes. Apresentaremos uma breve introdução a respeito dos algoritmos de Aprendizado de Máquina, depois uma seção a respeito dos algoritmos de *Random Forests* (Florestas Aleatórias), escolhidos como base desse trabalho, e uma seção sobre os hiperparâmetros desses algoritmos e os métodos utilizados para sua otimização.

Por último, o capítulo também conta com uma seção que propõe descrever a estrutura final das Florestas Aleatórias de Classificação de Objetos (FlACOs), que serão utilizadas para treinar e validar os modelos de classificação nos próximos capítulos.

2.1 Aprendizado de Máquina

O Aprendizado de Máquina (Machine Learning ou ML) é uma área da computação que busca desenvolver algoritmos e modelos capazes de gerar novas informações e entendimentos a partir de um conjunto de dados inicial, sem que haja uma programação explícita para tal. Essa geração de informação se baseia no reconhecimento de padrões e tendências existentes dentro dos dados, que podem ser utilizados por modelos de ML para realizar previsões.

Nos últimos anos, o aprendizado de máquina vem sendo aplicado em uma ampla gama de problemas, como reconhecimento de fala, extração de texto a partir de imagens, previsão de séries temporais, classificação de imagens e recomendações de produtos em lojas online, entre outros. Em todos esses casos, um dos grandes diferenciais dos modelos de ML é a sua capacidade de lidar com grandes quantidades de dados de maneira eficiente. Com o grande aumento da capacidade computacional disponível e a disponibilização de recursos de fácil acesso para o desenvolvimento de modelos ML, o número de aplicações possível desse tipo de solução tende a crescer muito no futuro.

De acordo com o método utilizado para desenvolver (ou treinar) um modelo de ML, eles podem ser divididos em dois grandes grupos:

- **Treinamento Supervisionado:** O modelo desenvolvido para prever uma certa propriedade é treinado a partir de um conjunto de exemplos dos quais já se conhece o valor da propriedade (exemplos rotulados);
- **Treinamento Não Supervisionado:** O modelo é desenvolvido para detectar padrões dentro de dados sem rótulo, e a partir desses padrões gerar classificações.

Dentro das classes definidas acima existem diversos algoritmos capazes de gerar modelos ML, cada um com suas próprias vantagens e desvantagens, que geralmente são consideradas a depender do problema que se procura resolver com o modelo. Alguns exemplos de algoritmos de treinamento supervisionado (foco desse trabalho) mais utilizados atualmente são:

- **Regressão Linear:** É um dos algoritmos mais simples disponível, e por isso também um dos mais populares. Se baseia em encontrar relações lineares entre a variável que se quer prever e as variáveis que se têm disponíveis para realizar a previsão;
- **Árvores de Decisão:** É um algoritmo que se baseia em uma sequência de regras que levam a uma previsão final. Pode ser utilizado tanto para classificação (previsão de um alvo binário) quanto para regressão (previsão de um alvo contínuo). Capaz de detectar padrões não-lineares dentro de um conjunto de dados, e com isso gerar modelos mais complexos do que a regressão linear;
- **Redes Neurais:** Algoritmo baseado no funcionamento do cérebro humano, que consiste em desenvolver uma rede de "neurônios"que interagem entre si para gerar uma previsão. Geralmente utilizado em problemas que demandam um grau de complexidade impossível de ser atingido por outros algoritmos mais simples, como o reconhecimento de fala e a análise de imagens;
- **Random Forest:** Algoritmo que considera os resultados de um conjunto de árvores de decisão para realizar sua previsão. Geralmente performa melhor do que uma única árvore, pois uma previsão equivocada de um membro do conjunto não prejudica a previsão final do modelo.

Com isso, a próxima seção traz uma descrição detalhada do funcionamento e da metodologia por trás do treinamento do Random Forest, que é o algoritmo base dos modelos utilizados neste trabalho.

2.2 Random Forest

O algoritmo de geração de modelos de aprendizado de máquina conhecido como Random Forest foi proposto de maneira definitiva em 2001 (BREIMAN, 2001), como a junção de di-

versas propostas anteriores focadas na introdução de aleatoriedade durante o treinamento de árvores de decisão.

2.2.1 Árvores de Decisão

Como o nome sugere, a peça principal de um modelo Random Forest são árvores de decisão. Por si só, cada uma das árvores já é um modelo de aprendizado de máquina, capaz de fazer previsões a partir de um conjunto de dados de entrada através de um conjunto de regras. No entanto, antes de tratar especificamente do funcionamento das árvores de decisão, é importante realizar algumas definições, que abaixo são exemplificadas com base em um problema de previsão da temperatura efetiva de uma estrela a partir de suas magnitudes:

- **Amostra:** Conjunto de dados utilizados pelo algoritmo de ML para gerar um modelo. No exemplo considerado, a amostra seria a base com as magnitudes de todas as estrelas e suas temperaturas efetivas correspondentes;
- **Objeto:** Membros que formam a amostra, dos quais o modelo quer prever uma certa propriedade. No exemplo considerado os objetos seriam as estrelas;
- **Variável:** Propriedades conhecidas dos objetos a partir das quais o modelo faz suas previsões. No exemplo considerado as variáveis são as magnitudes;
- **Alvo:** Propriedade que o modelo procura prever. No exemplo considerado, o alvo é a temperatura efetiva da estrela.

Numa árvore de decisão, as regras são etapas que consideram o valor X_A de uma variável A de um certo objeto e, a partir dele, enviam esse objeto para um caminho dentro do fluxo da árvore. Para decidir o caminho que será tomado, cada regra $\theta = (A, L)$ é definida a partir de uma variável A e um valor limite L:

$$\begin{cases} \text{Se } X_A < L, \text{ caminho direito} \\ \text{Se } X_A > L, \text{ caminho esquerdo} \end{cases} \tag{2.1}$$

Quanto aos caminhos possíveis dentro de uma árvore de decisão, esse objeto pode tanto ser enviado para uma outra regra, quanto receber uma previsão final. Sendo assim, é possível representar essa sequência de regras e previsões como na Figura 2.1, que se assemelha bastante a uma árvore, o que dá nome ao modelo.

Nesse contexto, as etapas dentro de uma árvore de decisão são chamadas de nós, e estão divididas em três tipos:

1. **Nó raiz:** Primeiro nó da árvore de decisão, onde o fluxo se inicia;

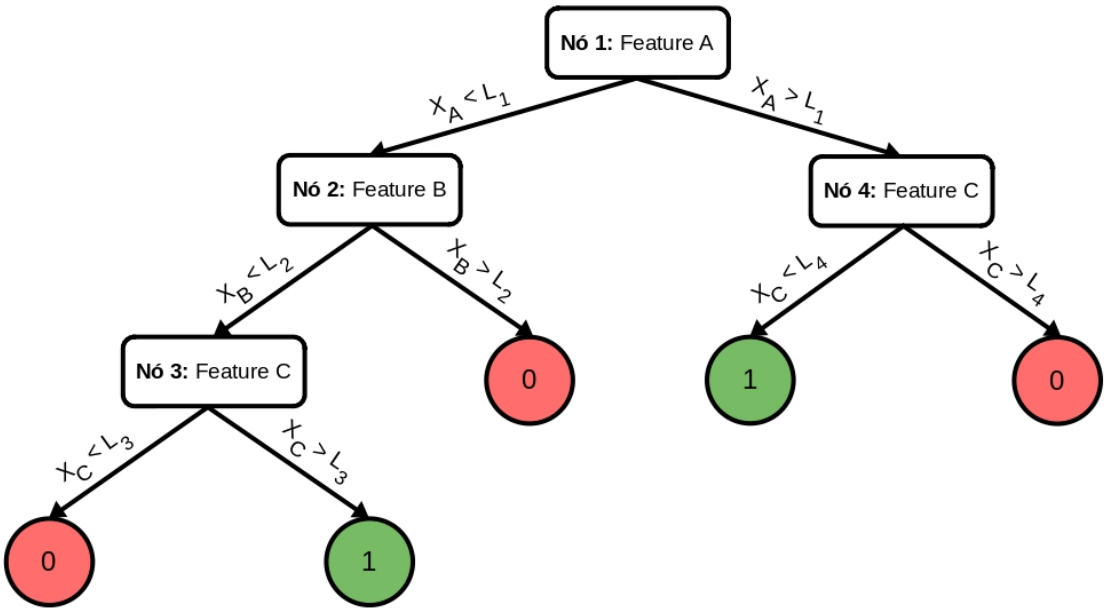

Figura 2.1: Exemplo de uma árvore de decisão de classificação, onde se podem observar os nós internos como retângulos e os nós de decisão (folhas) como círculos.

2. **Nó interno:** Nós após o nó raiz onde o fluxo da árvore se divide em dois caminhos, e uma regra é aplicada para se decidir qual seguir;

3. **Nó de decisão (folha):** Nó onde uma previsão final é realizada.

Como pode ser observado na Figura 2.1, o fluxo de previsão dentro de uma árvore de decisão sempre começa no nó raiz, passando pelos nós internos até se chegar em um nó de decisão, onde a previsão final é feita.

A geração, ou treinamento, de uma árvore de decisão consiste na criação de regras que separam uma amostra inicial de treinamento em subamostras. Por se tratar de um algoritmo de aprendizado supervisionado, esse processo necessita de uma amostra com alvos já definidos.

O primeiro passo do treinamento é encontrar a melhor regra para separar os objetos com alvos diferentes na amostra inicial. Para isso, uma das variáveis é escolhida e todas as regras possíveis de serem criadas com ela são testadas.

Apesar de existirem infinitos valores possíveis para o L de uma regra, as divisões que eles são capazes de criar na amostra inicial são finitas. No caso de uma variável A com n_A valores $X_{i,A}$ distintos, o tipo de regra definido dentro da árvore de decisão só consegue criar $n_A - 1$ divisões diferentes.

Essa questão fica clara na Figura 2.2, onde se pode observar a distribuição de valores para uma variável A de uma amostra exemplo, representados por pontos. O valor limite L escolhido nesse caso divide os pontos em duas subamostras, uma com três objetos e

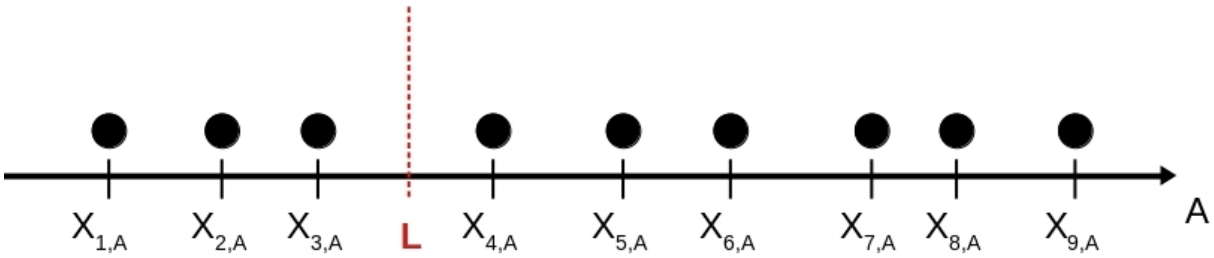

Figura 2.2: Exemplo visual da aplicação do valor limite L em uma variável.

outra com seis. No entanto, qualquer valor de L escolhido entre $X_{3,A}$ e $X_{4,A}$ retornaria as mesmas subamostras. Sendo assim, a cada dois valores consecutivos só é necessário se testar um L, que pode ser definido pelo valor médio entre os dois.

Com isso, o conjunto de valores limite que devem ser testados para uma variável A com n valores $X_{i,A}$ distintos pode ser calculado com:

$$L_i = \frac{X_{i,A} + X_{i+1,A}}{2}, \text{ para i} = 1, ..., n_A - 1 \tag{2.2}$$

Tendo definido todos os valores limites a serem testados, a qualidade das regras relacionadas a eles é definida a partir de uma função de impureza ou de perda H. A escolha dessa função depende diretamente do tipo de previsão que está sendo feita (ou seja, se o modelo é uma regressão ou uma classificação).

Para uma certa regra a ser testada $\theta = (A, L)$, que considera a variável A e o valor limite L e divide uma amostra Q_{ini} com n_{ini} objetos em duas subamostras $Q_{esq}(\theta)$ e $Q_{dir}(\theta)$ com n_{esq} e n_{dir} objetos, respectivamente, a impureza total é definida como:

$$G(Q_{ini}, \theta) = \frac{n_{esq}}{n_{ini}} H(Q_{esq}(\theta)) + \frac{n_{dir}}{n_{ini}} H(Q_{dir}(\theta)) \tag{2.3}$$

Após calcular a impureza de todo o conjunto de regras possíveis para todas as variáveis disponíveis, a que resulta na menor impureza é escolhida e a amostra original é dividida de acordo com ela. O restante do processo está baseado na repetição deste passo de criação de regras para dividir as subamostras resultantes até que não seja mais possível criar regras, ou até que uma profundidade (número de níveis dentro da árvore) máxima seja alcançada.

No que diz respeito à forma da função H, é possível citar algumas mais utilizadas para cada tipo de previsão. Para um problema de classificação, onde cada objeto pode pertencer a uma classe 0 ou a uma classe 1, pode-se utilizar a impureza de Gini ou a medida de entropia:

$$\text{- Gini: } H(Q) = p_0(1 - p_0) + p_1(1 - p_1), \tag{2.4}$$

$$\text{- Entropia: } H(Q) = -(p_0\, log(p_0) + p_1\, log(p_1)), \tag{2.5}$$

onde os valores p_i representam a proporção da classe i na amostra Q.

As subamostras que terminam o processo de treino sem serem divididas se transformam nos nós folha da árvore, onde as previsões finais são definidas pelo modelo.

Para problemas de classificação, onde o alvo desejado é binário (0 ou 1), os modelos de árvore de decisão geralmente atribuem aos objetos dentro de um certo nó folha uma probabilidade de ser da classe 0 (ou 1) igual à fração de objetos da classe 0 (ou 1) dentro desse nó.

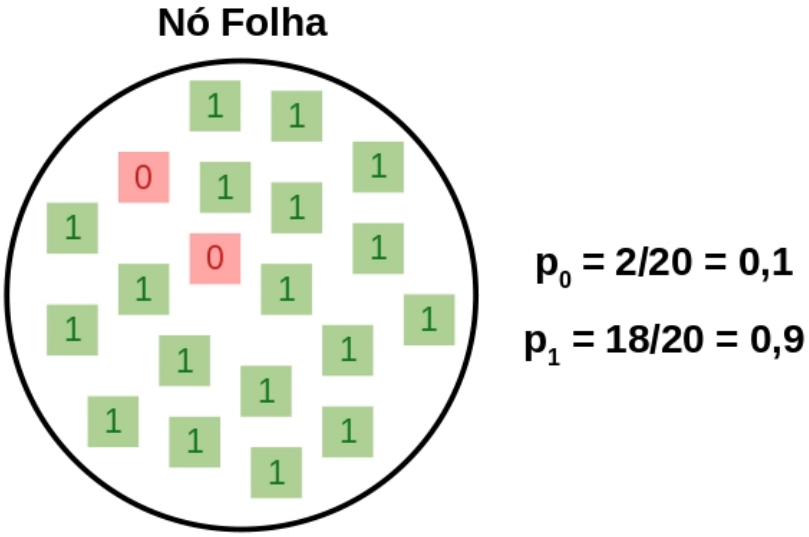

Figura 2.3: Exemplo de um nó folha em uma árvore de decisão. Na figura o nó é representado pelo círculo e os objetos contidos nele por quadrados com o valor do alvo em seu interior.

Esse processo é exemplificado na Figura 2.3, onde se considera um nó folha em uma árvore treinada para classificar objetos entre a classe 0 e a classe 1. Nesse caso as proporções das classes foram $p_0 = 0,1$ e $p_1 = 0,9$, de modo que todos os objetos dentro do nó receberão uma probabilidade de $0,9$ de pertencer à classe 1 e uma probabilidade de $0,1$ de pertencer à classe 0. Além disso, qualquer objeto novo que entrar no modelo e, depois de passar por todo o fluxo de regras, cair nesse nó, receberá essas mesmas probabilidades.

Com isso, é possível definir um ponto de corte entre 0 e 1 para uma classificação da classe 1 onde qualquer objeto com uma probabilidade de pertencer à classe 1 acima do corte será classificado como 1 pelo modelo.

Já para problemas de regressão, onde o alvo é uma variável contínua y, as formas mais comuns da função H são o erro médio quadrado (MSE) e o erro médio absoluto (MAE):

$$\text{- MSE: } H(Q) = \frac{1}{n}\sum(y - \bar{y})^2, \tag{2.6}$$

$$\text{- MAE: } H(Q) = \frac{1}{n}\sum|y - mediana(y)|, \tag{2.7}$$

onde $\bar{y}$ representa a média dos valores de y dentro da amostra Q, que também é geralmente utilizada como previsão final do modelo dentro dos nós folha de problemas de regressão.

Tendo definido os conceitos base por trás de uma árvore de decisão, é possível entender o processo por trás de como os Métodos de Ensemble (dos quais o Random Forest faz parte) operam.

2.2.2 Métodos de Ensemble

Em seu princípio, os métodos de Ensemble consistem em tomar as previsões de vários modelos de ML (denominados de estimadores base) e uni-las em uma única previsão final, que geralmente é mais precisa e confiável. A depender do processo por trás do treinamento dos estimadores base, é possível dividir esses algoritmos em dois tipos:

- **Métodos de Média**, nos quais os estimadores base são treinados independentemente e a média de suas previsões é considerada como a previsão final do modelo;
- **Modelos de Boosting**, nos quais os estimadores base são treinados em sequência, e cada um deles tenta diminuir o erro da previsão do conjunto.

O algoritmo de Random Forest se encaixa nos modelos de média, e o princípio por trás de seu funcionamento é adicionar fontes de aleatoriedade junto ao mecanismo de previsão das árvores de decisão.

Geralmente, o treinamento de uma árvore de decisão gera estimadores com uma considerável variância, que tendem a sofrer de *overfit* (situação na qual o modelo performa muito bem na base de treino, mas não mantém essa performance em bases novas) (GOODFELLOW *et al.*, 2016), e é exatamente essa variância que as fontes de aleatoriedade do *Random Forest* focam em mitigar.

Mais especificamente, dentro de um modelo *Random Forest* com n árvores de decisão distintas, cada uma delas é treinada com uma amostra gerada a partir da amostra de treino utilizando substituição (o mesmo valor pode se repetir mais de uma vez durante a amostragem), de modo que todas elas são treinadas com dados distintos. Além disso, ao se definir a melhor regra para um nó de uma árvore de decisão, apenas uma fração de todas as variáveis disponíveis é considerada e testada.

Com isso, cada uma das árvores dentro da floresta tem erros sistemáticos independentes entre si, e ao se tomar a média de todas as suas previsões, parte desses erros se cancela, resultando em um modelo melhor do que uma árvore de decisão solitária.

2.2.3 Importância de Variáveis

Uma das grandes vantagens dos modelos baseados em árvores de decisão é o fato de que eles podem ser facilmente representados através de fluxogramas como o da Figura 2.1, o

que aumenta consideravelmente sua explicabilidade.

Além disso, uma outra característica importante das árvores que facilita o entendimento de seu processo é o fato de sua construção possibilitar o cálculo da importância de cada uma das variáveis de entrada dentro do modelo. Essa importância pode ser interpretada como o impacto que cada variável teve nas previsões finais do modelo, e a partir desse conhecimento é possível diferenciar as variáveis realmente relacionadas com o alvo.

Matematicamente, a importância de uma certa variável A pode ser calculada como a diminuição total (ou normalizada) na impureza das amostras que passaram por todos os nós que utilizam A para realizar sua divisão.

Tomando um certo nó com a regra $\theta = (A, L)$ que divide uma amostra Q_{ini} com n_{ini} objetos em duas subamostras Q_{dir} e Q_{esq} com n_{dir} e n_{esq} objetos, respectivamente, a importância U desse nó será:

$$U = \frac{n_{ini}}{n}\left(H(Q_{ini}) - \frac{n_{esq}}{n_{ini}}H(Q_{esq}(\theta)) + \frac{n_{dir}}{n_{ini}}H(Q_{dir}(\theta))\right), \tag{2.8}$$

onde n é o número total de objetos na amostra de treinamento e H é a função de impureza escolhida para o problema.

Nessa equação o primeiro termo representa a impureza da amostra original ponderada pela fração de objetos nela, enquanto que o segundo e o terceiro termos representam a impureza total nas subamostras divididas pela regra θ. Com isso, os nós que começam com uma amostra de alta impureza e a dividem em subamostras de baixa impureza terão um valor de importância maior dentro da árvore.

Assim, a importância $U(A)$ de uma variável A que foi utilizada por m nós internos distintos dentro de uma árvore de decisão pode ser calculada como a soma da importância desses nós:

$$U(A) = \sum_{i}^{m} U_i \tag{2.9}$$

Para facilitar a comparação das importâncias de todas as variáveis dentro de uma árvore, geralmente é interessante utilizar os valores normalizados de importância $u(A)$:

$$u(A) = \frac{U(A)}{U(A) + U(B) + U(C) + ...}, \tag{2.10}$$

onde $[A, B, C, ...]$ são as variáveis de entrada do modelo.

No caso de um algoritmo de *Random Forest*, a aleatoriedade adicionada durante o treinamento das árvores faz com que cada uma delas tenha valores diferentes de importância para as variáveis. Assim, geralmente a importância de uma variáveis dentro da floresta é definida como a média ou a mediana das importâncias de todas as árvores.

2.2.4 Eliminação Recursiva de Variáveis

Além de facilitar a explicabilidade e entendimento do modelo, a importância das variáveis também possibilita que os métodos baseados em árvores de decisão sejam utilizados no pré-processamento de dados para excluir as variáveis não-informativas dentro de um conjunto de dados (KOHAVI e JOHN, 1997).

Nesse caso de aplicação, o objetivo do algoritmo é definir, dentro de um conjunto de M variáveis disponíveis, quais as m variáveis mais importantes para a previsão de um certo alvo y. Apesar dessa escolha poder ser realizada diretamente a partir da lista de importâncias dessas variáveis retornada por uma árvore de decisão (ou Random Forest), geralmente o processo é realizado recursivamente para melhorar seu resultado final.

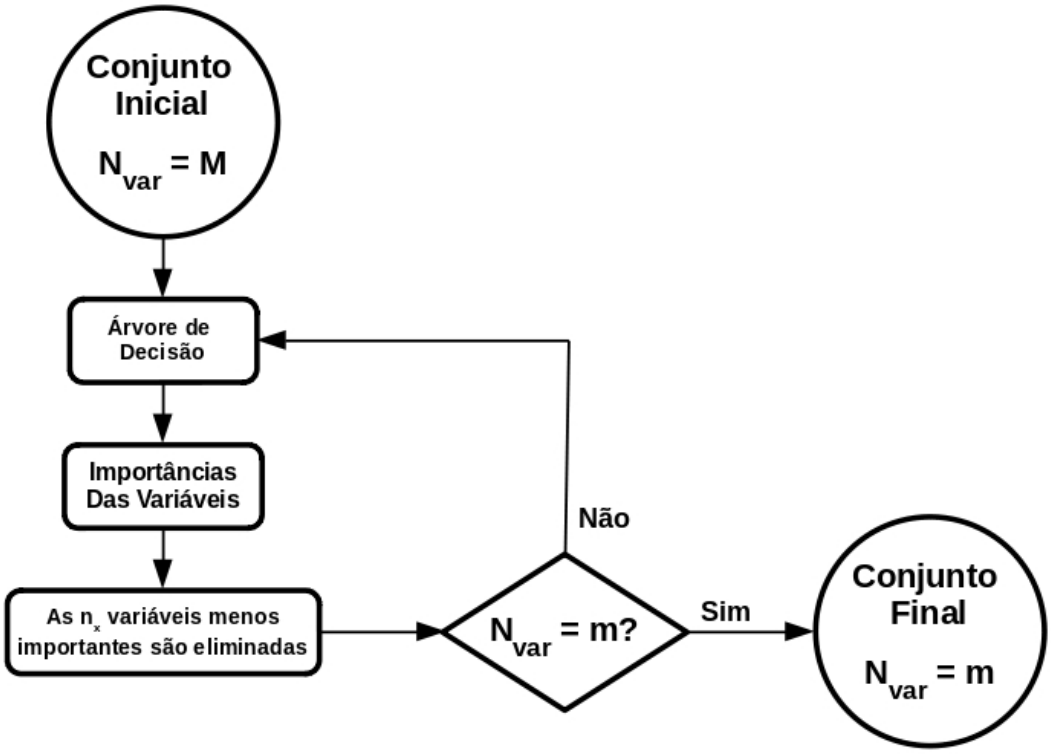

Figura 2.4: Fluxo de funcionamento de um algoritmo de eliminação recursiva de variáveis baseado em uma árvore de decisão.

Como é possível observar na Figura 2.4, que ilustra o fluxo de um algoritmo desse tipo, o conjunto inicial de M variáveis é utilizado como entrada para treinar uma árvore de decisão focada em prever o alvo y. Depois de realizado o treino, a lista de importâncias das variáveis gerada pela árvore é utilizada para se eliminar as n_x menos importantes entre elas. Depois disso, caso o conjunto filtrado ainda tenha mais do que m variáveis, os passos de treinamento e eliminação são repetidos até que ele esteja com o tamanho desejado.

Em relação ao número n_x de variáveis eliminadas em cada iteração do algoritmo, valores baixos dessa variável implicam em mais iterações e com isso um maior tempo gasto

para se obter o conjunto final. No entanto, apesar de valores altos de n_x possibilitarem que o algoritmo gere um resultado mais rápido, esse resultado não será tão confiável.

2.3 Parâmetros e Hiperparâmetros

Dentro de um modelo de aprendizado de máquina, os parâmetros são valores que definem e influenciam seu funcionamento. No caso de uma árvore de decisão, por exemplo, dois parâmetros que podem ser citados são as variáveis que cada nó considera e os valores limite L_i que dividem as amostras.

Quanto a sua origem, os parâmetros podem ser internos, quando seu cálculo é feito pelo algoritmo do modelo durante seu treinamento, ou externos - também conhecidos como hiperparâmetros -, quando eles precisam ser definidos antes do treinamento do algoritmo (YANG e SHAMI, 2020). Como exemplo no *Random Forest* é possível indicar o número de árvores da floresta, que não pode ser definido automaticamente através dos dados e por isso precisa ser passado para o algoritmo manualmente.

Dentro de um algoritmo de *Random Forest*, os hiperparâmetros em questão podem ser divididos em um grupo relacionado às árvores de decisão individuais e um grupo relacionado à floresta. Na Tabela 2.1, estão listados os mais importantes no que diz respeito à performance de um modelo *Random Forest* avaliada por VAN RIJN e HUTTER (2018) em um conjunto de diferentes amostras.

2.3.1 Otimização de Hiperparâmetros

Como os algoritmos de Aprendizado de Máquina não se encarregam de encontrar os valores ótimos de seus hiperparâmetros, geralmente é necessário realizar algum tipo de ajuste para se defini-los antes que o treinamento final seja realizado. Esse tipo de processo é conhecido como otimização de hiperparâmetros, e pode ser feito através de diferentes processos (YANG e SHAMI, 2020).

As metodologias mais comuns de otimização geralmente se iniciam com a consideração de cada hiperparâmetro como uma das dimensões de um hiperespaço. Nesse espaço cada ponto representa uma possível combinação de hiperparâmetros para o algoritmo, e a busca do ponto ótimo pode ser realizada de diferentes maneiras, onde as duas mais utilizadas são, de acordo com BERGSTRA e BENGIO (2012):

- **Busca em Grade**: Nesse método o hiperespaço é vasculhado ao longo de uma grade bem definida de valores. Para isso, são escolhidos conjuntos de valores para cada um dos hiperparâmetros, e cada combinação possível entre eles é testada separadamente;

- **Busca Aleatória**: Nesse método o hiperespaço é vasculhado aleatoriamente. Para isso um certo número de pontos é sorteado aleatoriamente dentro dos valores possíveis de cada hiperparâmetro, e então cada ponto é testado separadamente.

Hiperparâmetro	Descrição
bootstrap	Define se cada uma das árvores é construída com a amostra de treinamento completa (no caso False) ou com uma amostra *bootstrap* (com substituição) gerada a partir dela (no caso True)
max_features	Define a fração de variáveis a ser considerada no treinamento de cada árvore.
min_samples_leaf (msl)	Define o número mínimo de objetos em um nó folha para que ele seja considerado válido. Se durante o treinamento uma regra resultar numa divisão com menos do que msl objetos em um dos lados, ela é descartada automaticamente

Tabela 2.1: Hiperparâmetros com maior influência na performance final de um modelo Random Forest e uma descrição de seu funcionamento. Adaptado de VAN RIJN e HUTTER (2018)

Para ilustrar os dois métodos descritos acima, é possível considerar um caso geral de um modelo com dois hiperparâmetros a serem otimizados; um que pode assumir qualquer valor inteiro denominado HP_A e outro que pode assumir qualquer valor no intervalo $[0, 1]$ denominado HP_B.

Para uma busca em grade, o primeiro passo é a definição de um conjunto de valores para cada hiperparâmetro. Essa escolha pode ser baseada num conhecimento prévio do próprio algoritmo sendo otimizado, e geralmente é possível encontrar na literatura os intervalos razoáveis para cada hiperparâmetro. Para o exemplo considerado, foram escolhidos os valores $[5, 10, 15, 20, 25]$ para HP_A e $[0, 0, 25, 0, 50, 0, 75, 1, 0]$ para HP_B, o que resulta em 25 pontos distintos a serem testados.

Já para uma busca aleatória com o mesmo número de pontos são sorteados 25 pontos distintos dentro do hiperespaço formado por HP_A e HP_B. Esse sorteio pode ser limitado dentro de um certo intervalo, caso se conheça os valores razoáveis dos hiperparâmetros sendo otimizados. Para o exemplo considerado, os pontos foram escolhidos dentro do intervalo $[5, 25]$ para HP_A e dentro do intervalo $(0, 1]$ para HP_B.

Observando a Figura 2.5 é possível notar as diferentes coberturas do hiperespaço que cada tipo de busca gera para o exemplo considerado. No caso da busca em grade é possível

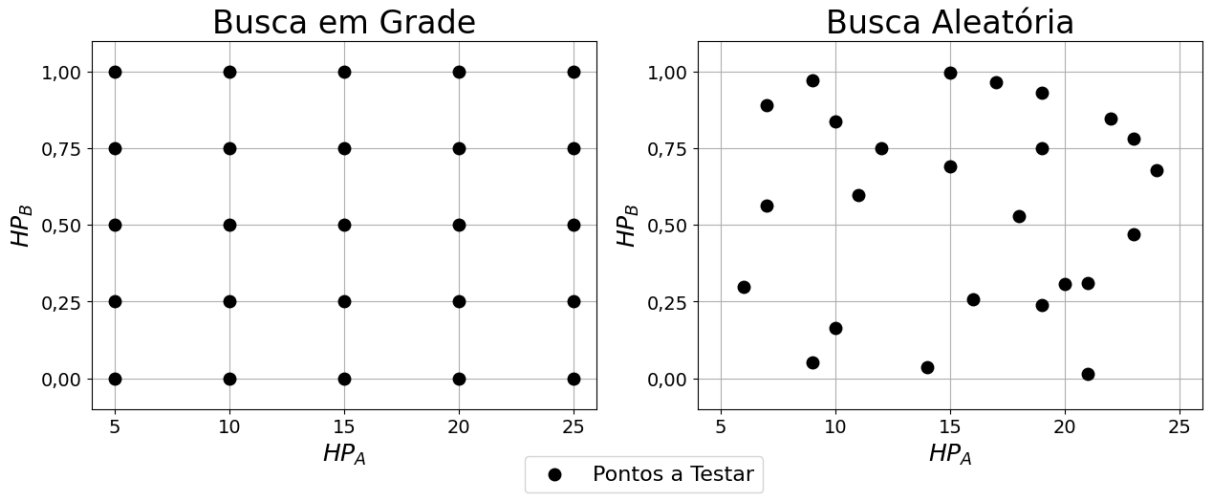

Figura 2.5: Exemplo das cobertura do hiperespaço de parâmetros obtido pela busca em grade, na esquerda, e pela busca aleatória, na direita.

garantir que o espaço será vasculhado de maneira consistente, mas os espaços internos da grade são ignorados no processo. Já para a busca aleatória, os espaços entre os pontos da grade têm uma certa chance de serem testados, mas existe a possibilidade de algumas áreas do espaço serem bem mais vasculhadas do que outras.

Em ambos os casos, após os pontos serem definidos cada uma das combinações definidas por eles é utilizada para gerar um modelo que é então treinado numa amostra de treino e sua performance é avaliada numa amostra de validação baseando-se em uma certa métrica definida previamente. A partir desses resultados é possível então escolher a combinação de hiperparâmetros otimizada para o problema em questão.

Em seguida, a performance do modelo escolhido precisa ser avaliada numa terceira amostra nunca antes vista por ele, conhecida como amostra de teste, para evitar qualquer viés no valor final calculado da métrica.

2.3.1.1 Otimização por Validação Cruzada

Para o processo de otimização ser realizado num certo conjunto de dados é necessário que a amostra seja dividida em três partes distintas:

- **Amostra de Treino**: Utilizada para treinar o algoritmo, gerando os modelos para cada uma das combinações de hiperparâmetros que se deseja testar. No caso das árvores de decisão, esse treinamento consiste em definir seu conjunto de regras;
- **Amostra de Validação**: Utilizada para avaliar a performance dos modelos treinados baseando-se em uma certa métrica escolhida;
- **Amostra de Teste**: Utilizada para avaliar a performance final do modelo otimizado com a melhor combinação de hiperparâmetros.

Em casos onde uma grande quantidade de dados estão disponíveis, a divisão do conjunto em três amostras não é tão impactante no número de objetos em cada uma delas. No entanto, quando não há uma grande quantidade de dados à disposição, esse tipo de divisão pode fazer com que restem poucos objetos em cada uma das amostras, o que impacta diretamente na performance do modelo (HASTIE *et al.*, 2001). Em casos como esses é possível utilizar algumas técnicas para aumentar a efetividade do seu uso, e um desses métodos é a validação cruzada *k-fold*.

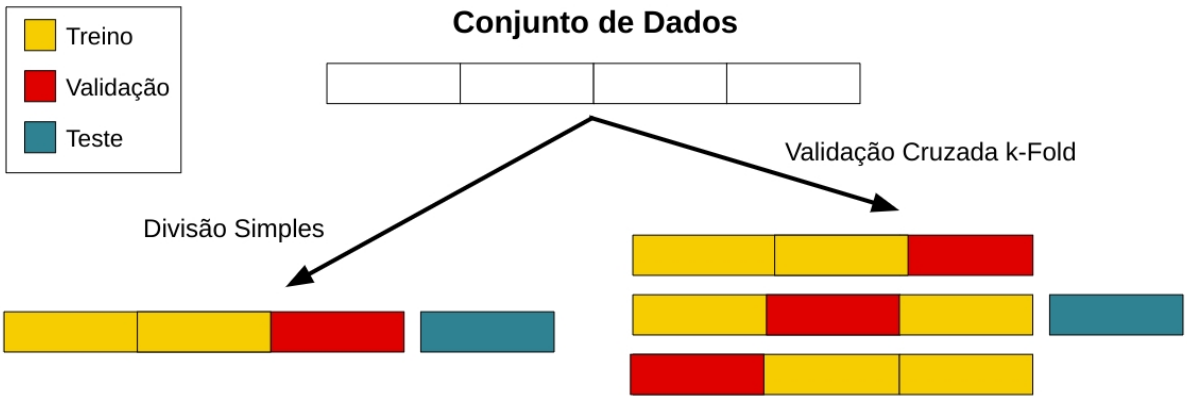

Figura 2.6: Esquemas de divisão do conjunto de dados em amostras de treino/validação/teste de acordo com o método padrão, na esquerda, e de acordo com a validação cruzada k-fold, na direita.

Como pode se observar na Figura 2.6, onde estão esquematizados o esquema padrão de divisão da amostra e o esquema da validação cruzada k-Fold, a maior diferença entre os dois está na maneira como as amostras de validação e treino são consideradas.

No caso da divisão simples as três amostras são separadas inicialmente e todos os objetos delas são utilizados apenas uma vez. Para a situação ilustrada na Figura 2.6 isso significa que os diferentes modelos seriam treinados com a fração indicada em amarelo (0,5), validados na fração indicada em vermelho (0,25) e por fim testados na fração indicada em azul (0,25).

Já na validação cruzada k-fold, os objetos nas amostras de treino e validação são reutilizados k vezes, sendo $k-1$ vezes como um objeto de treino, e 1 vez como objeto de validação. Isso garante que em modelagens com poucos dados disponíveis o algoritmo terá uma quantidade razoável de observações nas quais basear suas previsões (HASTIE *et al.*, 2001).

Para a validação cruzada k-fold primeiro se divide a massa de objetos que serão utilizados nas amostras de treino/validação em k partes iguais. Em seguida, todas as combinações possíveis onde uma das partes é utilizada como amostra de validação e o restante como amostra de treino são geradas. A partir disso, cada uma das k combinações pode ser utilizada para treinar e validar todos os modelos.

Com isso, uma outra grande diferença entre os dois métodos está na maneira como a métrica escolhida é calculada. Na divisão simples esse cálculo é baseado na performance dos modelos em uma única amostra de validação, o que pode trazer viéses associados. Enquanto isso, na validação cruzada k-fold, esse cálculo se baseia na performance dos modelos em k amostras de validação diferentes (geralmente a métrica final é considerada como a média das métricas nas k amostras) o que implica numa avaliação de performance mais robusta e menos suscetível a viéses dos dados (HASTIE *et al.*, 2001).

Para problemas onde se deseja reaproveitar ainda mais os dados disponíveis é possível utilizar a validação cruzada k-fold repetida. Nesse caso, após os k passos de validação serem realizados os objetos nas amostras de treino/validação são misturados aleatoriamente e toda a metodologia é realizada novamente, gerando outras k combinações nas quais os modelos podem ser validados. Esse mistura pode ser então repetida n vezes até que se alcance um número satisfatório de validações para a amostra em questão (VANWINCKELEN *et al.*, 2012).

2.4 Florestas Aleatórias de Classificação de Objetos

2.4.1 Estrutura Geral

Baseando-se em todos os conceitos explanados anteriormente, propomos um modelo de Aprendizado de Máquina capaz de tomar como entrada dados de observações astronômicas de estrelas e a partir deles prever se um certo objeto é membro ou não de um certo grupo. Esse tipo de modelo pode ser utilizado para classificar grandes quantidades de observações de maneira rápida e eficiente, gerando listas de potenciais candidatas que podem ser então refinadas e validadas através de observações mais específicas.

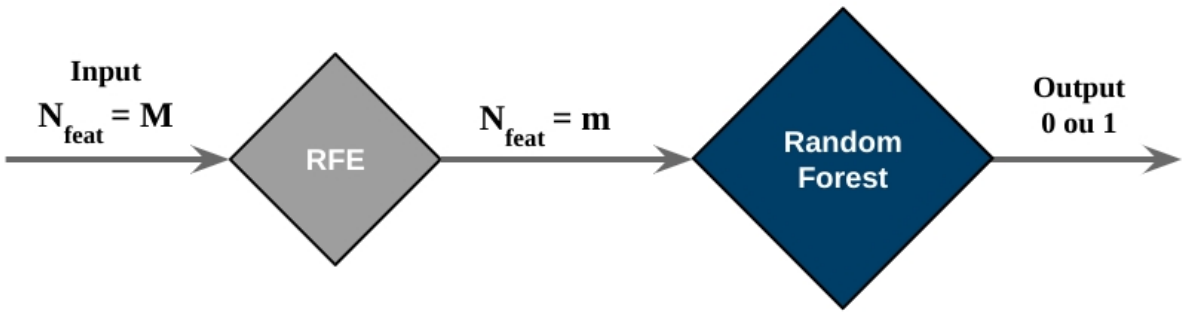

Figura 2.7: Estrutura geral das Florestas Aleatórias de Classificação de Objetos (FlACOs), formados por um primeiro passo de seleção de variáveis através de um RFE e um segundo passo que recebe as variáveis mais importantes e os utiliza dentro de um modelo Random Forest para prever se o objeto em questão faz parte ou não de um certo grupo.

No contexto desse trabalho, um modelo como esse pode ser aplicado em grandes levantamentos astronômicos para identificar, dentro dos milhões de objetos observados, uma

lista de candidatas mais promissoras à serem subanãs quentes. Com isso, é possível expandir os catálogos já existentes desse tipo de objeto, algo extremamente valioso para todas as pesquisas atuais nesse campo (CULPAN *et al.*, 2022).

Os modelos propostos, denominados Florestas Aleatórias de Classificação de Objetos (FlACOs), são formados pelos dois passos indicados na Figura 2.7, que mostra o fluxo do modelo desde a entrada até sua saída final:

1. ***Recursive Feature Elimination***: O primeiro passo do modelo recebe uma lista de M variáveis de um certo objeto e filtra as m mais importantes baseando-se na metologia descrita na subseção 2.2.4. Assim, garantimos que apenas as variáveis realmente informativos são passados para o segundo passo do modelo, onde as previsões são efetivamente realizadas;

2. ***Random Forest***: O segundo passo do modelo recebe as m variáveis filtradas pelo RFE e os utiliza para realizar uma previsão baseada no algoritmo de Random Forest cujo funcionamento está descrito nas seções 2.2.1 e 2.2.2.

O fluxo de treinamento desse tipo de modelo envolve primeiramente a otimização dos três hiperparâmetros de maior influência dentro do Random Forest indicados na Tabela 2.1, bem como o hiperparâmetro do número m de variáveis filtradas pelo RFE. Além disso, o número de árvores dentro da floresta (n_trees) e o ponto de *cutoff* para a definição de um objeto como candidato também são considerados, totalizando assim 6 hiperparâmetros otimizados. Depois que a combinação ideal de hiperparâmetros é definida, o modelo final é treinado e então testado numa amostra nunca vista durante a otimização.

2.4.2 Validação de Modelos

Como a classe de subanãs quentes é bastante rara, e os catálogos compilados atualmente não têm um número considerável de objetos, já se espera que o treinamento e os resultados obtidos durante o teste dos modelos de classificação propostos acima não tenham uma significância estatística muito alta.

Assim, para melhor validar a capacidade desse tipo de modelo, a mesma metodologia é também aplicada na predição de parâmetros estelares de uma amostra de geral de estrelas. Para esse fim, o Random Forest classificador no passo 2 do modelo é substituído por um Random Forest de regressão, capaz de prever valores contínuos em sua saída.

Os parâmetros estelares selecionados para o treinamento desses modelos de regressão foram a temperatura efetiva (T_{eff}), o logaritmo da gravidade superficial (logg) e a metalicidade ([Fe/H]), e a escolha dos três se baseou no fato deles serem foco de diversos estudos já desenvolvidos na área de Aprendizado de Máquina na astronomia (WHITTEN *et al.* 2019; PLACCO *et al.* 2021; WHITTEN *et al.* 2021; ANDRÉS GALARZA *et al.* 2021; YANG *et al.* 2022; WANG *et al.* 2022).

Além disso, os catálogos disponíveis para esses três parâmetros (como o LAMOST (ZHAO *et al.*, 2012), utilizado neste trabalho) permitem a geração de amostras muito mais numerosas de objetos para o treinamento de modelos ML, de modo que metodologia proposta pode ser avaliada em um ambiente sem a problemática de poucos dados.

Capítulo 3

Surveys

Nesse capítulo apresenta-se uma descrição geral dos principais levantamentos utilizados como base para a geração dos conjuntos de dados de desenvolvimentos dos FlACOs. Nesse processo foram considerados dois grandes levantamentos fotométricos (J-PLUS e S-PLUS) e um catálogo de subanãs compilado por GEIER *et al.* (2017) e atualizado periodicamente desde então (com a atualização mais recente ocorrendo em 2022). Além disso, também são apresentados os detalhes gerais do levantamento LAMOST, utilizado no desenvolvimento dos estimadores de parâmetros estelares.

No caso dos levantamentos fotométricos citados, ambos são projetos estruturantes do Observatório Nacional em colaboração com outras instituições brasileiras e espanholas, no caso do J-PLUS, e com instituições brasileiras e chilenas no caso do S-PLUS. Além disso, ambos já estão em atividade e tiveram múltiplos data releases (DR) liberados, fazendo com que grande parte dos seus dados já estejam disponíveis para toda a comunidade astronômica.

3.1 Javalambre Photometric Local Universe Survey (J-PLUS)

O Javalambre Photometric Local Universe Survey (J-PLUS) é um levantamento astronômico iniciado em 2015 e atualmente em andamento no Observatório Astronômico de Javalambre (OAJ), sendo realizado com o telescópio *Javalambre Auxiliary Survey Telescope* (JAST80), de $80 - cm$. Seu objetivo final é observar uma área de cerca de 8500 graus^2 do hemisfério celestial norte, o que resultaria na obtenção de dados fotométricos em 12 filtros para cerca de 35 milhões de estrelas e 24 milhões de galáxias distintas (CENARRO *et al.*, 2019).

3.1.1 Espectroscopia e Fotometria

Diferente de observações espectroscópicas, onde o perfil espectral contínuo de um objeto é obtido com uma alta resolução - o que demanda tempo e instrumentos específicos -, observações fotométricas como as do J-PLUS são mais simples pelo fato de utilizarem filtros fotométricos para detectar toda a luz de um objeto dentro de um certo intervalo de comprimento de onda.

A grande diferença de tempo necessário para a realização desses dois tipos de observação está no fato de ser possível acoplar o filtro fotométrico a uma câmera e gerar dados para todos os objetos no campo de visão do equipamento (desde que tenham intensidade suficiente para serem detectados). Enquanto isso, para obter o perfil espectral contínuo de um objeto, é necessário passar sua luz por um elemento dispersivo e então registrar o resultado, o que adiciona um passo intermediário no processo e limita os objetos que podem ser observados.

Apesar de algumas aplicações científicas necessitarem do nível de detalhe e resolução de uma análise espectroscópica de seus objetos, algumas são capazes de trabalhar apenas com magnitudes e fluxos fotométricos, e essas se beneficiam fortemente da imensa quantidade e qualidade de dados e observações que levantamentos como o J-PLUS e o S-PLUS são capazes de gerar.

3.1.2 Conjunto de Filtros

Os 12 filtros utilizados pelo J-PLUS foram escolhidos de maneira a cobrir a região do óptico e capturar também algumas linhas espectrais importantes.

Filtro	**Comprimento de Onda Central (Å)**	**Largura à Meia Altura (Å)**	**Linhas**
u	3485	508	
J0378	3785	168	[OII]
J0395	3950	100	Ca H+K
J0410	4100	200	Hδ
J0430	4300	200	Banda G
g	4803	1409	
J0515	5150	200	Tripleto de Mgb
r	6254	1388	
J0660	6600	138	Hα
i	7668	1535	
J0861	8610	400	Tripleto de Ca II
z	9114	1409	

Tabela 3.1: Lista de filtros utilizados pelo levantamento J-PLUS, descrevendo o nome de cada filtro, o comprimento central do seu perfil, sua largura à meia altura - ambos indicados em Angstroms - e as principais linhas presentes em cada filtro estreito. Fonte: CENARRO *et al.* (2019)

De acordo com a Tabela 3.1, onde estão indicadas os comprimentos de onda centrais e larguras à meia altura de todos os filtros, quatro deles são filtros em comum com o SDSS (g, r, i, z) (YORK *et al.*, 2000b), e junto com o u são filtros de banda larga. Em contraste, sete deles são filtros estreitos (todos os filtros de nome iniciando com J0), desenvolvidos com o objetivo de capturar linhas espectrais específicas em absorção e emissão.

A distribuição desses filtros pode ser observada na Figura 3.1, onde são exibidas suas curvas de transmissão gerais, considerando os efeitos do céu, espelhos, lentes e câmera (CENARRO *et al.*, 2019).

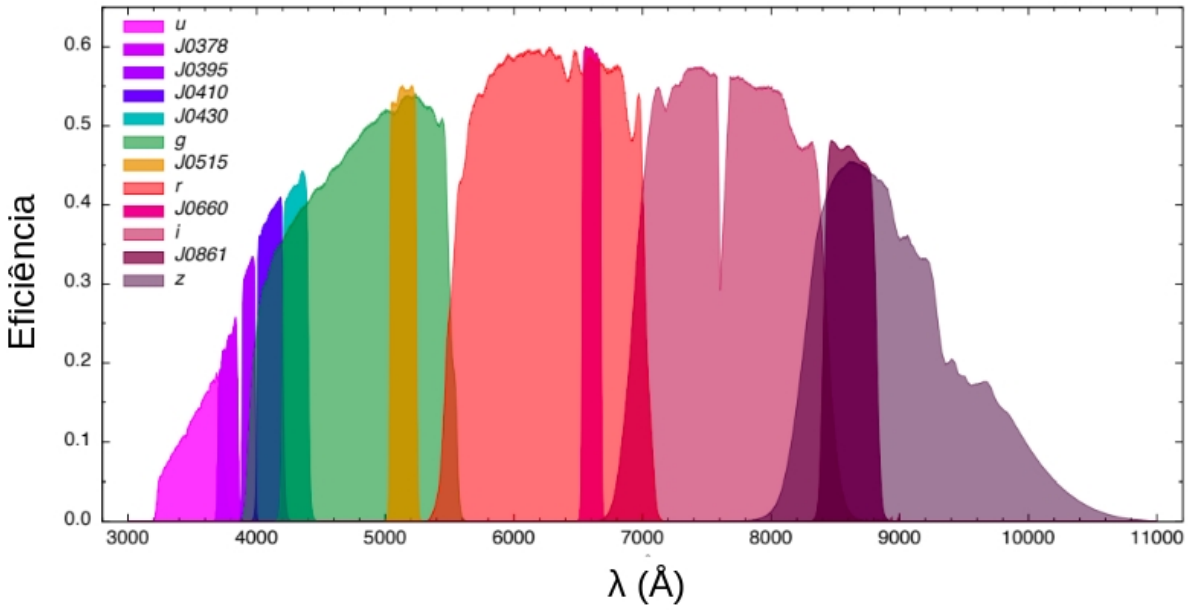

Figura 3.1: Curvas de transmissão para os 12 filtros utilizados pelo levantamento J-PLUS. Fonte: CENARRO *et al.* (2019)

3.1.3 Área do Levantamento e Data Releases

Apesar do objetivo final do J-PLUS ser a cobertura de uma área com cerca de 8500 graus^2 no hemisfério celestial norte, a estratégia de disponibilização de dados escolhida pela equipe responsável consiste em realizar liberações de dados periodicamente (Data Releases) com novos objetos e novos campos adicionados a cada nova liberação.

Data Release	**Ano**	**Área Coberta (graus^2)**
EDR	2017	36
DR1	2018	1022
DR2	2020	2176
DR3	2022	3192

Tabela 3.2: Dados gerais dos Data Releases do J-PLUS ao longo de sua realização, descrevendo o ano de cada liberação e a área coberta por cada um deles.

Como se pode observar na Tabela 3.2, até novembro de 2022 já foram liberados quatro Data Releases pela equipe do J-PLUS. O primeiro deles, denominado Early Data Release (EDR), foi realizado com o intuito de fornecer um referencial preliminar para o levantamento em relação às profundidades fotométricas (em torno de $21,25$ mag em cada banda), qualidade de imagens e calibrações, e por isso a quantidade de campos cobertos foi baixa em relação às outras liberações.

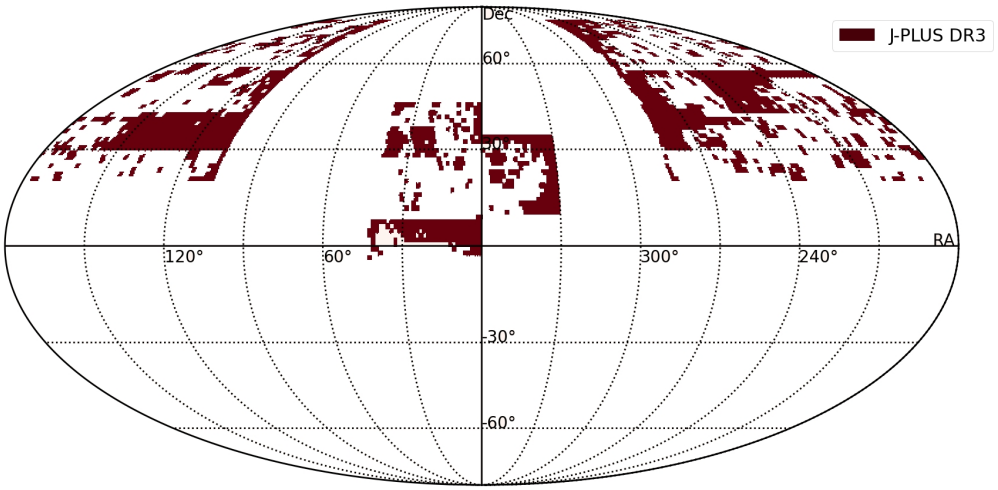

Figura 3.2: Distribuição dos campos presentes no DR3 do J-PLUS no referencial ICRS.

Após a liberação do EDR, os outros DRs passaram a adicionar cada vez mais campos observados, além de melhorar a qualidade fotométrica do levantamento. O último Data Release liberado já cobre cerca de 38% da área total que o levantamento final busca observar, e os quase 30 milhões de objetos presentes nele estão distribuídos nas regiões indicadas na Figura 3.2.

3.2 Southern Photometric Local Universe Survey (S-PLUS)

O Southern Photometric Local Universe Survey (S-PLUS) é um levantamento astronômico bastante similar ao J-PLUS descrito na seção anterior. Também focado em gerar dados fotométricos para os mesmos 12 filtros, o S-PLUS foi iniciado em 2016 com o objetivo de observar uma área de cerca de 9300 graus^2 do hemisfério celestial sul, de modo a completar o conjunto de dados sendo construído pelo J-PLUS para esse hemisfério (MENDES DE OLIVEIRA *et al.*, 2019).

Além de realizarem observações nos mesmos 12 filtros, tanto o telescópio quanto a câmera que formam o sistema observacional desses levantamentos são idênticos, de modo que os dados desses levantamentos podem, em teoria, ser utilizados em conjunto no futuro.

3.2.1 Conjunto de Filtros

O S-PLUS utiliza o mesmo conjunto de filtros do J-PLUS, que tem como objetivo cobrir a região do óptico com filtros largos e detectar regiões específicas do espectro com filtros estreitos. Os detalhes gerais desse conjunto de filtros podem ser encontrados na Tabela 3.1, e uma representação visual de seus perfis na Figura 3.3.

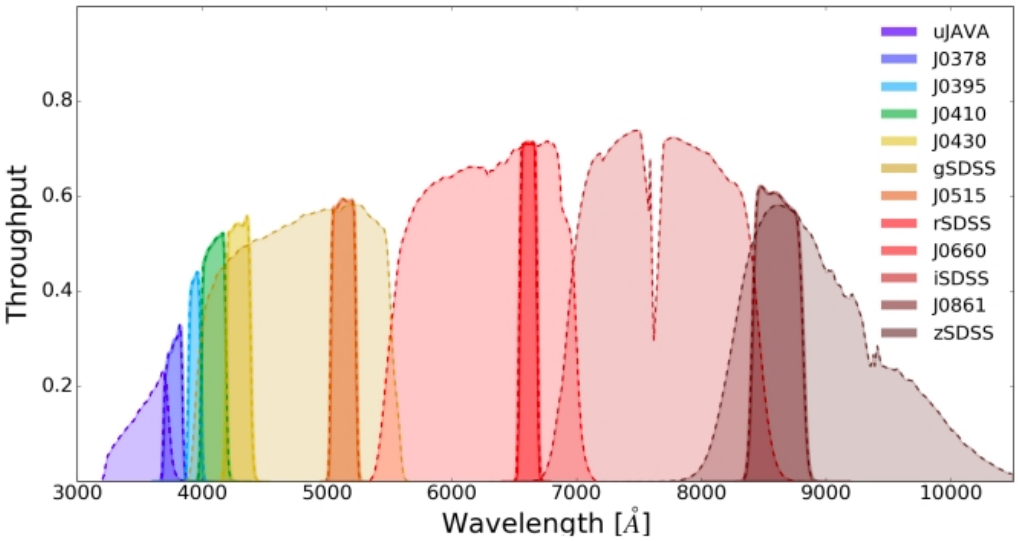

Figura 3.3: Curvas de transmissão para os 12 filtros utilizados pelo levantamento S-PLUS. Fonte: MENDES DE OLIVEIRA *et al.* (2019).

3.2.2 Área do Levantamento e Data Releases

Assim como é o caso para o J-PLUS, a equipe do S-PLUS adota um modelo de trabalho baseado em liberar os dados obtidos pelo levantamento em Data Releases periódicos.

Data Release	**Ano**	**Área Coberta (graus^2)**
DR1	2019	336
DR2	2021	950,5
DR3	2021	1818,8
DR4	2023	3000

Tabela 3.3: Dados gerais dos Data Releases do S-PLUS ao longo de sua realização, descrevendo o ano de cada liberação e a área coberta por cada uma delas.

Como se pode observar na Tabela 3.3, o S-PLUS começou a liberar seus dados no ano 2019, dois anos depois do primeiro DR do J-PLUS. Como o S-PLUS foi iniciado um ano após o J-PLUS, essa relação é esperada, e devido a isso ele não se encontra em um estágio tão avançado quanto o J-PLUS quando se considera a fração da área final já observada.

Na Figura 3.4 está indicada a área presente no terceiro Data Release do S-PLUS, onde é possível ver a área do hemisfério sul sendo observada pelo levantamento, bem como

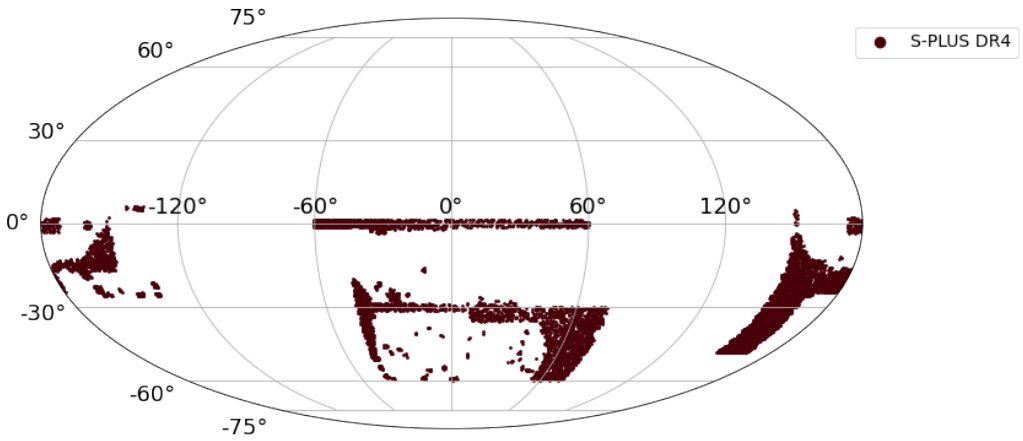

Figura 3.4: Distribuição dos campos presentes no DR4 do S-PLUS no referencial ICRS.

a área da Stripe 82 no centro do gráfico presente desde o DR1 e utilizada inicialmente para demonstrar a qualidade dos dados fotométricos do S-PLUS e exemplificar algumas possíveis aplicações do levantamento.

3.3 Ciência com J-PLUS e S-PLUS

Como citado anteriormente, os dados fotométricos fornecidos pelos 12 filtros do J-PLUS e do S-PLUS podem ser aplicados em diversos campos da astronomia que não necessitam de perfis espectrais de alta resolução para realizar seus estudos. Além disso, a grande área de cobertura desses filtros, e a presença tanto de filtros largos quanto de filtros estreitos aumentam ainda mais os possíveis usos desses dados.

Alguns exemplos de casos de aplicação desses levantamentos que podem ser citados são a estimativa de parâmetros estelares (WHITTEN *et al.* 2021; YANG *et al.* 2022; WANG *et al.* 2022), a classificação de objetos como estrelas e galáxias (LÓPEZ-SANJUAN *et al.* 2019; COSTA-DUARTE *et al.* 2019; WANG, C. *et al.* 2022), a identificação de novos membros e estudo de aglomerados (SAN ROMAN *et al.* 2019; BONATTO *et al.* 2019; MOLINO *et al.* 2019; BRITO-SILVA *et al.* 2021; BUZZO *et al.* 2021; CHIES-SANTOS *et al.* 2022) e a identificação de estrelas pobres em metais (WHITTEN *et al.* 2019; PLACCO *et al.* 2021; ANDRÉS GALARZA *et al.* 2021).

Dentre essas aplicações citadas, é interessante notar que diversas delas se baseiam em métodos de ML e no treinamento de modelos capazes de captar padrões a partir dos dados fotométricos (WHITTEN *et al.* 2021; YANG *et al.* 2022; WANG *et al.* 2022). O diferencial do uso do J-PLUS e do S-PLUS para esse tipo de ciência é a grande quantidade de dados de alta qualidade gerados por eles, o que possibilita o desenvolvimento de modelos robustos e precisos.

3.4 Catálogo de Subanãs Quentes

3.4.1 Versões Disponíveis

Compilado inicialmente em 2017 como um catálogo preliminar de subanãs quentes presentes no levantamento Gaia que começou a ter suas observações liberadas no mesmo ano, o "*The population of hot subdwarf stars studied with Gaia*"se mostra como o mais completo catálogo desse tipo de objeto disponível atualmente (CULPAN *et al.*, 2022).

Versão	Ano	Número de Objetos	Fonte
V1	2017	5613	GEIER *et al.* (2017)
V2	2020	5874	GEIER (2020)
V3	2022	6616	CULPAN *et al.* (2022)

Tabela 3.4: Histórico de versões do catálogo de subanãs quentes e dados gerais de cada uma delas.

Em sua primeira versão esse catálogo, como se pode observar na Tabela 3.4, foram listadas 5613 subanãs conhecidas e candidatas obtidas da literatura e de conjuntos de dados ainda não publicados (GEIER *et al.*, 2017). Três anos depois, em 2020, uma nova versão do catálogo foi publicada a partir da retirada dos objetos da primeira versão adicionados erroneamente e da classificação de 528 subanãs novas (GEIER, 2020). Por fim, em 2022 a versão mais recente do catálogo foi publicada com 742 objetos adicionados aos da anterior (CULPAN *et al.*, 2022).

3.4.2 Área do Catálogo

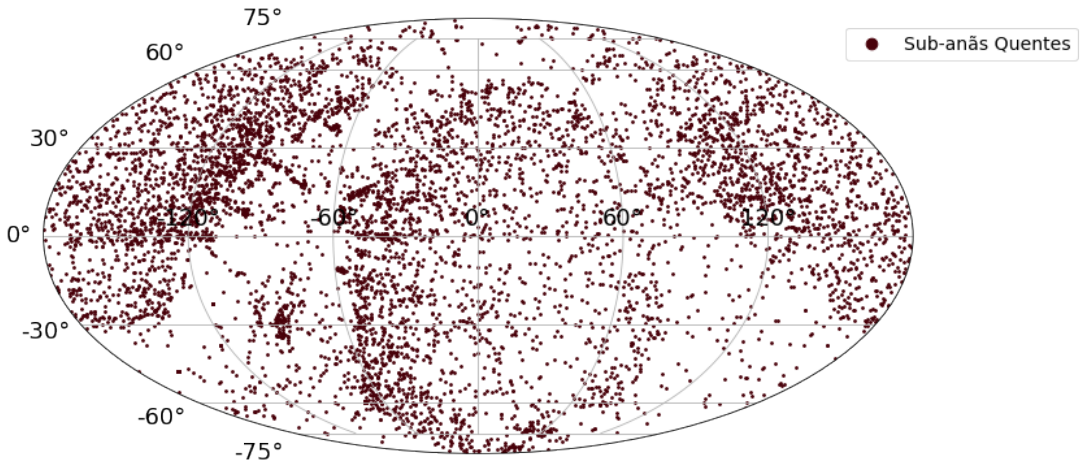

Figura 3.5: Distribuição dos objetos listadas em CULPAN *et al.* (2022) no referencial ICRS.

Em relação às áreas cobertas pelo catálogo, é possível observar na Figura 3.5 que a distribuição das subanãs confirmadas se concentra em áreas fora do plano galáctico. Essa

questão está relacionada com a dificuldade de se observar e confirmar a classificação desse tipo de objeto quando ele se encontra em áreas de alta concentração estelar ou com grande extinção interestelar (CULPAN *et al.*, 2022).

3.5 LAMOST

O Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) é um levantamento espectroscópico iniciado em 2011 com o objetivo de captar espectros de milhões de estrelas e galáxias, trazendo assim grandes contribuições para os campos da astrofísica estelar, galáctica e extragaláctica, bem como da cosmologia e do estudo de exoplanetas (YAN *et al.*, 2022).

O telescópio responsável por realizar esse levantamento também é denominado LAMOST, possui um diâmetro de $4-m$ e, com seus dezesseis espectrógrafos que acomodam 250 fibras cada um, é capaz de captar dados espectroscópicos de até 4000 objetos simultaneamente. A magnitude limitante dessas observações é de $r = 19$ onde se obtém uma resolução de $R = 1800$, e para objetos mais brilhantes a resolução pode chegar até $R = 10000$. Com isso, o LAMOST é capaz de produzir espectros de alta qualidade para até dezenas de milhares de objetos em uma única noite de observação (ZHAO *et al.*, 2012).

3.5.1 Data Releases

As liberações de dados do LAMOST seguem um modelo de inicialmente serem disponibilizadas internamente para os membros da colaboração, e posteriormente serem liberadas para o público geral. Com isso, como se pode observar na Tabela 3.5 apesar de o último DR do levantamento ser o DR10, o mais recente DR com acesso público é o DR8.

Data Release	Ano	Número de Espectros	Status
DR1	2013	2,2 milhões	Público
DR2	2014	4,1 milhões	Público
DR3	2015	5,7 milhões	Público
DR4	2016	7,6 milhões	Público
DR5	2017	9,0 milhões	Público
DR6	2018	9,9 milhões	Público
DR7	2019	10,4 milhões	Público
DR8	2020	10,6 milhões	Público
DR9	2021	11,2 milhões	Interno
DR10	2022	11,8 milhões	Interno

Tabela 3.5: Dados gerais dos Data Releases do LAMOST ao longo de sua realização, descrevendo o ano de cada liberação, o número de espectros e o status de cada uma delas.

Por fim, além de trazerem os espectrais dos objetos observados, os DRs do LAMOST

também disponibilizam um catálogo de parâmetros estelares (mais especificamente T_{eff}, log(g) e [Fe/H]) calculados utilizando o LAMOST Stellar Parameter Pipeline (LASP) (WU *et al.*, 2014).

No DR8 esse catálogo contém os parâmetros de cerca de 6,7 milhões de estrelas distribuídas na área indicada na Figura 3.6, escolhidas com base no sinal-ruído de seu espectro na banda g (SN $>$ 6 para noites escuras e SN $>$ 15 para noites claras).

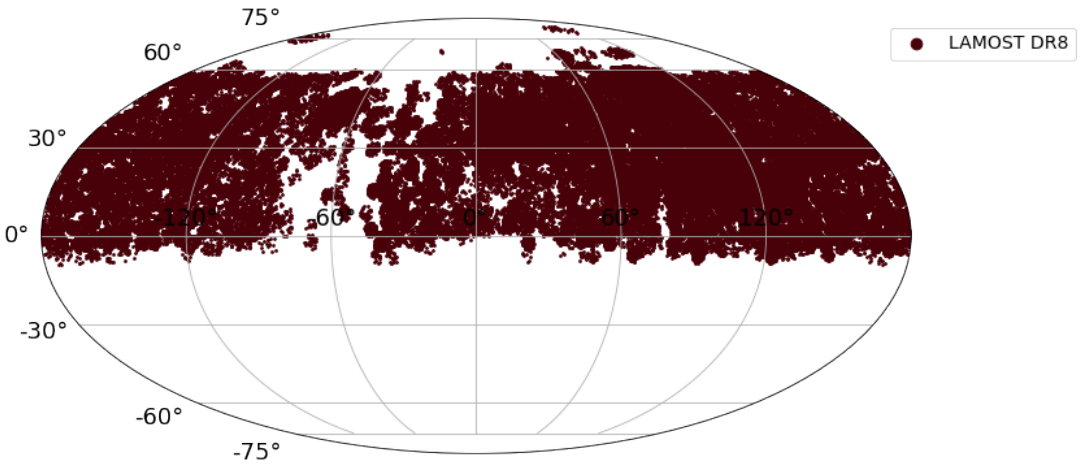

Figura 3.6: Distribuição dos objetos com parâmetros estelares calculados presentes no DR8 do LAMOST no referencial ICRS, disponível para consulta no site do levantamento [2].

3.5.2 Previsão de Parâmetros a partir do LAMOST

Por disponibilizar parâmetros estelares calculados a partir dos espectros de uma grande quantidade de estrelas em seus catálogos, o LAMOST é um ótimo candidato a ser utilizado como base no desenvolvimento de modelos de ML focados na previsão desses parâmetros. Devido à isso, já existem diversos trabalhos que exploram essa possibilidade, trazendo resultados promissores com o uso de diferentes algoritmos como base para modelos de previsão de parâmetros dentro do J-PLUS (YANG *et al.* 2022; DE CARVALHO 2022). Em relação ao S-PLUS, apesar de não haver nenhum resultado concreto em relação à modelos de previsão de parâmetros estelares baseados no LAMOST, é de se esperar que a qualidade dos modelos do J-PLUS seja replicável.

Além disso, até mesmo quando comparado com outros levantamentos que fornecem parâmetros estelares (como o SEGUE e o GALAH), o LAMOST se mostra o mais promissor para o uso com o J-PLUS no que diz respeito à consistência dos dados, o número de objetos em comum e a qualidade dos modelos gerados (DE CARVALHO, 2022). Baseando-se em todos esses fatores, o LAMOST foi escolhido como fonte dos parâmetros estelares utilizados no desenvolvimento de modelos ao longo desse trabalho.

Capítulo 4

Resultados e Discussão

Nesse capítulo estão compilados os resultados obtidos a partir das metodologias descritas no Capítulo 2 e dos dados indicados no Capítulo 3.

A primeira seção se inicia com uma explicação mais aprofundada em relação à criação das bases de treino, validação e teste utilizadas no desenvolvimento dos FlACOs. Em seguida, são apresentados os resultados da otimização de hiperparâmetros e da performance final dos modelos escolhidos. Por último, um estudo do impacto de cada variável de entrada é realizado, e as mais importantes são apontadas.

Na segunda seção são apresentados os resultados da aplicação dos FlACOs para a criação de listas de candidatas à subanãs quentes a partir dos dados do J-PLUS e do S-PLUS, bem como uma análise do grupo de candidatas obtidas.

E por fim, a terceira seção traz os resultados do uso da metodologia apresentada no treinamento de modelos para a previsão de parâmetros estelares de estrelas gerais em ambos os levantamentos. Além disso, também são descritos os métodos por trás da criação de catálogos de parâmetros estelares para as estrelas do J-PLUS e do S-PLUS.

4.1 FlACOs para Subanãs Quentes

Com o objetivo de criar listas de candidatas à subanãs quentes, foram treinados FlACOs para a classificação desse tipo de objeto baseando-se nos dados fotométricos observados pelo J-PLUS e pelo S-PLUS e no catálogo de subanãs disponibilizado por CULPAN *et al.* (2022).

Cada um dos levantamentos foi utilizado para otimizar e treinar um modelo de classificação, e foi avaliada sua performance e capacidade de identificar subanãs quentes dentro de uma amostra de teste não considerada durante o treinamento. Por fim, cada um dos modelos foi utilizado para gerar uma lista de candidatas dentro de seu respectivo levantamento, e as características dos objetos em cada lista foram analisadas.

4.1.1 Amostras de Desenvolvimento

Para realizar o desenvolvimento dos modelos foram criadas duas amostras de desenvolvimento distintas (uma baseada no J-PLUS e uma no S-PLUS), mas com ambas sendo geradas a partir da mesma metodologia. Como detalhado na seção 2, o treinamento de um modelo de classificação é feito a partir de uma amostra que contém tanto exemplos da classe que se quer identificar (positivos) quanto exemplos das demais classes presentes na população (negativos).

Sendo assim, cada um desses grupos foi montado da seguinte forma:

- **Positivos:** Objetos do catálogo de subanãs quentes que foram observados pelo levantamento (J-PLUS ou S-PLUS);
- **Negativos:** Objetos amostrados aleatoriamente entre todos os objetos observados pelo levantamento (J-PLUS ou S-PLUS), além de uma amostra de anãs brancas também observada pelo levantamento. Devido à amostragem aleatória existe a possibilidade de existirem subanãs quentes não-catalogadas na amostra de negativas, o que pode prejudicar a performance do modelo.

Para criar os grupos de positivos foi realizado o cruzamento do catálogo de subanãs detalhado na seção 3.4 com ambos os levantamentos fotométricos. No caso do J-PLUS essa referência cruzada resultou em 319 objetos em comum, enquanto que para o S-PLUS esse número foi de 377.

A partir dessas duas amostras iniciais de objetos positivos, foram aplicados dois filtros para refinar sua qualidade. O primeiro deles desconsiderou as subanãs com qualquer um dos erros nos valores de suas 12 magnitudes observadas nos levantamentos maior do que 0,3. Isso é feito com o intuito de retirar objetos que pudessem adicionar um viés ao modelo através de valores de magnitude mal calculados.

Já o segundo filtro foi aplicado na classe espectral dos objetos informada pelo catálogo de subanãs. Para evitar que o modelo considerasse objetos em sistemas binários com companheiras cujos espectros contaminassem o espectro final do sistema, foram desconsideradas todas as subanãs com estrelas companheiras na sequência principal. Com isso, após a aplicação dos dois filtros descritos acima as amostras finais de positivos do J-PLUS e S-PLUS foram formadas por 271 e 308 objetos, respectivamente.

Como citado anteriormente, para criar os grupos de negativos foram utilizados dois tipos de objetos diferentes. O primeiro deles foi formado inicialmente por 50 mil estrelas amostradas aleatoriamente dentro de cada levantamento, nas quais foi aplicado também o mesmo filtro de 0,3 nos erros das magnitudes utilizado nas subanãs quentes. Após a aplicação desse filtro, as amostras de estrelas gerais do J-PLUS e do S-PLUS possuíam, respectivamente, 16267 objetos e 16822 objetos.

Um ponto importante de ser levantando em relação à esses dois grupos amostrados aleatoriamente dentro dos levantamentos é que, para todos os efeitos de treinamento do modelo, esses objetos são considerados como 'não subanãs' (negativas). Devido à isso, qualquer subanã não catalogada incluída nesses dois grupos pode adicionar um certo viés à performance do modelo por ter a classe errada durante o treinamento. No entanto, como as subanãs são objetos relativamente raros, não se espera que exista um grupo expressivo desses objetos dentro das 50 mil estrelas amostradas aleatoriamente em cada um dos levantamentos.

Por fim, o segundo grupo que forma as negativas nas amostras de treinamento são de anãs brancas observadas em cada um dos levantamentos. Para encontrar essas estrelas, foram realizados os cruzamentos do J-PLUS e do S-PLUS com os objetos catalogados por GENTILE FUSILLO *et al.* (2021) a partir do GAIA. Além de também aplicar o filtro de erros de magnitudes utilizado nos outros objetos, aqui também são consideradas apenas estrelas com uma probabilidade de serem anãs brancas maior do que 90%. Para que haja uma quantidade comparável de anãs brancas e subanãs na amostra final, após a aplicação dos filtros são amostradas aleatoriamente 500 anãs brancas que são então adicionadas às estrelas gerais para formar as negativas.

O papel das anãs brancas na amostra de negativas está relacionado com o fato desses objetos ocuparem uma região similar à das subanãs quentes no diagrama HR (como pode se observar na Figura 1.1). Devido à isso, é de se esperar que essas duas classes de objeto tenham parâmetros físicos (temperatura, gravidade superficial) comparáveis, o que pode fazer com que elas tenham um certo grau de semelhança no que diz respeito às 12 magnitudes observadas pelos levantamentos considerados.

No entanto, como as anãs brancas também são objetos relativamente raros, a sua representatividade na amostra de estrelas gerais não será muito alta, e os modelos podem não considerar objetos suficientes para detectar os padrões e diferenças entre uma subanã quente e uma anã branca. Com isso, a adição das anãs brancas na amostra de negativas procura melhorar a capacidade dos modelos de distinguir entre essas duas classes de objetos.

Com esses três grupos de objetos (subanãs quentes, estrelas gerais e anãs brancas) foi possível montar as amostras de desenvolvimento dos modelos do J-PLUS e do S-PLUS. A distribuição das quantidades de cada objeto, bem como sua representatividade percentual, podem ser encontradas na Tabela 4.1.

A respeito das amostras obtidas, é necessário considerar o fato de que elas são bastante desbalanceadas em relação ao número de subanãs quentes (aproximadamente 50:1). Essa questão influencia diretamente na performance dos modelos desenvolvidos, podendo fazer com que eles sejam muito 'otimistas' em relação à probabilidade de um objeto pertencer à classe positiva.

Em seguida, tendo se definido os objetos que seriam utilizados nas amostras, as va-

	J-PLUS		S-PLUS	
Objeto	**Quantidade**	**Porcentagem**	**Quantidade**	**Porcentagem**
Subanãs Quentes	271	1,6%	308	1,8%
Estrelas Gerais	16263	95,5%	16822	95,4%
Anãs Brancas	500	2,9%	500	2,8%
Total	17034	100%	17630	100%

Tabela 4.1: Distribuição das quantidades de objetos nas amostras de desenvolvimento de cada um dos modelos criados

riáveis consideradas como entrada para os modelos foram formadas pelas 12 magnitudes descritas nas subseções 3.1.2 e 3.2.1 com correções de extinção baseadas em mapas de extinção de SCHLAFLY e FINKBEINER (2011), junto com todas as combinações de cores possíveis de serem calculadas a partir delas. Já que cor é definida como a diferença entre duas magnitudes, o número de cores possível de ser calculado a partir de 12 magnitudes é 66, de modo que ao todo foram utilizados 78 variáveis em cada amostra de desenvolvimento.

Por fim, como o processo de desenvolvimento dos modelos necessita de amostras tanto de treino/validação quanto de teste, os objetos descritos na Tabela 4.1 foram divididos utilizando uma fração de 1/3 para teste e 2/3 para treino/validação. Com isso, foi possível realizar todos os passos de desenvolvimento dos FlACOs descritos a seguir.

4.1.2 Otimização de Hiperparâmetros

Para a otimização dos hiperparâmetros o método utilizado foi o da validação cruzada k-fold repetida descrita na subseção 2.3.1.1. Nesse caso o número de amostras escolhido foi de $k = 3$ e o método foi repetido 3 vezes, de modo que cada uma das combinações de hiperparâmetros testada foi treinada e validada em 9 combinações de amostra distintas.

4.1.2.1 Valores Testados

Hiperparâmetro	**Valores Testados**
bootstrap	[True, False]
max_features	[0,25, 0,50, 0,75, 1,0]
min_samples_leaf (msl)	[1, 5, 10, 20]
n_trees	[25, 50, 100]
m	[20, 40, 60, 78]
cutoff	[0,3, 0,5, 0,7, 0,9]

Tabela 4.2: Lista de hiperparâmetros otimizados e os seus valores considerados durante o desenvolvimento dos FlACOs para classificação de subanãs quentes nos levantamentos J-PLUS e S-PLUS.

Em relação aos hiperparâmetros otimizados foram considerados os seis indicados na

Tabela 4.2, onde também podem ser encontrados os valores testados de cada um através de uma busca em grade (explanada na seção 2.3.1 e exemplificada na Figura 2.5).

É importante ressaltar que, para os hiperparâmetros que não demonstram ter um impacto considerável na performance do modelo final (VAN RIJN e HUTTER, 2018), os valores escolhidos foram os padrões da biblioteca *scikit-learn* em sua versão 1.3 (PEDREGOSA *et al.*, 2011). Um desses hiperparâmetros foi o da função de impureza H, que para todos os modelos treinados neste trabalho foi considerada como a impureza de Gini descrita pela Equação 2.4.

Com isso, cada um dos dois modelos finais (um para o J-PLUS e outro para o S-PLUS) foi escolhido dentro de um conjunto de 1536 modelos, gerados a partir de todas as combinações possíveis entre os valores indicados na Tabela 4.2.

4.1.2.2 Métrica de Otimização

Para avaliar esses modelos a métrica escolhida foi o Score F1, que se baseia no *recall* e da precisão do modelo na amostra de validação. Para calcular esses dois valores, a performance do modelo é analisada a partir de quatro valores principais:

- **Verdadeiros Positivos (*True Positives*, TP)**: Número de objetos da classe positiva (subanãs quentes, ou 1) classificados pelo modelo como positivos;
- **Falsos Positivos (*False Positives*, FP)**: Número de objetos da classe negativa (estrelas que não são subanãs quentes, ou 0) classificados pelo modelo como positivos;
- **Verdadeiros Negativos (*True Negatives*, TN)**: Número de objetos da classe negativa classificados pelo modelo como negativos;
- **Falsos Negativos (*False Negatives*, FN)**: Número de objetos da classe positiva classificados pelo modelo como negativos.

É possível visualizar esses quatro valores numa representação denominada matriz de confusão, que é uma matriz $2x2$ onde as colunas representam as duas classes reais dos objetos, e as linhas representam as duas possíveis previsões do modelo. Esse tipo de representação é extremamente intuitiva, como se pode observar na Figura 4.1, e permite que a performance do modelo seja resumida e analisada facilmente.

Com isso, o *recall* e a precisão são definidos a partir das fórmulas:

$$recall = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{precisão} = \frac{TP}{TP + FP}. \tag{4.2}$$

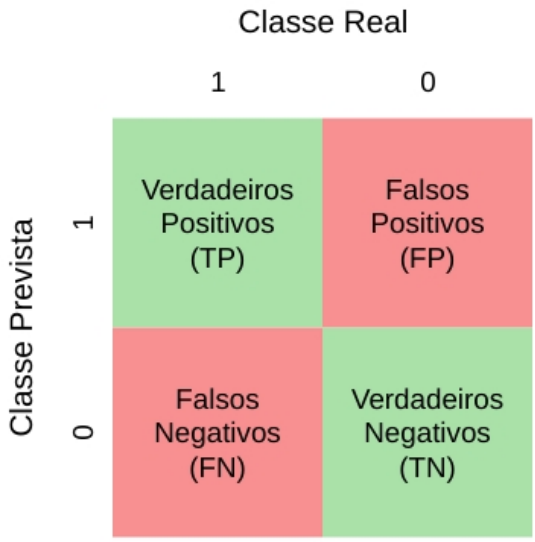

Figura 4.1: Matriz de confusão de um modelo de classificação, onde estão indicados os dois tipos de acertos (verdadeiros positivos e negativos) e os dois tipos de erros (falsos positivos e negativos).

Assim, o *recall* pode ser interpretado como a fração de subanãs da amostra de validação (TP + FN) que foram classificados corretamente pelo modelo (TP), enquanto que a precisão representa a fração de objetos classificados como subanãs pelo modelo (TP + FN) que realmente eram subanãs (TP).

A partir desses dois valores, é possível calcular o score F1:

$$F1 = \frac{2 \cdot \mathit{recall} \cdot \text{precisão}}{\mathit{recall} + \text{precisão}}. \tag{4.3}$$

Como se pode observar, o score F1 é definido como a média harmônica do *recall* e da precisão, de modo que a otimização de um modelo a partir dele é capaz de encontrar um balanço entre a capacidade do modelo de encontrar os casos positivos dentro da base (*recall*) e a capacidade de acertar suas previsões (precisão).

Além disso, como esse score é calculado a partir de duas métricas cujos valores variam entre 0 e 1, ele também sempre estará nesse intervalo, com valores próximos a 1 indicando os melhores modelos.

4.1.2.3 Análise de Performance Média

A partir da metodologia utilizada foram calculadas as performances dos 1536 modelos gerados a partir das combinações distintas de hiperparâmetros para cada um dos levantamentos. Uma primeira análise que pode ser gerada é a distribuição do score F1 de todos os modelos com um certo valor de um hiperparâmetro, o que permite que se observe o impacto individual desse hiperparâmetro na performance final do modelo.

Considerando o hiperparâmetro *bootstrap*, a Figura 4.2 indica que não houve uma

diferença entre seu impacto nos modelos do S-PLUS e do J-PLUS. Em ambos os casos os modelos com *bootstrap* = $False$ e com *bootstrap* = $True$ (768 modelos com cada um desses valores) performaram de maneira similar tanto em relação à distribuição de scores F1 quanto na posição da mediana.

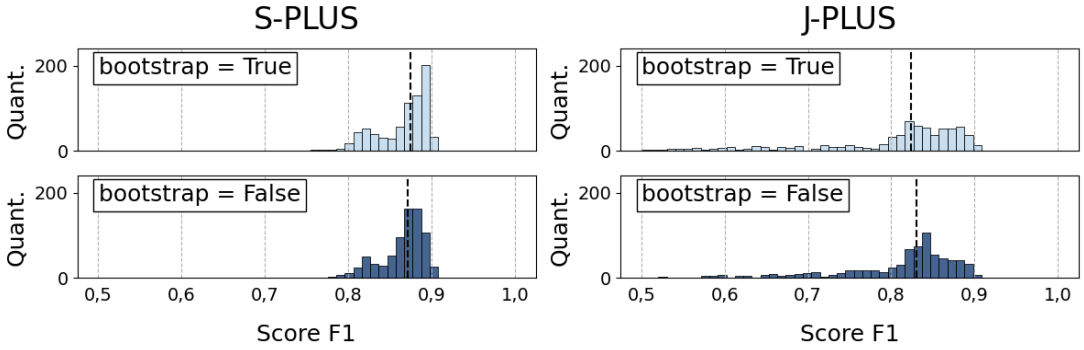

Figura 4.2: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro *bootstrap* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana do score F1 para cada distribuição.

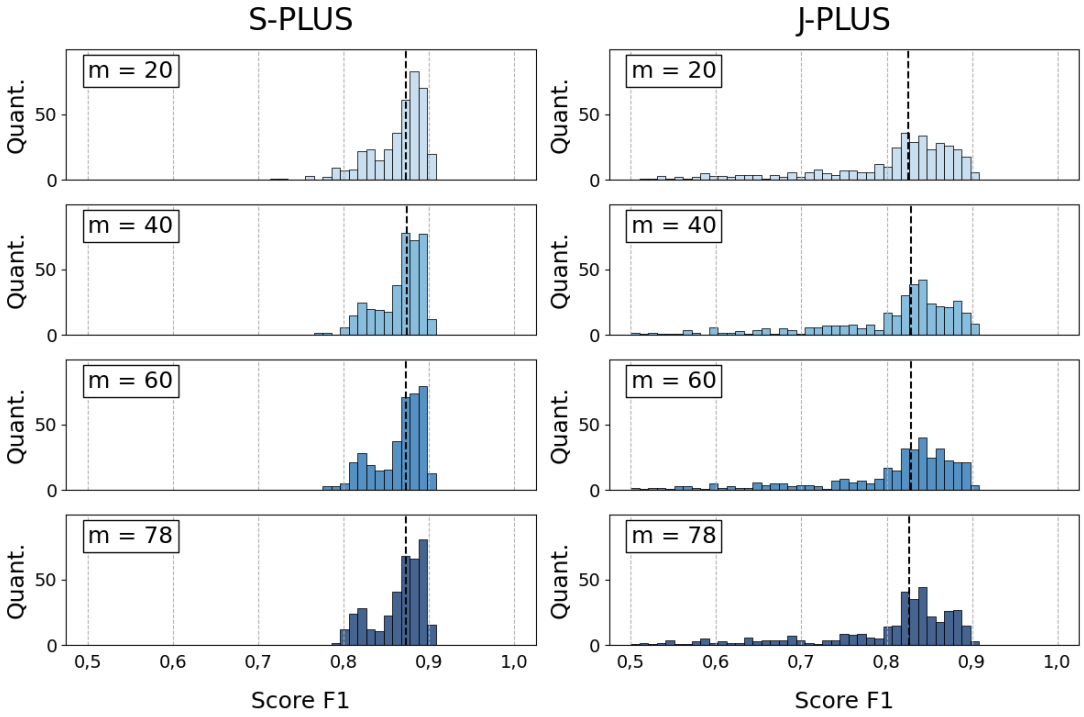

Figura 4.3: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro m para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana de score F1 para cada distribuição.

Além disso, é notável também o fato de que os scores F1 dos modelos treinados com dados do S-PLUS tiveram uma tendência de serem mais altos do que os scores do modelos do J-PLUS, algo que fica claro ao se comparar as medianas nos dois casos.

Em seguida, como se pode observar na Figura 4.3, tanto no J-PLUS quanto no S-PLUS não houve variação significativa de performance entre os modelos treinados com os quatro valores de m testados (20, 40, 60, 78).

Isso indica que as $m = 20$ variáveis mais importantes já trazem quase toda a informação possível para os modelos, e qualquer quantidade extra adicionada não interfere na performance. A conclusão que se pode extrair desse comportamento é que uma fração considerável das variáveis utilizadas como entrada não são informativas, ou que elas trazem as mesmas informações repetidas para a modelagem. Devido à maneira como as 66 cores foram adicionadas à amostra esse tipo de comportamento é esperado, já que todas as variáveis foram criadas a partir de apenas 12 magnitudes.

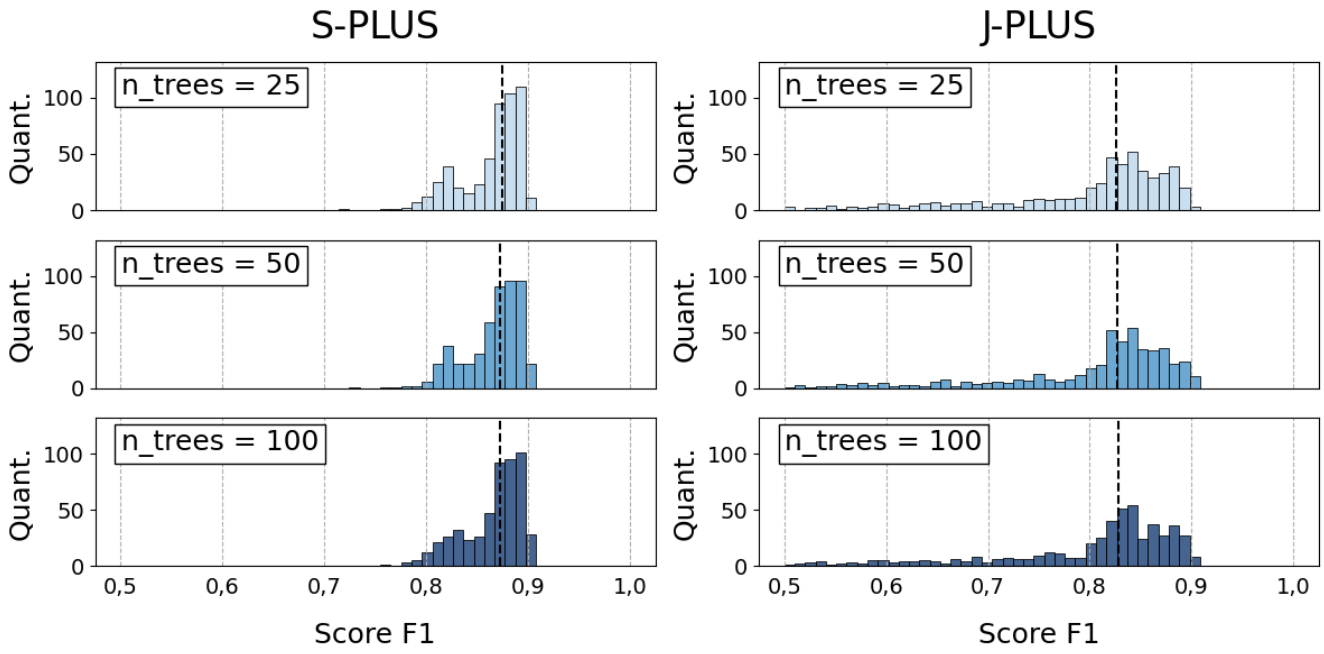

Figura 4.4: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro *n_trees* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana de score F1 para cada distribuição.

Em relação ao número de árvores dentro das florestas também fica claro, pela Figura 4.4, que não houve variação significativa entre os modelos treinados com 25, 50 e 100 árvores. Isso indica que a adição de mais árvores às florestas não melhora a performance dos modelos finais, de modo que a diferença entre um modelo com 100 árvores e um com 25 ou 50 é apenas o tempo de treinamento, que cresce junto com *n_trees*.

Já para o hiperparâmetro min_samples_leaf, como pode se observar na Figura 4.5, o impacto dos diferentes valores testados nos modelos finais foi mais notável. Em ambos os casos, mas principalmente no J-PLUS, a performance dos modelos foi inversamente proporcional ao valor de *min_samples_leaf*, de modo que os melhores modelos em ambos os casos foram aqueles treinados com $min\_samples\_leaf = 1$.

Como esse hiperparâmetro controla o número mínimo de objetos em uma folha da

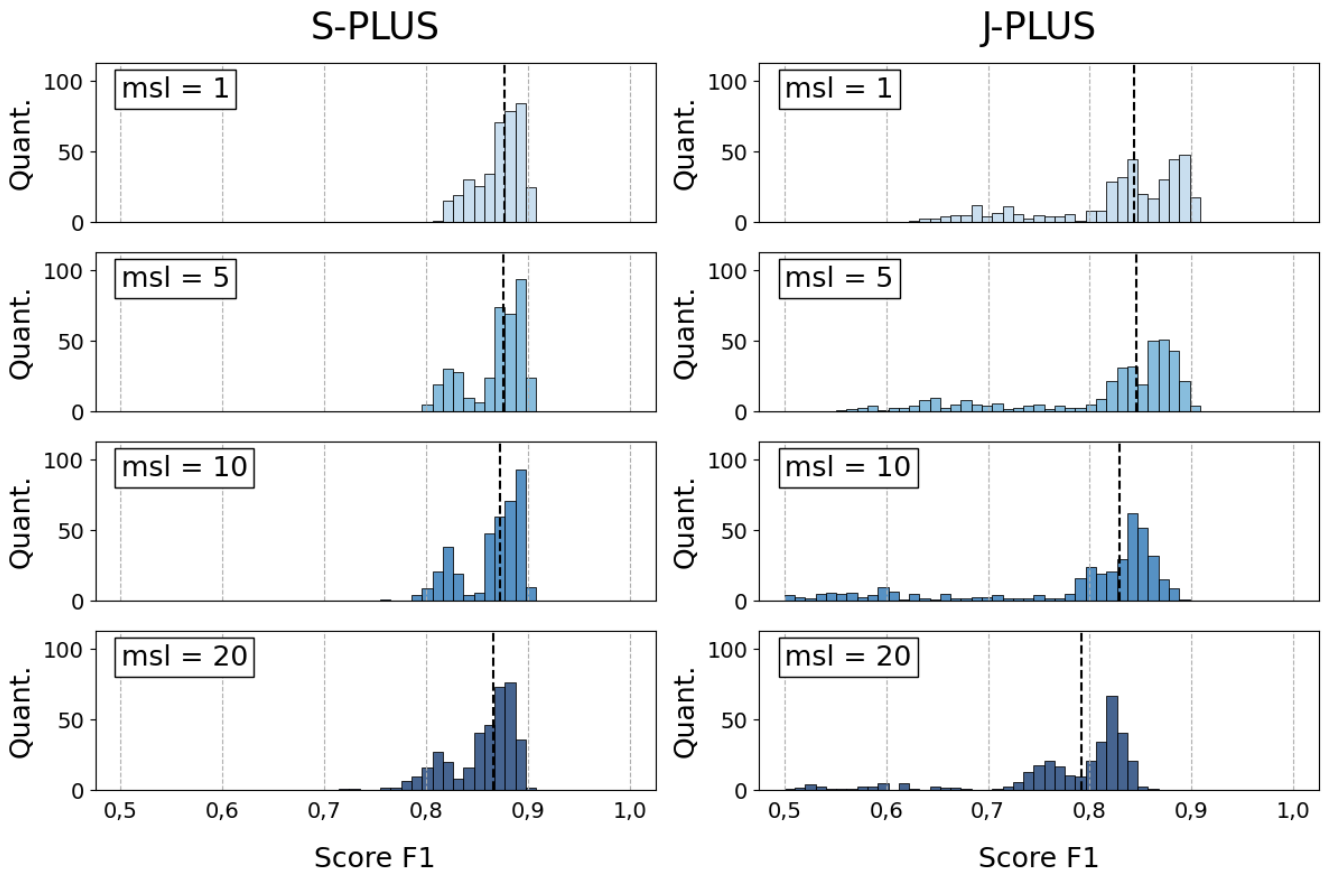

Figura 4.5: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro *min_samples_leaf* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana de score F1 para cada distribuição.

árvore para que ela seja considerada válida, um valor de $min\_samples\_leaf = 1$ resulta em um modelo onde a melhor divisão possível em seus nós é sempre a escolhida, independente da quantidade de objetos direcionados para cada folha.

Além disso, como os modelos foram treinados com uma quantidade pequena de subanãs quentes, valores altos de *min_samples_leaf* podem acabar impossibilitando a criação de folhas com uma fração alta de objetos positivos (já que elas necessariamente teriam poucos objetos no total), que são as responsáveis por classificar as subanãs no modelo final.

Já considerando o hiperparâmetro *max_features*, é possível observar na Figura 4.6 uma relação inversamente proporcional entre o valor de *max_features* e a dispersão de scores F1 dos modelos gerados a partir desse valor. Devido a isso, os modelos com $max\_features = 0,25$ apresentam uma maior variação em seus scores F1, de modo que tanto os melhores quanto os piores modelos foram treinados com esse valor de hiperparâmetro.

Essa relação está relacionada com o fato de *max_features* controlar a fração de variáveis que cada árvore da floresta considera durante seu treinamento. Para valores altos desse hiperparâmetro, cada uma das árvores vai observar uma fração alta das variáveis disponíveis, o que acaba gerando modelos com árvores mais similares entre si, e diminui

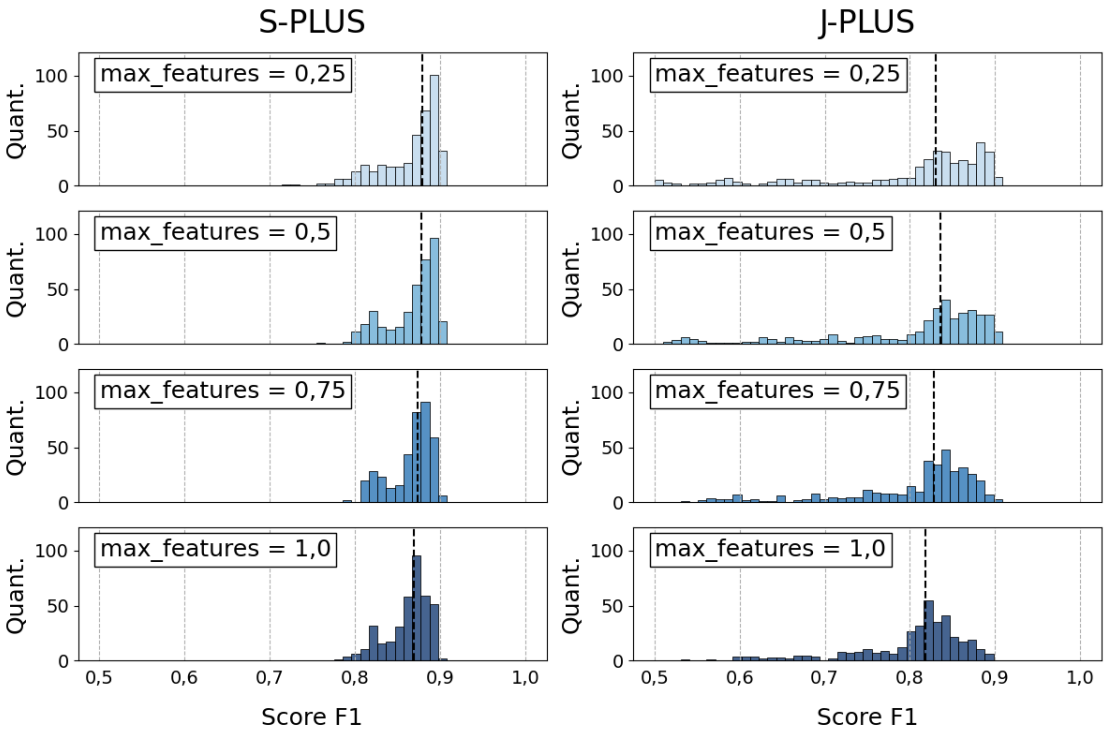

Figura 4.6: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro *max_features* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana de score F1 para cada distribuição.

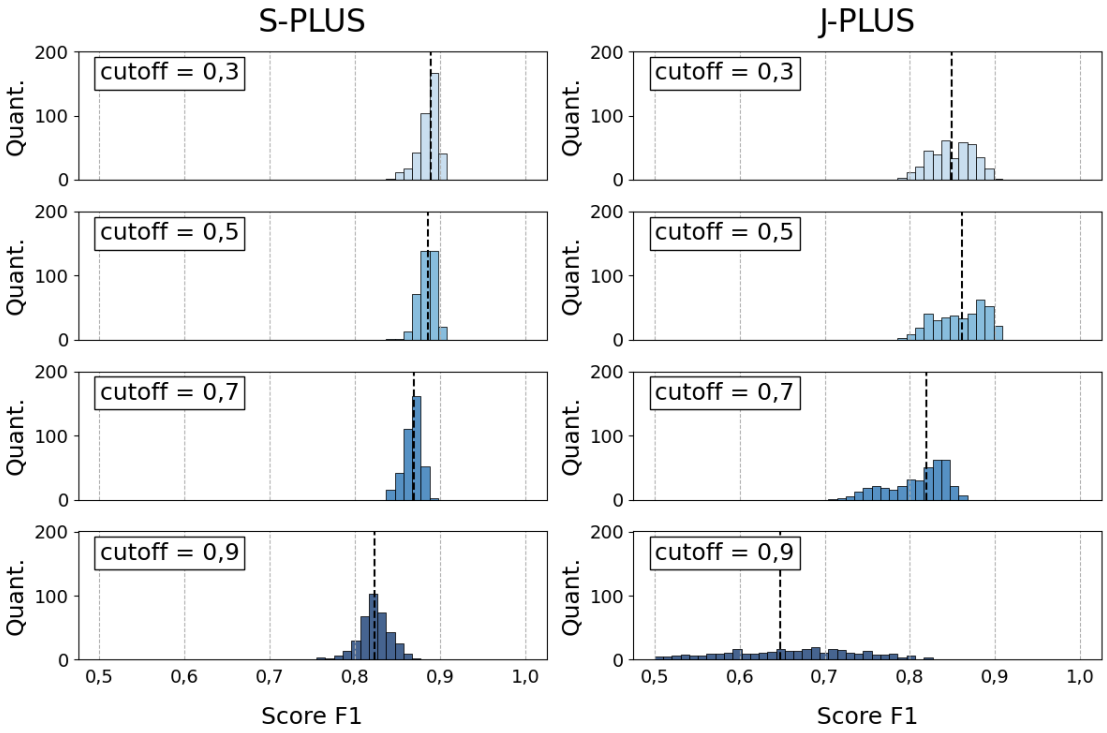

Figura 4.7: Distribuição dos scores F1 dos modelos de classificação em função do hiperparâmetro *cutoff* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam a mediana de score F1 para cada distribuição.

a variação de modelo para modelo.

De maneira inversa, para valores baixos de *max_features* cada árvore dentro das florestas é treinada com uma fração pequena das variáveis disponíveis. Isso favorece modelos com árvores distintas entre si, de modo que a variação entre modelos tende a aumentar.

Por fim, o último hiperparâmetro otimizado foi o *cutoff*, que representa o valor limite de probabilidade para um certo objeto ser classificado como subanã quente. Esse valor interfere diretamente nas classificações dadas pelo modelo, de modo que ele aparenta ser o parâmetro com maior impacto na performance, como pode se observar na Figura 4.7.

A grande importância da escolha correta do *cutoff* está no fato de que valores baixos para esse parâmetro fazem com que os modelos classifiquem muito mais objetos como subanãs quentes, o que aumenta o *recall* (pois mais subanãs são classificadas corretamente) mas diminui a precisão (pois mais estrelas são classificadas como subanãs incorretamente). No entanto, valores altos de *cutoff* aumentam a precisão (classificando como subanãs apenas os objetos com maior score) mas diminuem o *recall* (pois qualquer subanã com um score mais baixo será classificada erroneamente). Devido a isso, uma métrica como o score F1 permite que os modelos sejam otimizados com esse comportamento em mente.

Tanto para o J-PLUS quanto para o S-PLUS os modelos que em média apresentaram melhores resultados foram aqueles treinados com $cutoff = 0,3$ e $cutoff = 0,5$, de modo que esses pontos aparentam ser os que melhor conseguem equilibrar as perdas e ganhos associados a cada valor *cutoff*.

4.1.2.4 Análise de Performance Individual

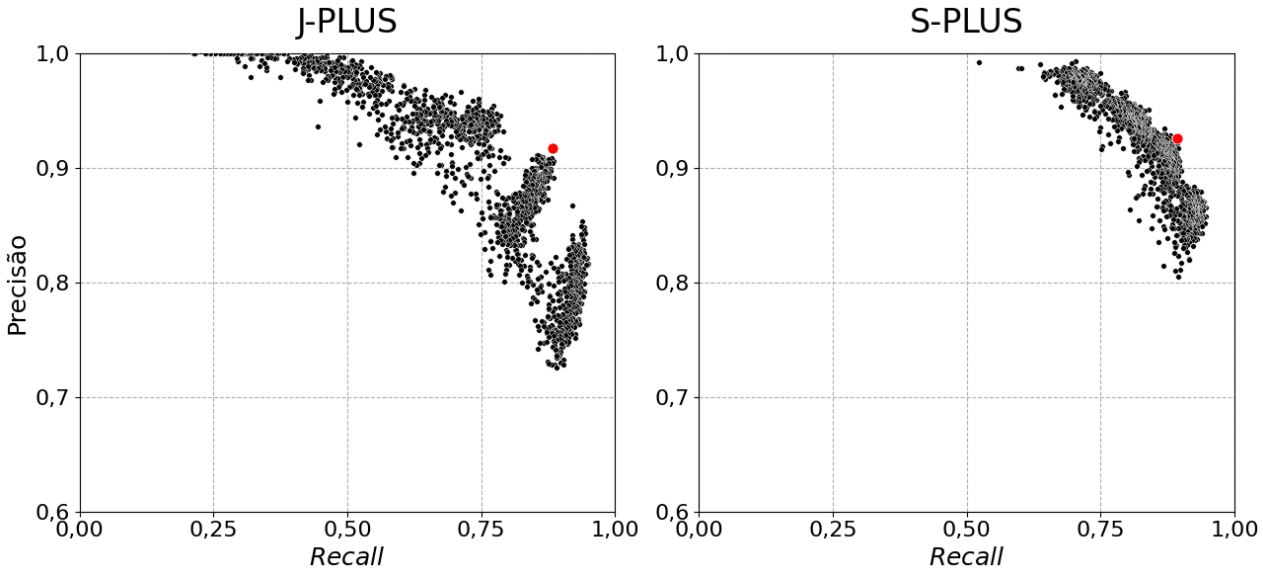

Figura 4.8: Distribuição de precisão e *recall* das 1536 combinações de hiperparâmetros testadas para cada um dos levantamentos. Em vermelho, estão indicadas as combinações com maior valor de score F1.

Apesar de ser possível fazer um estudo mais geral a partir das distribuições analisadas nos parágrafos anteriores, a escolha final do melhor modelo precisa ser feita a partir das performances individuais de cada combinação de hiperparâmetros.

Essas performances podem ser analisadas mais especificamente a partir da Figura 4.8, onde os valores de precisão e *recall* de cada uma das 1536 estão indicados. Como se pode observar, os modelos de maior score F1 (indicados em vermelho nos gráficos da Figura 4.8) ocupam regiões mais próximas do canto superior direito dos gráficos, que corresponde à valores altos tanto de *recall* quanto de precisão.

m	**n_trees**	**msl**	**bootstrap**	**max_features**	**cutoff**	**Score F1**
60	100	1	False	0.25	0.5	0.907
78	100	1	False	0.25	0.5	0.907
20	100	1	False	0.25	0.5	0.907
40	50	1	True	0.25	0.5	0.906
78	50	1	False	0.25	0.5	0.905
60	100	1	False	0.25	0.3	0.904
78	100	1	True	0.75	0.5	0.903
60	25	1	False	0.25	0.3	0.902
20	100	1	False	0.25	0.3	0.902
60	25	1	True	0.25	0.5	0.901

Tabela 4.3: Combinações de hiperparâmetros dos dez melhores modelos em relação ao seu score F1 na amostra de validação do J-PLUS.

Para o caso do J-PLUS, as dez combinações com os maiores valores de score F1 obtidos através da validação cruzada podem ser encontradas na Tabela 4.3. Com relação à performance desses modelos, todos eles obtiveram valores altos ($> 0,890$) de F1, o que indica que em todos esses casos eles foram capazes de capturar as informações relevantes dentro dos dados de treinamento e realizar boas previsões na amostra de validação.

Como esperado, os quatro hiperparâmetros que mais impactaram as distribuições de score F1 dos modelos treinados com o J-PLUS (*msl*, *bootstrap*, *max_features*, *cutoff*) foram aqueles que mais tiveram valores dominantes nas dez combinações de melhor performance, com:

- 10 dos 10 melhores modelos tendo $msl = 1$;
- 7 dos 10 melhores modelos tendo $bootstrap = False$;
- 9 dos 10 melhores modelos tendo $max\_features = 0,25$;
- 7 dos 10 melhores modelos tendo $cutoff = 0,5$.

Para *msl*, *bootstrap* e *cutoff* os valores dominantes indicados acima foram aqueles cujos modelos apresentaram uma maior média de score F1 na análise realizada anteriormente, o que corrobora as análises realizadas até aqui. Além disso, apesar dos modelos treinados com $max\_features = 0,25$ terem apresentado uma mediana de score F1 menor que os outros no J-PLUS, o fato deles terem uma maior variação entre si explica o fato de haver uma grande fração deles no grupo de melhores modelos.

Além disso, também de maneira similar ao observado durante a análise de distribuição de scores F1, houve um equilíbrio nos valores de m na Tabela 4.3. Por fim, apesar dos diferentes valores de n_trees não trazerem uma diferença perceptível na performance dos modelos, houve uma certa dominância do valor $n\_trees = 100$ na Tabela 4.3. Isso pode significar que apesar de não alterar o comportamento geral dos modelos de maneira considerável, valores mais altos desse hiperparâmetro trazem performances melhores.

m	n_trees	msl	bootstrap	max_features	cutoff	Score F1
40	100	1	False	0.25	0.3	0.909
20	25	1	True	0.25	0.5	0.906
60	100	1	True	0.50	0.3	0.906
60	50	1	True	0.25	0.5	0.905
78	50	1	False	0.50	0.3	0.904
60	100	5	True	0.25	0.3	0.904
20	25	10	True	0.25	0.3	0.904
20	50	1	False	0.25	0.3	0.904
60	100	5	False	0.25	0.3	0.903
20	25	10	False	0.25	0.5	0.903

Tabela 4.4: Combinações de hiperparâmetros dos dez melhores modelos em relação ao seu score F1 na amostra de validação do S-PLUS.

Considerando os 10 melhores modelos para o levantamento S-PLUS indicados na Tabela 4.4 nota-se que assim como observado durante a análise de performance média, os scores F1 obtidos com o S-PLUS foram maiores do que os obtidos com o J-PLUS. Isso indica que todos esses modelos também foram treinados com sucesso, e suas previsões são capazes de diferenciar as subanãs quentes do restante dos objetos na amostra.

Além disso, os três hiperparâmetros (*msl*, *max_features* e *cutoff*) que mais demonstraram impactar a performance média dos modelos do S-PLUS também tiveram valores dominantes na Tabela 4.4, com:

- 6 dos 10 melhores modelos tendo $msl = 1$;
- 8 dos 10 melhores modelos tendo $max\_features = 0,25$;

- 7 dos 10 melhores modelos tendo $cutoff = 0,3$.

Para *msl* e *cutoff* os valores dominantes indicados acima foram aqueles cujos modelos apresentaram uma maior média de score F1 na análise realizada anteriormente, o que corrobora as análises realizadas até aqui.

Além disso, apesar de não ter havido uma diferença expressiva entre as médias de score F1 dos modelos treinados com os diferentes valores de *max_features* testados, o fato dos modelos com $max\_features = 0,25$ terem uma maior variação entre si explica o fato de haver uma grande fração deles no grupo de maior performance.

Também de maneira similar ao observado durante a análise de distribuição de scores F1, houve um equilíbrio nos valores de *m* e *bootstrap* na Tabela 4.4. Como esses hiperparâmetros não aparentaram causar um impacto considerável nas distribuições, é de se esperar que nenhum dos seus valores seja dominante na lista de melhores combinações.

Por fim, assim como no caso do J-PLUS, os modelos treinados com $n\_trees = 100$ dominaram o grupo de maior performance. Aqui novamente é possível levantar a hipótese de que apesar de não alterar o comportamento geral dos modelos de maneira considerável, valores mais altos desse hiperparâmetro trazem performances levemente melhores.

Levantamento	m	n_trees	msl	bootstrap	max_features	cutoff
J-PLUS	60	100	1	False	0,25	0,5
S-PLUS	40	100	1	False	0,25	0,3

Tabela 4.5: Combinações de hiperparâmetros escolhidas para treinar os FlACOs do S-PLUS e do J-PLUS.

Com isso, através de uma análise dos melhores scores F1 de cada uma das combinações em ambos os levantamentos, as combinações escolhidas para treinar os modelos FlACOs finais foram as indicadas na Tabela 4.5.

Por fim, utilizando apenas as duas melhores combinações também foi testado o impacto do hiperparâmetro *class_weight*, responsável por atribuir um peso a cada classe dentro da amostra (positivos e negativos). Por padrão o algoritmo do sklearn considera cada classe como tendo o peso 1, mas caso $class\_weight = balanced$ o peso das classes da amostras é definido de acordo com sua representatividade, o que pode auxiliar em problemas com alto desbalanceamento. No entanto, o uso desse hiperparâmetro com seu valor "balanceado"não trouxe melhora na performance em nenhum dos dois casos.

4.1.3 Performance dos Modelos

A partir das combinações definidas na seção anterior, foram treinados FlACOs com as amostras de treino/validação de ambos os levantamentos considerados até aqui. Após esse treinamento, as performances desses modelos foram avaliadas nas amostras de testes que nenhum deles tinha considerado até então.

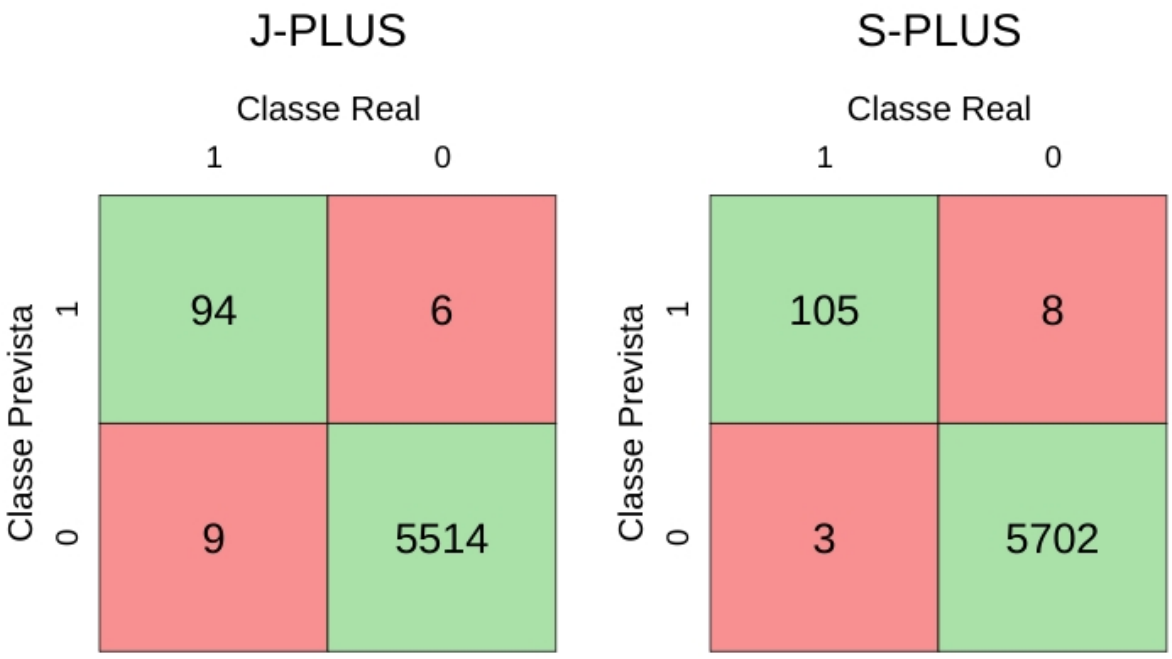

Figura 4.9: Matrizes de confusão dos modelos otimizados para os levantamentos J-PLUS e S-PLUS.

4.1.3.1 J-PLUS

Para o J-PLUS a amostra de teste era formada por 5623 objetos, dos quais 103 eram subanãs quentes confirmadas. Após o modelo calcular as probabilidades de cada uma das estrelas serem subanãs quentes, aqueles com probabilidade maior do que o *cutoff* escolhido de $0,5$ foram classificadas como candidatas à subanãs quentes.

Como pode se observar na Figura 4.9, entre as 103 subanãs da amostra de teste, o FlACOs foi capaz de classificar 94 corretamente, o que equivale a um *recall* de $94/103 = 0,91$. Além disso, dos 100 objetos que o modelo classificou como subanãs, 94 deles realmente eram dessa classe, o que equivale a uma precisão de $94/100 = 0,94$.

Em relação ao cálculo da precisão, é importante notar que apesar de 6 dos objetos classificados como subanãs quentes pelo modelo serem considerados 'erradas' pelo fato desses objetos não estarem no catálogo de subanãs quentes utilizados, é possível que eles façam parte desse grupo de estrelas, mas que ainda não tenham sido catalogados. Mesmo levando isso em conta, o valor de precisão reportado aqui é o mais conservador possível, e considera que qualquer objeto fora do catálogo não é uma subanã quente.

Substituindo esses dois valores na equação 4.3 obtem-se um valor para o score F1 de $0,93$, que está ligeiramente abaixo do obtido durante o passo de otimização dos hiperparâmetros.

Essa diferença pode estar relacionada com algum viés presente nas amostras de treino/validação, de modo que os modelos sofreram *overfit* (situação na qual o modelo performa muito bem na base de treino, mas não mantém essa performance em bases no-

vas) durante esse passo, algo que já é de certo modo esperado em casos onde se trabalha com tão poucos objetos (GOODFELLOW *et al.*, 2016).

No entanto, mesmo com essa variação de performance do modelo, a métrica de score F1 obtida ainda é satisfatória, e indica que o modelo pode ser aplicado nos objetos observados pelo J-PLUS para encontrar candidatas à subanãs dentro desse levantamento.

4.1.3.2 S-PLUS

Já para o S-PLUS, a amostra de teste era formada por 5818 objetos no total, dos quais 108 eram subanãs quentes confirmadas. Assim como no caso do J-PLUS, todos os objetos nessa amostra foram escorados pelo modelo, e aqueles com score maior do que $0,3$ foram classificados como subanãs.

Das 108 subanãs presentes na amostra de teste, 105 foram classificadas de maneira correta pelo modelo, o que equivale a um *recall* de $105/108 = 0,97$. Além disso, dentre os 113 objetos que o modelo classificou como subanãs quentes, 105 deles eram realmente dessa classe, o que equivale a uma precisão de $105/113 = 0,93$. Calculando o score F1 pela equação 4.3, obtem-se um valor de $0,95$.

Diferentemente do modelo do J-PLUS, o valor de F1 calculado para o modelo do S-PLUS foi maior do que o valor obtido durante a otimização dos hiperparâmetros, o que pode estar diretamente relacionado com o desbalanceamento das amostras de treino e teste. Com isso, o modelo desenvolvido para o S-PLUS também se mostra capaz de ser aplicado no restante dos objetos do levantamento para encontrar candidatas à subanãs quentes.

4.1.4 Importância das Variáveis

Variáveis J-PLUS	Variáveis S-PLUS
(J0395 - J0515)	**(J0395 - J0515)**
(J0410 - J0515)	**(J0430 - J0515)**
(J0430 - J0515)	**(J0410 - J0515)**
(J0395 - J0660)	(J0395 - rSDSS)
(J0430 - zSDSS)	(J0410 - gSDSS)

Tabela 4.6: Listas de 5 variáveis mais importantes em cada um dos dois FlACOs treinados por ordem de importância, com as variáveis em comum marcadas em negrito.

Como cada modelo foi treinado com as mesmas 78 variáveis (12 magnitudes e 66 cores calculadas), uma possível comparação a ser realizada envolve o cálculo da importância de cada variável nos dois diferentes modelos, através da metodologia descrita na subseção 2.2.3.

Ao comparar as 5 variáveis mais importantes em cada modelo nota-se que 3 delas apareceram nas duas listas, o que já é esperado considerando que ambos os modelos foram treinados para classificar o mesmo tipo de objeto, e que suas amostras de treino foram geradas com a mesma metodologia.

Como se pode observar na Tabela 4.6, todas as 5 variáveis mais importantes em cada um dos modelos são cores. Essa dominância vem do fato de cores fornecerem informação sobre a relação entre duas magnitudes, e as diferenças entre duas classes de objetos se manifestam muito mais nessas relações do que em magnitudes individuais (que variam também com a distância do objeto até o observador).

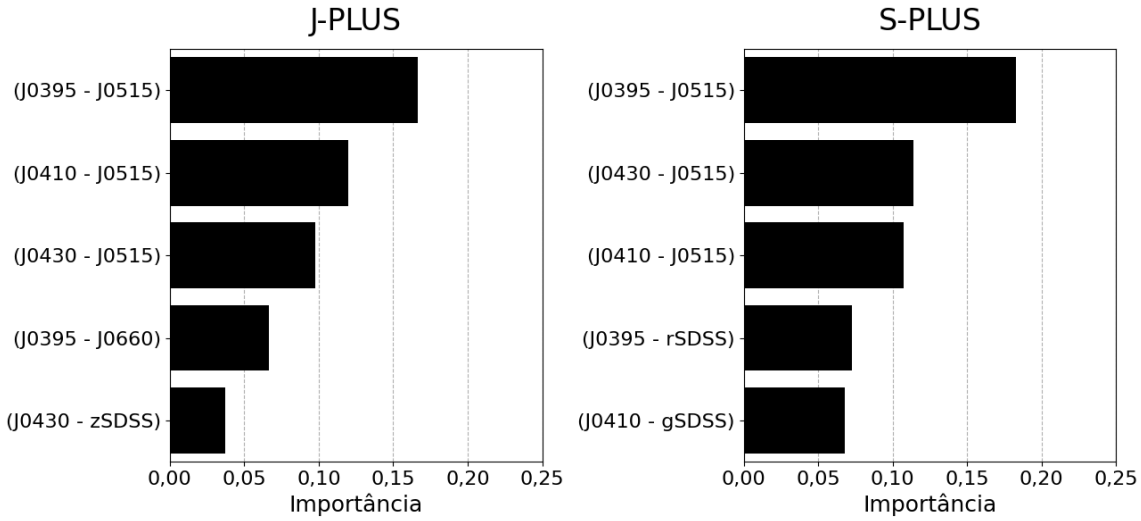

Figura 4.10: Importâncias das 10 melhores variáveis dentro de cada um dos modelos treinados.

Através da metodologia descrita na seção 2.2.3 foi possível calcular as importâncias de cada uma das 20 variáveis de entrada nas florestas aleatórias treinadas para o FlACOs do S-PLUS e do J-PLUS, e as 10 mais importantes de cada modelo podem ser encontradas na Figura 4.10.

Um primeiro ponto a ser levantado a respeito desses *rankings* é o fato de para ambos os modelos as três variáveis mais importantes são as mesmas, mais especificamente as cores (J0410 - J0515), (J0395 - J0515) e (J0430 - J0515). Essa é uma observação que demonstra concretamente a importância dessas variáveis para a identificação de estrelas subanãs, já que dois modelos distintos treinados com dados de levantamentos distintos as apontaram como cruciais em suas classificações.

Para explicar melhor o impacto das cores na classificação de uma subanã quente, a Figura 4.11 mostra o espectro observado pelo levantamento SDSS (YORK *et al.*, 2000b) para estrelas da sequência principal de diferentes classes espectrais e para os dois tipos de subanãs quentes.

Como se pode observar, as regiões do espectro correspondente aos três filtros envolvidos no cálculo das duas cores mais importantes dos modelos (J0395, J0410 e J0515, indicados

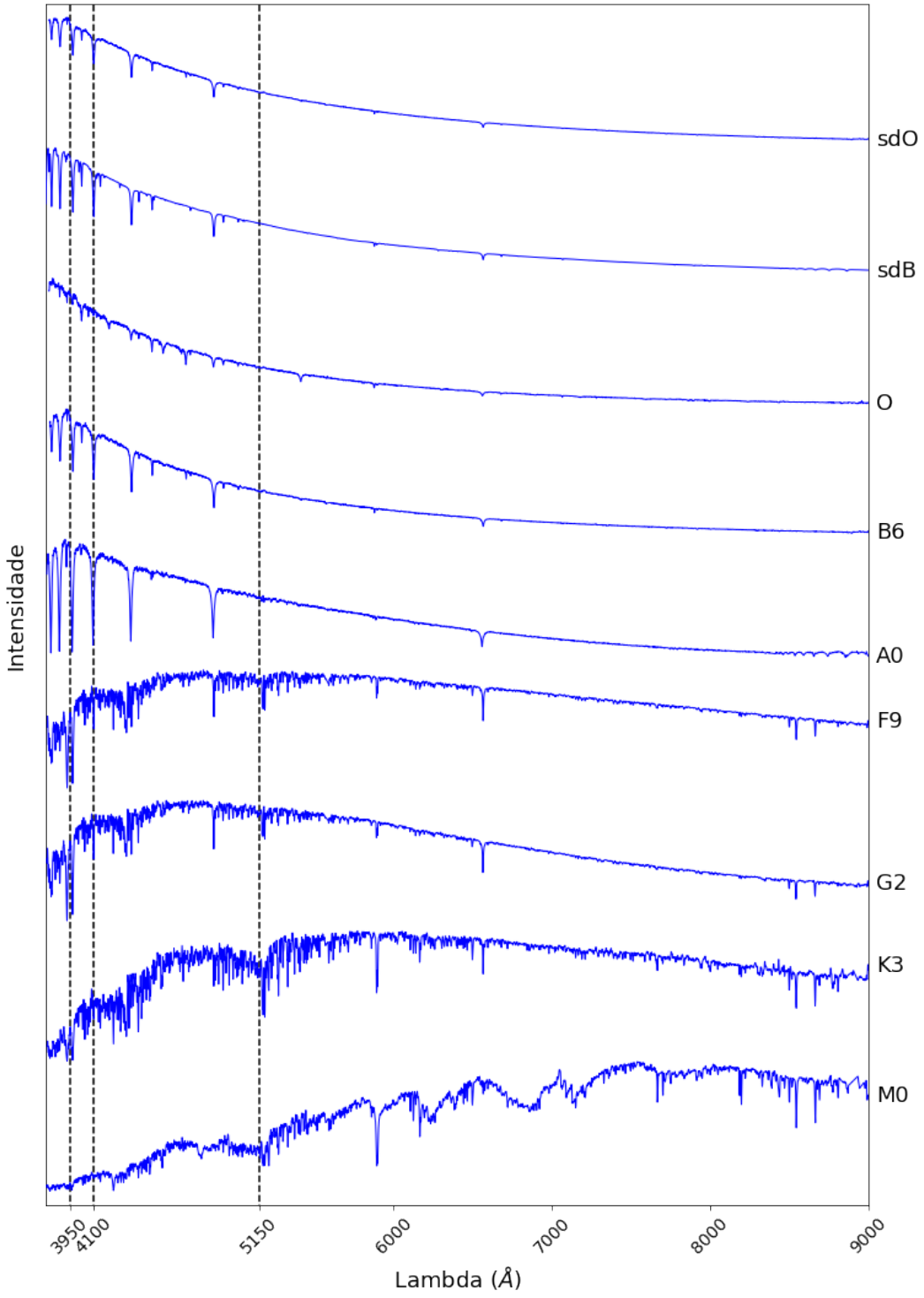

Figura 4.11: Exemplos de espectros corrigidos para a extinção do SDSS (Levantamentos Legacy, SEGUE e BOSS: YORK *et al.* 2000a; YANNY *et al.* 2009; DAWSON *et al.* 2013) para estrelas da sequência principal de diferentes classes espectrais. As linhas pontilhadas indicam a região central dos três filtros mais importantes para ambos os modelos FlACOs (J0395, J0410 e J0515).

pelas linhas pontilhadas na figura) apresentam um comportamento relacionado à classe espectral dos objetos:

- Para objetos mais quentes (sdO, sdB, O, B e A) as regiões correspondentes aos filtros J0395 e J0410 são mais intensas que a região correspondente ao filtro J0515;
- Para o restante dos objetos (F, G, K, M) as regiões dos filtros J0395 e J0410 têm intensidades iguais ou menores do que a região do filtro J0515.

Com isso, nota-se que as subanãs quentes (que como citado anteriormente, são das classes O e B) têm mais intensidade nos filtros J0395 e J0410 do que no filtro J0515. Como a magnitude fotométrica tem uma relação de proporção (não-linear) com o fluxo observado, e quanto maior o fluxo menor a magnitude fotométrica, essas relações entre os filtros nas subanãs serão:

$$J0395 < J0515 \rightarrow (J0395 - J0515) < 0 \tag{4.4}$$

$$J0410 < J0515 \rightarrow (J0410 - J0515) < 0 \tag{4.5}$$

Já para a amostra de estrelas gerais escolhidas aleatoriamente dentro dos levantamentos, espera-se que haja uma grande fração de objetos dos tipos espectrais F, G, K e M, de modo que, baseando-se nos exemplos de espectros desses objetos na Figura 4.11, o grupo terá muito mais objetos com valores positivos das cores (J0395 - J0515) e (J0410 - J0515). Para analisar essa hipótese, as distribuições dessas cores para as subanãs quentes e os demais objetos das amostras de desenvolvimento podem ser analisadas.

Como se pode observar nas Figuras 4.12 e 4.13, as subanãs quentes se concentram em valores negativos das cores (J0410 - J0515) e (J0395 - J0515), enquanto que o restante da amostra tem valores mais distribuídos e majoritariamente positivos para as duas cores. Esse comportamento ocorre tanto no J-PLUS quanto no S-PLUS, de modo que a hipótese levantada a partir da análise das classes espectrais e representada pelas Equações 4.4 e 4.5 é corroborada.

No entanto, é necessário também considerar que as relações descritas acima não são as únicas por trás da classificação realizada pelos FlACOs, e que os valores de outras cores e magnitudes podem influenciar a probabilidade calculada pelo modelo para um certo objeto ser classificado como uma subanã quente. Devido à isso, é possível que uma estrela com cores que não obedeçam as Equações 4.4 e 4.5 seja classificada como subanã quente pelo modelo, algo que pode ser observado principalmente nos gráficos do J-PLUS nas Figuras 4.12 e 4.13.

Assim, a partir do estudo das métricas de performance dos dois modelos e das importâncias das variáveis dentro de cada um, foi possível validá-los, comprovar sua capacidade de classificação e trazer um certo grau de explicabilidade para o processo por trás de suas previsões. Com isso, se conclui a construção de uma base para justificar a aplicação dos FlACOs na identificação de candidatas à subanãs quentes dentro de ambos os levantamentos considerados.

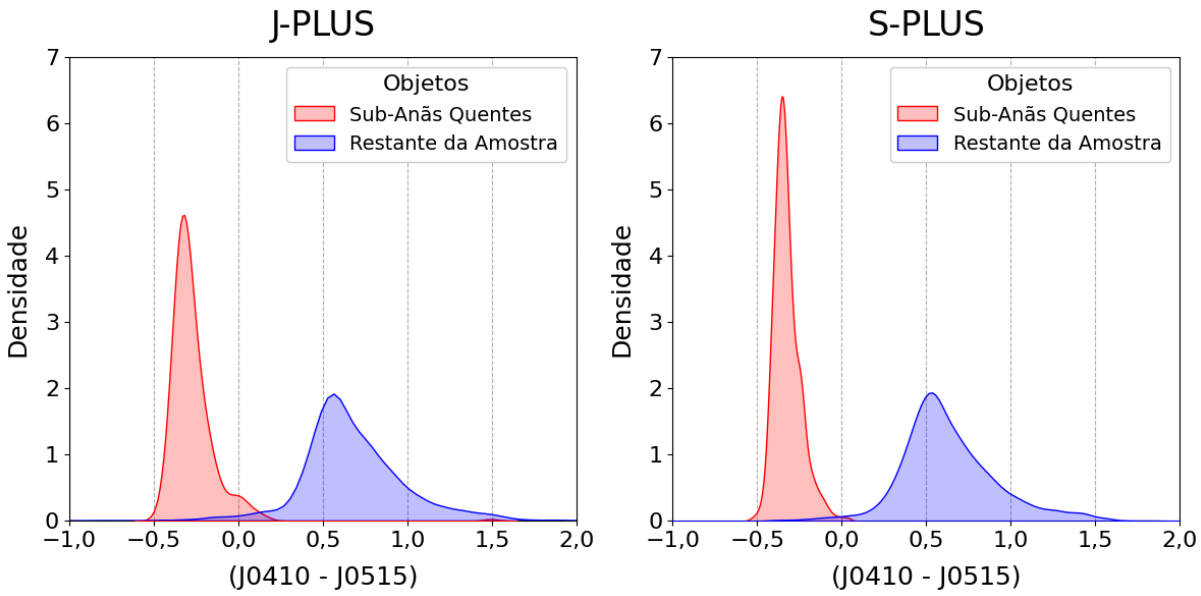

Figura 4.12: Distribuição das subanãs quentes e do restante da amostra de desenvolvimento em função dos seus valores na cor (J0410 - J0515).

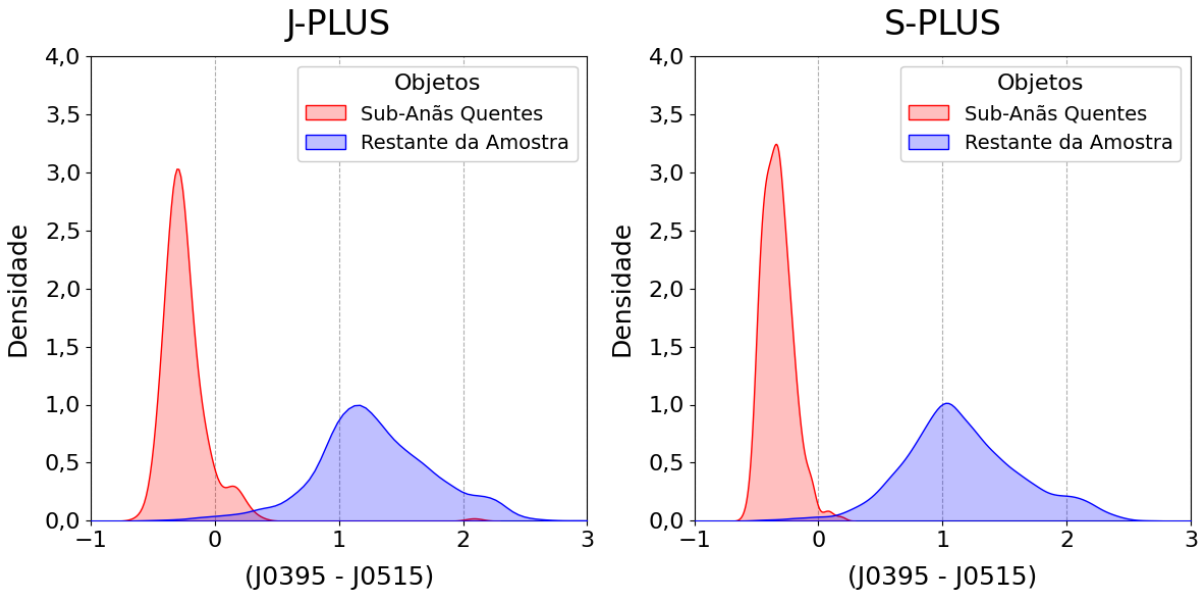

Figura 4.13: Distribuição das subanãs quentes e do restante da amostra de desenvolvimento em função dos seus valores na cor (J0395 - J0515).

4.2 Listas de Candidatas

Com todas as validações a respeito da metodologia proposta e análises realizadas a partir dos resultados dos modelos nas amostras de teste, os FlACOs foram utilizado para identificar candidatas à subanãs quentes ainda não catalogadas no S-PLUS e J-PLUS.

Para isso partiu-se das amostras completas de objetos liberados nos Data Releases mais recentes de ambos os levantamentos (DR3 para o J-PLUS, DR4 para o S-PLUS) e aplicou-se o mesmo filtro de qualidade de magnitudes utilizado na criação das amostras

de desenvolvimento dos modelos (ignorando assim qualquer objeto com um erro maior do que 0,3 em qualquer magnitude).

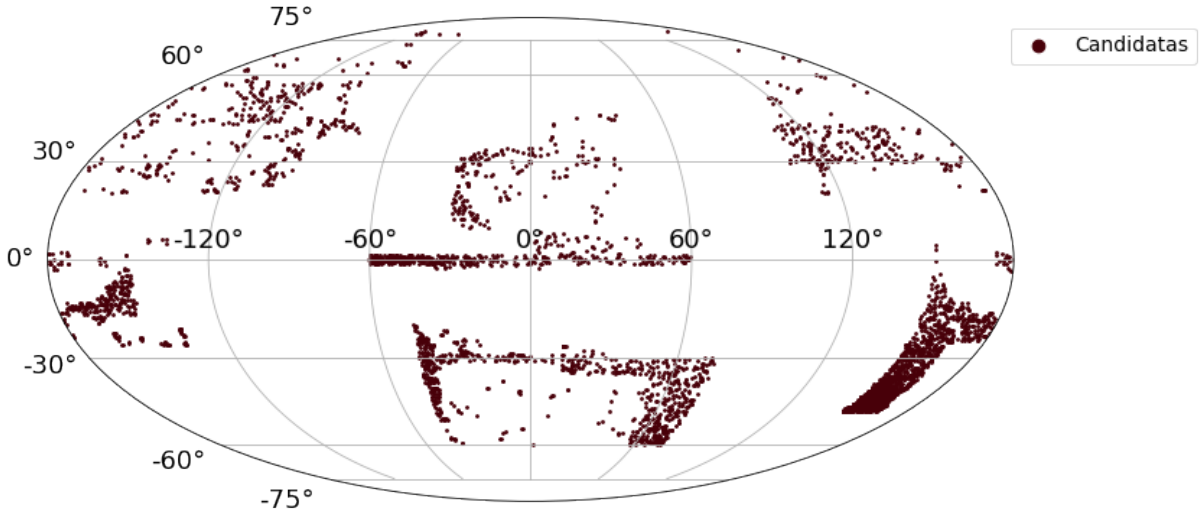

Figura 4.14: Distribuição das candidatas à subanãs quentes identificadas em dados do S-PLUS e J-PLUS.

Após as filtragens foram obtidas amostras com cerca de 5 milhões de objetos no S-PLUS e 3 milhões do J-PLUS. A probabilidade de cada um desses objetos ser uma subanã quente foi então calculada, e qualquer um com uma probabilidade maior do que o *cutoff* definido para o levantamento durante a otimização de hiperparâmetros (0,5 para o J-PLUS e 0,3 para o S-PLUS) foi considerado como uma candidata à subanã quente.

Esse processo resultou em 2857 candidatas no S-PLUS e 987 no J-PLUS. Sua distribuição espacial pode ser observada na Figura 4.14, onde se constata que foram identificadas candidatas em todas as áreas observadas pelos levantamentos, indicadas nas Figuras 3.2 e 3.4.

Em relação à sua catalogação, das 2857 candidatas apontadas pelo modelo de classificação dentro do S-PLUS, 339 delas estão no catálogo de subanãs quentes compilado por CULPAN *et al.* (2022). Assim, pode-se concluir que o FlACOs foi capaz de identificar 2518 novas candidatas à subanãs quentes. Além disso, das 308 subanãs utilizadas no treinamento/teste do modelo, 305 foram classificadas corretamente durante esse processo, o que representa 99% desses objetos.

Ao se aplicar a mesma análise nas candidatas do J-PLUS, conclui-se que 227 das 987 estrelas indicadas pelo modelo fazem parte do catálogo de subanãs quentes considerado durante o treinamento dos modelos. Assim, o número de novas candidatas identificadas nesse caso foi de 760. Considerando as 271 subanãs presentes nas amostras de treinamento/teste, 262 delas foram classificadas corretamente nesse processo, o que representa 97% desses objetos.

Assim, os dois grupos de novas candidatas foram unidos e a lista completa com as

ID	LEVANTAMENTO	RA	DEC	PROB_SUBANA_QUENTE
S-02099	S-PLUS	336.948	-30.550	0.360
S-01031	S-PLUS	73.600	-34.664	0.660
J-00527	J-PLUS	224.172	42.369	0.120
S-00719	S-PLUS	47.806	-41.561	0.830
S-00205	S-PLUS	152.323	-42.158	1.000
S-01396	S-PLUS	151.473	-46.724	0.530
S-01192	S-PLUS	160.585	-46.025	0.600
J-00077	J-PLUS	338.641	26.796	0.950
J-00107	J-PLUS	3.883	34.319	0.900
S-00615	S-PLUS	153.916	-23.248	0.890

Tabela 4.7: Amostra da lista de candidatas à subanãs no J-PLUS e S-PLUS desenvolvida durante esse trabalho.

3126 novas candidatas à subanãs quentes já se encontra disponível para consulta[1]. Uma amostra dessa lista de candidatas pode ser encontrada na Tabela 4.7, formada por uma coluna de identificação própria desse trabalho (ID), uma coluna listando o survey de origem da candidata (LEVANTAMENTO), duas colunas com a posição do objeto (RA e DEC) e uma coluna com a probabilidade calculada pelo FlACOs do objeto ser uma subanã quente (PROB_SUBANA_QUENTE).

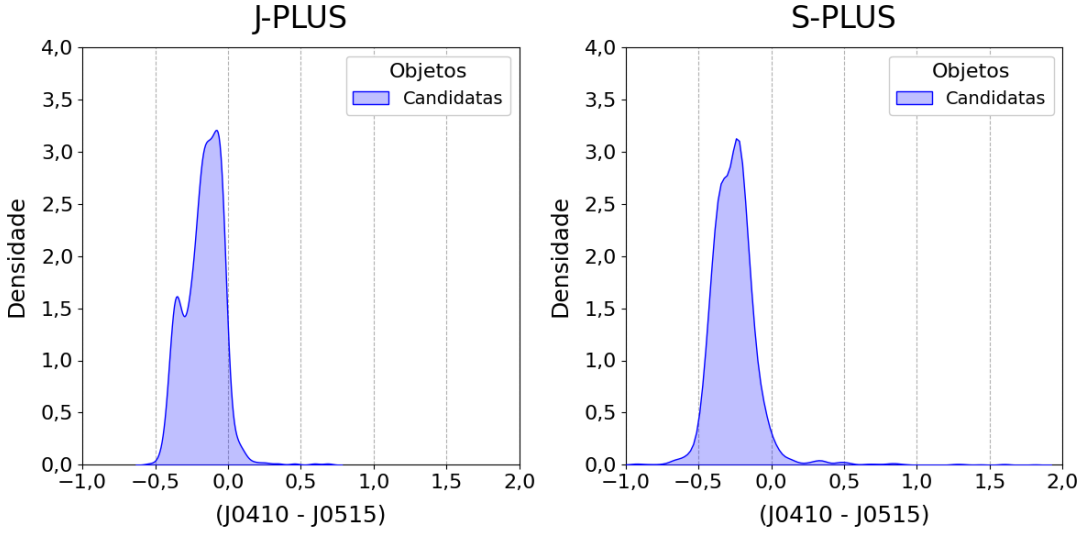

Figura 4.15: Distribuição das candidatas à subanãs quentes em função dos seus valores na cor (J0410 - J0515).

É possível também analisar as distribuições de valores para as cores mais importantes

[1] https://drive.google.com/drive/folders/1vAX2m1PaQNIY5vS5unqOi55W2l_zrbON

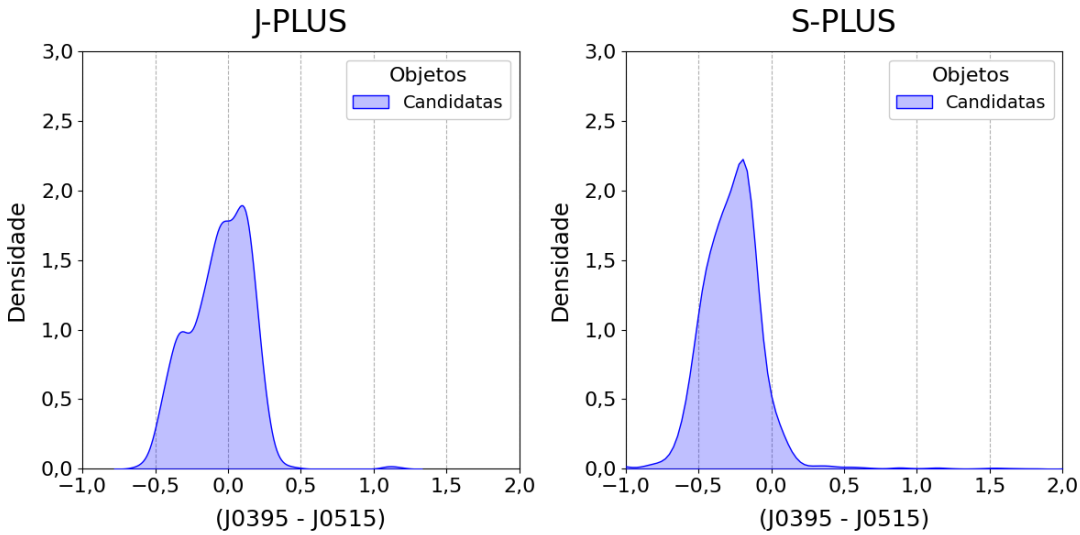

Figura 4.16: Distribuição das candidatas à subanãs quentes em função dos seus valores na cor (J0395 - J0515).

no processo de classificação dos modelos, a (J0410 - J0515) e a (J0395 - J0515), indicadas nas Figuras 4.15 e 4.16.

Em ambos os casos, é possível notar nas candidatas os mesmos picos em valores negativos das duas cores observados nas amostras de subanãs quentes já catalogadas (Figuras 4.12 e 4.13).

Além disso, observa-se também que uma fração das candidatas do J-PLUS possui valores positivos dessas duas cores, algo que ocorreu em menor escala na amostra de treinamento mas se tornou mais expressivo nas candidatas. Isso pode estar relacionado ao fato já citado anteriormente de que essas duas cores não serem as únicas consideradas na probabilidade final calculada pelo modelo, de modo que outras cores e magnitudes podem ter influenciado na classificação dessa candidatas.

Considerando todos os pontos, isso indica que as candidatas apresentam, ao menos nessas duas cores, um comportamento similar às subanãs quentes catalogadas.

Além disso, com o intuito de confirmar que as candidatas selecionadas não são falsos positivos como estrelas variáveis cataclísmicas (CVs) ou simbióticas, as duas listas de candidatas foram cruzadas com catálogos desses dois tipos de objetos (BELCZYŃSKI *et al.* 2000; COPPEJANS *et al.* 2016). Através desses cruzamentos, foram encontradas sete candidatas em comum com o catálogo de variáveis cataclísmicas e nenhuma em comum com o catálogo de estrelas simbióticas (indicadas na Tabela 4.8). Assim, é possível atestar o fato de que em sua maioria, as candidatas selecionadas durante este trabalho não são falsos positivos de outros tipos de estrelas.

Por fim, é necessário se apontar o fato de que a identificação de um objeto como candidata à subanã quente não pode ser utilizada como confirmação dessa classificação. Do contrário, essa classificação só pode ser realizada a partir de observações espectroscópi-

ID	SURVEY	RA	DEC	PROB_HOT_SD	CRTS
S-00790	S-PLUS	155.176	-33.834	0.790	CRTS_J102042.1-335002
S-00949	S-PLUS	348.882	-30.814	0.690	CRTS_J231531.8-304848
S-01345	S-PLUS	352.252	-29.779	0.540	CRTS_J232900.4-294646
S-01680	S-PLUS	345.965	1.114	0.450	CRTS_J230351.6+010651
S-01706	S-PLUS	154.723	-40.112	0.450	CRTS_J101853.5-400643
S-01770	S-PLUS	351.464	-1.673	0.430	CRTS_J232551.4-014023
S-01879	S-PLUS	151.894	-20.292	0.410	CRTS_J100734.6-201732

Tabela 4.8: Candidatas à subanãs quentes presentes no catálogo de variáveis cataclísmicas compilado por COPPEJANS *et al.* 2016, identificadas tanto pelo ID da tabela de candidatas selecionadas neste trabalho (ID) quanto pelo ID do catálogo de CVs (CRTS).

cas dos objetos, e a lista de candidatas pode ser aplicada para definir de maneira mais assertiva os alvos desse tipo de observação.

4.3 Modelos para Previsão de Parâmetros Estelares

Como indicado na seção 2.4.2, um segundo passo de validação para a metodologia apresentada neste trabalho foi executado, nesse caso considerando dados de parâmetros de estrelas gerais do J-PLUS e do S-PLUS. Para isso, foram treinados três tipos de modelos de regressão para cada um dos levantamentos, com cada um desses modelos sendo capaz de prever um dos seguintes parâmetros estelares: T_{eff}, log(g), [Fe/H].

Além de possibilitar uma validação mais concreta da metodologia, o valor desses tipos de modelos também está no fato deles permitirem que se calculem parâmetros estelares para um grande quantidade de estrelas de maneira rápida e eficiente.

4.3.1 Amostras de Desenvolvimento

Assim como nos FlACOs, o processo de desenvolvimento desses modelos necessita de amostras com objetos dos quais os alvos já são conhecidos. Para gerar essas amostras foram utilizados dados de parâmetros estelares disponibilizados pelo Data Release 8 do levantamento LAMOST, além de dados de magnitude do J-PLUS e do S-PLUS.

Após a definição dos levantamentos a serem utilizados, a geração das amostras de desenvolvimento foi feita a partir de seu cruzamento. No que diz respeito aos filtros de qualidade aplicados, foram considerados tanto o erro nas magnitudes quanto os erros nos parâmetros calculados pelo LAMOST. Com isso, foram considerados nas amostras finais de desenvolvimento apenas os objetos que satisfaziam as seguintes condições:

- Erros nas 12 magnitudes do J-PLUS/S-PLUS menores do que 0,1;

- Razão sinal-ruído (SNR) do LAMOST maior do que 10 nos filtros g, i e z e maior do que 20 no filtro r;
- Erro nos parâmetros calculados pelo LAMOST menor do que 300K para T_{eff}, menor do que 0,4 para log(g) e menor do que 0,4 para [Fe/H].

Em relação à primeira condição aplicada, é interessante apontar o fato do erro nas magnitudes considerado aqui ser menor do que o erro considerado durante o desenvolvimento dos FlACOs (0,1 e 0,3, respectivamente). Essa escolha foi feita baseada no fato de não estarmos lidando com uma pequena quantidade de objetos durante o treinamento dos estimadores de parâmetros estelares, de modo que se pode considerar um erro menor e ainda assim continuar com uma quantidade considerável de objetos.

Esse processo resultou numa amostra de desenvolvimento denominada J1 para o J-PLUS e uma denominada S1 para o S-PLUS, todos eles com dados de 12 magnitudes e 3 parâmetros estelares.

Além disso, com o intuito de analisar a possibilidade do uso das distâncias dos objetos na construção dos modelos, foram geradas as amostras S2 e J2 a partir do cruzamento das amostras S1 e J1, respectivamente, com os dados de distância calculados por BAILER-JONES *et al.* (2021) a partir do levantamento GAIA (GAIA COLLABORATION *et al.*, 2016).

Essas amostras com dados de distâncias foram então filtradas para considerar apenas objetos com valores de *parallax_over_error* superiores à 5, de modo a remover os objetos cujo erro na paralaxe fosse maior do que 20% do seu valor. Como a paralaxe é o parâmetro principal utilizado para o cálculo da distância, essa filtragem garante a qualidade dos dados nas amostras S2 e J2.

Amostra	**Levantamento Base**	**Obj. em Comum com LAMOST**	**Variáveis**	**Alvos**
J1	J-PLUS	211766	12 Magnitudes	T_{eff}, log(g) e [Fe/H]
S1	S-PLUS	66278	12 Magnitudes	T_{eff}, log(g) e [Fe/H]
J2	J-PLUS	181407	12 Magnitudes Distância	T_{eff}, log(g) e [Fe/H]
S2	S-PLUS	49072	12 Magnitudes Distância	T_{eff}, log(g) e [Fe/H]

Tabela 4.9: Características gerais das amostras de desenvolvimento dos modelos de previsão de parâmetros estelares.

Por fim, se aplicou um filtro na variável RUWE (*Renormalised Unit Weight Error*)

calculada e fornecida pelo GAIA como um indicador da qualidade da astrometria dos objetos FABRICIUS *et al.* (2021). Para esse trabalho foram desconsideradas das amostras S2 e J2 os objetos com RUWE $>$ 1,4, de modo a também garantir a qualidade dos dados utilizados na entrada dos modelos (LINDEGREN, 2018).

A quantidade de objetos nas quatro amostras descritas pode ser consultada na Tabela 4.9, junto com as variáveis presentes em cada uma delas. A diferença entre o número de objetos nas amostras do J-PLUS em comparação com as amostras do S-PLUS pode ser atribuída ao fato do LAMOST ser um levantamento do hemisfério norte galáctico, como se pode observar na Figura 3.6, que é o mesmo hemisfério coberto pelo J-PLUS, como se pode observar na Figura 3.2.

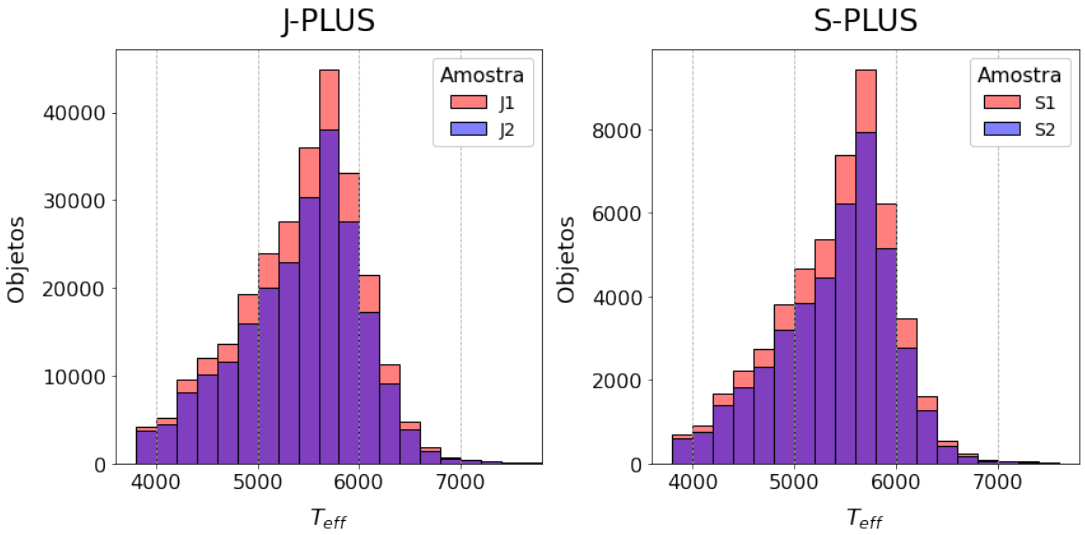

Figura 4.17: Distribuição dos valores de T_{eff} nas amostras J1, J2, S1 e S2.

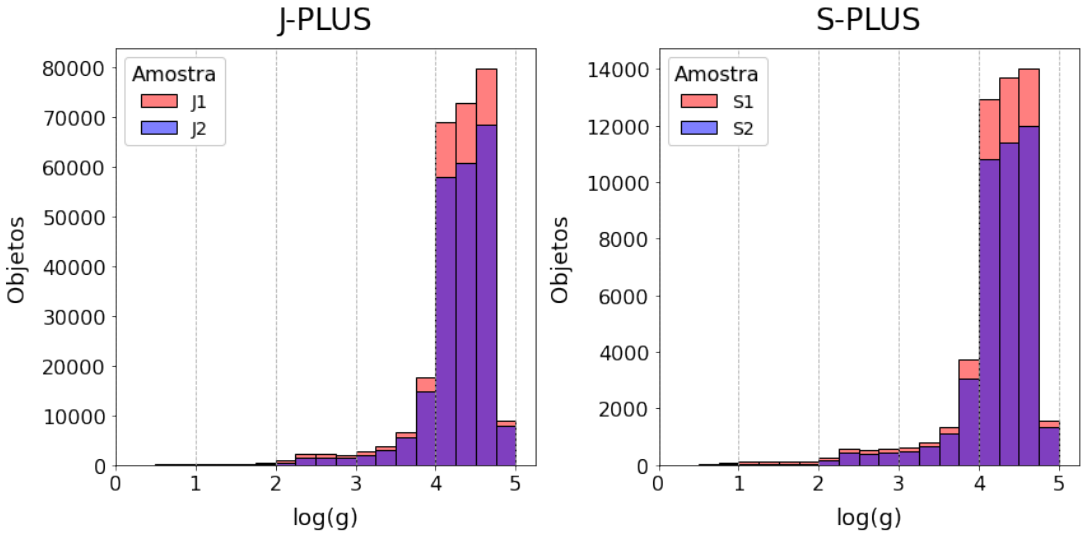

Figura 4.18: Distribuição dos valores de log(g) nas amostras J1, J2, S1 e S2.

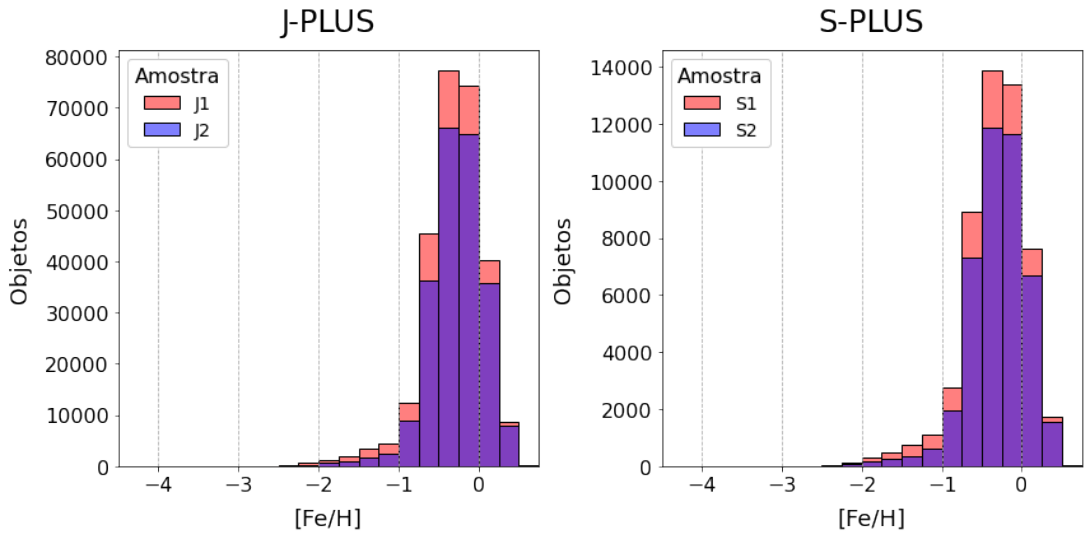

Figura 4.19: Distribuição dos valores de [Fe/H] nas amostras J1, J2, S1 e S2.

Nas Figuras 4.17, 4.18 e 4.19 é possível observar as distribuições de cada um dos parâmetros estelares nas quatro amostras. Esse é um ponto importante de ser analisado pelo fato dos modelos de Random Forest não serem capazes de fazer previsões fora do intervalo de dados utilizados em seu treinamento. Dessa forma, por exemplo, o modelo de regressão treinado com a amostra J1 para prever a temperatura efetiva de estrelas do J-PLUS só será capaz de prever valores entre ∼4000K e ∼7000K.

4.3.2 Magnitudes Absolutas

Com o intuito de analisar o impacto do uso da distância no processo de previsão dos parâmetros de uma estrela, o método selecionado para incluí-las nos dados de entrada do modelo foi o cálculo das magnitudes absolutas.

A partir dos valores de magnitude aparentes (m_{app}) corrigidas para extinção disponibilizados pelos levantamentos e as distâncias (d), é possível calcular as magnitudes absolutas (M_{abs}) a partir da equação:

$$m_{app} - M_{abs} = 5\, log\, \left(\frac{d}{10}\right) \tag{4.6}$$

A principal diferença entre esses dois valores de magnitude é o fato da aparente variar de acordo com a distância do objeto até o observador, enquanto que a absoluta se mantém constante independente da posição do observador (por definição, a magnitude absoluta é a magnitude aparente reportada por um observador à 10 parsecs do objeto).

Assim, as amostras J2 e S2 foram utilizadas para treinar modelos de previsão dos três parâmetros estelares em questão seguindo a estrutura indicada na Figura 2.7 (com o Random Forest classificador trocado por um Random Forest regressor). Uma primeira versão desses modelos foi treinada a partir das 12 magnitudes aparentes reportadas pelos

levantamentos, enquanto que uma segunda versão foi treinada com as 12 magnitudes absolutas calculadas a partir da equação 4.6.

Além disso, a mesma etapa do cálculo das 66 cores utilizada no desenvolvimento dos FlACOs também foi aplicada aqui, de modo que os modelos foram treinados com 78 variáveis de entrada. Como as cores são obtidas a partir da diferença entre duas magnitudes, e todas as magnitudes sofrem a mesma transformação entre aparente e absoluta, os valores de cores são os mesmos nas duas versões dos modelos.

Para esses modelos não foi realizada otimização de hiperparâmetros, e os valores utilizados foram os padrões da biblioteca utilizada (sklearn), e seus valores estão indicados na Tabela 4.10.

Devido à isso não foi necessária a criação de uma amostra de validação, e os objetos foram divididos apenas em uma amostra de treino com 75% delas e uma amostra de teste com 25%.

Hiperparâmetro	**Valor Padrão**
m	78
n_trees	100
msl	1
bootstrap	True
max_features	1,0

Tabela 4.10: Valores padrão dos hiperparâmetros dos pacotes utilizados.

Por fim, a métrica escolhida para avaliar a performance dos modelos foi o desvio mediano absoluto (*median absolute deviation*, MAD). O cálculo dessa métrica é realizado a partir dos erros E das previsões do modelo:

$$E_i = X_{i,r} - X_{i,p}, \tag{4.7}$$

onde E_i é o erro da i-ésima previsão, $X_{i,r}$ é o i-ésimo valor real e $X_{i,p}$ é o i-ésimo valor previsto. Após terem sido obtidos os erros do modelo em todos os objetos da amostra, a MAD é calculada com a equação:

$$MAD = mediana\left(|E_i|\right). \tag{4.8}$$

Ou seja, a MAD é definida como a mediana de todos os erros. Geralmente, essa métrica se mostra mais robusta do que as alternativas que utilizam a média em seus cálculos, pois a mediana é menos afetada por *outliers* e valores extremos dentro das amostras.

Como se pode observar na Tabela 4.11, tanto para a temperatura efetiva quando para a metalicidade, o uso das magnitudes absolutas trouxe uma melhora na performance de todos os modelos considerados. No entanto, em ambos os casos essa melhora não se mostrou significativa (melhora média de 4% para a T_{eff} e $7,5\%$ para [Fe/H]) o suficiente para justificar a criação de modelos específicos para a previsão desses hiperparâmetros

Levantamento	Parâmetro	MAD (app)	MAD (abs)	Variação
J-PLUS	T_{eff}	45 K	43 K	-4%
J-PLUS	log(g)	0,08 dex	0,06 dex	-30%
J-PLUS	[Fe/H]	0,07 dex	0,07 dex	-7%
S-PLUS	T_{eff}	50 K	48 K	-4%
S-PLUS	log(g)	0,09 dex	0,06 dex	-28%
S-PLUS	[Fe/H]	0,09 dex	0,08 dex	-8%

Tabela 4.11: Performances dos modelos de previsão de parâmetros estelares treinados e testados para o J-PLUS e o S-PLUS utilizando as magnitudes aparentes e absolutas.

utilizando as magnitudes absolutas, principalmente quando se considera o fato de que o uso das magnitudes absolutas adiciona tanto um passo extra de pré-processamento dos dados de entrada quanto uma fonte extra de possíveis erros nos valores que são fornecidos ao modelo.

Por fim, em ambos os levantamentos a utilização de magnitudes absolutas se mostrou impactante no treinamento de modelos para a previsão do log(g), com uma melhora média de 29% em relação aos modelos treinados com a magnitude aparente. Isso indica que o fato de a magnitude aparente de uma estrela variar com sua distância também pode mascarar informações importantes para o processo de previsão do logaritmo de sua gravidade superficial.

Levantamento	Parâmetro	Intervalo	MAD (app)	MAD (abs)	Variação
J-PLUS	log(g)	(1,0, 3,5)	0,15 dex	0,09 dex	-40%
J-PLUS	log(g)	(3,5, 5,0)	0,08 dex	0,05 dex	-29%
S-PLUS	log(g)	(1,0, 3,5)	0,16 dex	0,09 dex	-44%
S-PLUS	log(g)	(3,5, 5,0)	0,08 dex	0,06 dex	-25%

Tabela 4.12: Performances segmentadas dos modelos de previsão de log(g) treinados e testados para o J-PLUS e o S-PLUS utilizando as magnitudes aparentes e absolutas.

Para entender mais profundamente essa relação é possível realizar o cálculo da MAD em intervalos distintos de log(g), de modo a analisar se essa melhora na performance ocorre de maneira igual em todo o intervalo de valores ou se concentra em um intervalo específico.

Considerando então os intervalos de log(g) correspondentes às gigantes (de 1,0 à 3,5 dex) e às estrelas da sequência principal (de 3,5 à 5,0 dex), é possível obter os resultados indicados na Tabela 4.12.

Nesse caso fica claro que a melhora na performance dos modelos é ainda maior no intervalo de log(g) correspondente às estrelas gigantes. Esse comportamento pode indi-

car que o uso de magnitudes aparentes no treinamento dos modelos pode dificultar sua capacidade de diferenciar entre uma estrela gigante e uma estrela de sequência principal, e que o uso das magnitudes absolutas ajuda nessa diferenciação.

4.3.3 Otimização de Hiperparâmetros

Levando em conta os resultados obtidos na sub-seção anterior, são propostos quatro modelos distintos para cada um dos levantamentos em questão:

- **P-TEFF:** Previsor de temperatura efetiva de estrelas a partir de dados de magnitude aparente;
- **P-FEH:** Previsor de metalicidade de estrelas a partir de dados de magnitude aparente;
- **P-LOGG-APP:** Previsor de log(g) de estrelas a partir de dados de magnitude aparente;
- **P-LOGG-ABS:** Previcos de log(g) de estrelas a partir de dados de magnitude absoluta.

A partir desses modelos, os parâmetros de temperatura efetiva e metalicidade de qualquer estrela dentro do J-PLUS ou do S-PLUS podem ser previstos a partir de seus dados de magnitude aparente e índices de cor. Enquanto isso, o logaritmo da gravidade superficial de estrelas pode ser previsto pelo modelo P-LOGG-ABS caso existam dados de distância confiáveis, ou pelo modelo P-LOGG-APP caso não existam.

4.3.3.1 Valores Testados

Hiperparâmetro	**Valores Testados**
bootstrap	[True, False]
max_features	[0,25, 0,50, 0,75, 1,0]
min_samples_leaf (msl)	[1, 5, 10, 20]
n_trees	[25, 50, 100]
m	[20, 40, 60, 78]

Tabela 4.13: Lista de hiperparâmetros otimizados e os seus valores considerados durante o desenvolvimento dos modelos de previsão de parâmetros estelares nos levantamentos J-PLUS e S-PLUS.

Tendo definido os modelos que serão desenvolvidos, é possível passar para a etapa de otimização de hiperparâmetros. Para isso utilizou-se a metodologia da busca em grade

descrita na subseção 2.3.1 com os valores de hiperparâmetros indicados na Tabela 4.13, que são os mesmos testados para os FlACOs com a exceção do *cutoff*, que não se faz necessário em problemas de regressão como esse.

Com isso, foram testadas 384 combinações distintas de hiperparâmetros para cada um dos quatro modelos propostos anteriormente, e as performances dessas combinações foram comparadas a partir do desvio mediano absoluto, ou MAD. Os resultados dessa etapa podem ser encontrados nas sub-seções a seguir.

4.3.3.2 Análise de Performance Média

Iniciando-se a análise pelos modelos de previsão de temperatura efetiva de objetos do J-PLUS e do S-PLUS, são gerados os gráficos de distribuição de MAD em função dos valores testados de cada hiperparâmetro.

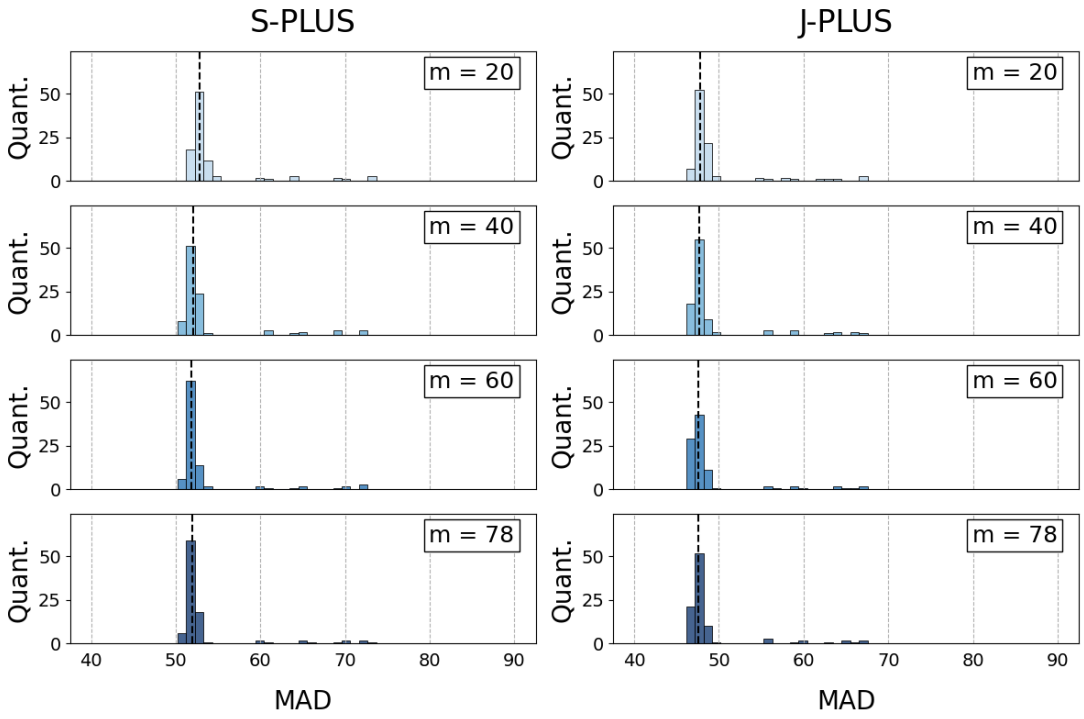

Figura 4.20: Distribuição dos MADs dos modelos de previsão de T_{eff} em função do hiperparâmetro m para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

Na Figura 4.20, onde estão indicadas as distribuições do desvio mediano absoluto dos modelos treinados com diferentes valores de m, é possível notar que esse hiperparâmetro não tem um impacto muito expressivo na performance final do modelo. Apesar disso, em ambos os casos, os modelos treinados com $m = 20$ tem medianas das MADs (como indicado pelas linhas pontilhadas) maiores do que os modelos treinados com os outros três valores de m, o que indica que o uso de $m = 20$ implica em uma pior performance.

Considerando o hiperparâmetro *min_samples_leaf*, que define o número mínimo de

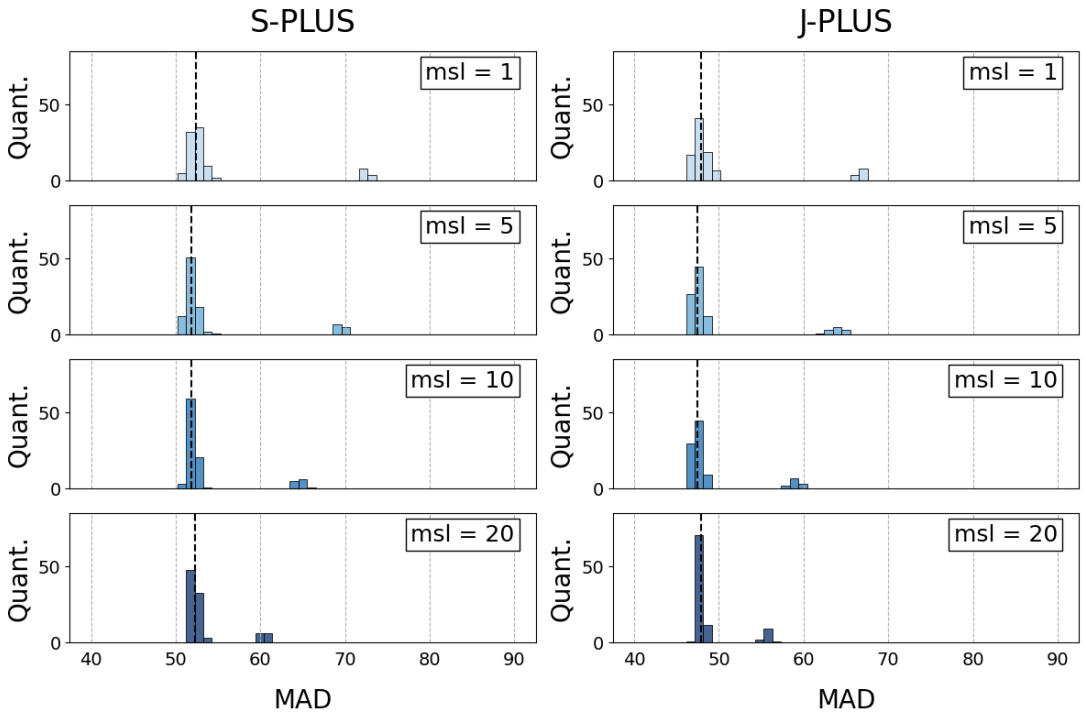

Figura 4.21: Distribuição dos MADs dos modelos de previsão de T_{eff} em função do hiperparâmetro *min_samples_leaf* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

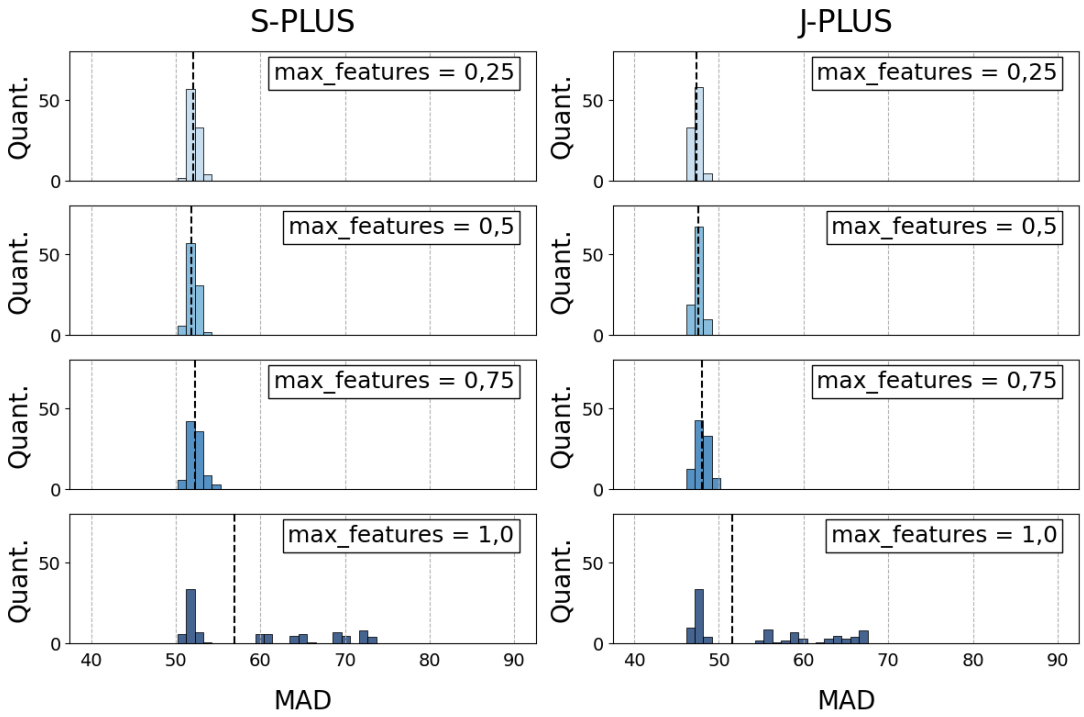

Figura 4.22: Distribuição dos MADs dos modelos de previsão de T_{eff} em função do hiperparâmetro *max_features* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

objetos em um nó folha pra que ele seja considerado válido, é possível concluir a partir da Figura 4.21 que esse hiperparâmetro também não teve uma influência considerável nos modelos treinados, apesar de valores maiores de *msl* resultarem numa distribuição mais concentrada.

Já em relação ao *max_features*, a Figura 4.22 mostra que esse hiperparâmetro teve uma certa influência nos modelos treinados em ambos os levantamentos. Nesse caso, apesar dos modelos treinados com $max\_features = 0,25$, $0,5$ e $0,75$ apresentarem distribuições similares, os modelos treinados com $max\_features = 1,0$ se diferenciaram dos demais tanto por apresentarem grupos de outliers quanto por terem uma mediana consideravelmente mais alta.

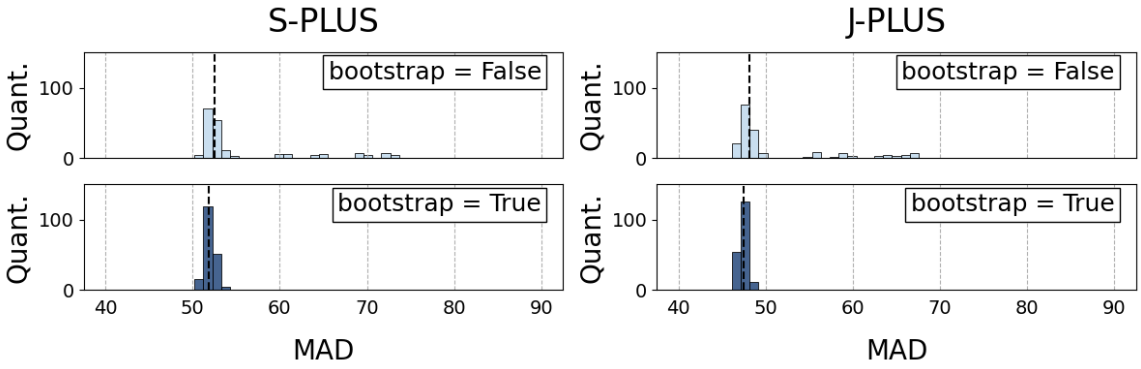

Figura 4.23: Distribuição dos MADs dos modelos de previsão de T_{eff} em função do hiperparâmetro *bootstrap* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

De maneira similar, o hiperparâmetro *bootstrap* mostra uma leve influência na performance geral dos modelos, com aqueles treinados utilizando $bootstrap = True$ performando ligeiramente melhor do que os treinados com $bootstrap = False$, como se pode observar pelas medianas da Figura 4.23 (indicadas pelas linhas pontilhadas).

Por fim, a análise do hiperparâmetro *n_trees* mostra que o número de árvores em cada uma das florestas não causou um impacto considerável na performance dos modelos, de modo que todas as distribuições da Figura 4.24 se comportaram de maneira similar tanto em relação à sua forma quanto em relação às medianas.

O comportamento descrito e analisado dos hiperparâmetros testados nos modelos de previsão de temperatura efetiva foram observados também nos outros três tipos de modelo considerados, como se pode observar no Apêndice A.

Com isso, a etapa de otimização de hiperparâmetros pode ser considerada concluída, e passamos para a análise de performance individual e escolha das melhores combinações de hiperparâmetros para cada um dos modelos.

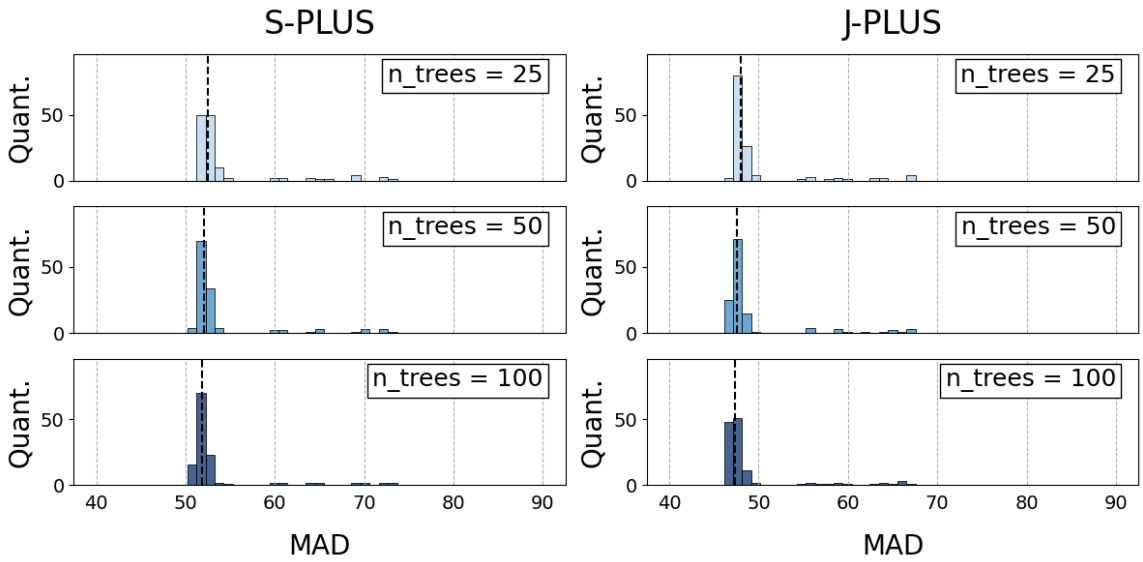

Figura 4.24: Distribuição dos MADs dos modelos de previsão de T_{eff} em função do hiperparâmetro *n_trees* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

4.3.3.3 Combinações Escolhidas

Baseando-se nas melhores combinações de hiperparâmetros de cada modelo testado, encontradas no Apêndice B, as combinações finais foram escolhidas e listadas na Tabela 4.14. Como se pode observar, em todos os casos o número de árvores escolhidos foi de *n_trees* = 100.

Levantamento	Modelo	m	n_trees	msl	bootstrap	max_features
J-PLUS	P-TEFF	60	100	5	True	0,75
J-PLUS	P-LOGG-APP	40	100	1	False	0,50
J-PLUS	P-LOGG-ABS	78	100	10	True	0,25
J-PLUS	P-FEH	40	100	1	False	0,50
S-PLUS	P-TEFF	40	100	5	True	1,00
S-PLUS	P-LOGG-APP	60	100	5	True	1,00
S-PLUS	P-LOGG-ABS	40	100	5	True	0,50
S-PLUS	P-FEH	40	100	5	False	0,25

Tabela 4.14: Combinações de hiperparâmetros escolhidas para treinar os modelos de previsão de parâmetros estelares do S-PLUS e do J-PLUS.

Em relação ao hiperparâmetro *bootstrap*, os modelos de previsão de gravidade superficial a partir das magnitudes absolutas, metalicidade e temperatura efetiva tiveram os mesmos valores escolhidos tanto no J-PLUS quanto no S-PLUS (True para P-LOGG-ABS, False para P-FEH e True para P-TEFF). De maneira contrária, os modelos P-LOGG-APP

tiveram valores de False para o levantamento J-PLUS e True para o levantamento S-PLUS.

Para o hiperparâmetro m que controla o número de variáveis filtrado pela etapa de RFE dentro dos modelos, todos os modelos tiveram valores escolhidos de $m = 60$, $m = 40$ ou $m = 78$, com o modelo P-FEH tendo os mesmos valores tanto no J-PLUS quanto no S-PLUS e os modelos P-TEFF, P-LOGG-APP e P-LOGG-ABS tendo valores diferentes para cada levantamento.

Considerando então o hiperparâmetro *max_features*, da fração de variáveis que cada árvore dentro da floresta considera durante seu treinamento, nenhum dos modelos escolhidos apresentaram o mesmo valor para o J-PLUS e o S-PLUS. Por fim, em relação aos valores do hiperparâmetro *min_samples_leaf* (msl), apenas os modelos P-TEFF tiveram valores iguais no S-PLUS e no J-PLUS.

4.3.4 Performance dos Modelos

Assim, tendo definido a combinação final de cada um dos modelos de previsão de parâmetros estelares, suas versões finais foram treinadas e sua performance testada.

Parâmetro	**Intervalos**
T_{eff}	[3500 K, 5000 K] [5000 K. 6000 K] [6000 K, 8500 K]
log(g)	[1,0, 3,5] [3,5, 5,0]
[Fe/H]	[-2,5, -0,5] [-0,5, 0,0] [0,0, 1,0]

Tabela 4.15: Intervalos de parâmetros considerados durante a análise de performance dos modelos finais de previsão de parâmetros estelares do S-PLUS e do J-PLUS.

Para analisar cada caso, a métrica de desvio foi calculada tanto para todos os objetos da amostra de teste, quanto para intervalos específicos de parâmetros (de maneira similar ao que foi feito para os modelos de log(g) na seção 4.3.2) indicados na Tabela 4.15. Dessa maneira, é possível entender como a performance do modelo se comporta para diferentes grupos de objetos.

4.3.4.1 Modelos P-TEFF

O teste de performance dos modelos para a previsão da temperatura efetiva de objetos no J-PLUS e no S-PLUS resultou nas métricas listadas na Tabela 4.16.

Um primeiro ponto a ser levantado é o fato do desvio ser menor no modelo do J-PLUS. Isso foi observado durante o processo de otimização de hiperparâmetros, e o fato

de ocorrer também nos modelos finais corrobora o comportamento. Como os alvos de ambos os modelos são calculados pelo mesmo levantamento (LAMOST), não é possível confirmar a hipótese de que as temperaturas efetivas utilizadas pelo modelo do J-PLUS são mais precisas do que as usadas pelo modelo do S-PLUS.

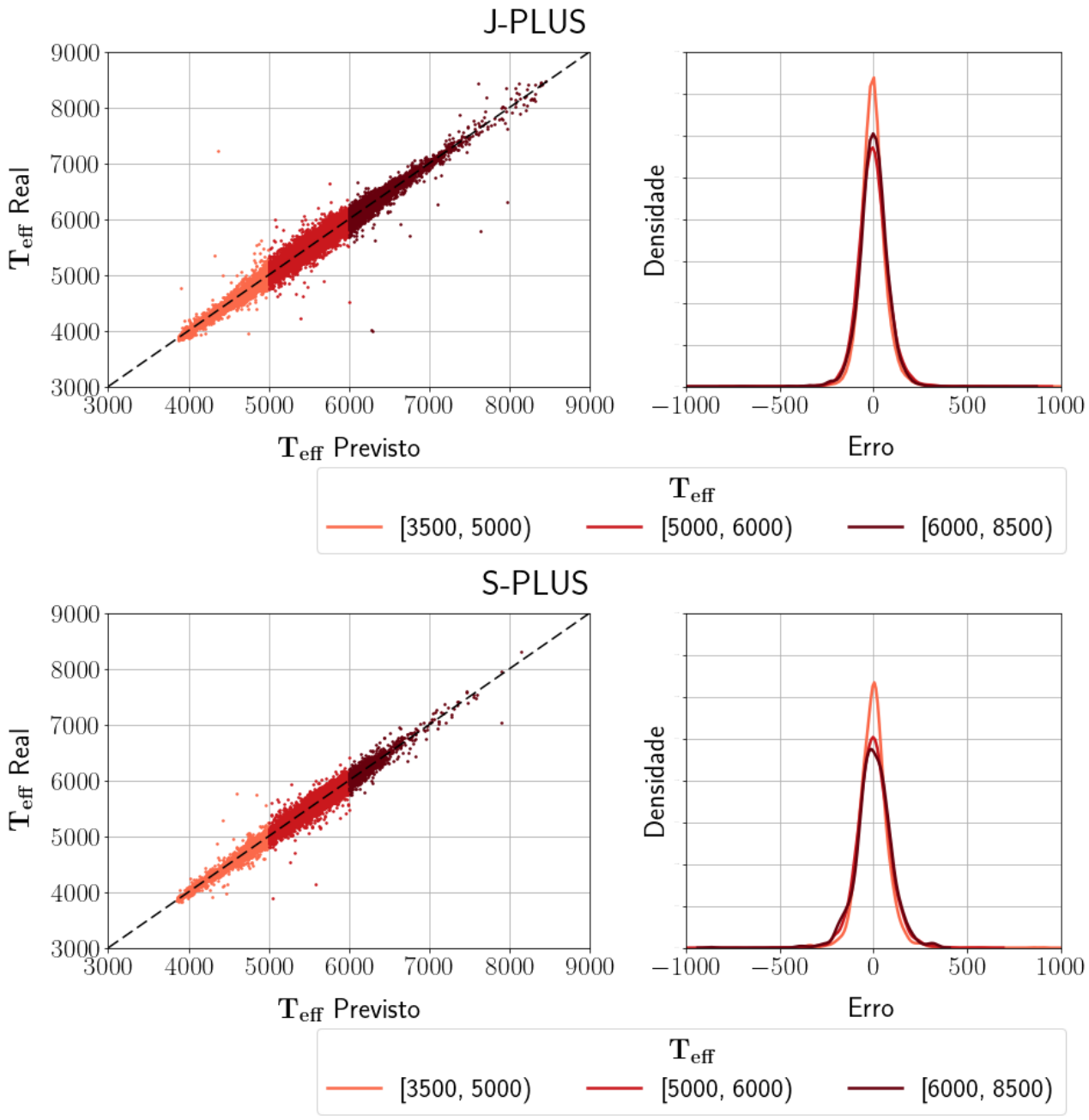

Figura 4.25: Resultados das previsões e dos erros dos modelos finais de previsão de temperatura efetiva no J-PLUS e no S-PLUS.

No que diz respeito ao desvio calculado nos diferentes intervalos considerados, é possível observar na Tabela 4.16 que em ambos os levantamentos a melhor performance dos modelos ocorreu para objetos com temperatura efetiva entre 3500 K e 5000 K.

Os comportamentos descritos nessa seção podem ser observados mais claramente na

J-PLUS		S-PLUS	
Intervalo	**MAD**	**Intervalo**	**MAD**
[3500 K, 5000 K]	36 K	[3500 K, 5000 K]	42 K
[5000 K, 6000 K]	47 K	[5000 K, 6000 K]	53 K
[6000 K, 8500 K]	43 K	[6000 K, 8500 K]	53 K
Completo	**45 K**	**Completo**	**51 K**

Tabela 4.16: Resultados de performance dos modelos finais de previsão de temperatura efetiva no J-PLUS e no S-PLUS.

Figura 4.25, onde cada ponto nos gráficos da esquerda representa um objeto da amostra de teste plotado em função de seu parâmetro real e seu parâmetro previsto. Nesse caso, objetos mais próximos da linha pontilhada (gerada a partir da equação $x = y$) indicam previsões melhores do modelo.

Além disso, nos gráficos da direita podem ser observadas as distribuições dos erros nas previsões dentro de cada um dos intervalos considerados. Nesse caso, quanto mais concentradas próximo ao ponto $erro = 0$, melhores essas previsões.

4.3.4.2 Modelos P-LOGG-APP

Já o teste de performance dos modelos para a previsão do logaritmo da gravidade superficial a partir das magnitudes aparentes de objetos no J-PLUS e no S-PLUS resultou nas métricas listadas na Tabela 4.17.

J-PLUS		S-PLUS	
Intervalo	**MAD**	**Intervalo**	**MAD**
[1,0, 3,5]	0,15	[1,0, 3,5]	0,18
[3,5, 5,0]	0,07	[3,5, 5,0]	0,08
Completo	**0,08**	**Completo**	**0,08**

Tabela 4.17: Resultados de performance dos modelos finais de previsão de log(g) a partir das magnitudes aparentes no J-PLUS e no S-PLUS.

Para os modelos de previsão de log(g), diferente do que foi observado na temperatura efetiva, as melhores performances não ocorrem no S-PLUS. Nesse caso os valores de desvio obtidos nos dois levantamentos são comparáveis tanto na amostra completa quanto nos dois intervalos considerados.

Além disso, assim como foi observado durante a otimização de hiperparâmetros (Tabela 4.12), a performance dos modelos no intervalo de valores correspondente às estrelas gigantes ([1,0, 3,5]) foi pior do que a performance deles no intervalo correspondente às estrelas da sequência principal ([3,5, 5,0]), e esse comportamento corrobora o fato do log(g)

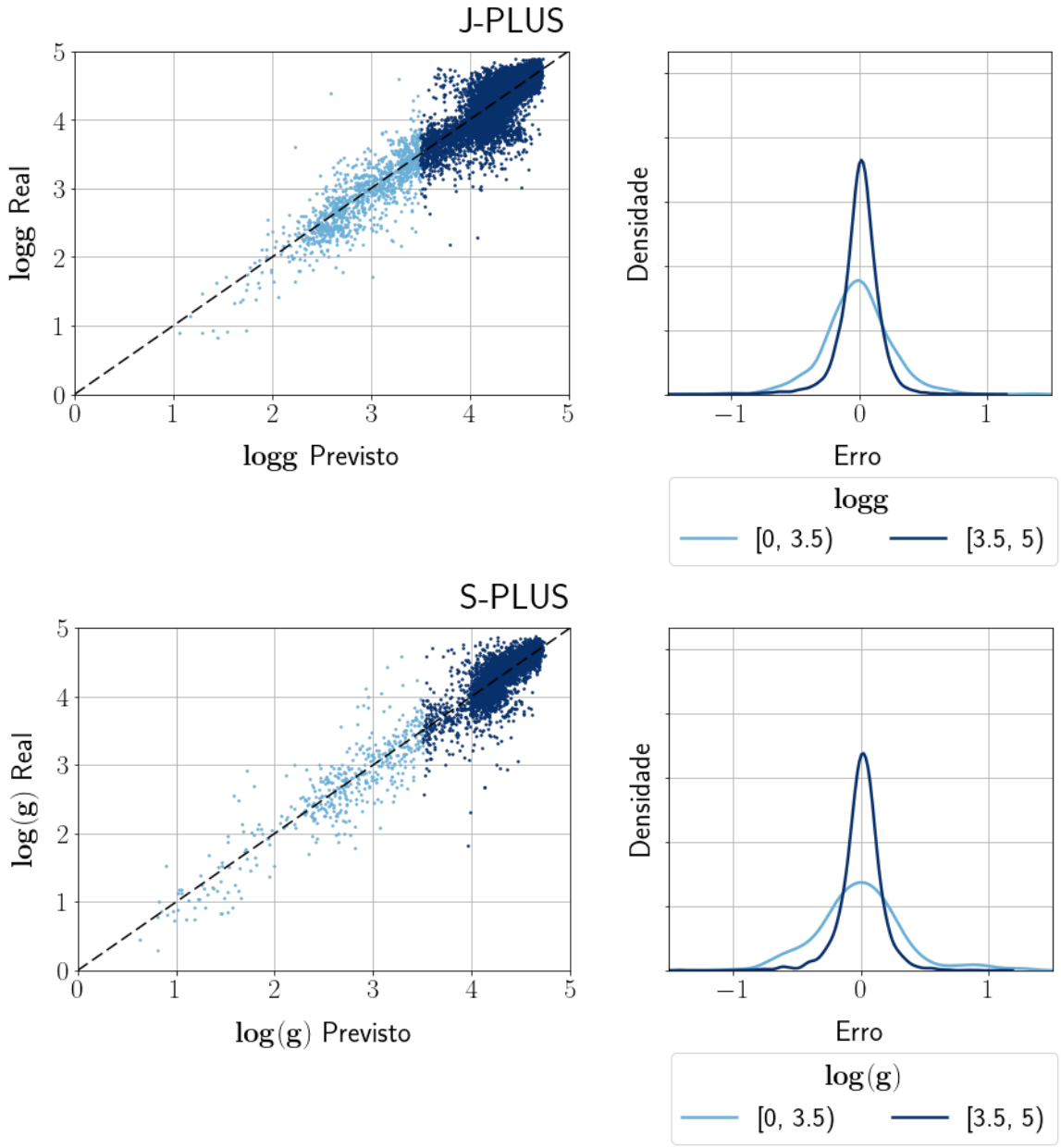

Figura 4.26: Resultados das previsões e dos erros dos modelos finais de previsão do logaritmo da gravidade superficial a partir das magnitudes aparentes no J-PLUS e no S-PLUS.

das estrelas gigantes ter um erro maior associado à sua previsão.

Para explicar essa diferença, pode-se levantar o fato das amostras de treino possuírem uma fração menor de objetos dentro do intervalo [1,0, 3,5] de log(g) (6% dos objetos do J-PLUS e 7% do S-PLUS), de modo que o processo de treinamento dos modelos acaba priorizando a diminuição do desvio dentro do intervalo com um maior número de objetos (como o método foca no intervalo completo de valores de log(g), é mais interessante para ele minimizar o erro do intervalo com mais objetos).

Novamente, esses comportamentos podem ser observados também na Figura 4.26, que indica os resultados das previsões individuais dos modelos e a distribuição dos erros.

4.3.4.3 Modelos P-LOGG-ABS

Considerando o teste de performance dos modelos para a previsão do logaritmo da gravidade superficial a partir das magnitudes absolutas de objetos no J-PLUS e no S-PLUS, foram obtidas as métricas listadas na Tabela 4.18.

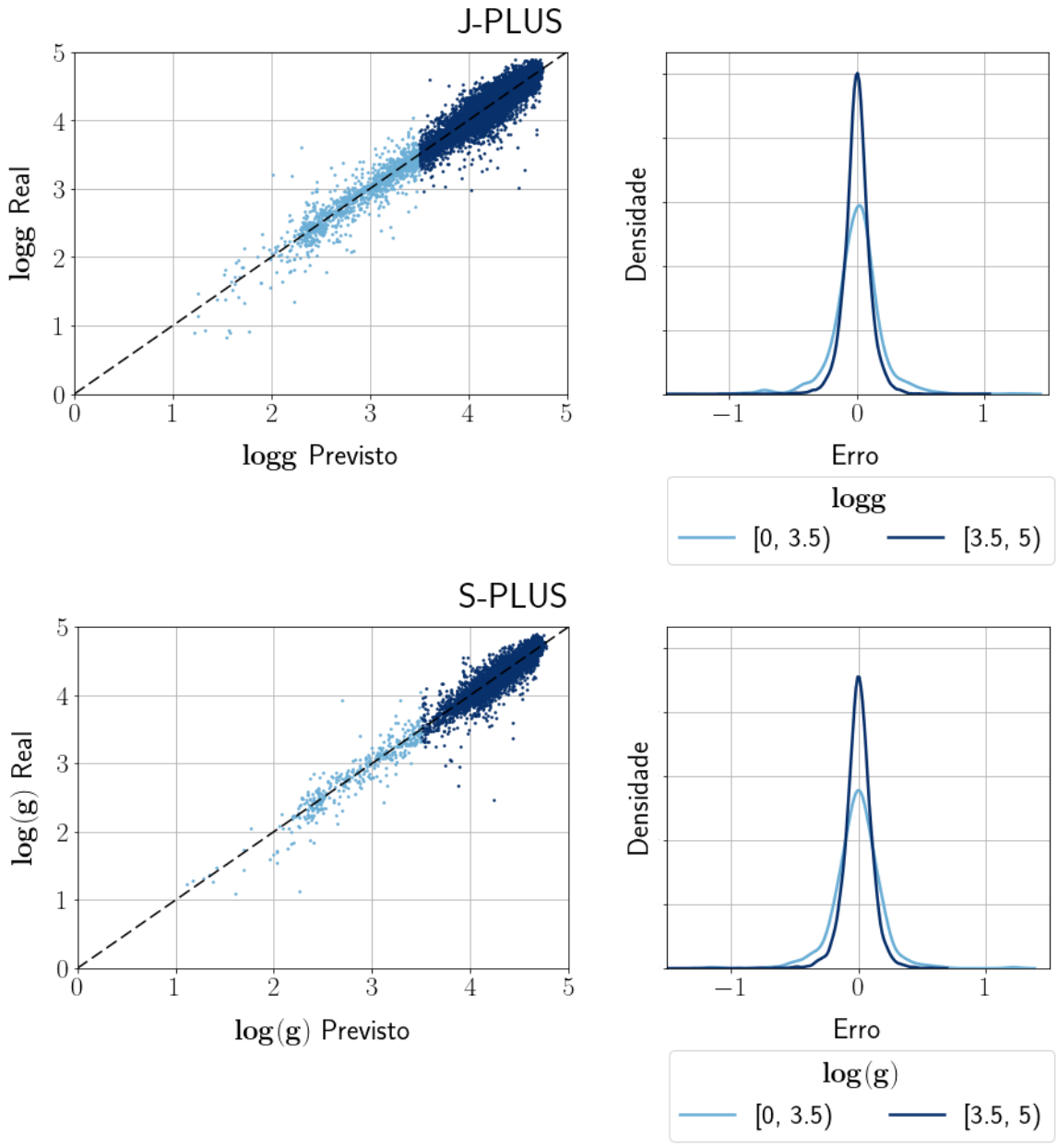

Figura 4.27: Resultados das previsões e dos erros dos modelos finais de previsão do logaritmo da gravidade superficial a partir das magnitudes absolutas no J-PLUS e no S-PLUS.

Em relação à comparação dos desvios em ambos os levantamentos, o comportamento desse caso foi o mesmo observado nos modelos de previsão de log(g) a partir das magnitudes aparentes, onde a performance do modelo do J-PLUS foi melhor do que a do S-PLUS, apesar dos valores serem próximos o suficiente para serem comparáveis.

J-PLUS		S-PLUS	
Intervalo	**MAD**	**Intervalo**	**MAD**
[1,0, 3,5]	0,08	[1,0, 3,5]	0,09
[3,5, 5,0]	0,05	[3,5, 5,0]	0,06
Completo	**0,05**	**Completo**	**0,06**

Tabela 4.18: Resultados de performance dos modelos finais de previsão do logaritmo da gravidade superficial a partir das magnitudes absolutas no J-PLUS e no S-PLUS.

Além disso, assim como nos modelos de previsão de log(g) a partir de magnitudes aparentes, os modelos de previsão de log(g) a partir de magnitudes absolutas também performaram melhor no intervalo de valores correspondente às estrelas da sequência principal. Nesse caso, o comportamento também pode ser atribuído ao fato das estrelas gigantes formarem uma fração pequena da amostra de desenvolvimento (6% no J-PLUS e 6% no S-PLUS).

Por fim, tanto a Tabela 4.18 quanto a Figura 4.27 mostram a melhora na performance de um modelo de previsão do log(g) que usa as magnitudes absolutas no lugar das magnitudes aparentes. Esse fato, já observado nos resultados da seção 4.3.2, corrobora o valor de se ter dois modelos distintos para a previsão do log(g).

4.3.4.4 Modelos P-FEH

O quarto e último tipo de modelo desenvolvido foi o de previsão de metalicidade, e suas performances podem ser encontradas tanto na Tabela 4.19 quanto na Figura 4.28.

J-PLUS		S-PLUS	
Intervalo	**MAD**	**Intervalo**	**MAD**
[-2,5, -0,5]	0,09	[-2,5, -0,5]	0,11
[-0,5, 0,0]	0,07	[-0,5, 0,0]	0,09
[0,0, 1,0]	0,05	[0,0, 1,0]	0,06
Completo	**0,07**	**Completo**	**0,09**

Tabela 4.19: Resultados de performance dos modelos finais de previsão da metalicidade no J-PLUS e no S-PLUS.

Para as metalicidades, é possível observar a partir da Tabela 4.19 que os modelos de ambos os levantamentos tiveram uma performance consideravelmente melhor nos objetos dentro do intervalo [0,0, 1,0], que representa as estrelas com uma quantidade de Ferro maior do que o Sol.

Nesse caso, as frações de objetos em cada um dos intervalos estão todas na mesma ordem de grandeza, com uma distribuição de 23%, 56% e 21% no J-PLUS e de 27%, 54%

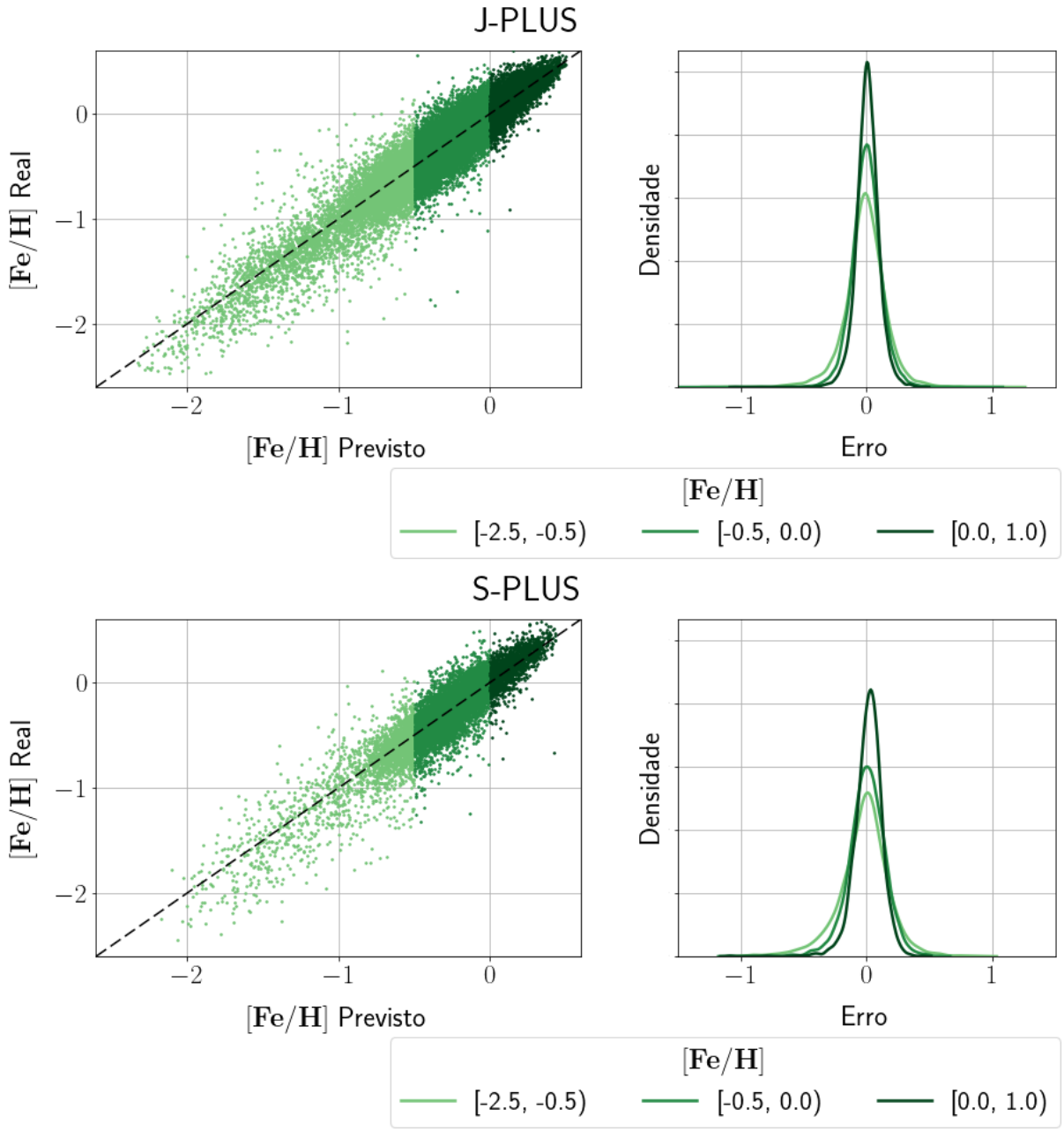

Figura 4.28: Resultados das previsões e dos erros dos modelos finais de previsão de [Fe/H] a partir das magnitudes absolutas no J-PLUS e no S-PLUS.

e 19% no S-PLUS para os intervalos de [-2,5, -0,5], [-0,5, 0,0] e [0,0, 1,0] respectivamente.

Assim, a hipótese de que o modelo teria priorizado a diminuição do erro dentro do intervalo com o maior número de objetos não pode ser aplicada para a metalicidade, e a explicação para esse caso pode estar ligada à correlação entre as variáveis de entrada do modelo e a metalicidade, algo que pode ser mais forte nos intervalos com menor erro. Além disso, é possível também relacionar essa melhor performance nas metalicidades mais altas pela melhor definição fotométrica das linhas espectrais nesses casos.

Essa diferença de performance entre os intervalos fica clara nos gráficos de distribuição dos erros presentes na Figura 4.28, onde também pode ser encontrada a distribuição das

previsões do modelo na amostra de teste.

4.3.4.5 Performance e Extinção

Para validar as correções de magnitude devido à extinção utilizadas durante o pré-processamento dos dados, a amostra de teste do J-PLUS foi dividida em objetos com pouca extinção (média das 12 correções menor ou igual à 0,15) e objetos com alta extinção (média das 12 correções maior do que 0,15). Esse processo foi realizado apenas para os objetos do J-PLUS pelo fato de serem os únicos com valores de extinção fornecidos separadamente.

Parâmetro	**Magnitudes**	**MAD (Pouca Extinção)**	**MAD (Alta Extinção)**
Teff	APP	45 K	50 K
log(g)	APP	0,08 dex	0,08 dex
log(g)	ABS	0,05 dex	0,06 dex
[Fe/H]	APP	0,07 dex	0,07 dex

Tabela 4.20: Resultados de performance dos quatro modelos finais no J-PLUS dentro da amostra com pouca extinção e da amostra de objetos com alta extinção.

As performances dos modelos para cada um dos grupos de extinção se encontram na Tabela 4.20, onde se pode observar que tanto para a metalicidade quanto para o logaritmo da gravidade superficial a variação de performance foi mínima, enquanto que para a temperatura efetiva é possível observar uma variação maior. Apesar disso, a diferença de performance não se mostra significativa o suficiente para impactar a efetividade geral dos modelos.

4.3.4.6 Comparações com a Literatura

Tendo realizado o treinamento dos modelos com a melhor combinação de hiperparêmetros encontrada durante a otimização e o teste de sua performance, é possível comparar esses resultados com outros modelos de previsão de hiperparâmetros encontrados na literatura.

Modelo	**MAD (J-PLUS)**	**MAD (S-PLUS)**
P-TEFF	45 K	51 K
P-LOGG-APP	0,08 dex	0,08 dex
P-LOGG-ABS	0,05 dex	0,06 dex
P-FEH	0,07 dex	0,09 dex

Tabela 4.21: Resultados de performance dos quatro modelos finais no J-PLUS e no S-PLUS.

Em relação aos modelos desenvolvidos neste trabalho, a Tabela 4.21 traz os desvios obtidos para cada um dos quatro modelos treinados, em ambos os levantamentos considerados.

Em relação à modelos de previsão de parâmetros estelares a partir de dados do J-PLUS, os desenvolvidos por ANDRÉS GALARZA *et al.* (2021) reportaram erros médios de $\sim$ 41K para T_{eff}, $\sim 0,11$ dex para log(g) e $\sim 0,09$ dex para [Fe/H] em intervalos de parâmetros similares aos considerados neste trabalho. Uma comparação desses valores de erro com os indicados na Tabela 4.19 mostra que tanto o modelo P-LOGG-APP quanto o P-LOOG-ABS performaram melhor do que o desenvolvido por ANDRÉS GALARZA *et al.* (2021), enquanto que os modelos P-FEH e P-TEFF performaram de maneira similar.

Também desenvolvidos para previsão de parâmetros no J-PLUS, os modelos de YANG *et al.* (2022) obtiveram erros de $\sim$ 55K para T_{eff}, $\sim 0,15$ dex para log(g) e $\sim 0,07$ dex para [Fe/H]. Novamente, uma comparação com os valores da tabela 4.19 mostra que os erros obtidos neste trabalho são comparáveis com os modelos P-TEFF e P-FEH, tendo uma performance similar aos seus correspondentes, e os modelos P-LOGG-APP e P-LOGG-ABS apresentam um erro muito menor do que o obtido por YANG *et al.* (2022).

Além da comparação de métricas gerais, é possível também realizar uma análise dos objetos em comum entre este trabalho e o catálogo de parâmetros disponibilizado por YANG *et al.* (2022), resultando nas Figuras 4.29, 4.31 e 4.30.

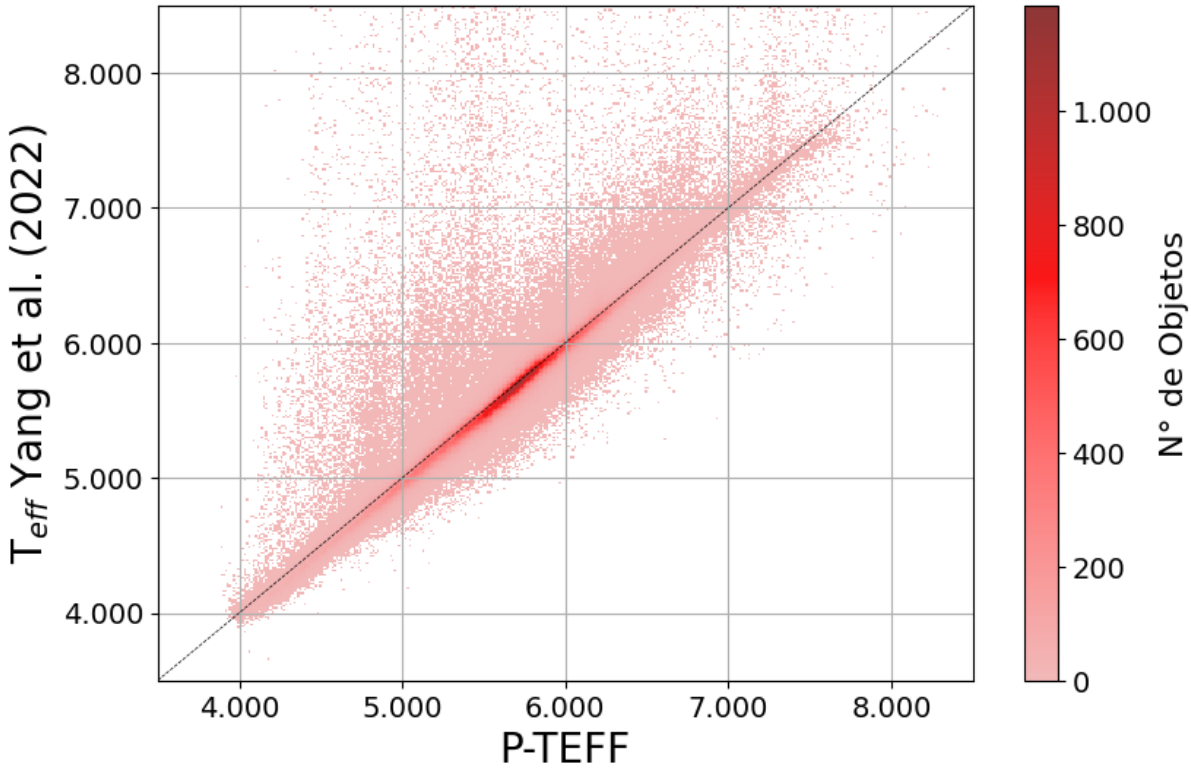

Figura 4.29: Distribuição de temperaturas efetivas dos 677.035 objetos filtrados em comum entre este trabalho e YANG *et al.* (2022).

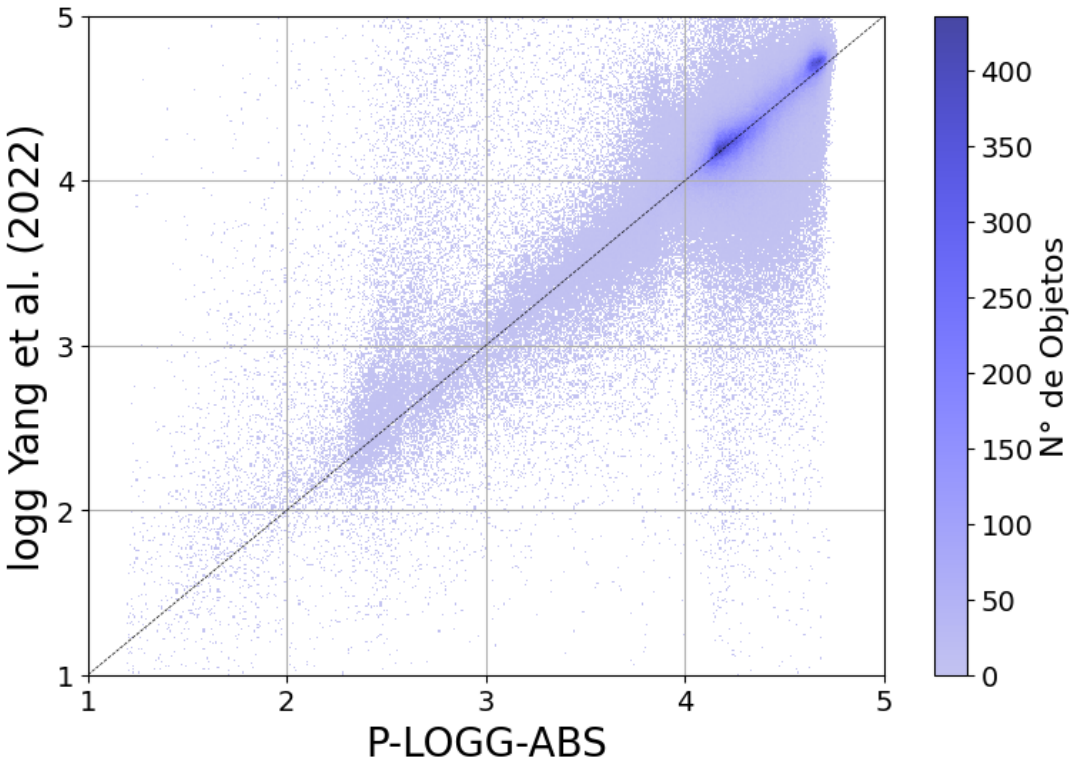

Figura 4.30: Distribuição de log(g) dos 534.149 objetos filtrados em comum entre este trabalho e YANG *et al.* (2022).

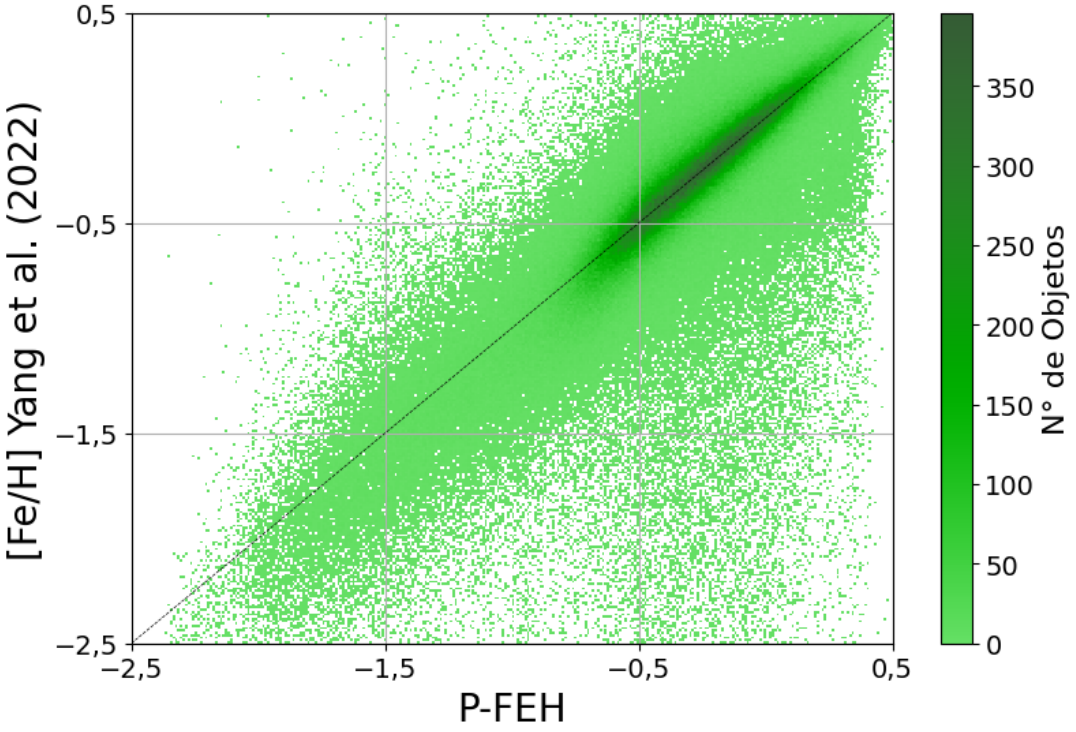

Figura 4.31: Distribuição de metalicidades dos 677.264 objetos filtrados em comum entre este trabalho e YANG *et al.* (2022).

Para essas comparações, foram considerados apenas os objetos com os erros nas 12 magnitudes do J-PLUS menores do que 0,1 e com flags de qualidade (q_Teff, q_logg e $q\_[Fe/H]$) fornecidas para os parâmetros de YANG *et al.* (2022) iguais a 0 (confiáveis). Além disso, devido ao fato das Florestas Aleatórias não serem capazes de calcular parâmetros fora de seus intervalos de efetividade, foram consideradas apenas as estrelas com parâmetros de YANG *et al.* (2022) dentro dos intervalos [3500 K; 8500 K] para T_{eff}, [1,0; 5,0] para log(g) e [-2,5; 1,0] para [Fe/H].

Considerando então a distribuição de temperaturas efetivas na Figura 4.29, é possível observar que a grande maioria dos objetos se concentra próximo a área da reta x = y, onde as previsões do modelo deste trabalho e do modelo de YANG *et al.* (2022) são similares. No entanto, é possível observar também um grupo de objetos concentrados na região acima dessa reta, o que indica que para esses objetos a estimativa de T_{eff} realizada pelo modelo deste trabalho está abaixo do valor estimado por YANG *et al.* (2022).

De maneira similar, a Figura 4.30 mostra que os objetos se concentram próximo à região da reta onde as estimativas dos dois trabalhos são similares. Nesse caso, é possível notar que no intervalo de P-LOGG-ABS entre 1 e 4, os objetos com maior diferença entre os dois trabalhos se concentram na região acima da reta (P-LOGG-ABS $<$ logg YANG *et al.* (2022)), enquanto que no intervalo com P-LOGG-ABS entre 4 e 5, o comportamento desses objetos com maior diferença aparenta ser o inverso (P-LOGG-ABS $>$ logg YANG *et al.* (2022)).

Já com base na Figura 4.31 indica que, apesar de haver uma concentração de objetos na região da reta x = y, a dispersão destes indica que a metalicidade foi o parâmetro com maior diferença entre os dois trabalhos.

A partir dessa amostra de objetos em comum, também é possível calcular as métricas presentes na Tabela 4.22, onde os MADs são calculados como a mediana das diferenças absolutas entre os valores dos parâmetros estimados neste trabalho e os valores estimados por YANG *et al.* (2022).

Parâmetro	**Número de Objetos em Comum**	**MAD**
Teff	677.035	45 K
log(g)	534.149	0,09 dex
[Fe/H]	677.264	0,07 dex

Tabela 4.22: Comparação de parâmetros estelares dos objetos em comum entre este trabalho e YANG *et al.* (2022).

Por fim, uma outra comparação que pode ser feita é em relação aos modelos desenvolvidos por WHITTEN *et al.* (2019) para a identificação de estrelas pobres em metais no J-PLUS. Utilizando algoritmos baseados em redes neurais e parâmetros estelares calculados pelos levantamentos BOSS, SEGUE e Legacy, WHITTEN *et al.* (2019) foram

capazes de desenvolver modelos para a previsão de temperatura efetiva e metalicidade. Para T_{eff}, esses modelos foram treinados com dados do DR1 do J-PLUS e apresentaram um desvio padrão de $\sim$ 91 K para estrelas com temperaturas entre 4500 K e 8500 K. Já para [Fe/H], foram treinadas redes neurais utilizando-se uma mistura de magnitudes do DR1 do J-PLUS e magnitudes sintéticas geradas a partir de espectros dos três levantamentos citados anteriormente, e esses modelos obtiveram desvios padrões de $\sim 0,2$ dex em relação aos valores fornecidos pelos levantamentos de metalicidade para estrelas com [Fe/H] $<$ -0,5.

Esses dois valores de desvio de WHITTEN *et al.* (2019) podem ser comparados aos erros reportados nas Tabelas 4.16 e 4.19 (especificamente os dois primeiros intervalos), e apesar de não se referirem ao mesmo cálculo de erro (desvio padrão x desvio mediano absoluto), é possível concluir que os modelos desenvolvidos neste trabalho obtiveram uma performance comparável aos de WHITTEN *et al.* (2019).

Nesse caso, é possível considerar também que a diferença de performance entre os modelos de WHITTEN *et al.* (2019) e os modelos desenvolvidos neste trabalho pode ser atribuída tanto aos levantamentos utilizados como fonte dos parâmetros estelares (SEGUE, BOSS, Legacy x LAMOST) quanto ao número de estrelas utilizadas no treinamento dos modelos (na casa de milhares em WHITTEN *et al.* (2019) e na casa de dezenas/centenas de milhares neste trabalho).

Assim, é possível validar os modelos desenvolvidos neste trabalho e confirmar que sua performance é similar ao restante da literatura para a previsão de temperatura efetiva e metalicidade, e que sua performance na previsão de log(g) (principalmente quando se consideram as magnitudes absolutas) é superior aos outros modelos tomados para comparação.

4.3.5 Importância das Variáveis

Utilizando a metodologia descrita na seção 2.2.3, foram calculadas as importâncias das variáveis dentro de cada um dos modelos desenvolvidos nesse estudo. Através de uma análise dos resultados, é possível entender melhor quais variáveis impactam de maneira significativa nas previsões finais.

4.3.5.1 Modelos P-TEFF

Na Figura 4.32, onde estão indicadas as 10 variáveis mais importantes dos modelos P-TEFF é possível notar que, das cinco mais importantes em cada um dos levantamentos, três delas estão em comum (sendo elas as cores (J0430 - J0861), (J0430 - zSDSS) e (J0430 - J0861)). Assim como foi o caso com os modelos FlACOs de classificação de subanãs quentes, isso corrobora a importância dessas cores na previsão da temperatura efetiva, já que essas variáveis foram selecionadas e utilizadas de maneira extensiva por dois modelos

treinados com dados de levantamentos fotométricos diferentes.

Além disso, considerando as dez variáveis mais importantes em cada levantamento, oito delas aparecem em comum nos dois. Apesar delas não terem um impacto tão grande quanto as já indicadas, o fato delas aparecerem em ambos os modelos também corrobora sua importância.

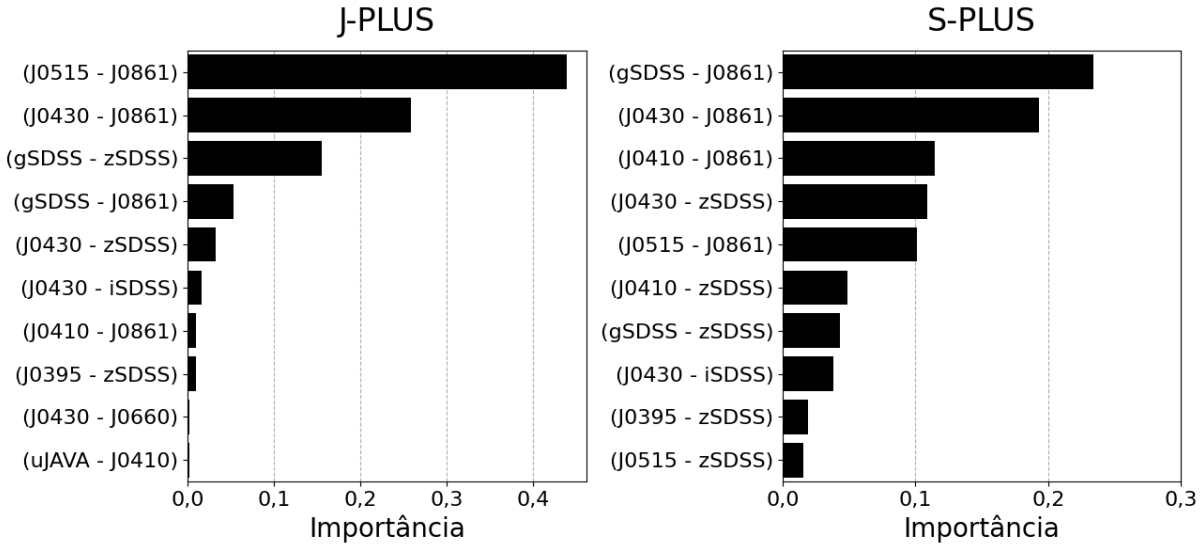

Figura 4.32: Importância das dez variáveis de maior impacto dentro dos modelos de previsão de temperatura efetiva no J-PLUS e no S-PLUS.

Por fim, é possível notar também que para ambos os levantamentos, todas as dez variáveis mais importantes para a previsão da temperatura efetiva foram cores. Isso mostra que a T_{eff} de uma estrela está mais correlacionada com suas cores do que com suas magnitudes.

4.3.5.2 Modelos P-LOGG-APP

Para os modelos de previsão de log(g) a partir de magnitudes aparentes, a relação de mesmas variáveis no topo das duas listas de maior importância se repetiu, como se pode observar na Figura 4.33. Nesse caso, as três variáveis mais importantes foram as mesmas tanto para o J-PLUS quanto para o S-PLUS, de modo que elas podem ser indicadas como realmente impactantes na previsão de log(g).

Dentre as dez variáveis mais importantes, apenas cinco delas se repetiram nos dois levantamentos, e o impacto das cinco restantes em ambos os casos não pode ser corroborado com tanta significância.

Por fim, de maneira similar ao modelo P-TEFF, as listas de variáveis mais importantes na previsão do log(g) são dominadas por cores, e uma hipótese similar em relação à correlação desse parâmetro com cores e magnitudes pode ser levantada aqui.

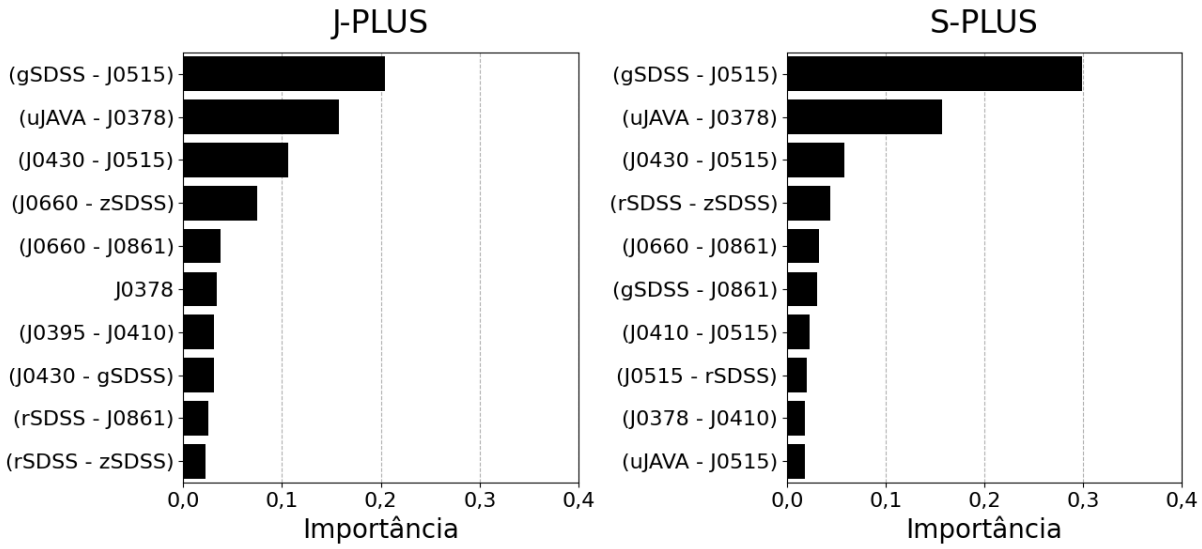

Figura 4.33: Importância das dez variáveis de maior impacto dentro dos modelos de previsão do logaritmo da gravidade superficial a partir de magnitudes aparentes no J-PLUS e no S-PLUS.

4.3.5.3 Modelos P-LOGG-ABS

Considerando então os modelos de previsão de log(g) a partir de magnitudes absolutas, as variáveis mais importantes de ambos os levantamentos podem ser encontradas na Figura 4.34.

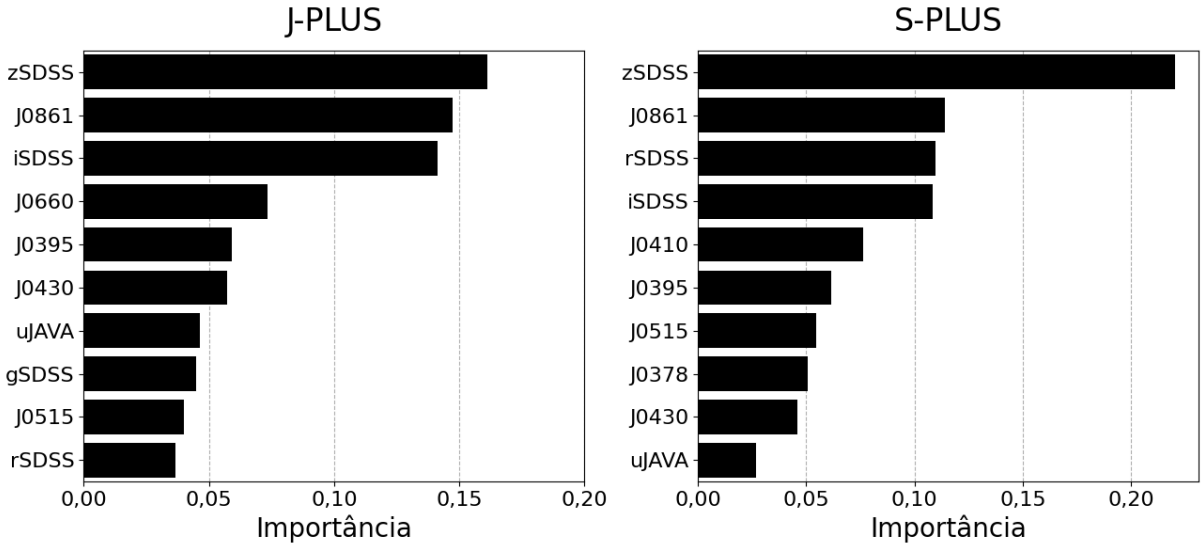

Figura 4.34: Importância das dez variáveis de maior impacto dentro dos modelos de previsão do logaritmo da gravidade superficial a partir de magnitudes absolutas no J-PLUS e no S-PLUS.

Nesses modelos é possível notar que as duas variáveis mais importantes em ambos os levantamentos foram as mesmas (J0861 e zSDSS), e considerando as dez mais importantes no J-PLUS e no S-PLUS, oito delas estão em comum. Esse fato corrobora fortemente a

hipótese de que todas elas realmente trazem informações correlacionadas com o valor do log(g) das estrelas.

Diferentemente do modelo de previsão de log(g) a partir das magnitudes aparentes, a lista de variáveis mais impactantes nesse caso foi dominada por magnitudes absolutas. Considerando que a diferença de performance entre os modelos de previsão do log(g) treinados com magnitudes aparentes e os treinados com magnitudes absolutas só podem ser fruto do tipo de magnitude utilizado (já que as cores são as mesmas nos dois casos), é de se esperar que essa diferença ocorra.

Além disso, considerando que quase todas as variáveis mais importantes são magnitudes absolutas, e que o modelo P-LOGG-ABS teve uma performance consideravelmente melhor do que o P-LOGG-APP, suporta a hipótese de que as magnitudes absolutas têm uma alta correlação com o valor do log(g) de uma estrela.

Para entender melhor essa situação é necessário considerar que duas estrelas com o mesmo valor de magnitude aparente podem possuir luminosidades diferentes, contanto que a com maior luminosidade esteja a uma distância maior. Considerando que a luminosidade de uma estrela está diretamente relacionada com seu raio (e com isso sua gravidade superficial), essa questão implica em estrelas com valores de log(g) completamente diferente entrando no modelo com os mesmos dados de entrada (magnitudes).

No entanto, ao se utilizar a magnitude absoluta o modelo é capaz de quebrar essa "degenerescência", pois seu valor é calculado a uma distância fixa de 10 parsecs. Com isso, estrelas de com luminosidades e raios diferentes passam a entrar no modelo com dados de entrada diferentes, o que pode explicar o aumento na precisão das estimativas finais.

4.3.5.4 Modelos P-FEH

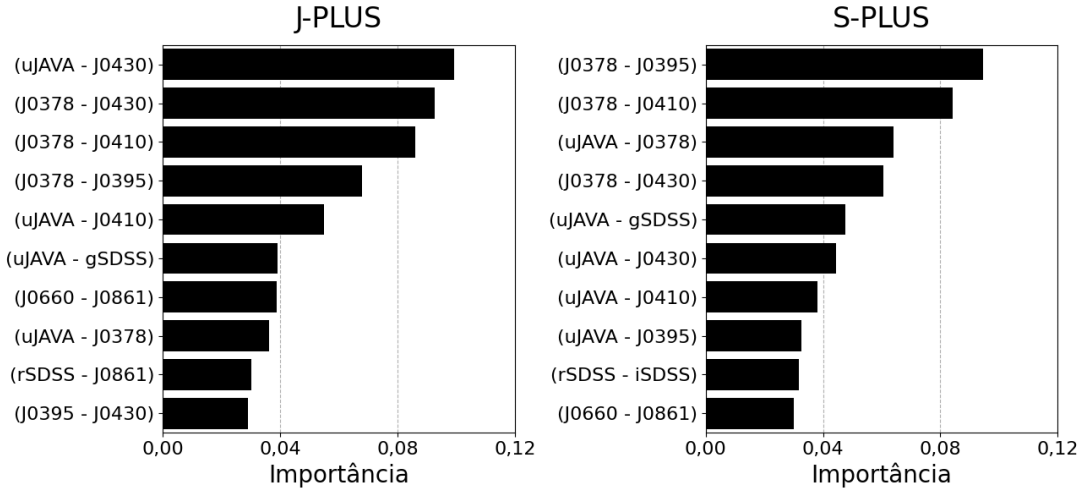

Figura 4.35: Importância das dez variáveis de maior impacto dentro dos modelos de previsão da metalicidade no J-PLUS e no S-PLUS.

Por fim, as variáveis mais importantes para os modelos de previsão de metalicidade podem ser encontradas na Figura 4.35.

Nesse caso, observa-se que três das cinco variáveis mais importantes em ambos os levantamentos estão em comum ((J0378 - J0395), (J0378 - J0410) e (J0378 - J0430)). Além disso, oito das dez mais importantes de cada modelo também estão em comum, o que corrobora sua importância na previsão da metalicidade.

Além disso, baseando-se no fato de todas as variáveis de maior impacto de ambos os modelos serem cores, é possível corroborar o fato de que elas são mais correlacionadas com o valor de [Fe/H] do que as magnitudes aparentes utilizadas.

4.3.6 Catálogos de Parâmetros Estelares

Tendo finalizado o treinamento dos quatro modelos descritos anteriormente, passamos para a criação de um catálogo de parâmetros estelares para as estrelas observadas pelo J-PLUS e o S-PLUS.

Para isso, são consideradas as amostras completas de objetos liberados nos Data Releases mais recentes de ambos os levantamentos (DR3 para o J-PLUS, DR4 para o S-PLUS). Em relação à qualidade das magnitudes dos objetos, optou-se por uma filtragem menos conservadora do que a utilizada durante o desenvolvimento dos modelos. Para a criação das amostras de desenvolvimento, foram desconsiderados todos os objetos que tivessem um erro maior do que 0,1 em qualquer uma de suas magnitudes. Já para a criação do catálogo de parâmetros estelares, esse valor foi alterado para 0,3.

Essa escolha foi feita com a intenção de aumentar o número de objetos no catálogo final, mas é necessário apontar o fato de que as métricas reportadas nos capítulos anteriores só dizem respeito à objetos com erros de magnitude menores do que 0,1, e que a performance dos modelos em objetos com erros maiores pode sofrer impactos.

Para possibilitar a diferenciação entre as estrelas com erros menores do que 0,1 e as com erro menores do que 0,3, o catálogo criado conta com uma variável denominada *good_magnitudes*, que indica a quantidade de magnitudes com erros menores do que 0,1 para cada um dos objetos. Com isso, as melhores estimativas de parâmetros serão as dos objetos que possuem $good\_magnitudes = 12$.

Além disso, todos os objetos do catálogo tiveram seus parâmetros previstos pelos modelos P-TEFF, P-FEH e P-LOGG-APP, enquanto que aqueles com distâncias fornecidas pelo GAIA e *parallax_over_error* superior à 5 também tiveram seu log(g) previsto pelo modelo P-LOGG-ABS. Assim, para os objetos com valores de log(g) calculados tanto pelo modelo P-LOGG-APP quanto pelo P-LOGG-ABS, é recomendado o uso do valor previsto a partir do modelo P-LOGG-ABS, que se demonstrou mais preciso durante os testes e validações realizados neste trabalho.

Por fim, também foram calculados os erros relacionados aos parâmetros presentes no

catálogo através do método Monte Carlo. Para isso, a estimativa dos parâmetros de cada uma das estrelas foi repetida $n_{MC} = 50$ vezes, onde cada uma das estimativas foi feita com um termo aleatório de erro somado às magnitudes m do objeto, calculadas a partir da Equação 4.9:

$$m_{MC} = m + (m_{err} * N(0,1)) \,, \tag{4.9}$$

onde m é o valor da magnitude corrigido para extinção, m_{err} é o valor do erro reportado para a magnitude m, e $N(0,1)$ é um valor amostrado aleatoriamente de uma distribuição normal com média igual à 0 e desvio padrão igual à 1. Com isso, o erro de cada um dos parâmetros pôde ser calculado como o desvio padrão das n_{MC} estimativas distintas para ele.

No total, o catálogo de parâmetros estelares do levantamento J-PLUS conta com estimativas de T_{eff}, log(g) e [Fe/H] para 3 milhões de estrelas, enquanto que o do S-PLUS possui estimativas para 5 milhões, e ambos estão disponíveis para utilização e consulta [2].

Amostras de cada um dos catálogos podem ser encontradas nas Tabelas 4.23 e 4.24. Em ambos os casos os catálogos trazem duas colunas de posição do objeto (RA e DEC), uma coluna com o número de magnitudes com erro abaixo de $0,1$ ($good\_magnitudes$), quatro colunas com as estimativas dos parâmetros estelares feitas pelos modelos (P-TEFF, P-FEH, P-LOGG-APP e P-LOGG-ABS) e quatro colunas com os erros dessas estimativas (P-TEFF-ERR, P-FEH-ERR, P-LOGG-APP-ERR e P-LOGG-ABS-ERR, respectivamente).

[2] https://drive.google.com/drive/folders/11w1q3boHfUFnyFsevLbGhhtMeFotKMNS

RA	DEC	good_magnitudes	P-TEFF	P-TEFF-ERR	P-FEH	P-FEH-ERR	P-LOGG-APP	P-LOGG-APP-ERR	P-LOGG-ABS	P-LOGG-ABS-ERR
4.187	-3.028	9	5205	391	-0.39	0.15	4.48	0.07	4.64	0.03
2.945	-3.021	12	5252	243	-0.31	0.24	4.41	0.12	4.60	0.03
3.128	-3.020	12	5818	564	-0.76	0.18	4.03	0.04	4.16	0.02
4.120	-3.017	12	5810	492	-0.56	0.09	4.12	0.04	4.15	0.01
3.297	-3.020	12	5214	367	-1.48	0.27	3.81	0.12	nan	nan
4.125	-3.012	10	5334	344	-0.68	0.05	4.49	0.10	4.51	0.02
4.019	-3.010	9	5896	622	-0.32	0.16	4.22	0.03	nan	nan
2.814	-3.012	12	4866	283	-0.37	0.21	4.62	0.15	nan	nan
3.261	-3.007	7	6015	623	-1.31	0.29	4.08	0.05	nan	nan
3.487	-3.004	12	6003	590	-1.85	0.21	4.02	0.05	nan	nan
3.812	-3.004	12	5292	243	0.03	0.26	4.38	0.13	nan	nan
3.980	-3.003	12	5920	409	-0.60	0.16	4.17	0.03	nan	nan
2.912	-3.013	12	5182	284	-0.15	0.11	4.59	0.08	nan	nan
2.913	-3.008	12	5628	386	-0.46	0.10	4.36	0.05	4.38	0.01
3.416	-3.014	12	5533	427	0.25	0.07	4.17	0.03	4.14	0.01
3.381	-3.001	12	5772	531	-0.63	0.15	4.26	0.02	4.29	0.01
3.754	-2.995	12	5310	343	-0.11	0.13	4.55	0.14	4.55	0.02
4.118	-2.989	12	5844	641	-0.85	0.08	4.13	0.07	nan	nan
3.437	-2.991	12	5727	503	-0.42	0.12	4.28	0.03	4.17	0.01
3.921	-2.988	12	5554	333	-0.49	0.09	4.29	0.06	4.45	0.01
3.815	-2.986	12	5747	301	-0.52	0.08	4.24	0.03	nan	nan
3.377	-2.983	12	6112	551	-0.91	0.19	4.14	0.06	4.13	0.01
3.548	-2.981	12	6111	495	-1.66	0.30	4.04	0.08	4.09	0.03
3.312	-2.978	9	4730	321	-1.16	0.26	4.55	0.16	nan	nan
2.811	-2.974	9	6215	545	-1.60	0.24	4.03	0.06	nan	nan

Tabela 4.23: Amostra de objetos no catálogo de parâmetros estelares para estrelas do J-PLUS desenvolvido durante esse trabalho. Os valores NaN estão presentes nas estimativas de P-LOGG-ABS dos objetos que não possuem distâncias confiáveis, e que por esse motivo não foram passados para o modelo de previsão de log(g) a partir das magnitudas absolutas.

RA	DEC	good_magnitudes	P-TEFF	P-TEFF-ERR	P-FEH	P-FEH-ERR	P-LOGG-APP	P-LOGG-APP-ERR	P-LOGG-ABS	P-LOGG-ABS-ERR
150,403	-24,577	12	5049	7	0,11	0,03	4,64	0,01	4,64	0,00
150,606	-24,571	12	5473	45	-0,43	0,09	4,46	0,01	4,52	0,02
150,614	-24,572	12	5579	7	0,03	0,09	4,31	0,04	4,26	0,02
150,416	-24,566	10	4468	15	-0,34	0,04	4,65	0,03	NaN	NaN
150,659	-24,570	12	4982	20	0,02	0,12	3,67	0,13	3,64	0,01
150,609	-24,564	8	4568	34	-0,04	0,11	4,60	0,29	4,73	0,01
150,555	-24,560	6	5335	77	-1,73	0,09	4,30	0,05	NaN	NaN
151,200	-24,562	7	5877	245	-0,85	0,19	4,14	0,32	NaN	NaN
150,174	-24,566	12	5304	36	0,07	0,06	4,50	0,03	4,63	0,02
151,171	-24,562	9	4247	12	-0,18	0,40	4,65	0,19	4,66	0,01
150,595	-24,585	10	3874	67	-1,18	0,02	0,95	0,16	NaN	NaN
150,786	-24,562	12	5538	16	-0,65	0,13	4,37	0,16	NaN	NaN
151,466	-24,559	11	5649	37	-0,41	0,10	4,20	0,06	NaN	NaN
150,864	-24,581	12	6462	22	-0,15	0,01	4,22	0,01	4,08	0,00
150,115	-24,564	12	7247	26	-0,39	0,02	4,18	0,00	4,18	0,00
149,994	-24,558	7	5823	96	-0,89	0,08	4,14	0,27	NaN	NaN
151,313	-24,564	11	4463	15	-0,19	0,13	4,65	0,03	4,72	0,01
150,509	-24,567	12	5774	15	-0,50	0,05	4,19	0,02	NaN	NaN
150,023	-24,563	12	4839	39	0,19	0,07	4,62	0,01	NaN	NaN
150,219	-24,559	9	5239	81	-0,23	0,10	4,59	0,17	NaN	NaN
150,753	-24,574	12	6380	8	-0,32	0,02	4,22	0,01	4,20	0,00
150,673	-24,568	12	5705	34	-0,30	0,05	4,25	0,06	4,36	0,02
150,493	-24,566	12	6206	34	-1,40	0,32	4,10	0,02	4,14	0,01
151,452	-24,559	12	5817	77	-0,58	0,10	4,20	0,03	NaN	NaN
150,905	-24,566	12	5448	9	0,30	0,05	4,33	0,03	NaN	NaN

Tabela 4.24: Amostra de objetos no catálogo de parâmetros estelares para estrelas do S-PLUS desenvolvido durante esse trabalho. Os valores NaN estão presentes nas estimativas de P-LOGG-ABS dos objetos que não possuem distâncias confiáveis, e que por esse motivo não foram passados para o modelo de previsão de log(g) a partir das magnitudas absolutas.

Capítulo 5

Conclusões

Em conclusão, utilizando-se como base os dados dos levantamentos fotométricos J-PLUS e S-PLUS foi possível desenvolver diversos modelos de ML tanto para a identificação de subanãs quentes dentro desses levantamentos, quanto para a estimativa de parâmetros estelares dos seus objetos. Isso comprova a qualidade dos dados fornecidos por esses dois levantamentos.

A estrutura de algoritmo baseada em florestas aleatórias proposta neste trabalho foi capaz de gerar modelos de classificação de subanãs quentes e previsão de parâmetros estelares com boas performances em suas amostras de teste, e ao todo foram desenvolvidos cinco modelos para cada um dos dois levantamentos considerados (J-PLUS e S-PLUS).

Além disso, através da otimização de hiperparâmetros baseada no método da busca em grade foi possível não só escolher a melhor combinação de hiperparâmetros para cada um dos modelos desenvolvidos, mas também analisar a maneira como a performance do modelo era impactada por cada uma das combinações.

Com relação à classificação de subanãs quentes, os modelos foram desenvolvidos utilizando um catálogo compilado por CULPAN *et al.* (2022) e obtiveram valores de score F1 satisfatórios tanto para o levantamento J-PLUS (0,93) quanto para o S-PLUS (0,95). Foi possível verificar também que a maioria das variáveis mais informativas para a classificação das subanãs quentes foram cores. Isso indica a importância de se considerar múltiplas informações fotométricas para a classificação de objetos astronômicos.

Apesar das variações de performance observadas nos modelos de identificação de subanãs quentes, especialmente devido ao uso de amostras com poucos objetos, os resultados obtidos indicam que esses modelos podem ser aplicados para identificação de subanãs quentes. A partir dessa conclusão, foram geradas listas de candidatas tanto para o J-PLUS (com 2604 objetos ainda não catalogados) quanto para o S-PLUS (com 226 objetos não catalogados).

Esses resultados são de grande importância para a área da astronomia, uma vez que fornecem uma forma rápida e eficiente de identificar candidatas à subanãs quentes dentro de uma grande amostra de objetos com base em dados fotométricos. Além disso, a meto-

dologia proposta para o desenvolvimento desses modelos pode ser aplicada na identificação de outras classes de estrelas.

Já no que diz respeito aos modelos de previsão de parâmetros estelares, desenvolvidos a partir de dados do levantamento LAMOST (ZHAO *et al.*, 2012), suas performances foram extremamente satisfatórias quando comparadas com a literatura existente.

Durante o desenvolvimento desses modelos, validou-se também a possibilidade do uso das magnitudes absolutas como variáveis de entrada dos modelos ao invés das magnitudes aparentes fornecidas pelos levantamentos. No caso do log(g), o uso das magnitudes absolutas se mostrou bastante benéfico para a performance do modelo, de modo que se optou pelo desenvolvimento de dois modelos distintos para a previsão desse parâmetro, onde seu uso depende da qualidade da determinação da distância do objeto.

Os modelos de previsão de T_{eff} obtiveram valores de desvio mediano absoluto (MAD) entre 45 K (J-PLUS) e 51 K (S-PLUS). Já os de previsão de log(g) obtiveram MADs entre 0,05 dex (J-PLUS) e 0,06 dex (S-PLUS), com os modelos que utilizaram magnitudes absolutas obtendo um resultado melhor do que aqueles que utilizaram magnitudes aparentes. Por fim, os modelos de previsão de metalicidade obtiveram MADs entre 0,07 dex (J-PLUS) e 0,09 dex (S-PLUS).

Também há uma grande importância nesses resultados, que mostram a possibilidade do uso da metodologia proposta baseada em Florestas Aleatórias para desenvolver modelos capazes de estimar os parâmetros estelares de um grande número de estrelas com base em dados fotométricos, tendo performances comparáveis e até melhores do que outros algoritmos utilizados na literatura (como o XGBoost e as Redes Neurais).

Em relação às perspectivas futuras, já se planeja a escrita de dois artigos com base nos resultados desenvolvidos neste trabalho: Um relacionado aos classificadores de subanãs quentes, e um relacionado aos estimadores de parâmetros estelares.

Além disso, os catálogos de parâmetros estelares para estrelas do S-PLUS e J-PLUS disponibilizados aqui podem ser utilizados como base para diversas aplicações, como o estudo de evolução da Galáxia, a identificação de estrelas pobres em metais, entre outros. De maneira similar, a lista de candidatas à subanãs quentes pode ser utilizada para um *follow-up* espectroscópico e a possível confirmação e expansão dos catálogos de objetos conhecidos dessa classe de estrela.

Por fim, é importante apontar o fato de que os modelos de classificação de subanãs quentes desenvolvidos durante esse trabalho podem ser melhorados significativamente caso sejam utilizadas amostras mais numerosas de objetos para seu treinamento. Outro ponto que pode ser explorado nesse sentido é o uso de diferentes algoritmos de ML ou técnicas de processamento de dados durante o desenvolvimento dos modelos.

No que diz respeito aos modelos de estimativa de parâmetros estelares, é possível levantar a possibilidade do uso de magnitudes e variáveis disponibilizadas por outros levantamentos para melhorar a precisão dos modelos, de maneira similar ao que a distância

(através das magnitudes absolutas) foi capaz de fazer para o modelo de previsão do log(g).

Além disso, considerando-se que os modelos treinados só são capazes de prever parâmetros dentro dos intervalos de valores presentes em suas amostras de desenvolvimento, é possível também apontar a necessidade de se obterem amostras capazes de cobrir intervalos mais extensos. No que diz respeito à temperatura efetiva e o log(g), os modelos treinados durante esse trabalho só são capazes de prever valores entre (3500 K, 8500 K) e (1,0, 5,0), respectivamente. Apesar de englobarem uma quantidade considerável de objetos, esses intervalos não permitem que se produza estimativas confiáveis de temperatura efetiva e log(g) para as candidatas à subanãs identificadas neste trabalho, por elas apresentarem temperaturas e log(g) mais elevados. No entanto, como as metalicidades desses objetos não estão fora do intervalo de funcionamento dos modelos de previsão de [Fe/H], as estimativas desse parâmetro podem ser utilizadas.

Bibliografia

AMARO-SEOANE, P., AUDLEY, H., BABAK, S., et al., 2017, "Laser Interferometer Space Antenna", *arXiv e-prints*, art. arXiv:1702.00786. doi: 10.48550/arXiv.1702.00786.

ANDRÉS GALARZA, C., DAFLON, S., PLACCO, V. M., et al., 2021, "J-PLUS: Searching for very metal-poor star candidates using the SPEEM pipeline", *arXiv e-prints*, art. arXiv:2109.11600. doi: 10.48550/arXiv.2109.11600.

BAILER-JONES, C. A. L., RYBIZKI, J., FOUESNEAU, M., et al., 2021, "VizieR Online Data Catalog: Distances to 1.47 billion stars in Gaia EDR3 (Bailer-Jones+, 2021)", *VizieR Online Data Catalog*, art. I/352.

BELCZYŃSKI, K., MIKOŁAJEWSKA, J., MUNARI, U., et al., 2000, "A catalogue of symbiotic stars", Astron. Astrophys. Suppl., v. 146 (nov.), pp. 407–435. doi: 10.1051/aas:2000280.

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

BONATTO, C., CHIES-SANTOS, A. L., COELHO, P. R. T., et al., 2019, "J-PLUS: A wide-field multi-band study of the M 15 globular cluster. Evidence of multiple stellar populations in the RGB", Astron. Astrophys., 622:A179. doi: 10.1051/0004-6361/201732441.

BREIMAN, L., 2001, "Random Forests", *Machine Learning*, v. 45, n. 1 (Oct), pp. 5–32. ISSN: 1573-0565. doi: 10.1023/A:1010933404324.

BRITO-SILVA, D., COELHO, P., CORTESI, A., et al., 2021, "J-PLUS: Detecting and studying extragalactic globular clusters – the case of NGC 1023", *arXiv e-prints*, art. arXiv:2110.04423. doi: 10.48550/arXiv.2110.04423.

BUZZO, M. L., CORTESI, A., FORBES, D. A., et al., 2021, "A new method to detect globular clusters with the S-PLUS survey", *Monthly Notices of the Royal Astronomical Society*, v. 510, n. 1 (12), pp. 1383–1392. ISSN: 0035-8711. doi: 10.1093/mnras/stab3489.

CATELAN, M., 2007, "Structure and Evolution of Low-Mass Stars: An Overview and Some Open Problems". In: Roig, F., Lopes, D. (Eds.), *Graduate School in Astronomy: XI Special Courses at the National Observatory of Rio de Janeiro (XI CCE)*, v. 930, *American Institute of Physics Conference Series*, pp. 39–90, set. doi: 10.1063/1.2790333.

CENARRO, A. J., MOLES, M., CRISTÓBAL-HORNILLOS, D., et al., 2019, "J-PLUS: The Javalambre Photometric Local Universe Survey", Astron. Astrophys., 622: A176. doi: 10.1051/0004-6361/201833036.

CHARPINET, S., VAN GROOTEL, V., FONTAINE, G., et al., 2011, "Deep asteroseismic sounding of the compact hot B subdwarf pulsator KIC02697388 from Kepler time series photometry", Astron. Astrophys., 530:A3. doi: 10.1051/0004-6361/201016412.

CHIES-SANTOS, A. L., DE SOUZA, R. S., CASO, J. P., et al., 2022, "J-PLUS: a catalogue of globular cluster candidates around the M81/M82/NGC3077 triplet of galaxies", *Monthly Notices of the Royal Astronomical Society*, v. 516, n. 1 (09), pp. 1320–1338. ISSN: 0035-8711. doi: 10.1093/mnras/stac2002.

COPPEJANS, D. L., KOERDING, E. G., KNIGGE, C., et al., 2016, "VizieR Online Data Catalog: Outburst catalogue of cataclysmic variables (Coppejans+, 2016)", *VizieR Online Data Catalog*, art. J/MNRAS/456/4441.

COSTA-DUARTE, M. V., SAMPEDRO, L., MOLINO, A., et al., 2019, "The S-PLUS: a star/galaxy classification based on a Machine Learning approach", *arXiv e-prints*, art. arXiv:1909.08626.

CULPAN, R., GEIER, S., REINDL, N., et al., 2022, "The population of hot subdwarf stars studied with Gaia. IV. Catalogues of hot subluminous stars based on Gaia EDR3", Astron. Astrophys., 662:A40. doi: 10.1051/0004-6361/202243337.

DAWSON, K. S., SCHLEGEL, D. J., AHN, C. P., et al., 2013, "The Baryon Oscillation Spectroscopic Survey of SDSS-III", Astron. J., 145(1):10. doi: 10.1088/0004-6256/145/1/10.

D'CRUZ, N. L., DORMAN, B., ROOD, R. T., et al., 1996, "The Origin of Extreme Horizontal Branch Stars", Astrophys. J., v. 466 (jul.), pp. 359. doi: 10.1086/177515.

DE CARVALHO, L. M., 2022, *Caracterização de Estrelas Fracas da missão KEPLER com base em dados do J-PLUS*. Tese de Mestrado. Divisão de Programas de Pós-Graduação - DIPPG.

FABRICIUS, C., LURI, X., ARENOU, F., et al., 2021, "Gaia Early Data Release 3. Catalogue validation", Astron. Astrophys., 649:A5. doi: 10.1051/0004-6361/202039834.

FEIGE, J., 1958, "A Search for Underluminous Hot Stars." Astrophys. J., v. 128 (set.), pp. 267. doi: 10.1086/146541.

GAIA COLLABORATION, PRUSTI, T., DE BRUIJNE, J. H. J., et al., 2016, "The Gaia mission", Astron. Astrophys., 595:A1. doi: 10.1051/0004-6361/201629272.

GEIER, S., 2020, "The population of hot subdwarf stars studied with Gaia. III. Catalogue of known hot subdwarf stars: Data Release 2", Astron. Astrophys., 635:A193. doi: 10.1051/0004-6361/202037526.

GEIER, S., FÜRST, F., ZIEGERER, E., et al., 2015, "The fastest unbound star in our Galaxy ejected by a thermonuclear supernova", *Science*, v. 347, n. 6226 (mar.), pp. 1126–1128. doi: 10.1126/science.1259063.

GEIER, S., ØSTENSEN, R. H., NEMETH, P., et al., 2017, "The population of hot subdwarf stars studied with Gaia. I. The catalog of known hot subdwarf stars", Astron. Astrophys., 600:A50. doi: 10.1051/0004-6361/201630135.

GENTILE FUSILLO, N. P., TREMBLAY, P. E., CUKANOVAITE, E., et al., 2021, "A catalogue of white dwarfs in Gaia EDR3", Mon. Not, Roy. Astron. Soc., v. 508, n. 3 (dez.), pp. 3877–3896. doi: 10.1093/mnras/stab2672.

GOODFELLOW, I. J., BENGIO, Y., COURVILLE, A., 2016, *Deep Learning*. Cambridge, MA, USA, MIT Press.

HALL, P. D., JEFFERY, C. S., 2016, "Hydrogen in hot subdwarfs formed by double helium white dwarf mergers", Mon. Not, Roy. Astron. Soc., v. 463, n. 3 (dez.), pp. 2756–2767. doi: 10.1093/mnras/stw2188.

HAN, Z., PODSIADLOWSKI, P., MAXTED, P. F. L., et al., 2002, "The origin of subdwarf B stars – I. The formation channels", *Monthly Notices of the Royal Astronomical Society*, v. 336, n. 2 (10), pp. 449–466. ISSN: 0035-8711. doi: 10.1046/j.1365-8711.2002.05752.x.

HAN, Z., PODSIADLOWSKI, P., MAXTED, P. F. L., et al., 2003, "The origin of subdwarf B stars - II", Mon. Not, Roy. Astron. Soc., v. 341, n. 2 (maio), pp. 669–691. doi: 10.1046/j.1365-8711.2003.06451.x.

HAN, Z., PODSIADLOWSKI, P., LYNAS-GRAY, A., 2010, "The formation of hot subdwarf stars and its implications for the UV-upturn phenomenon of elliptical galaxies", Astrophys. Space Sci., v. 329, n. 1-2 (out.), pp. 41–48. doi: 10.1007/s10509-010-0348-4.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., 2001, *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA, Springer New York Inc.

HEBER, U., 1986, "The atmosphere of subluminous B stars. II. Analysis of 10 helium poor subdwarfs and the birthrate of sdB stars." Astron. Astrophys., v. 155 (jan.), pp. 33–45.

HEBER, U., 2016, "Hot Subluminous Stars", Publ. Astron. Soc. Pacific, v. 128, n. 966 (ago.), pp. 082001. doi: 10.1088/1538-3873/128/966/082001.

HEBER, U., 2009, "Hot Subdwarf Stars", Ann. Rev. Astron. Astrohys., v. 47, n. 1 (set.), pp. 211–251. doi: 10.1146/annurev-astro-082708-101836.

JUSTHAM, S., PODSIADLOWSKI, P., HAN, Z., 2011, "On the formation of single and binary helium-rich subdwarf O stars", Mon. Not, Roy. Astron. Soc., v. 410, n. 2 (jan.), pp. 984–993. doi: 10.1111/j.1365-2966.2010.17497.x.

KOHAVI, R., JOHN, G. H., 1997, "Wrappers for feature subset selection", *Artificial Intelligence*, v. 97, n. 1, pp. 273–324. ISSN: 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00043-X.

KUPFER, T., GEIER, S., HEBER, U., et al., 2015, "Hot subdwarf binaries from the MUCHFUSS project. Analysis of 12 new systems and a study of the short-period binary population", Astron. Astrophys., 576:A44. doi: 10.1051/0004-6361/201425213.

KUPFER, T., KOROL, V., SHAH, S., et al., 2018, "LISA verification binaries with updated distances from Gaia Data Release 2", Mon. Not, Roy. Astron. Soc., v. 480, n. 1 (out.), pp. 302–309. doi: 10.1093/mnras/sty1545.

LATOUR, M., RANDALL, S. K., CALAMIDA, A., et al., 2018, "SHOTGLAS. I. The ultimate spectroscopic census of extreme horizontal branch stars in ω Centauri", Astron. Astrophys., 618:A15. doi: 10.1051/0004-6361/201833129.

LINDEGREN, L., 2018, "Re-normalising the astrometric chi-square in Gaia DR2", .

LÓPEZ-SANJUAN, C., VÁZQUEZ RAMIÓ, H., VARELA, J., et al., 2019, "J-PLUS: Morphological star/galaxy classification by PDF analysis", Astron. Astrophys., 622:A177. doi: 10.1051/0004-6361/201732480.

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

MILLER BERTOLAMI, M. M., ALTHAUS, L. G., UNGLAUB, K., et al., 2008, "Modeling He-rich subdwarfs through the hot-flasher scenario", Astron. Astrophys., v. 491, n. 1 (nov.), pp. 253–265. doi: 10.1051/0004-6361:200810373.

MOLINO, A., COSTA-DUARTE, M. V., MENDES DE OLIVEIRA, C., et al., 2019, "J-PLUS: On the identification of new cluster members in the double galaxy cluster A2589 and A2593 using PDFs", Astron. Astrophys., 622:A178. doi: 10.1051/0004-6361/201731348.

MOLL, R., RASKIN, C., KASEN, D., et al., 2014, "Type Ia Supernovae from Merging White Dwarfs. I. Prompt Detonations", Astrophys. J., 785(2):105. doi: 10.1088/0004-637X/785/2/105.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., et al., 2011, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, v. 12, pp. 2825–2830.

PELISOLI, I., VOS, J., GEIER, S., et al., 2020, "Alone but not lonely: Observational evidence that binary interaction is always required to form hot subdwarf stars", Astron. Astrophys., 642:A180. doi: 10.1051/0004-6361/202038473.

PELISOLI, I., NEUNTEUFEL, P., GEIER, S., et al., 2021, "A hot subdwarf-white dwarf super-Chandrasekhar candidate supernova Ia progenitor", *Nature Astronomy*, v. 5 (jul.), pp. 1052–1061. doi: 10.1038/s41550-021-01413-0.

PLACCO, V. M., ROEDERER, I. U., LEE, Y. S., et al., 2021, "SPLUS J210428.01-004934.2: An Ultra Metal-poor Star Identified from Narrowband Photometry", Astrophys. J. Lett., 912(2):L32. doi: 10.3847/2041-8213/abf93d.

PODSIADLOWSKI, P., HAN, Z., LYNAS-GRAY, A. E., et al., 2008, "Hot Subdwarfs in Binaries as the Source of the Far-UV Excess in Elliptical Galaxies". In: Heber, U., Jeffery, C. S., Napiwotzki, R. (Eds.), *Hot Subdwarf Stars and Related Objects*, v. 392, *Astronomical Society of the Pacific Conference Series*, p. 15, jan. doi: 10.48550/arXiv.0808.0574.

SAN ROMAN, I., SÁNCHEZ-BLÁZQUEZ, P., CENARRO, A. J., et al., 2019, "J-PLUS: Two-dimensional analysis of the stellar population in NGC 5473 and NGC 5485", Astron. Astrophys., 622:A181. doi: 10.1051/0004-6361/201832894.

SARGENT, W. L. W., SEARLE, L., 1968, "A Quantitative Description of the Spectra of the Brighter Feige Stars", Astrophys. J., v. 152 (maio), pp. 443. doi: 10.1086/149561.

SCHLAFLY, E. F., FINKBEINER, D. P., 2011, "Measuring Reddening with Sloan Digital Sky Survey Stellar Spectra and Recalibrating SFD", Astrophys. J., 737(2): 103. doi: 10.1088/0004-637X/737/2/103.

SCHWAB, J., 2018, "Hot subdwarfs formed from the merger of two He white dwarfs", Mon. Not, Roy. Astron. Soc., v. 476, n. 4 (jun.), pp. 5303–5311. doi: 10.1093/mnras/sty586.

TILLICH, A., HEBER, U., GEIER, S., et al., 2011, "The Hyper-MUCHFUSS project: probing the Galactic halo with sdB stars", Astron. Astrophys., 527:A137. doi: 10.1051/0004-6361/201015539.

TOONEN, S., NELEMANS, G., PORTEGIES ZWART, S., 2012, "Supernova Type Ia progenitors from merging double white dwarfs. Using a new population synthesis model", Astron. Astrophys., 546:A70. doi: 10.1051/0004-6361/201218966.

VAN GROOTEL, V., CHARPINET, S., FONTAINE, G., et al., 2010, "Structural and core parameters of the hot B subdwarf KPD 0629-0016 from CoRoT g-mode asteroseismology", Astron. Astrophys., 524:A63. doi: 10.1051/0004-6361/201015437.

VAN RIJN, J. N., HUTTER, F., 2018, "Hyperparameter Importance Across Datasets". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, p. 2367–2376, New York, NY, USA, out. Association for Computing Machinery. ISBN: 9781450355520. doi: 10.1145/3219819.3220058.

VANWINCKELEN, G., BLOCKEEL, H., DE BAETS, B., et al., 2012. "On estimating model accuracy with repeated cross-validation". ISSN: 978-94-6197-044-2.

WANG, C., BAI, Y., YUAN, H., et al., 2022, "J-PLUS: Support vector regression to measure stellar parameters", *A&A*, v. 664, pp. A38. doi: 10.1051/0004-6361/202243130.

WANG, C., BAI, Y., LÓPEZ-SANJUAN, C., et al., 2022, "J-PLUS: Support vector machine applied to STAR-GALAXY-QSO classification", *A&A*, v. 659, pp. A144. doi: 10.1051/0004-6361/202142254.

WEBBINK, R. F., 1984, "Double white dwarfs as progenitors of R Coronae Borealis stars and type I supernovae." Astrophys. J., v. 277 (fev.), pp. 355–360. doi: 10.1086/161701.

WHITTEN, D. D., PLACCO, V. M., BEERS, T. C., et al., 2019, "J-PLUS: Identification of low-metallicity stars with artificial neural networks using SPHINX", *A&A*, v. 622, pp. A182. doi: 10.1051/0004-6361/201833368.

WHITTEN, D. D., PLACCO, V. M., BEERS, T. C., et al., 2021, "The Photometric Metallicity and Carbon Distributions of the Milky Way's Halo and Solar Neighborhood from S-PLUS Observations of SDSS Stripe 82", Astrophys. J., 912(2): 147. doi: 10.3847/1538-4357/abee7e.

WU, Y., DU, B., LUO, A., et al., 2014, "Automatic stellar spectral parameterization pipeline for LAMOST survey". In: Heavens, A., Starck, J.-L., Krone-Martins, A. (Eds.), *Statistical Challenges in 21st Century Cosmology*, v. 306, pp. 340–342, maio. doi: 10.1017/S1743921314010825.

YAN, H., LI, H., WANG, S., et al., 2022, "Overview of the LAMOST survey in the first decade", *The Innovation*, v. 3, n. 2 (Mar). ISSN: 2666-6758. doi: 10.1016/j.xinn.2022.100224.

YANG, L., SHAMI, A., 2020, "On hyperparameter optimization of machine learning algorithms: Theory and practice", *Neurocomputing*, v. 415, pp. 295–316. ISSN: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.07.061.

YANG, L., YUAN, H., XIANG, M., et al., 2022, "J-PLUS: Stellar parameters, C, N, Mg, Ca, and [α/Fe] abundances for two million stars from DR1", Astron. Astrophys., 659:A181. doi: 10.1051/0004-6361/202142724.

YANNY, B., ROCKOSI, C., NEWBERG, H. J., et al., 2009, "SEGUE: A Spectroscopic Survey of 240,000 Stars with g = 14-20", Astron. J., v. 137, n. 5 (maio), pp. 4377–4399. doi: 10.1088/0004-6256/137/5/4377.

YORK, D. G., ADELMAN, J., ANDERSON, JOHN E., J., et al., 2000a, "The Sloan Digital Sky Survey: Technical Summary", Astron. J., v. 120, n. 3 (set.), pp. 1579–1587. doi: 10.1086/301513.

YORK, D. G., ADELMAN, J., ANDERSON, JOHN E., J., et al., 2000b, "The Sloan Digital Sky Survey: Technical Summary", Astron. J., v. 120, n. 3 (set.), pp. 1579–1587. doi: 10.1086/301513.

YU, J., ZHANG, X., LÜ, G., 2021, "Post-merger evolution of double helium white dwarfs and distribution of helium-rich hot subdwarfs", *Monthly Notices of the Royal*

Astronomical Society, v. 504, n. 2 (05), pp. 2670–2674. ISSN: 0035-8711. doi: 10.1093/mnras/stab1063.

ZHANG, X., JEFFERY, C. S., 2012, "The Origin and Evolution of Helium-rich Hot Subdwarfs". In: Kilkenny, D., Jeffery, C. S., Koen, C. (Eds.), *Fifth Meeting on Hot Subdwarf Stars and Related Objects*, v. 452, *Astronomical Society of the Pacific Conference Series*, p. 13, mar.

ZHAO, G., ZHAO, Y., CHU, Y., et al., 2012, "LAMOST Spectral Survey", *arXiv e-prints*, art. arXiv:1206.3569. doi: 10.48550/arXiv.1206.3569.

Apêndice A

Otimização de Hiperparâmetros

A.1 Modelo P-LOGG-APP

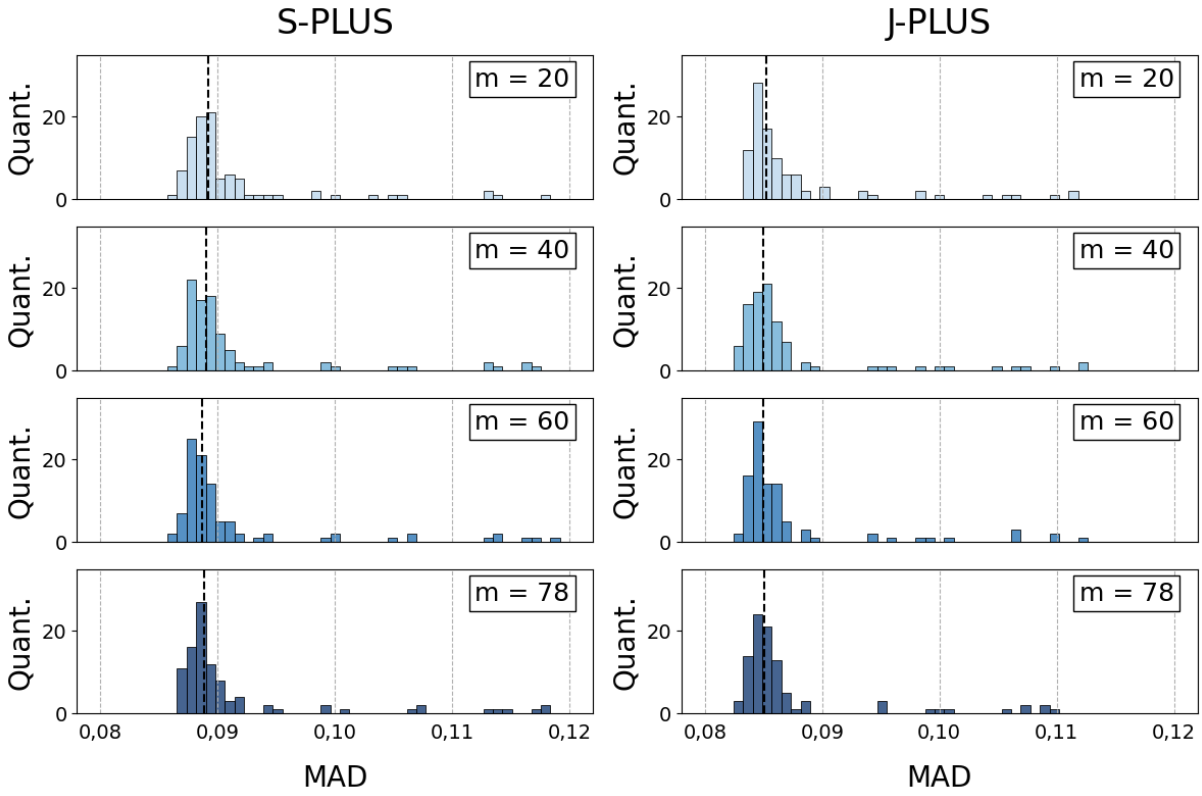

Figura A.1: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes aparentes em função do hiperparâmetro m para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

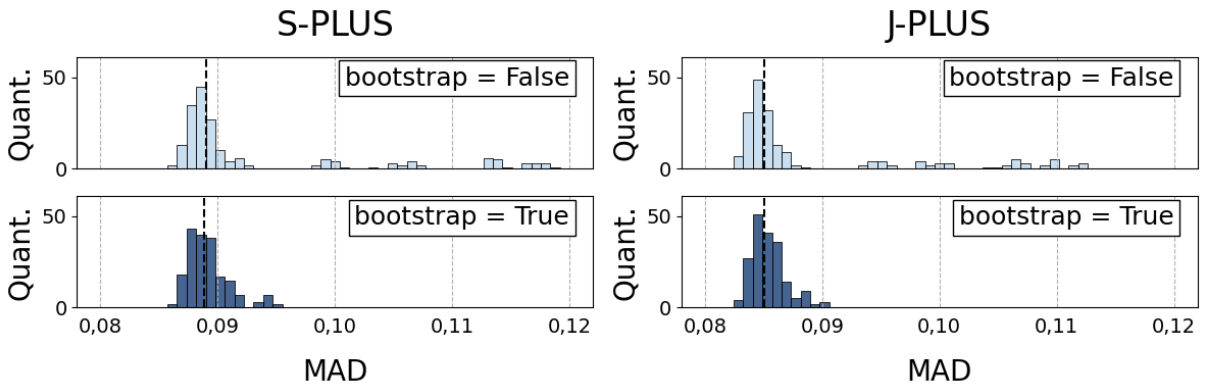

Figura A.2: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes aparentes em função do hiperparâmetro *bootstrap* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

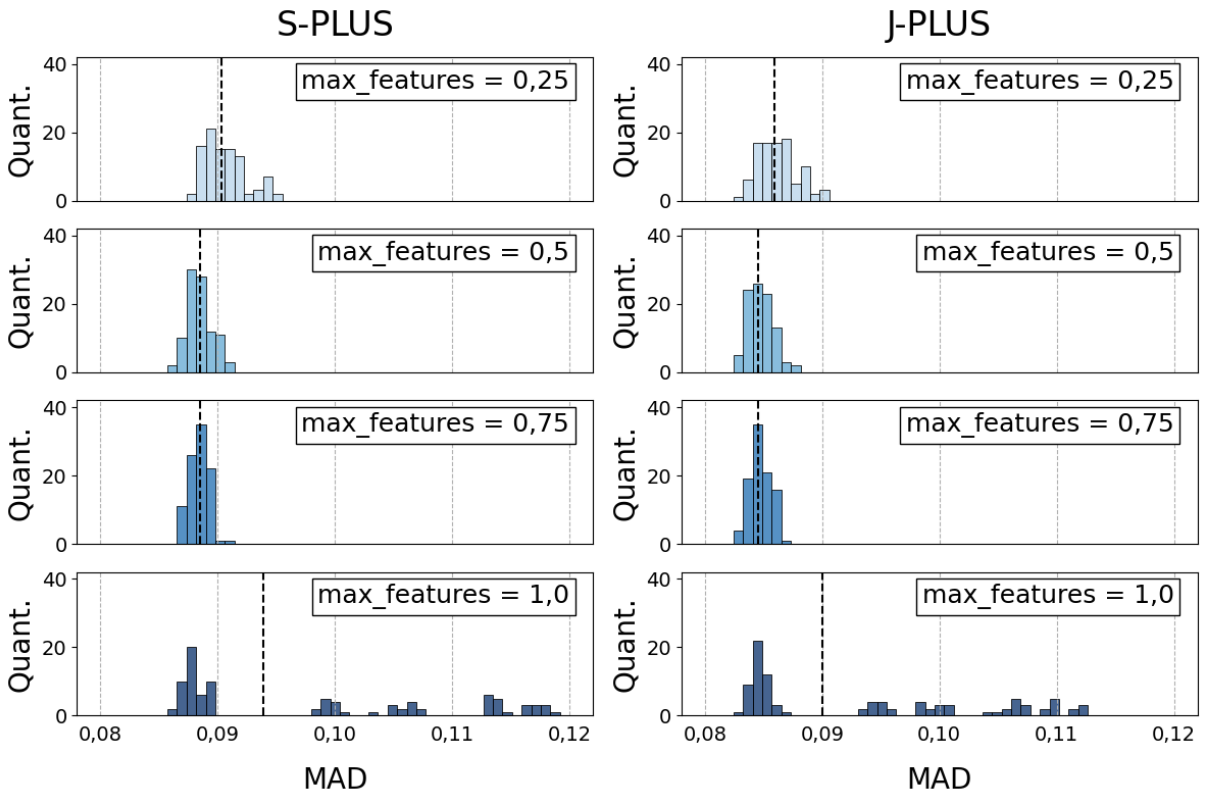

Figura A.3: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes aparentes em função do hiperparâmetro *max_features* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

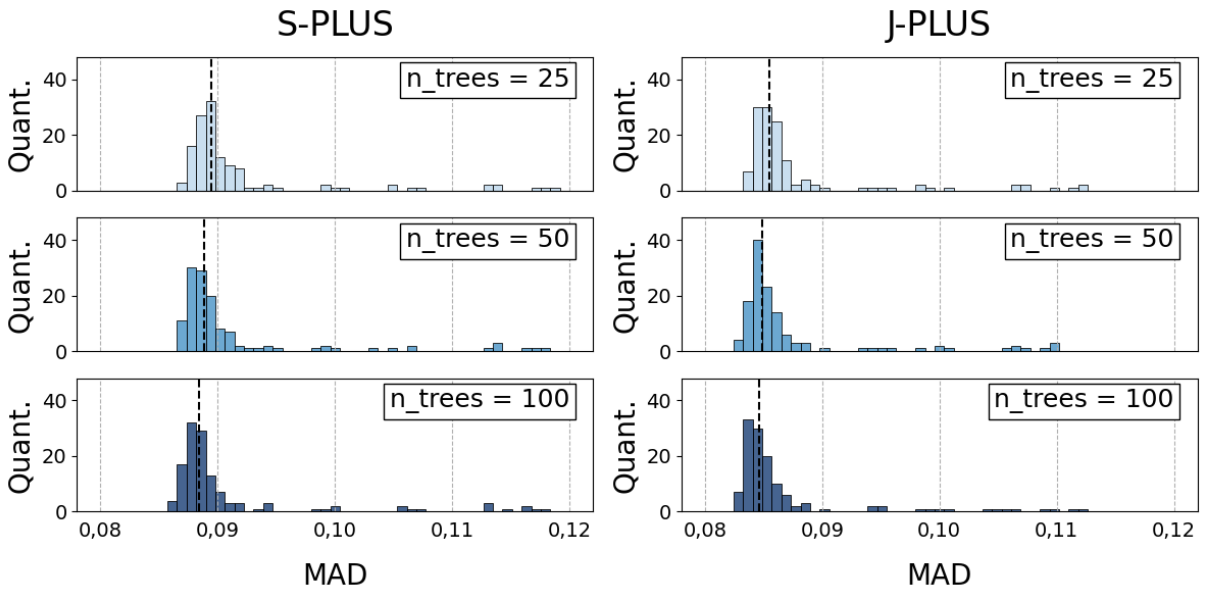

Figura A.4: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes aparentes em função do hiperparâmetro *n_trees* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

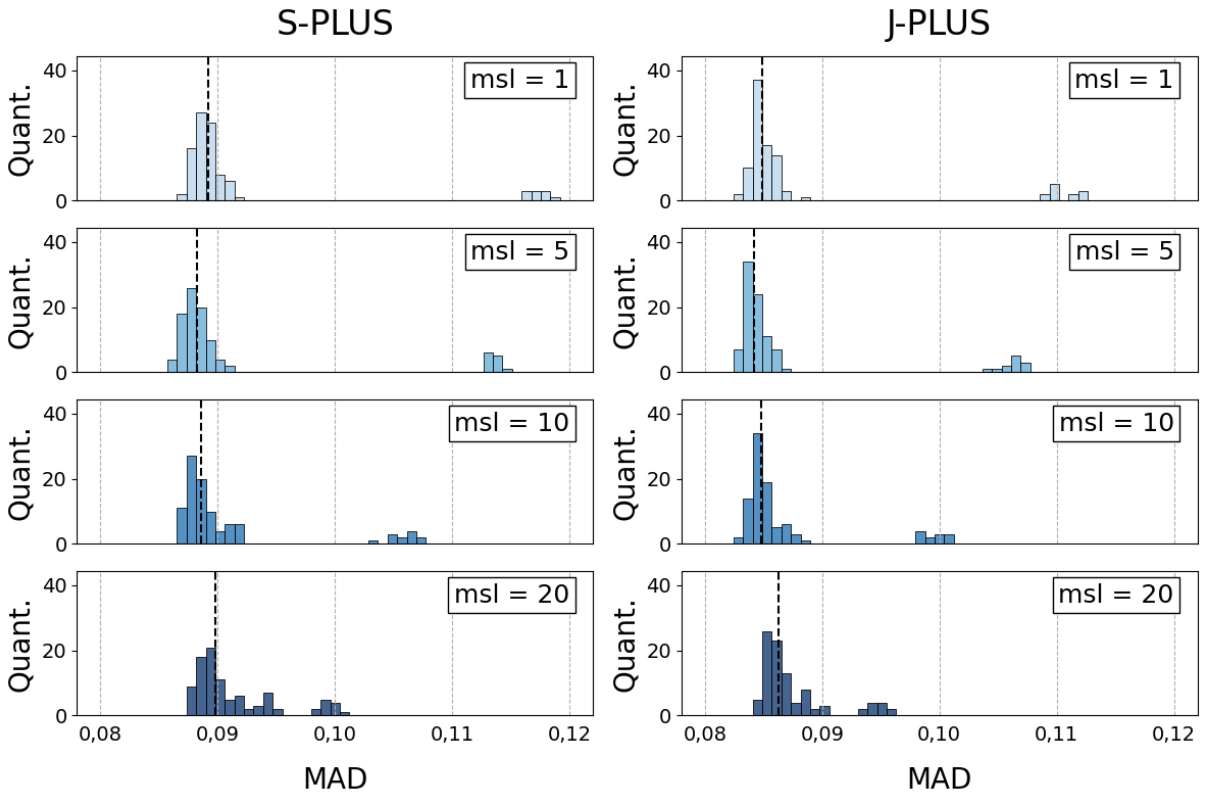

Figura A.5: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes aparentes em função do hiperparâmetro *min_samples_leaf* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

A.2 Modelo P-LOGG-ABS

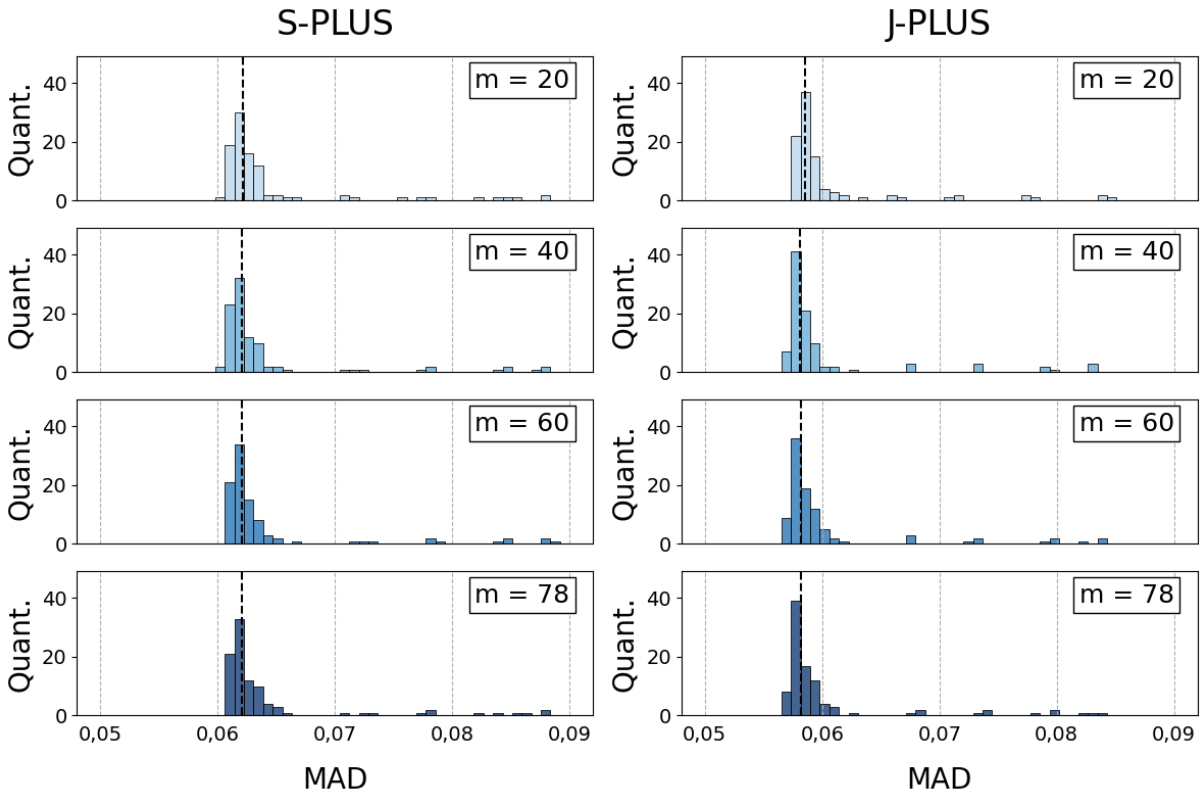

Figura A.6: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes absolutas em função do hiperparâmetro m para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

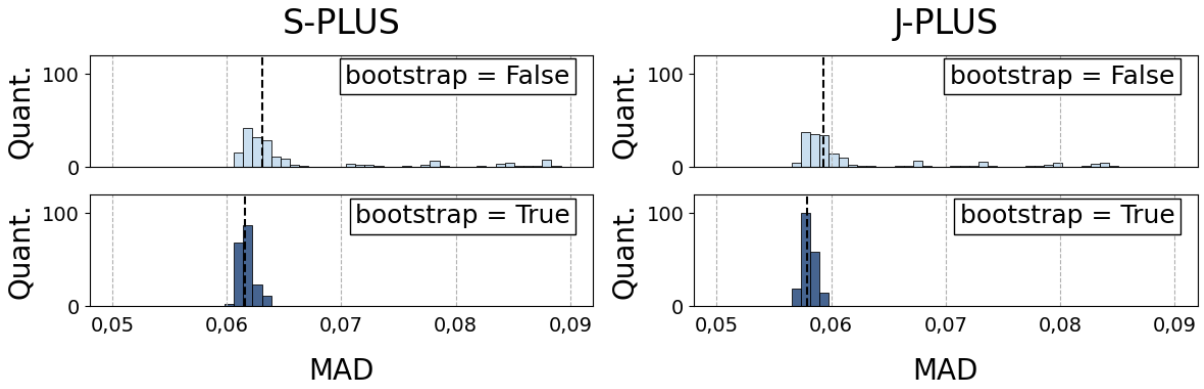

Figura A.7: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes absolutas em função do hiperparâmetro *bootstrap* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

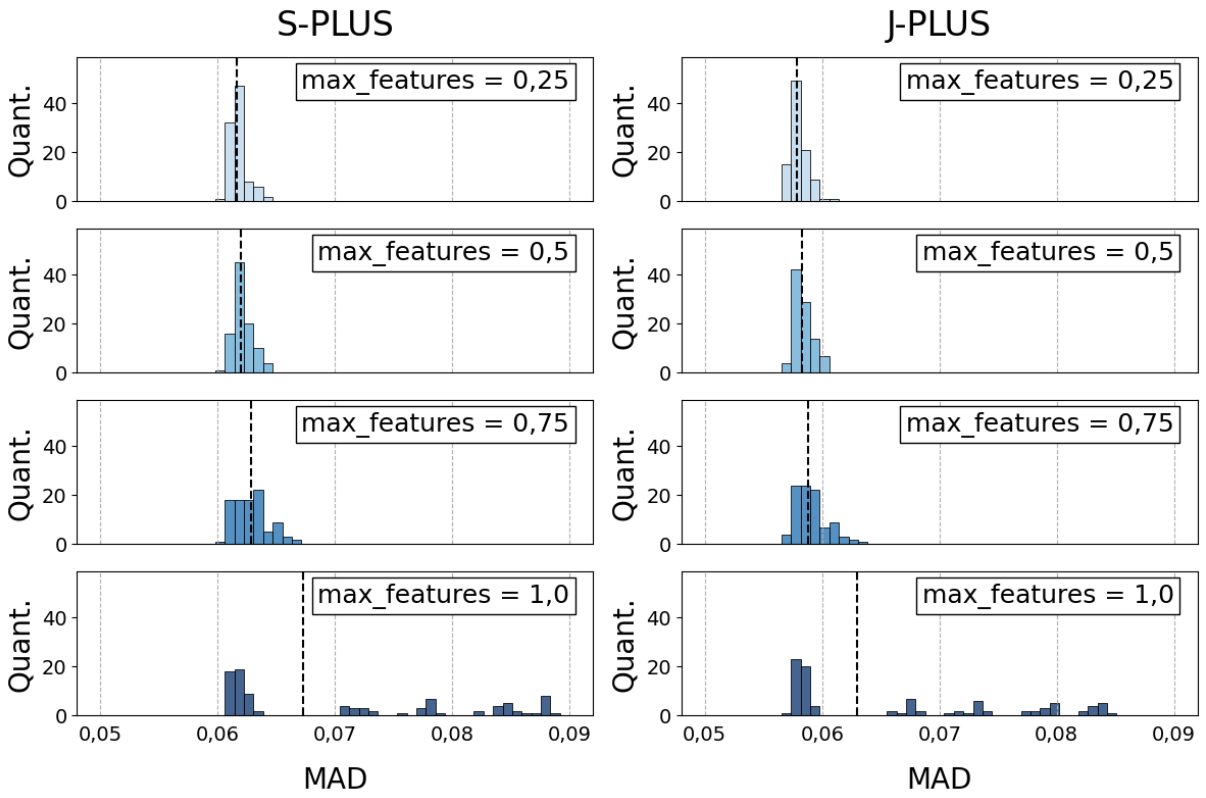

Figura A.8: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes absolutas em função do hiperparâmetro *max_features* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

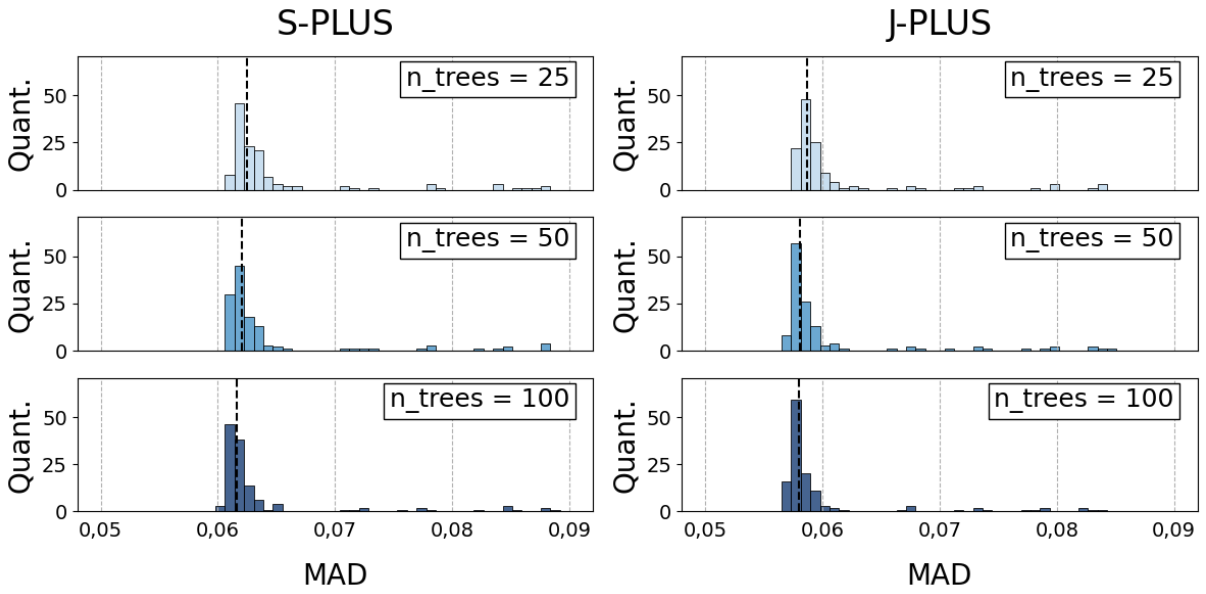

Figura A.9: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes absolutas em função do hiperparâmetro *n_trees* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

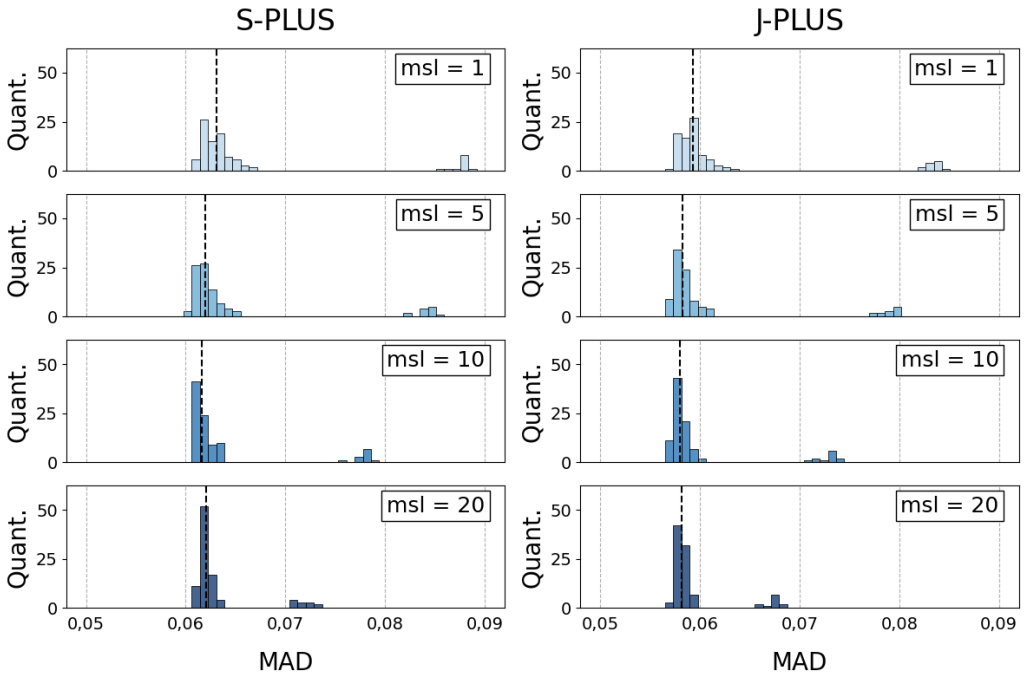

Figura A.10: Distribuição dos MADs dos modelos de previsão de log(g) a partir de magnitudes absolutas em função do hiperparâmetro *min_samples_leaf* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

A.3 Modelo P-FEH

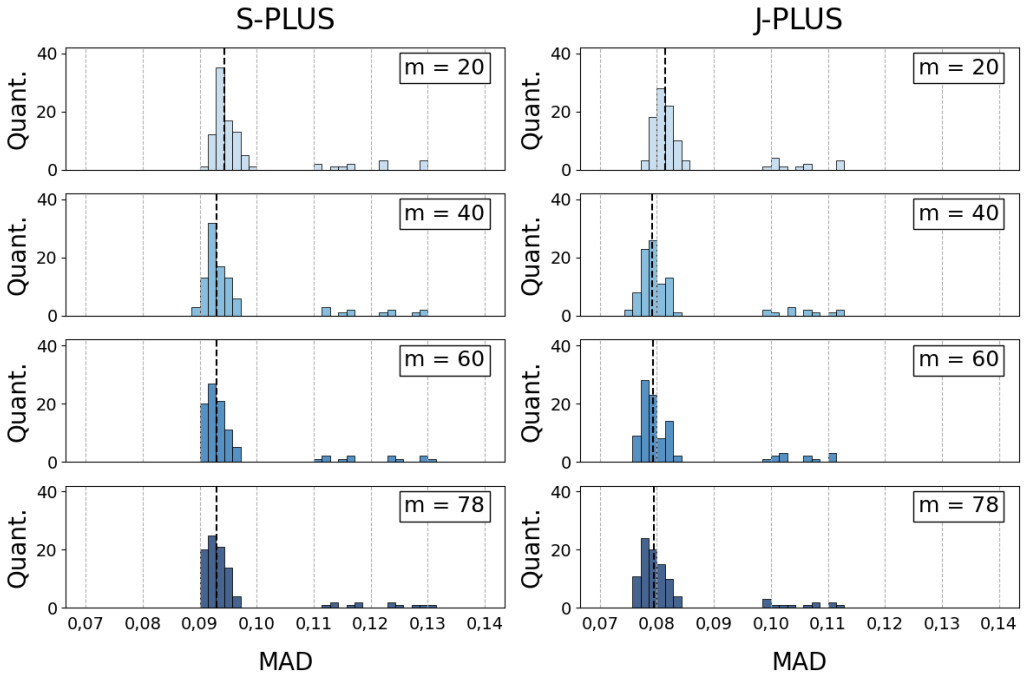

Figura A.11: Distribuição dos MADs dos modelos de previsão de [Fe/H] em função do hiperparâmetro m para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

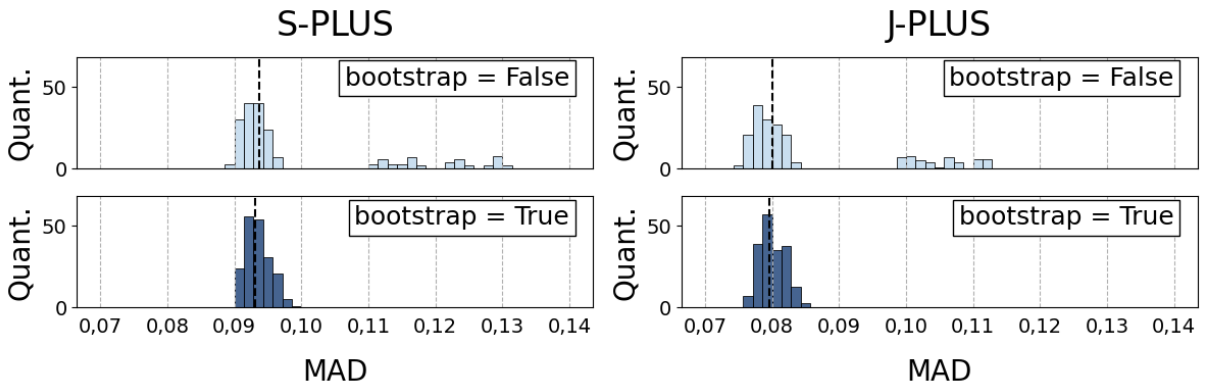

Figura A.12: Distribuição dos MADs dos modelos de previsão de [Fe/H] em função do hiperparâmetro *bootstrap* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

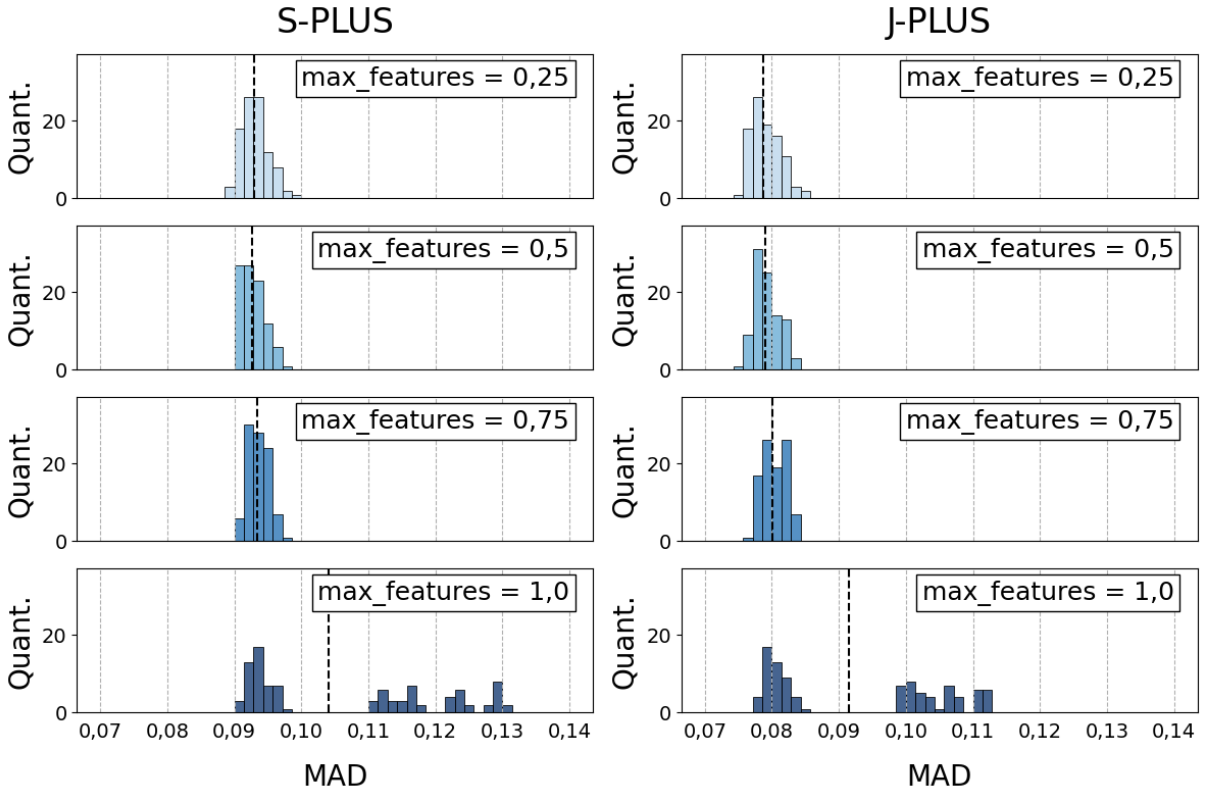

Figura A.13: Distribuição dos MADs dos modelos de previsão de [Fe/H] em função do hiperparâmetro *max_features* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

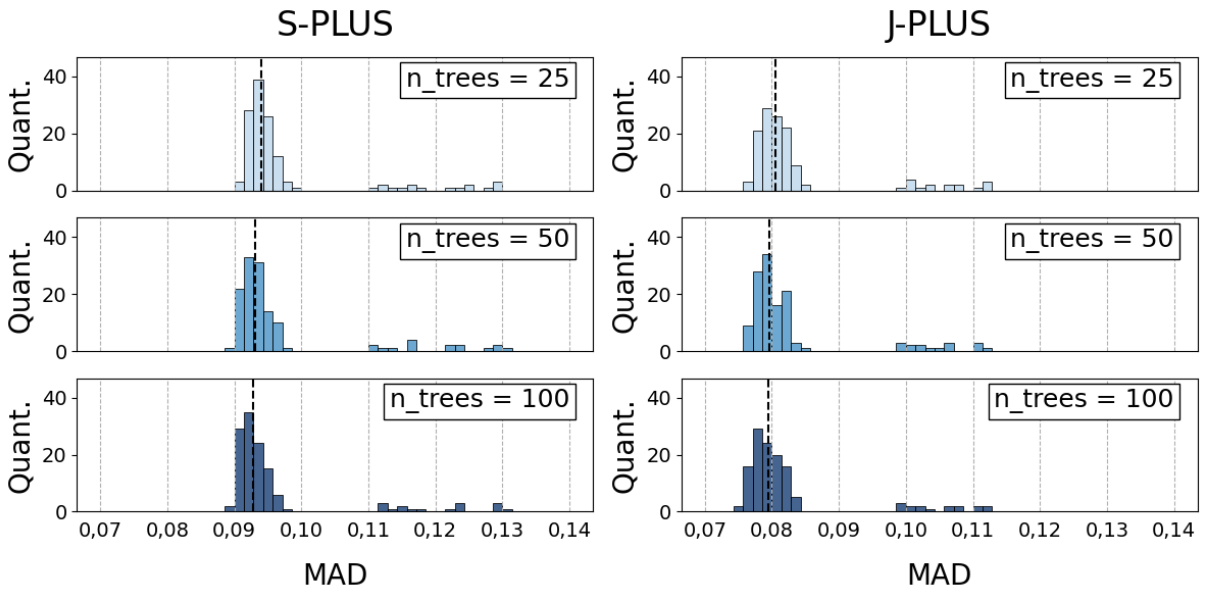

Figura A.14: Distribuição dos MADs dos modelos de previsão de [Fe/H] em função do hiperparâmetro *n_trees* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

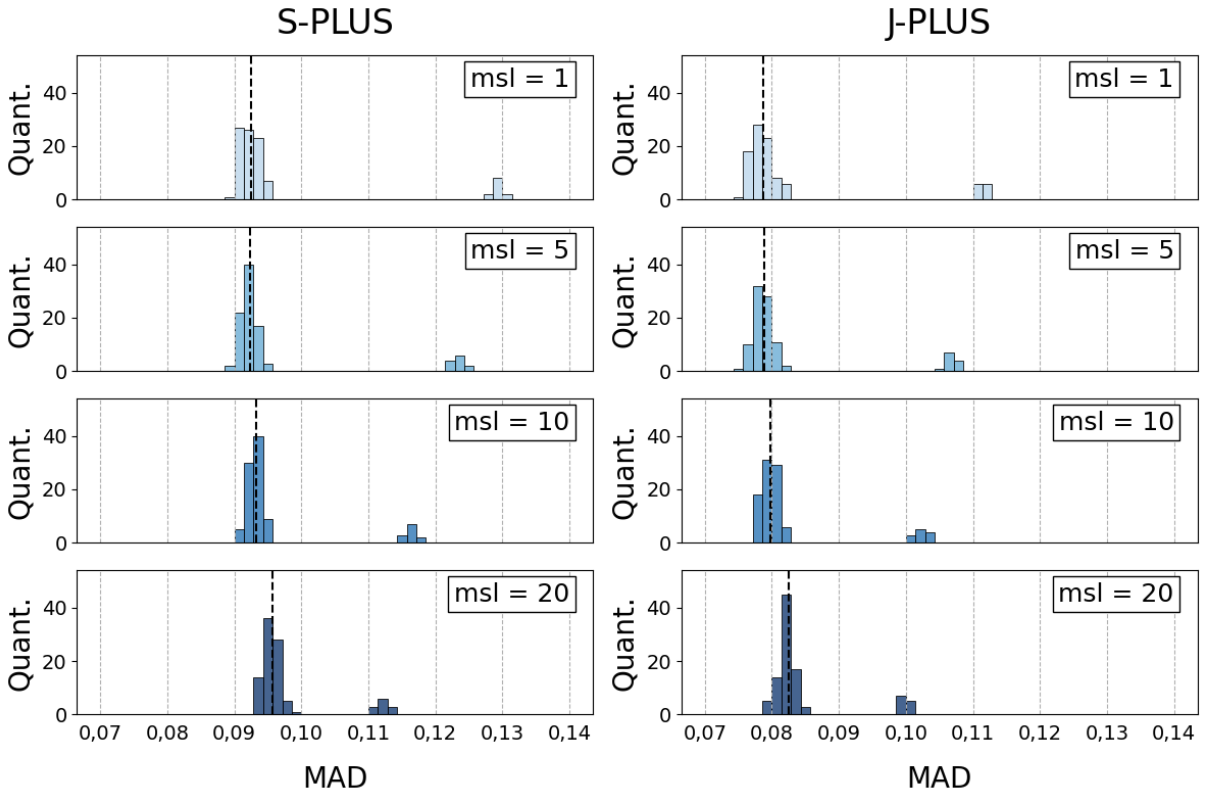

Figura A.15: Distribuição dos MADs dos modelos de previsão de [Fe/H] em função do hiperparâmetro *min_samples_leaf* para a amostra do J-PLUS e do S-PLUS. As linhas pontilhadas representam as medianas dos MADs para cada distribuição.

Apêndice B

Combinações de Hiperparâmetros

B.1 Modelo P-TEFF

B.1.1 S-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
40	100	5	True	1.00	51
78	50	5	True	1.00	51
78	100	5	True	0.50	51
60	100	10	True	1.00	51
40	100	5	True	0.50	51
40	100	5	False	0.25	51
60	100	5	False	0.25	51
78	100	5	True	1.00	51
40	100	1	True	1.00	51
40	100	1	True	0.50	51

Tabela B.1: Combinações de hiperparâmetros dos dez melhores modelos de previsão de T_{eff} em relação ao seu MAD na amostra de validação do S-PLUS.

B.1.2 J-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
60	100	5	True	0.75	46
78	100	5	True	0.25	46
78	100	1	True	0.50	46
78	100	5	False	0.25	46
60	100	1	True	0.50	46
20	100	5	False	0.25	46
60	100	10	True	1.00	46
78	100	5	True	0.50	46
60	100	5	True	0.25	47
78	100	5	True	0.75	47

Tabela B.2: Combinações de hiperparâmetros dos dez melhores modelos de previsão de T_{eff} em relação ao seu MAD na amostra de validação do J-PLUS.

B.2 Modelo P-LOGG-APP

B.2.1 S-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
60	100	5	True	1.00	0.09
60	100	5	False	0.50	0.09
40	100	5	False	0.50	0.09
20	100	5	True	1.00	0.09
78	50	5	False	0.50	0.09
40	50	10	False	0.75	0.09
78	100	10	True	1.00	0.09
20	50	5	True	1.00	0.09
78	100	5	True	0.75	0.09
20	100	5	False	0.500	0.09

Tabela B.3: Combinações de hiperparâmetros dos dez melhores modelos de previsão de log(g) a partir de magnitudes aparentes em relação ao seu MAD na amostra de validação do S-PLUS.

B.2.2 J-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
40	100	1	False	0.50	0.08
40	100	10	False	0.50	0.08
40	50	5	False	0.50	0.08
40	100	5	True	0.75	0.08
60	50	5	False	0.50	0.08
78	100	5	True	1.00	0.08
78	100	5	False	0.25	0.08
40	100	1	True	0.75	0.08
40	100	5	False	0.75	0.08
78	50	10	False	0.50	0.08

Tabela B.4: Combinações de hiperparâmetros dos dez melhores modelos de previsão de log(g) a partir de magnitudes aparentes em relação ao seu MAD na amostra de validação do J-PLUS.

B.3 Modelo P-LOGG-ABS

B.3.1 S-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
40	100	5	True	0.50	0.06
40	100	5	True	0.75	0.06
20	100	5	True	0.25	0.06
40	100	10	True	0.25	0.06
78	50	10	True	1.00	0.06
78	50	5	True	0.75	0.06
60	100	5	True	1.00	0.06
78	100	10	True	0.25	0.06
60	100	10	True	0.25	0.06
40	100	10	True	1.00	0.06

Tabela B.5: Combinações de hiperparâmetros dos dez melhores modelos de previsão de log(g) a partir de magnitudes absolutas em relação ao seu MAD na amostra de validação do S-PLUS.

B.3.2 J-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
78	100	10	True	0.25	0.06
78	100	5	True	0.25	0.06
60	50	5	True	0.75	0.06
40	50	10	True	0.25	0.06
40	100	5	True	0.75	0.06
78	100	5	True	1.00	0.06
78	100	1	True	0.25	0.06
78	50	10	False	0.25	0.06
60	100	5	True	0.75	0.06
60	100	10	True	0.25	0.06

Tabela B.6: Combinações de hiperparâmetros dos dez melhores modelos de previsão de log(g) a partir de magnitudes absolutas em relação ao seu MAD na amostra de validação do J-PLUS.

B.4 Modelo P-FEH

B.4.1 S-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
40	100	5	False	0.25	0.09
40	50	5	False	0.25	0.09
40	100	1	False	0.25	0.09
40	100	5	False	0.50	0.09
78	50	1	False	0.25	0.09
60	100	1	False	0.25	0.09
78	100	1	False	0.25	0.09
40	50	1	False	0.25	0.09
60	50	5	False	0.25	0.09
40	100	1	False	0.50	0.09

Tabela B.7: Combinações de hiperparâmetros dos dez melhores modelos de previsão de [Fe/H] em relação ao seu MAD na amostra de validação do S-PLUS.

B.4.2 J-PLUS

m	n_trees	msl	bootstrap	max_features	MAD
40	100	1	False	0.50	0.08
40	100	5	False	0.25	0.08
60	100	5	False	0.25	0.08
78	100	1	False	0.25	0.08
40	50	5	False	0.25	0.08
78	100	5	False	0.25	0.08
60	50	1	False	0.25	0.08
40	100	1	False	0.25	0.08
60	100	1	False	0.25	0.08
40	50	1	False	0.25	0.08

Tabela B.8: Combinações de hiperparâmetros dos dez melhores modelos de previsão de [Fe/H] em relação ao seu MAD na amostra de validação do J-PLUS.