Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Astronomia

Lucas Miguel Fanti

Machine Learning em Astronomia: Uma Aplicação em Populações Estelares

São Paulo
2022

Lucas Miguel Fanti

Machine Learning em Astronomia: Uma Aplicação em Populações Estelares

Monografia apresentada ao Departamento de Astronomia do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Astronomia.

Orientador: Prof. Dr. Alex Cavalieri Carciofi

São Paulo

2022

—Dedico este trabalho àqueles que sempre buscam os novos horizontes de conhecimentos. Àqueles que nunca estão satisfeitos com respostas prontas e que não tem medo de errar tentando chegar um passo mais longe.

Agradecimentos

Ao meu orientador, Prof. Dr. Alex Cavalieri Carciofi, por me auxiliar na elaboração do trabalho e por sempre me guiar na direção correta da produção científica;

Ao meu colega e amigo Pedro Ticiani, que teve papel fundamental para tanto gerar o material de análise deste trabalho como na revisão do texto;

A meus pais, que sempre me incentivaram a seguir minhas aspirações e desde pequeno me estimularam a ser curioso e sempre questionar. Educação não é encher um balde: é acender uma chama;

À companheira da minha vida, Mirella. A pessoa que sempre me apoiou, sempre me levantou quando caí e me celebrou quando venci. A pessoa que a todos os dias me faz querer ser alguém melhor e que me ensina sempre sobre uma vida feliz e repleta de amor;

Esse trabalho não seria possível sem cada um de vocês.

"*Em algum lugar, algo incrível está esperando para ser descoberto.*"

Carl Sagan

"*Os filósofos apenas interpretaram o mundo de diferentes maneiras; o que importa é transformá-lo.*"

Karl Marx

Resumo

O presente estudo busca compreender ferramentas modernas e atuais como *machine learning* e entender de que formas estas ferramentas podem ser úteis para resolver problemas em astronomia, em particular, problemas de populações estelares. Para isso, foram gerados aglomerados sintéticos, dadas algumas hipóteses sobre as características intrínsecas e extrínsecas da população, por meio de um processo de fotometria sintética, utilizando o conjunto de filtros Javalambre do levantamento fotométrico S-PLUS, em espectros de modelos realísticos de estrelas B e Be da grade BeAtlas, produzidos através do código de transferência radiativa HDUST, que serviram como base de testes para modelos de *machine learning* focados em fazer classificações visando identificar estrelas Be.

Modelos do tipo *Support Vector Machines* mostraram os melhores resultados para dados sintéticos. Ao aplicar o modelo para dados reais do aglomerado NGC 330, foram geradas estimativas de grandes quantidades de estrelas candidatas a Be. Este número de estrelas pode ser reduzido ao calibrar o modelo, criando um valor de corte de 99.9% de probabilidade de classificação. Além disso, estudos de *Permutation Feature Importance* mostram quais filtros fotométricos são os mais importantes para a classificação, sendo encontrados para o modelo os filtros i, J395, J660 e J861. Foram construídos diagramas a partir das classificações feitas pelo modelo, fornecendo uma visualização de quais regiões se encontram as estrelas Be.

Abstract

The present study seeks to understand modern and state-of-the-art tools, such as machine learning, in order to understand how these tools can be useful to solve problems in astronomy, in particular in stellar populations. Synthetic clusters were generated, given some assumptions about the intrinsic and extrinsic characteristics of the population, through a synthetic photometry process using the Javalambre filter set from the S-PLUS photometric survey, on realistic model spectra of B and Be stars from the BeAtlas grid, produced using the radiative transfer code HDUST, which served as the basis of tests for machine learning models focused on making classifications aimed at identifying Be stars.

Support Vector Machines models showed the best results for synthetic data. When applying the model to real data for the open stellar cluster NGC 330, estimates of large numbers of Be candidate stars were generated. This number of stars could be reduced by calibrating the model and creating a cutoff value of 99.9 percent probability of classification. In addition, Permutation Feature Importance studies show which photometric filters are the most important for classification, with the i, J395, J660 and J861 filters being found for the model. Diagrams were constructed from the classifications made by the model, providing a visualization of which regions the Be stars are found.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 *Noções gerais de* machine learning

Conforme a humanidade aprimora seu potencial tecnológico, também cresce sua capacidade de armazenar e de coletar dados. Dados possibilitam a visão crítica e objetiva da natureza, pois, ao utilizar-se técnicas estatísticas, é possível perceber tendências dentro dos conjuntos de dados e, assim, extrair conclusões sobre a natureza da entidade, fenômeno ou objeto nos quais os dados se baseiam. Técnicas modernas e sofisticadas podem ser utilizadas para avaliar grandes quantidades de dados, com o auxílio de computadores e linguagens de programação, e é nesse encontro entre a estatística e a computação que surge a ciência de dados. A partir de quantidades colossais de dados, muitas vezes em formas não facilmente acessíveis para a exploração humana convencional, grandes computadores conseguem encontrar relações de forma rápida e eficiente (Dhar, 2013). Uma das principais formas de explorar o potencial das gigantescas bases de dados é a inteligência artificial.

A inteligência artificial pode ser definida como o atributo da capacidade de tomada de decisões próprias de forma artificial, não-natural. Uma máquina de lavar roupas que saiba em que momento e sob quais circunstâncias deve iniciar seu ciclo de lavagem, sem a necessidade de um ser humano dar explicitamente o comando, é um exemplo de inteligência artificial.

O aprendizado de máquina, mais conhecido por seu termo em inglês *machine learning*, é uma subcategoria de inteligência artificial. De maneira frequente, embora erroneamente, usado como sinônimo de inteligência artificial, este é definido pelo que chamamos de programação implícita: não é preciso definir de forma clara e explícita quais são as condições necessárias para que a máquina tome sua decisão. À luz do supracitado exemplo da

máquina de lavar roupas, não se considera *machine learning* se esta for programada para sempre iniciar seu ciclo às nove horas da manhã. Para tal, é preciso o fornecimento de grandes quantidades de registros de comportamentos humanos, além das informações de como e quando tais pessoas utilizavam suas máquinas de lavar roupa. Desta maneira, a partir desses dados, o algoritmo encontra padrões e relações e define as melhores condições para seu funcionamento (Campesato, 2020).

Isto posto, vê-se que definir manualmente os critérios para que uma decisão seja tomada não é prático para muitos problemas do mundo moderno. Isso se deve ao fato de que os critérios acabam sendo muito específicos a um problema em particular e uma ligeira variação pode exigir a mudança das condições. Além disso, a lógica para a tomada de decisão pode ser muito complexa para certos casos ou até mesmo ser difícil de ser interpretada para a linguagem de uma máquina. Como exemplo, sabe-se que a forma que o cérebro humano vê imagens é bem diferente da forma que um computador enxerga *pixels*. Por conta disso, o problema de reconhecimento de rostos por computadores era considerado sem solução até tão recentemente quanto 2001 (Müller e Guido, 2016). Hoje, os algoritmos de *machine learning* precisam apenas de milhões de imagens com rostos e o computador se encarrega de fazer o processo de classificação.

A fim de tratar da classificação dos algoritmos de *machine learning*, estes se dividem em três tipos: algoritmos supervisionados, algoritmos não-supervisionados e algoritmos de reforço. A presente pesquisa foca especificamente em algoritmos supervisionados, os quais são os mais indicados para resolver os problemas propostos. Estes métodos são chamados desta forma, visto que a base de dados contém explicitamente os valores da propriedade de interesse, tornando assim possível que o algoritmo calcule com precisão o seu grau de certeza em suas previsões. Ainda, os valores corretos são chamados de rótulos (*labels*) e, por conseguinte, a base de dados, de rotulada (Géron, 2022).

Adiante, dois outros conceitos fundamentais são os de *features* e *targets*. *Features* são os valores individuais de cada dado na presente base, juntamente à identificação de qual característica é levada em conta, por exemplo, uma distância igual a 50 parsec, enquanto os *targets* são os valores almejados de serem previstos a partir das *features*, as quais também chamadas de preditores (Müller e Guido, 2016). É importante frisar que os modelos de *machine learning* são quase sempre de característica preditiva, ou seja, quer-se prever um resultado de acordo com suas *features*. Ainda, a preditividade é uma ênfase na comuni-

dade de estudiosos de ciência de dados e modelos que não são preditivos, portanto, são comumente vistos com ceticismo (Dhar, 2013).

Ainda, os algoritmos supervisionados podem ser de dois tipos: regressões ou classificações. Uma regressão busca prever uma variável contínua e numérica, como por exemplo o valor do dólar em um dado momento. Neste exemplo, o *target* é o valor do dólar e as *features* são dados tais como a variação no dia anterior e o dia da semana. Por outro lado, uma classificação visa categorizar um dado em uma classe. Dessa forma, os *targets* são elementos de um conjunto discreto. A imagem de um cachorro ou um gato, por exemplo, é passível desta classificação. Neste caso, o *target* é "cachorro" ou "gato" e as *features* são as posições dos *pixels* da imagem e seus valores de cor (Géron, 2022).

Dessa forma, agora que os conceitos gerais de *machine learning* foram definidos, busca-se a análise de quais são exatamente os componentes principais que definem um algoritmo de *machine learning*. Estes são quatro: uma base de dados, um modelo, uma função objetiva e um algoritmo de otimização.

A base de dados alimenta o algoritmo. É dentro dela que ele vai encontrar padrões e relações para fornecer as previsões. Dessa forma, a base de dados precisa ser robusta e grande, pois assim as eventuais aleatoriedades inerentes vão ser apagadas e as verdadeiras relações podem ser descobertas. Antes de se trabalhar com os dados, um pré-processamento é comumente realizado. Neste processo, a base de dados sofre transformações, como padronizações para evitar que o algoritmo crie preferências por algumas *features* com ordens de grandeza muito diferentes, ou interpolações para que se lide com bases de dados com valores incompletos.

O segundo componente é o modelo em si, que é caracterizado pelo conjunto de operações matemáticas que vai guiar como os valores das *features* se transformam até o valor de um *target* contínuo, no caso de uma regressão, ou um valor de probabilidade de ser uma determinada classe, no caso de uma classificação. O modelo pode ser de diversos tipos, desde modelos mais simples e lineares até longos e complexos algoritmos de redes neurais. Os valores das *features* são transformados de acordo com os parâmetros do modelo, valores numéricos que são alterados ao longo do processo de treinamento.

A função objetiva é uma função matemática que informa ao código o quão preciso ele está sendo em sua previsão. Em aprendizados de máquina supervisionados, esta função é chamada de função de perda (*loss function*). A partir dos parâmetros do modelo, o

algoritmo calcula a *loss function*, comparando os valores previstos com os *targets* que já são conhecidos, pois trata-se de um modelo supervisionado. O valor da *loss function* é tanto maior quanto menor for a precisão do algoritmo em fazer suas previsões.

O algoritmo de otimização é a forma na qual os parâmetros do modelo são atualizados para que a *loss function* seja reduzida e a precisão das previsões seja mais alta. Devido à sua maior facilidade de implementação, o método mais comum de otimização é o chamado descida do gradiente (*gradient descent*). Neste método, o gradiente da *loss function* é calculado e os parâmetros são alterados de forma que o gradiente se torne zero e, por consequência, a *loss function* chegue a um valor de mínimo. Chama-se de época (*epoch*) a fase da análise em que os parâmetros são atualizados uma vez e este processo é denominado como treinamento de modelo.

1.2 Estrelas Be

Estrelas Be constituem um subgrupo estelar, composto principalmente por estrelas não supergigantes de tipo espectral B, e que têm como característica a presença de linhas de emissão no seu espectro (e que introduz então o sufixo "e" no tipo espectral, para descrever emissão) devido à presença de um disco de decréscimo viscoso, ejetado pela própria estrela.

Essas estrelas possuem uma rotação muito acima do observado para estrelas de tipo B normais, com taxa de rotação W[1] > 0.50 (Rivinius et al., 2013), o que as torna objetos com trajetórias evolutivas muito mais complexas que estrelas de baixa rotação, e traz aspectos observacionais como achatamento, que influencia no brilho observado dependendo do ângulo de visada e também escurecimento gravitacional (inicialmente descrito por von Zeipel 1924).

Com respeito às linhas de emissão, as mais importantes são as linhas da série de Balmer do Hidrogênio (sendo Hα a principal), o que reflete a principal composição do disco cujo surgimento está intimamente relacionado com a fotosfera da estrela central. Porém há também emissão em outras linhas, como linhas de Hélio ou linhas metálicas, como o Ferro.

Além da emissão, estas estrelas apresentam excesso na região do infravermelho e polarização linear após o reprocessamento da luz pelo disco. Todas estas informações obser-

[1] $W = v_{\rm rot}/v_{\rm orb}$, onde $v_{\rm orb}$ é a velocidade orbital circular Kepleriana no equador, e $v_{\rm rot}$ é a velocidade de rotação no equador. A rotação crítica se dá quando $W = 1$.

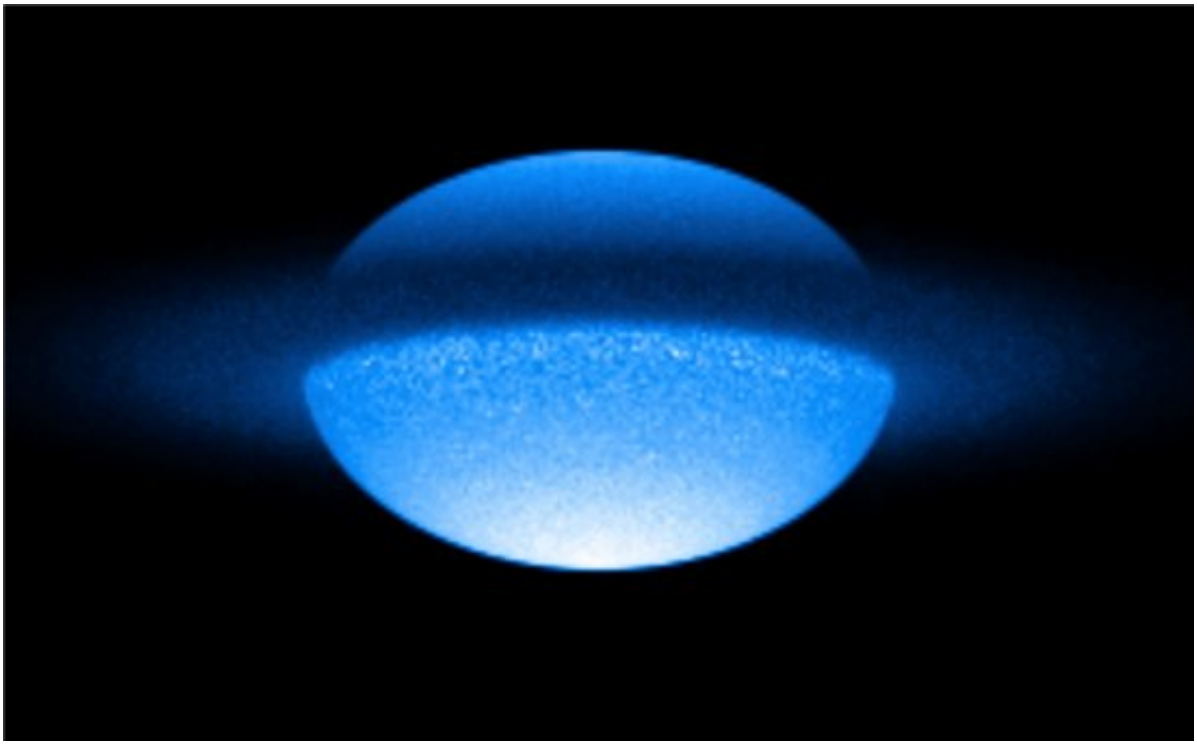

Figura 1.1: Modelo físico da estrela Be Achernar, com um ângulo de inclinação de 80°. O disco de decréscimo viscoso no equador e o achatamento da estrela são duas das principais peculiaridades da figura. Crédito: Dr. Daniel Faes Moser.

vacionais, retiradas de fotometria, interferometria, espectroscopia ou polarimetria, estão sujeitas a uma variabilidade que pode ser de horas até anos, sem regularidade entre os diversos objetos Be.

Esta classe de estrelas é investigada há muitas décadas, com um avanço expressivo no conhecimento sobre elas desde a primeira observação de uma Be (γ Cassiopeiae) relatada em Secchi (1866), porém com inúmeras questões em aberto. Atualmente, além da alta rotação, o mecanismo principal responsável pela ejeção de matéria pela estrela central são as pulsações não radiais (Baade et al., 2016). As propriedades das pulsações foram intensivamente estudadas com a ajuda dos dados do satélite TESS, que observou múltiplas estrelas Be já classificadas na literatura (Labadie-Bartz et al., 2022).

Já para o mecanismo do disco, o modelo que melhor reproduz as observações é o modelo VDD, que trata o disco de decréscimo como um disco fino viscoso Kepleriano, no qual a cisalhamento viscoso é o processo físico responsável por transportar a matéria ejetada pela estrela para orbitais maiores, fazendo assim o disco crescer. Este modelo foi primeiro apresentado por Lee et al. (1991), com ideias que datam desde Lynden-Bell e Pringle (1974).

O surgimento destes objetos faz parte das questões ainda não compreendidas. Em observações nas últimas duas décadas, notou-se que há um aumento na fração Be/(B+Be) em populações de menor metalicidade (Martayan et al., 2007), levando a questões evolutivas muito importantes e dilemas sobre como a fração se altera em ambientes extremamente pobres em metais.

Estrelas Be podem surgir de canais binários, através da transferência de massa e momento angular em algum momento da vida de um sistema binário, ou através do canal evolutivo, na qual a estrela adquire a sua alta taxa de rotação durante a sua evolução enquanto na sequência principal. Devido ao quão comum estes objetos são em populações estelares, estrelas Be são também um ótimo laboratório astrofísico em estudos de classificação, podendo servir como um pilar para estudos futuros com outros objetos peculiares.

1.3 Fotometria

Na astronomia, observações e relatos do céu começaram há muito tempo, porém houve a necessidade de quantificar, de encontrar alguma maneira de medir o fluxo luminoso que chega até nós, vindo de um astro. Um exemplo interessante é o catálogo de Hiparcos, completo em 129 A.C., no qual as estrelas observadas eram classificadas em "muito brilhante" até "fracas". Mesmo sendo um avanço tremendo, foram necessários dois milênios para que os astrônomos encontrassem maneiras de quantificar o brilho de um objeto.

A magnitude, medida importante na fotometria, foi primeiro definida em 137 D.C., a partir de uma escala subjetiva, de 1 a 6, baseado na visibilidade das estrelas no crepúsculo, onde as estrelas mais brilhantes possuíam magnitude igual a 1, e as mais fracas igual a 6 (para mais detalhes, ver Miles 2007).

Apenas com o surgimento dos primeiros telescópios que se notou que existiam inúmeros objetos mais fracos do que se imaginava. No fim do século XIX, foi sugerida uma simplificação matemática baseada na resposta visual do olho humano, na qual a sensibilidade varia logaritmicamente. Até hoje, esta última escala proposta é a utilizada. A magnitude de um objeto é definida por:

$$m = -2.5 \log(F) + K, \tag{1.1}$$

onde F é o fluxo da estrela e K é uma constante. O fator 2.5 vem da imposição de que uma diferença de 5 magnitudes seja equivalente a um fator de 100 no brilho ($\sqrt[5]{100} \approx 2.512$).

Para duas intensidades distintas, a diferença se dá por:

$$m_1 - m_2 = -2.5 \log(F_1/F_2), \tag{1.2}$$

onde o sinal negativo recupera a ordem na qual estrelas mais brilhantes possuem menores magnitudes.

Na fotometria, é comum o uso de filtros que permitem separar o fluxo luminoso em faixas espectrais. Um sistema de filtros muito conhecido é o UBV, o primeiro a ser padronizado com relação ao estabelecido sistema de classificação espectral MK, com comprimentos de onda indo do ultravioleta próximo até o fim do óptico (Johnson e Morgan, 1953).

Com o uso de filtros fotométricos, uma estrela pode ter múltiplos valores de magnitudes a depender do filtro, possibilitando o cálculo da diferença de magnitudes entre os diferentes filtros. Essa diferença, conhecida como índice de cor, é extremamente importante na astronomia, pois se relaciona com a temperatura da estrela e também com a extinção interestelar na linha de visada.

A fotometria historicamente é muito importante para estrelas Be, tanto em monitoramentos contínuos em levantamentos como na missão OGLE (Soszyński et al., 2013), na qual eventos de construção e dissipação de disco podem ser observados e modelados (Ghoreyshi et al., 2018), quanto em observações pontuais que servem como um retrato de uma estrela ou população naquela noite. Alguns aglomerados apresentam uma separação na sequência principal, com uma sequência mais avermelhada (provavelmente originada pelas estrelas Be em alta rotação, vistas a partir de ângulos de visada *edge-on*), como reportado por Milone et al. (2018).

Dependendo do sistema fotométrico, a detecção fotométrica de estrelas Be com alta emissão em Hα ou com excessos em comprimentos de onda grandes é trivial (ver McSwain e Gies 2005 para o caso do uso de filtro estreito centralizado em Hα). A questão que se apresenta é que a separação fotométrica entre uma estrela Be (com disco) e um B normal (sem disco) nem sempre é simples.

A observação de Be's utilizando apenas filtros largos se mostrou ineficiente no papel de encontrar candidatas a Be, mesmo se em conjunto com um filtro estreito centralizado no comprimento de Hα, visto que nem toda estrela Be possui forte emissão. Estes objetos podem estar com um disco tênue ou com uma dissipação em andamento na hora da observação, o que ainda leva a um viés observacional em direção à estrelas do tipo B.

Um outro fator importante é o ciclo de trabalho destes objetos, que define o fração de tempo em que uma estrela está perderndo massa (e alimentando seu disco) vs. a fração em que a estrela está quiescente. Sem conhecer esse ciclo de trabalho podem surgir viéses observacionais na determinação da fração real de estrelas Be.

Recentes levantamentos fotométricos de grandes porções do céu, como o S-PLUS (Mendes de Oliveira et al., 2019) e o J-PAS (Benitez et al., 2014), apresentam múltiplos filtros estreitos, que melhor resolvem partes específicas do espectro eletromagnético de uma estrela e possibilitam combinações de cores e pseudo cores ainda não exploradas no estudo de estrelas Be.

1.4 Objetivos

O principal objetivo deste trabalho foi utilizar aglomerados estelares sintéticos produzidos com modelos realísticos de estrelas normais (A7 até O9) e estrelas Be em conjunto com métodos de classificação binária supervisionados nos filtros do S-PLUS (os detalhes dos filtros estão disponíveis na Tabela 1.1) para estudar a sensibilidade de combinações específicas de cores na separação entre estrelas Be ativas e estrelas normais.

A importância deste trabalho é que pela primeira vez está sendo usado um conjunto grande de filtros que cobrem uma faixa expressiva do espectro eletromagnético (filtros S-PLUS, ver Fig. 1.2) em combinação com técnicas de machine learning para se abordar o problema complexo da determinação da fração real de estrelas Be em determinada população estelar.

Tabela 1.1 - Sistema Javalambre do S-PLUS

Filtro	λ_{eff} [Å]	$\Delta\lambda$ [Å]	Característica[3]
u	3574	330	Filtro largo u'
J378	3771	151	[OII]
J395	3941	103	Ca H+K
J410	4094	201	Hδ
J430	4292	200	Banda G
g	4756	1536	Filtro largo g'
J515	5133	207	Tripleto de Mgb
r	6260	1462	Filtro largo r'
J660	6614	147	Hα
i	7692	1504	Filtro largo i'
J861	8611	408	Tripleto de Ca
z	8783	1072	Filtro largo z'

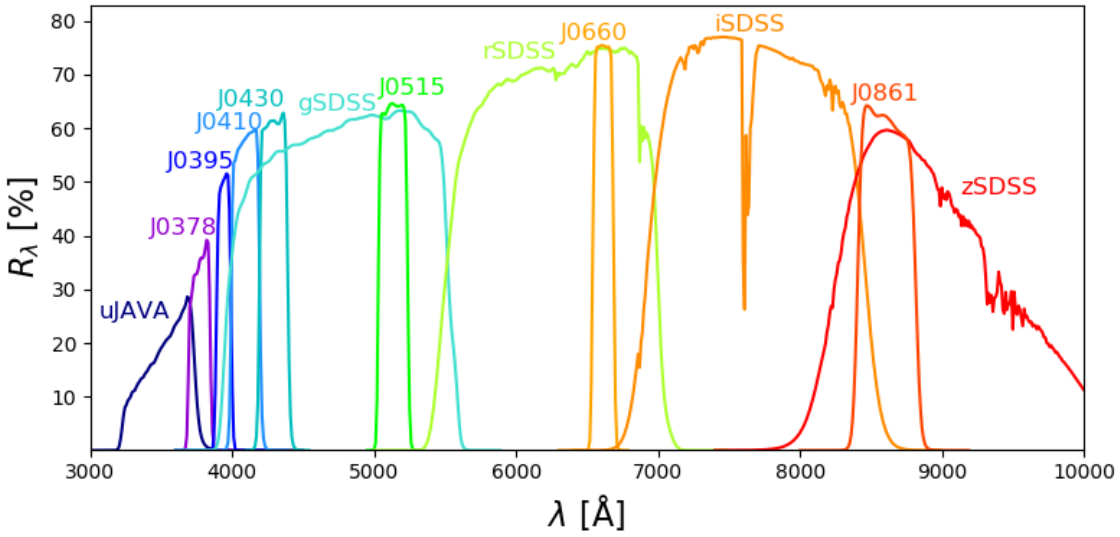

Figura 1.2: Curvas de transmissão (transmitância percentual R_λ *vs* comprimento de onda em Å) ao longo da cobertura do sistema fotométrico de 12 filtros do S-PLUS. Crédito: Dr. Fábio Rafael Herpich.[2]

[2] Retirado do GitHub da colaboração através de: `https://github.com/splus-github/splus-filters`. Acesso em 11 de Novembro de 2022.

[3] Adaptação da Tabela 2 de Mendes de Oliveira et al. (2019). Os filtros u'g'r'i'z' são semelhantes ao sistema ugriz adotado no SDSS (Loveday, 1996).

Capítulo 2

Metodologia

2.1 BeAtlas e a geração de aglomerados sintéticos

A estratégia principal foi construir diferentes populações de estrelas normais de tipo A (*early type*) até estrelas do tipo O (*late type*), incluindo as estrelas Be. A população deve ser criada com hipóteses sobre a função de massa inicial, idade, distribuição de taxas de rotação, metalicidade, entre outros parâmetros.

Para produção dos observáveis fotométricos de cada população, foi utilizada a grade de modelos BeAtlas para obter espectros sintéticos para cada estrela de dada população (o atlas é bem detalhado na tese de Mota 2019, como também na dissertação de Rubio 2019, com um artigo detalhado atualmente em produção).

A grade contém uma divisão entre modelos puramente fotosféricos (estrelas sem disco) e sistemas estrela+disco. A sub-grade de modelos fotosféricos possui uma atualização recente, já a sub-grade de modelos com disco está recebendo uma atualização - portanto, para este trabalho, a grade de disco original foi utilizada.

Todos os modelos da grade foram calculados utilizando o código de transferência radiativa HDUST (Carciofi e Bjorkman, 2006). Ambas as sub-grades consistem de espectros realistas computados desde o UV até o rádio, e permitem a comparação com os dados do S-PLUS através de uma convolução do espectro teórico com cada um dos 12 filtros Javalambre usados pela colaboração (Figura 1.2).

Até chegar em um modelo fotométrico de um aglomerado sintético, múltiplos passos devem ser tomados, desde as condições sobre a criação do aglomerado (i.e., quais as regras de seleção dos parâmetros da população) até o cálculo final das magnitudes para os diferentes filtros escolhidos. É essencial reunir diversas hipóteses, tanto na caracterização

das propriedades intrínsecas, quanto das extrínsecas. Além disso, para estrelas Be, temos como um parâmetro intrínseco exclusivo desta classe de estrelas, o já citado ciclo de trabalho, período na qual estrelas Be permanecem ativas além dos variados parâmetros do disco (descritos mais detalhamente abaixo).

A primeira hipótese é a função de massa inicial (*Initial Mass Function*, IMF) a ser utilizada. A IMF é uma função que estabelece a probabilidade da formação de uma estrela de massa M. Não há a identificação da IMF de uma população observada, e sim uma distribuição de massas iniciais é gerada a partir de uma IMF pré-definida. Com uma IMF escolhida, um valor de massa total do aglomerado deve ser indicado de forma a se obter o número total de estrelas para cada massa. A segunda hipótese para o aglomerado é sua idade. Assume-se a mesma idade para todas as estrelas, como em populações estelares simples (Tinsley, 1980).

Como dito acima, uma das grades do BeAtlas é a grade com disco, em que são representadas estrelas Be com discos típicos. O disco é representado por uma densidade superficial de base (Σ_0, definida com a densidade superficial no equador da estrela), um perfil de densidade (dado por uma lei de potências com índice n) e um raio externo (R_D, definida em raios estelares da estrela central). Além disso, o ângulo de inclinação é um parâmetro intrínseco importante, pois o espectro aparente de uma estrela Be varia muito com i. Nota-se que i é também um parâmetro da grade fotosférica.

Para a população disco/fotosférica, o limite inferior e superior são as menores e maiores massas da grade disco/fotosférica respectivamente. Um passo a passo genérico foi determinado para criação de um aglomerado sintético:

- A partir da IMF escolhida e da massa total do aglomerado, sorteia-se aleatoriamente uma estrela com massa M. Este sorteio deve, ao final, produzir a massa total desejada respeitando-se as proporções impostas pela IMF;

- Para cada estrela amostrada, uma taxa de rotação é escolhida aleatoriamente (distribuição uniforme) entre os limites inferior e superior da grade (W = 0.50 a 0.95 para a grade disco, e 0 a 0.99 para a grade fotosférica);

- Uma inclinação é escolhida aleatoriamente (distribuição uniforme) entre $i = 0°$ a $90°$;

- Para a grade fotosférica, a idade é escolhida a partir do desejado para o aglomerado. No caso disco, a grade antiga possui uma limitação da idade estar restrita a modelos

com a fração de hidrogênio no núcleo igual a 30%. Esta limitação será contornada quando a nova grade tiver sido computada.[1];

- Outro ponto importante é determinar como parâmetro intrínseco um valor para a fração B/(B+Be), na qual ajustaria o número de estrelas com disco da amostra condizente com a fração escolhida. Esta fração possui alguns valores empíricos calculados em alguns ambientes, mas sua determinação é algo bastante complexo, e que frequentemente diverge na literatura (e.g., Bodensteiner et al. 2020, Keller e Bessell 1998 e Feast 1972 sobre NGC 330). Como o foco deste trabalho não é a população sintética em si mas o método de *machine learning*, adota-se abaixo um valor típico para Be/(B+Be). De posse deste valor, pode-se determinar, usando números aleatórios, se a estrela amostrada no item acima é uma estrela Be.

- Caso a estrela amostrada seja uma Be, a densidade de base Σ_0 e o expoente radial n são escolhidos de aleatoriamente (distribuição uniforme) a partir dos limites inferior e superior da grade. A densidade de base Σ_0 é na verdade dependente do tipo espectral (Vieira e Carciofi, 2017), e existem limites superior e inferior empíricos para este parâmetro, sendo a amostragem aleatória a partir de uma distribuição uniforme algo não ideal, o que torna necessária uma reavaliação no futuro;

- A partir disto, os valores necessários de parâmetros estelares intrínsecos estão determinados, sendo possível interpolar em ambas grades o fluxo desejado para cada objeto. Com isso, para cada objeto, está atribuído um espectro com fluxo calibrado a 10 pc de distância ainda "limpo", sem interferências externas, como por exemplo o avermelhamento.

Daqui em diante, entram as considerações extrínsecas do aglomerado, na qual a distância, o avermelhamento e outros artefatos podem ser introduzidos. Para situar o aglomerado sintético a uma distância desejada, o fluxo obtido inicialmente é multiplicado pelo fator $(10^2/d^2)$, em que d é a distância desejada.

Para o avermelhamento, foi utilizada a lei de extinção de Cardelli et al. (1989), com $R_v = 3.1$, sendo E(B-V) o parâmetro livre. A partir daqui, o espectro de cada objeto já está

[1] Em relação a grade nova, a sub-grade fotosférica está completa, sendo necessária a computação apenas dos modelos disco. A sub-grade dos modelos disco está com um percentual de pouco mais de 30% do planejado completo, porém para o seu uso inicial não será necessário atingir a marca de 100%.

pronto para ser convoluído[2] com as curvas de transmissão do sistema fotométrico escolhido. Um ponto importante é que o ponto zero escolhido foi um espectro representativo da estrela Vega, com fluxo composto de observações e de um modelo atmosférico da grade ATLAS9 (Castelli e Kurucz, 2003) com $T_{\rm eff}$ = 9550K, $\log g = 3.95$ e metalicidade solar. Assim, a magnitude de um objeto é calculada seguindo:

$$m_X = -2.5 \log(F_X/F_{X,\rm Vega}), \tag{2.1}$$

onde m_X é a magnitude do objeto calculada em um filtro X, F é o fluxo integrado da estrela sintética e $F_{\rm Vega}$ o fluxo integrado de Vega (ambos convoluídos através da curva de transmissão do filtro X). Considerando o conjunto Javalambre (Fig 1.2, no final são produzidos 12 magnitudes para cada estrela da população sintética.

Alguns detalhes para que o aglomerado esteja de fato condizente com uma observação fotométrica real ainda podem ser introduzidos. O viés observacional de maior precisão para alvos mais brilhantes pode ser explorado, por exemplo, com uma dispersão que seja maior nas estrelas com maiores magnitudes, e que diminua em direção aos alvos mais brilhantes.

Uma das populações sintéticas criadas para o estudo está representada nas figuras 2.1 e 2.2, na qual foram utilizados alguns dos filtros mais importantes da classificação (a ser detalhada na Seção 3.1).

[2] A convolução fotométrica é realizada multiplicando, para cada comprimento de onda, o fluxo com a resposta do filtro, para que no fim o fluxo integrado possa ser obtido com a soma de cada termo.

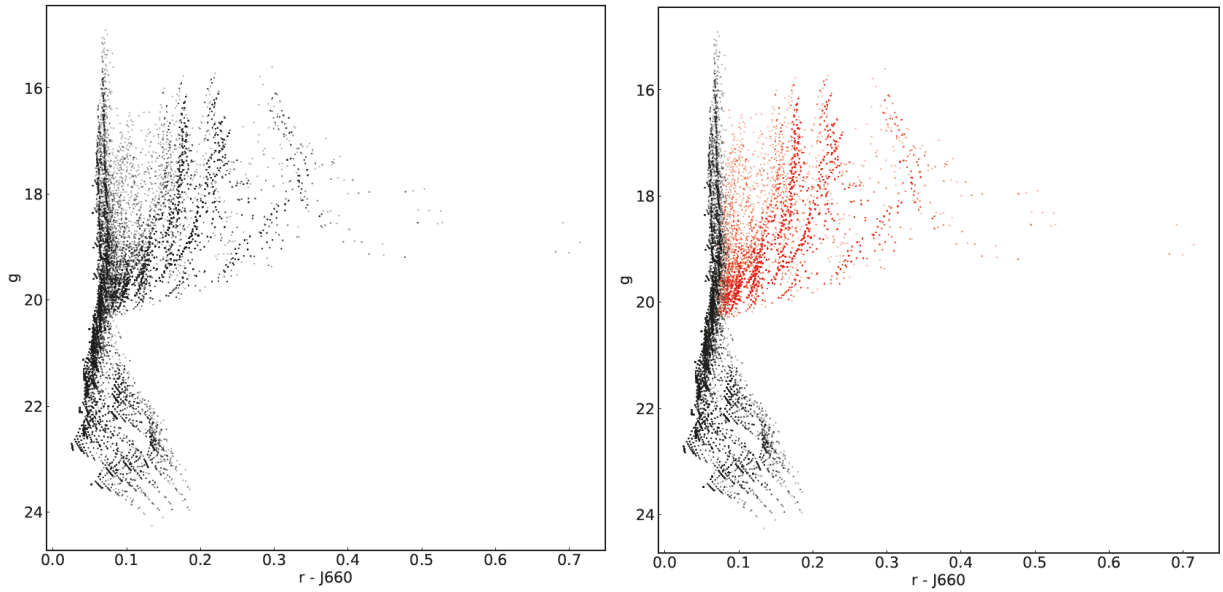

Figura 2.1: Diagrama cor-magnitude g *vs* (r-J660) de uma população sintética. À esquerda, toda a população está amostrada sem distinção. À direita, estrelas Be estão em vermelho.

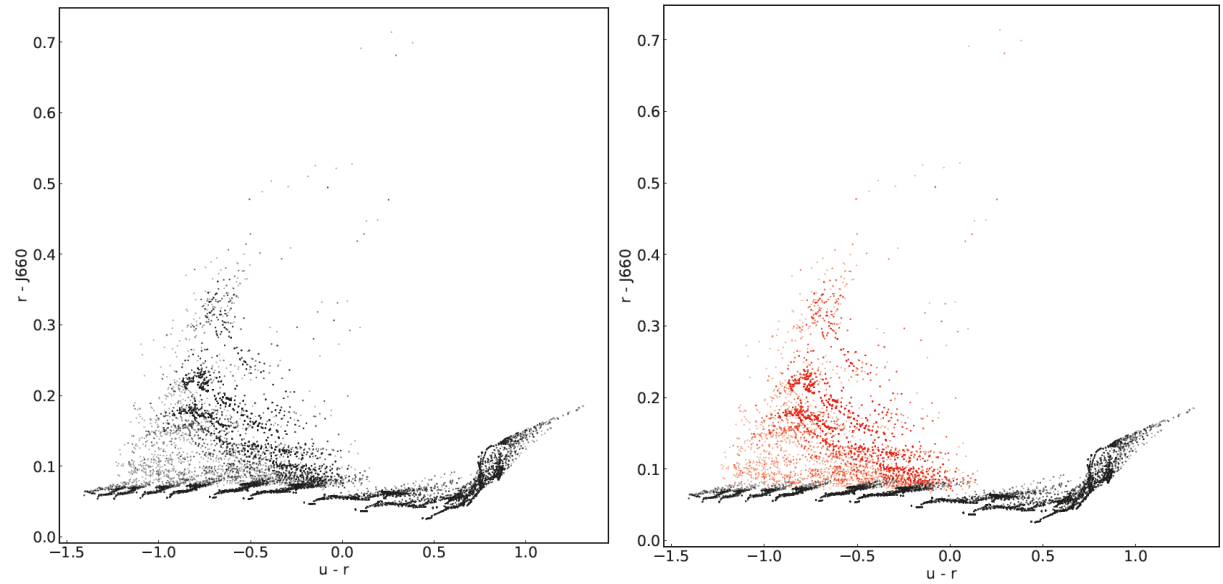

Figura 2.2: Diagrama cor-cor (r-J660) *vs* (u-r) de uma população sintética. Esquema de cores equivalente à Fig. 2.1.

2.2 Aplicação dos modelos de Aprendizado de Máquina

Busca-se então trabalhar com aglomerados sintéticos como uma forma eficiente e rápida de gerar dados sobre estrelas Be para alimentar algoritmos de *machine learning* e classificar aglomerados reais a partir deste modelo. Para tal, trabalha-se com três bases de dados: duas sintéticas e uma de observações reais.

A primeira base de dados, sintética, chamada de agora em diante como Cluster 1, é

usada como base de treinamento para o modelo. Esta base contém 19045 estrelas sem disco e 11551 estrelas com disco, para um total de 30596 *inputs*, com suas magnitudes em doze filtros diferentes: u, J378, J395, J410, J430, g, J515, r, J660, i, J861 e z. A segunda base de dados, chamada de Cluster 2, é ligeiramente menor e serve como base de testes iniciais. Esta contém 18918 estrelas sem disco e 5787 estrelas com disco, para um total de 24705 *inputs*, também nos mesmos doze filtros que o Cluster 1. A terceira base de dados, chamada de Cluster Real, é composta de estrelas do aglomerado aberto da Pequena Nuvem de Magalhães NGC 330 a partir de dados do S-PLUS. São totalizadas 7154 estrelas, incluindo objetos não membros do aglomerado e estrelas de campo, com 51 estrelas Be conhecidas (Bodensteiner et al., 2020).

É necessário fazer um pré-processamento das bases de dados para que o modelo possa fazer seu treinamento e previsões. Como o Cluster 1 está desbalanceado, ou seja, há um excesso de estrelas sem disco (19045) em comparação com estrelas Be (11551), é necessária a feitura do *downsampling* da base de dados, que consiste em selecionar uma quantidade de *inputs* menores da classe dominante para que, ao final, os números estudados nesta fase sejam iguais. Dessa forma, o modelo não estará enviesado, isto é, não escolherá uma classe em detrimento da outra, pois, caso contrário, escolher valores triviais apenas com a classe dominante já resultaria em uma acurácia superior a 50%. Para fins de avaliar todas as possibilidades, reserva-se uma base de testes com *downsampling* e uma sem *downsampling*.

O segundo passo necessário do pré-processamento é a padronização, ou *scaling*. Foi utilizada uma função dentro da biblioteca sklearn da linguagem Python para fazer esse procedimento, chamada de StandardScaler[3]. Essa função vai transformar todos os valores dos *inputs* em distribuições com média igual a zero e desvio padrão igual a 1, a fim de que não haja vieses de uma certa variável por ter uma ordem de grandeza maior que as demais e, desta maneira, pese mais durante a atualização dos parâmetros do modelo e influencie mais fortemente o valor da *loss function*.

Desse modo, 6 modelos diferentes são treinados nesta base de dados, com ou sem *downsampling*. A saber, Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, Naive Bayes e K-Nearest Neighbors, todos modelos de classificação. Para avaliar se estes modelos estão conseguindo realizar previsões de forma correta, utiliza-

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

se 4 métricas diferentes: acurácia, precisão, *recall* e F-Measure.

Ao realizar uma previsão de um algoritmo de classificação, pode-se criar uma matriz de confusão, isto é, uma matriz que contém o valor total de previsões corretas e erradas para cada uma das classes. No caso de uma classificação binária, como no problema em questão, a matriz de confusão é da forma da Tabela 2.1. Essa matriz cruza os dados entre todas as possibilidades de classificação. Isto posto, se o algoritmo verdadeiramente prevê, por exemplo, que determinada estrela é Be, isso conta como um *True Positive* e assim por diante.

Tabela 2.1 - Matriz de confusão para classificações binárias.

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Define-se, então, a acurácia como a razão entre as previsões corretas e todas as previsões feitas, de acordo com a seguinte fórmula:

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2.2}$$

onde seu valor sempre será entre 0 e 1. O caso em questão da classificação de uma estrela Be é chamado de desbalanceado, pois o número de classificações positivas na base de testes é muito menor que o número de classificações negativas. Para casos como esses, outras métricas podem fornecer análises mais significativas, como a precisão e o *recall*.

A precisão é definida como a razão entre o número de *True Positives* e o total de resultados cujo valor verdadeiro seria positivo. Com a precisão, vê-se qual a fração de valores verdadeiros foi corretamente prevista. No caso binário, vê-se pela fórmula:

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.3}$$

Recall, por outro lado, é a razão entre o número de *True Positives* e o total de previsões positivas. Avalia-se assim, dada uma previsão positiva, qual a probabilidade desta previsão ser verdadeira com:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.4}$$

Também é possível avaliar um algoritmo de classificação desbalanceada usando a métrica F-Measure, que nada mais é que a média harmônica entre a precisão e o *recall*. Essa medida, em geral, avalia o quão bem o modelo consegue prever a categoria desejada e se dá

por:

$$\text{F-Measure} = \frac{2 * \text{Precision} * \textit{recall}}{\text{Precision} + \textit{recall}}. \tag{2.5}$$

Capítulo 3

Análise

3.1 Avaliação dos resultados dos modelos treinados

Métricas, conforme explicado em outras partes deste trabalho, ajudam a avaliar o quão bom um modelo é em prever corretamente os resultados. Dessa forma, constrói-se uma tabela com os diferentes valores das métricas para cada um dos modelos de classificação, para a base de dados com e sem *downsampling*, visto na Tabela 3.1 e na Tabela 3.2.

Tabela 3.1 - Métricas para os diferentes modelos, sem *downsampling* da base de testes, testados no aglomerado sintético Cluster 2.

Modelo	Acurácia	Precisão	Recall	F-Measure
Logistic Regression	0.999902	1.000000	0.999740	0.999870
Support Vector Machines	1.000000	1.000000	1.000000	1.000000
Decision Trees	0.983284	0.978173	0.977665	0.977919
Random Forest	0.996755	0.993244	0.998172	0.995702
Naive Bayes	0.800541	0.893855	0.679842	0.772296
K-Nearest Neighbor	0.997640	0.994933	0.998826	0.996876

Importantes análises podem ser feitas a partir dos valores acima listados. Primeiramente, a acurácia de todos os modelos diminui com a feitura do *downsampling*. O número de previsões corretas cresce porquanto o algoritmo entende melhor o desbalanceamento entre o número bruto de estrelas com disco comparadas àquelas sem disco.

Em segundo lugar, vê-se que a precisão dos algoritmos é bastante alta, ou seja, as estrelas Be estão sendo encontradas e têm-se um baixo número de falsos negativos - um valor previsto como negativo muito dificilmente será errôneo. Também é observado que

Tabela 3.2 - Métricas para os diferentes modelos, com *downsampling* da base de testes, testados no aglomerado sintético Cluster 2.

Modelo	Acurácia	Precisão	Recall	F-Measure
Logistic Regression	0.949361	1.000000	0.881975	0.937287
Support Vector Machines	1.000000	1.000000	1.000000	1.000000
Decision Trees	0.937463	0.953358	0.889347	0.920241
Random Forest	0.961357	0.995193	0.910929	0.951198
Naive Bayes	0.803540	0.978173	0.662939	0.790280
K-Nearest Neighbor	0.983235	0.994543	0.962409	0.978212

Support Vector Machines se comporta melhor para fazer as previsões como um todo, tendo resultados perfeitos. Naive Bayes teve os piores resultados, apesar de ainda serem considerados bons caso estivessem sendo considerados dados reais. Assim, conclui-se que os aglomerados sintéticos estão consistentes entre si e algoritmos de *machine learning* mostram-se aptos a captar com alta precisão as relações subjacentes, bem como prever corretamente as estrelas que são do tipo Be.

O mesmo agora é feito para os modelos treinados e testados com os dados reais do Cluster Real, como visto na Tabela 3.3 e na Tabela 3.4.

Tabela 3.3 - Métricas para os diferentes modelos, sem *downsampling* da base de testes, testados no aglomerado real Cluster Real.

Modelo	Acurácia	Precisão	Recall	F-Measure
Logistic Regression	0.644674	0.843137	0.016686	0.032725
Support Vector Machines	0.560665	0.921569	0.014752	0.029039
Decision Trees	0.933743	0.156863	0.018223	0.032653
Random Forest	0.951076	0.019608	0.003322	0.005682
Naive Bayes	0.678502	0.803922	0.017589	0.034425
K-Nearest Neighbor	0.906206	0.058824	0.004792	0.008863

Primeiramente, é importante ressaltar algumas coisas a respeito de métricas e matrizes de confusão calculados para o Cluster Real: existem 51 estrelas Be confirmadas, porém, não é garantido que todas as outras estrelas não tenham disco. Assim, podemos pensar

Tabela 3.4 - Métricas para os diferentes modelos, com *downsampling* da base de testes, testados no aglomerado real Cluster Real.

Modelo	Acurácia	Precisão	Recall	F-Measure
Logistic Regression	0.522225	0.921569	0.013580	0.026765
Support Vector Machines	0.503215	0.921569	0.013066	0.025768
Decision Trees	0.927593	0.098039	0.010482	0.018939
Random Forest	0.941012	0.019608	0.002681	0.004717
Naive Bayes	0.618675	0.784314	0.014509	0.028490
K-Nearest Neighbor	0.890271	0.078431	0.005391	0.010088

nos valores *True Positive* como acertos do modelo e os valores *False Negative* como erros do modelo, pois são estrelas confirmadas como Be. Os *False Positive* nada mais são que candidatas a estrelas Be e *True Negative* estrelas que o modelo não espera que sejam Be.

Dessa forma, percebe-se que a métrica mais adequada para julgar estes modelos, com relação ao aglomerado Cluster Real, é a precisão, pois lida apenas com estrelas em que há confiança de sua classificação. Além disso, a questão do *downsampling* também se repete para os modelos testados com dados reais. As conclusões são as mesmas que as feitas para aglomerados sintéticos. Assim, se faz necessário olhar as matrizes de confusão para os modelos SVM e *Random Forest*, apresentadas nas tabelas 3.5 e 3.6.

Tabela 3.5 - Matriz de confusão para a classificação do modelo *Support Vector Machines*, treinado com dados sintéticos sem *downsampling*, testado em dados do Cluster Real.

TP = 47	FN = 4
FP = 3139	TN = 3964

Tabela 3.6 - Matriz de confusão para a classificação do modelo *Random Forest*, treinado com dados sintéticos sem *downsampling*, testado em dados do Cluster Real.

TP = 3	FN = 48
FP = 299	TN = 6804

Fica claro que o *Random Forest* tem uma alta acurácia, de 95.1%, pois conseguiu prever melhor o desbalanceamento entre as classes. Por outro lado, sua precisão foi baixíssima,

apenas 5.9%. Isso evidencia como nem sempre uma alta acurácia significa um bom modelo preditivo, principalmente ao se tratar de classes desbalanceadas. O modelo SVM, por outro lado, teve uma precisão de 92.2%, mostrando que conseguiu captar melhor quais relações entre os filtros representam uma estrela ser Be ou não.

Um algoritmo de *Support Vector Machines* busca encontrar qual o melhor hiperplano que separa as classes, ou seja, o hiperplano com a maior distância entre os elementos de cada classe (Saunders et al., 1998). Dessa forma, ele não retorna uma probabilidade de um certo elemento pertencer a uma categoria, e é preciso usar outras técnicas para encontrar qual a probabilidade de uma estrela ser Be, como por exemplo calibração.

Um classificador bem calibrado quantifica corretamente o nível de incerteza de cada classificação e denota uma probabilidade adequada de cada valor pertencer a uma categoria (Song et al., 2021). É usada então uma calibração logística, também chamada de calibração de Platt. A probabilidade é estimada a partir de uma curva sigmoide:

$$\mathrm{P}(y = 1|x) = \frac{1}{1 + \exp(\mathrm{A}f(x) + \mathrm{B})}, \tag{3.1}$$

onde A e B são parâmetros aprendidos pelo modelo utilizando do método da máxima verossimilhança (Platt et al., 1999). A partir disso, é possível delimitar um valor de corte no qual considera-se apenas candidatos a estrelas Be os dados que tenham probabilidade maior que esse valor de corte. Por exemplo, ao considerar apenas estrelas com mais que 99.9% de chance de serem Be, de acordo com a calibração, a matriz de confusão muda:

Tabela 3.7 - Matriz de confusão para a classificação do modelo *Support Vector Machines*, treinado com dados sintéticos sem *downsampling*, testado em dados do Cluster Real, após calibração logística.

TP = 43	FN = 8
FP = 1715	TN = 5388

A precisão segue alta, porém cai um pouco: 84.3%. Por outro lado, a acurácia sobe significativamente: 75.9%. A acurácia aumenta desta maneira pois reduziu-se muito o número de falsos positivos, que neste caso são estrelas candidatas.

Para identificar exatamente quais dos diferentes filtros foram mais influentes para a determinação, pode-se aplicar o procedimento conhecido como *Permutation Feature Importance* (PFI). Este método consiste em permutar aleatoriamente os valores de uma das diferentes *features* do modelo, no caso, um dos filtros. Isto é feito algumas vezes, e a cada

uma das aleatorizações, é calculada a métrica do modelo. Caso a variável seja importante para o poder explanatório do modelo, é esperado que a métrica fique menor caso seu valor seja aleatório. Como é um procedimento aleatório, pode-se extrair a média e o desvio padrão dos valores de importância a cada uma das repetições, e assim avaliar cada uma das *features*.

É primordial frisar também que, se um modelo não é bom, há pouca relevância para a análise de quais são suas *features* mais importantes. Uma *feature* irrelevante para um modelo ruim pode ser fundamental para um modelo bom que ajuste os mesmos dados (Géron, 2022).

O PFI foi feito com as *features* do modelo SVM treinado sem *downsampling* e testado nos valores do Cluster 2. Permutou-se 15 vezes cada *feature* e foi calculada a média da diminuição da métrica F-Measure para cada um dos filtros, assim como seus respectivos desvios padrão. Os resultados podem ser visualizados na Tabela 3.8.

Tabela 3.8 - Resultados do *Permutation Feature Importance*, com 15 permutações diferentes, avaliando a diminuição da métrica F-Measure do modelo *Support Vector Machines*, para dados sintéticos.

Filtro	λ_{eff} [Å]	$\Delta\lambda$ [Å]	Importância	Incerteza
u	3574	330	0.3218	0.0029
J378	3771	151	0.2712	0.0028
J395	3941	103	0.7158	0.0028
J410	4094	201	0.2494	0.0035
J430	4292	200	0.0033	0.0006
g	4756	1536	0.3168	0.0037
J515	5133	207	0.3240	0.0034
r	6260	1462	0.0251	0.0013
J660	6614	147	0.7472	0.0031
i	7692	1504	0.7332	0.0030
J861	8611	408	0.7462	0.0030
z	8783	1072	0.3652	0.0031

Também foi calculado o PFI para os valores reais do Cluster Real. Na tabela 3.9 fica explicitada a importância de cada filtro para a precisão do modelo SVM, junto com seu

respectivo erro.

Tabela 3.9 - Resultados do *Permutation Feature Importance*, com 15 permutações diferentes, avaliando a diminuição da métrica precisão do modelo *Support Vector Machines*, para dados reais.

Filtro	λ_{eff} [Å]	$\Delta\lambda$ [Å]	Importância	Incerteza
u	3574	330	0.0009	0.0005
J378	3771	151	0.0007	0.0005
J395	3941	103	0.0121	0.0009
J410	4094	201	0.00046	0.00037
J430	4292	200	0.00015	0.00023
g	4756	1536	0.0011	0.0005
J515	5133	207	0.0012	0.0006
r	6260	1462	-0.00009	0.00023
J660	6614	147	0.0126	0.0008
i	7692	1504	0.0113	0.0007
J861	8611	408	0.0118	0.0008
z	8783	1072	0.0024	0.0007

Desse modo, vê-se que os quatro filtros com maiores valores de significância para o modelo SVM são os filtros i, J395, J660 e J861. Isto vale tanto para os testes em dados sintéticos quanto reais. Além disso, para dados reais, estes são os únicos filtros que têm valores incompatíveis com zero para uma taxa de confiança de 3σ.

As figuras 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7 e 3.8 abaixo detalham o funcionamento de alguns modelos aplicados a um aglomerado sintético e aos dados reais de NGC 330. No caso do modelo SVM, não houve a necessidade de anexar figuras para o caso do aglomerado sintético devido à perfeita classificação.

Um dos pontos mais importantes e interessantes das figuras 3.1 e 3.2 é a clara demonstração visual da precisão e da acurácia do modelo *Random Forest*, o que traz a confiança de que o modelo SVM, na qual atingiu resultado perfeito, também está conseguindo separar de forma consistente as estrelas Be dos objetos normais naquela população, e de que a aplicação em dados reais é promissora.

Já as figuras 3.3 à 3.7 demonstram a mais simples aplicação destes modelos em um

aglomerado real, com poucas estrelas Be confirmadas, e sem um robusto tratamento prévio para conter apenas os objetos mais prováveis de serem membros. É possível reparar que um tratamento mais sólido é necessário, como fazer tanto a modelagem quanto a aplicação em dados reais em escalas absolutas de magnitude e em mesmo avermelhamento, para evitar ao máximo a falta de coesão entre modelo e dado real.

No caso das figuras 3.7 e 3.8, o interessante é visualizar inicialmente como o melhor modelo na aplicação sintética atua em dados reais, após a calibração que trouxe uma quantificação da probabilidade atribuída ao modelo para cada objeto em relação ao pertencimento ao grupo de Be's ou de estrelas normais. Além do claro acerto promissor das estrelas Be confirmadas pela literatura, é nítido o aumento visual de resolução da separação do modelo, no sentido de estar mais claro quais objetos de fato estão com alto nível de confiança de serem ou não estrelas Be, o que nos leva ao termo previamente citado de estrelas candidatas.

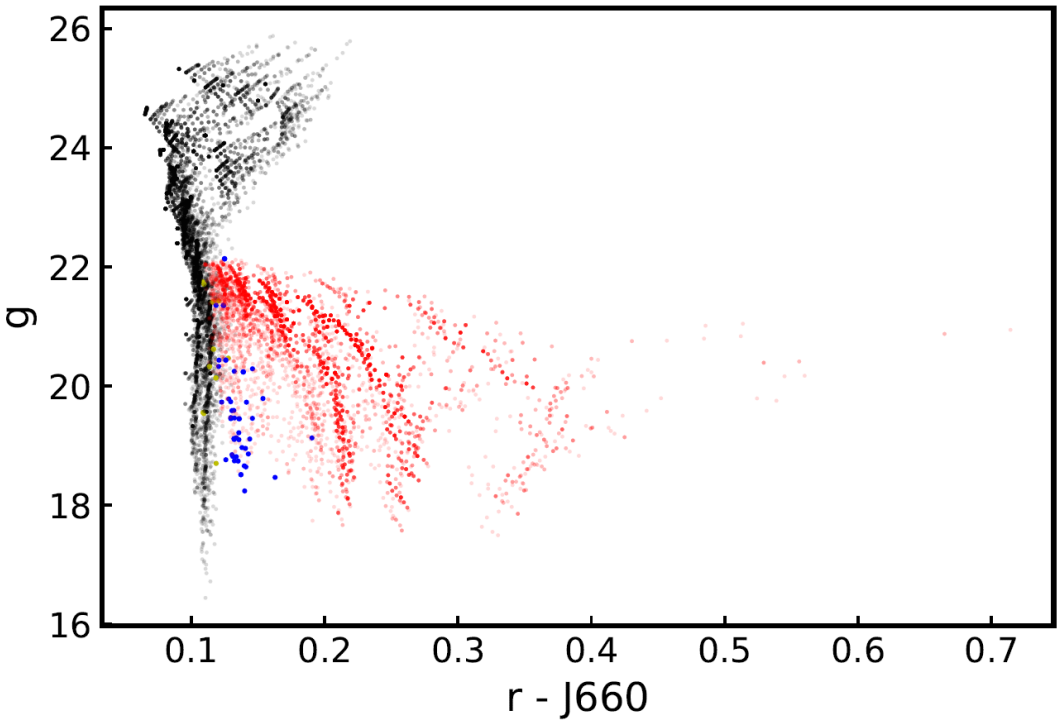

Figura 3.1: Diagrama cor-magnitude g *vs* (r-J660) de uma população sintética com a classificação realizada pelo modelo Random Forest. Em vermelho estão os *True Positives*, em preto *True Negatives*, em amarelo *False Positives* e em azul *False Negatives*.

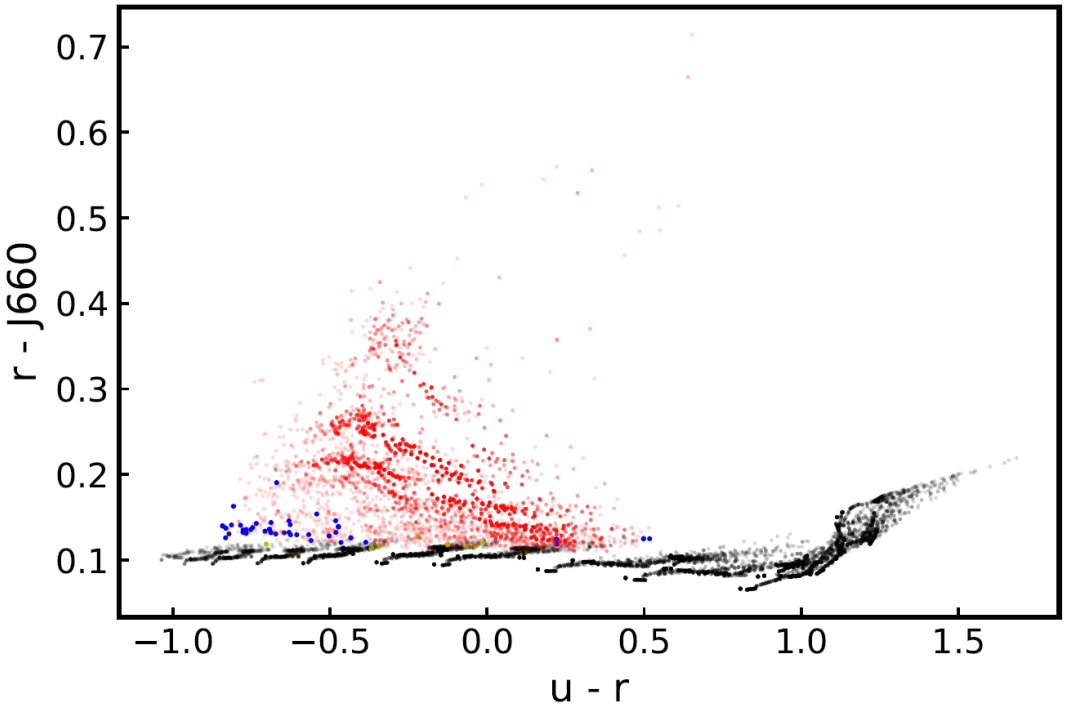

Figura 3.2: Diagrama cor-magnitude (r-J660) *vs* (u-r) de uma população sintética com a classificação realizada pelo modelo Random Forest. Esquema de cor idêntico à Fig. 3.1.

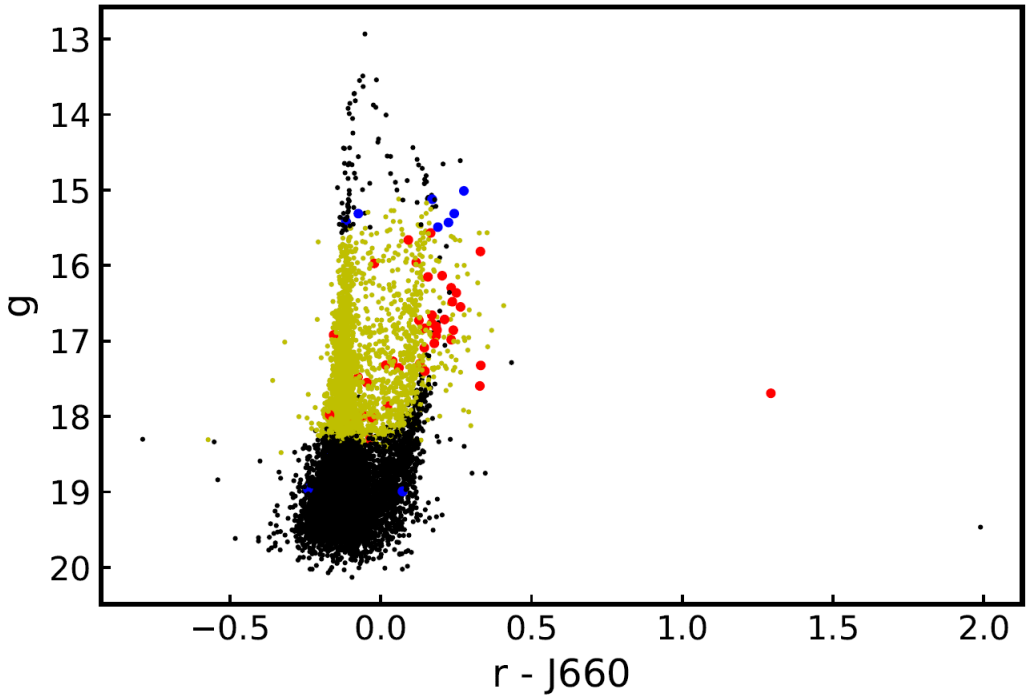

Figura 3.3: Diagrama cor-magnitude g *vs* (r-J660) de NGC 330 com a classificação realizada pelo modelo Naive-Bayes. Esquema de cor idêntico às figuras anteriores. Como os dados são reais com múltiplas estrelas não-membras do aglomerado e com estrelas Be classificadas de forma limitada, estrelas em amarelo são **candidatas**, em vermelho as Be's da literatura que o modelo acertou, e em azul estrelas Be's da literatura que o modelo não acertou.

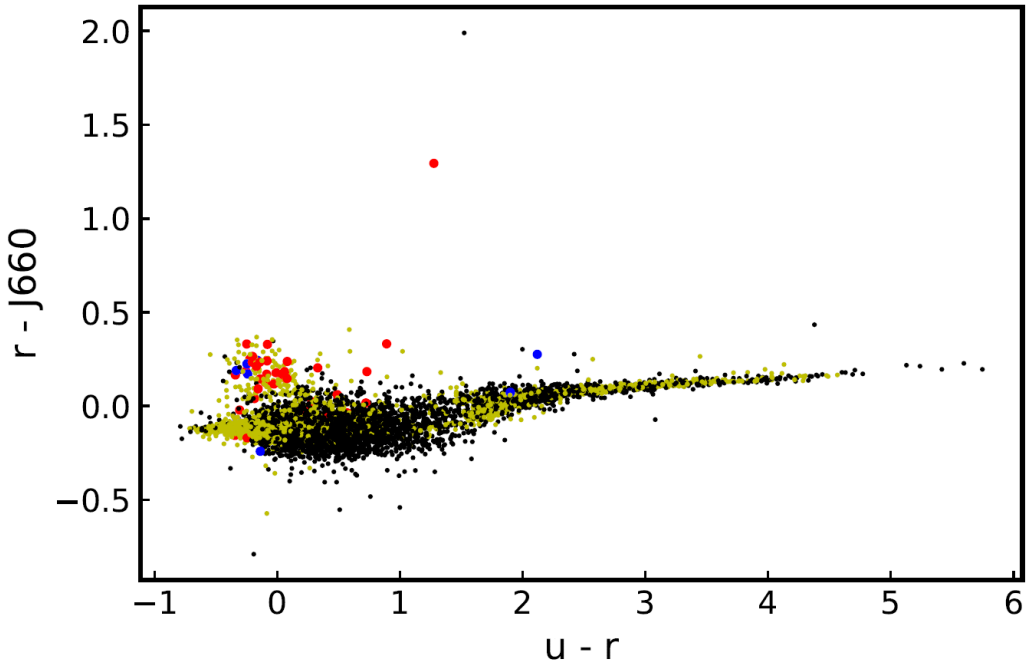

Figura 3.4: Diagrama cor-magnitude (r-J660) *vs* (u-r) de NGC 330 com a classificação realizada pelo modelo Naive-Bayes. Esquema de cor idêntico à Fig. 3.3.

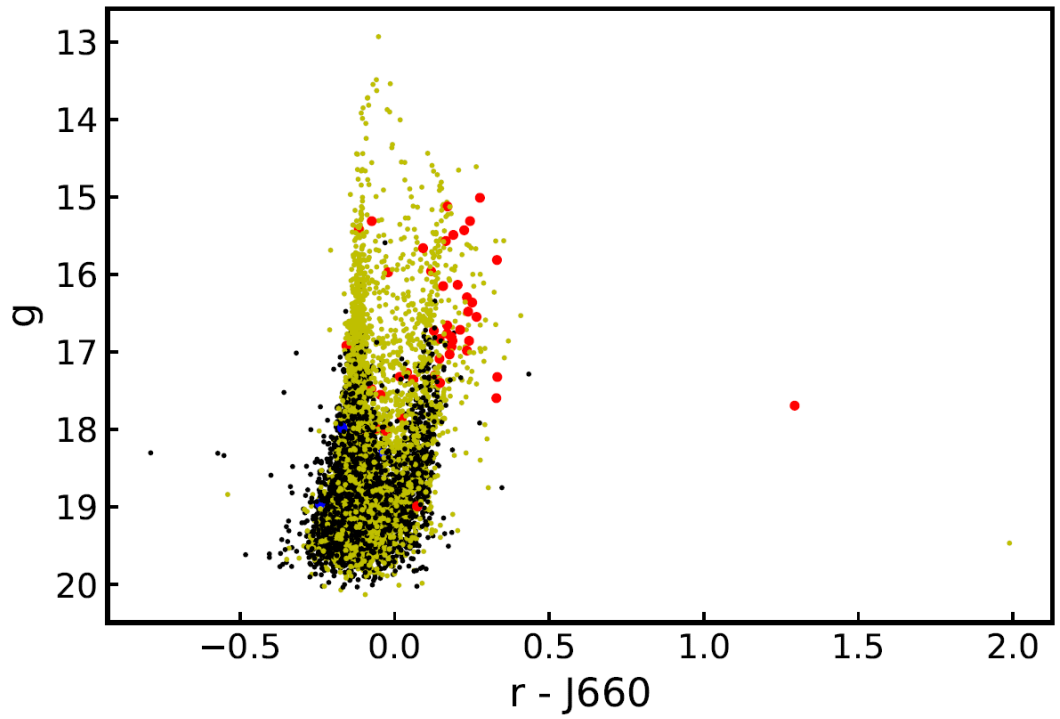

Figura 3.5: Diagrama cor-magnitude g *vs* (r-J660) de NGC 330 com a classificação realizada pelo modelo SVM pré calibração. Esquema de cor idêntico à Fig. 3.3.

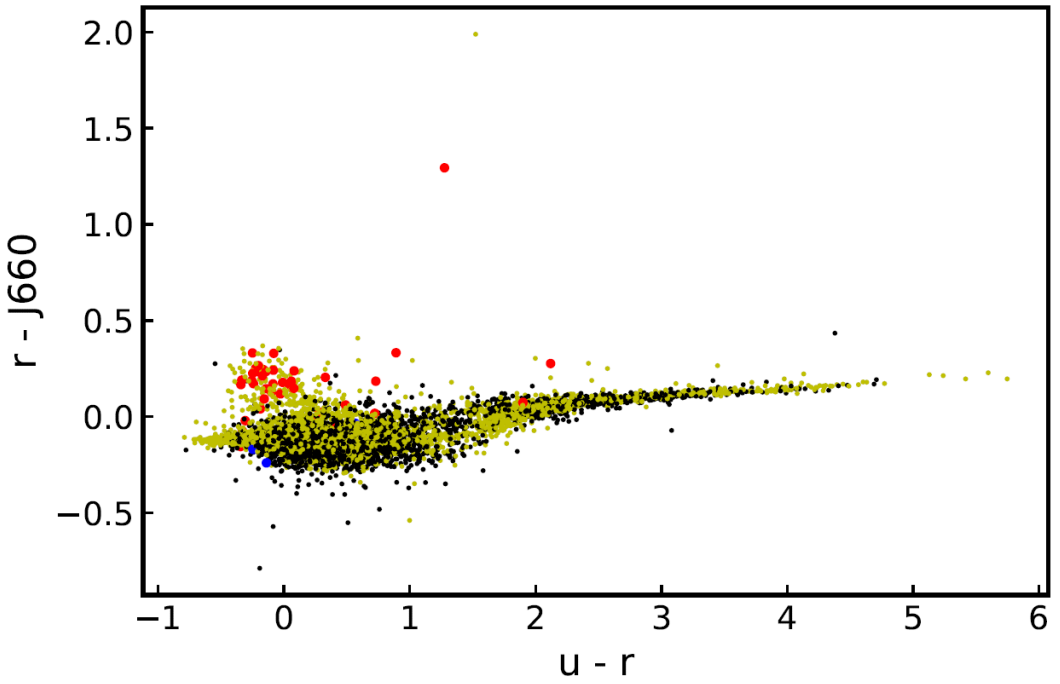

Figura 3.6: Diagrama cor-magnitude (r-J660) *vs* (u-r) de NGC 330 com a classificação realizada pelo modelo SVM. Esquema de cor idêntico à Fig. 3.3.

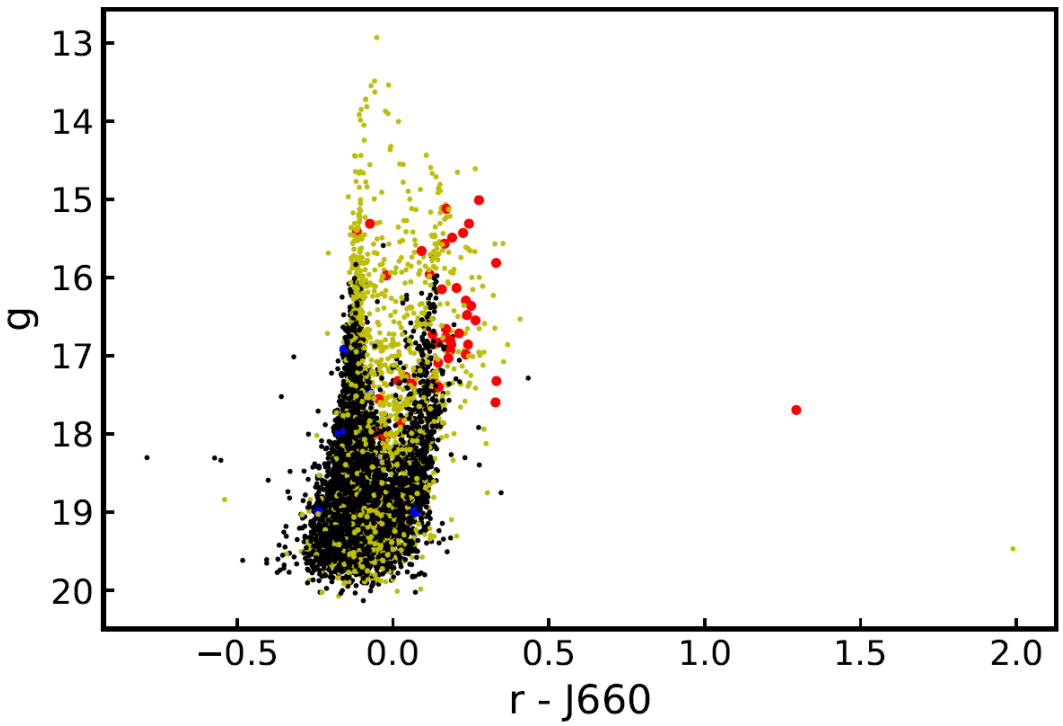

Figura 3.7: Diagrama cor-magnitude g *vs* (r-J660) de NGC 330 com a classificação realizada pelo modelo SVM pós calibração. As estrelas em amarelo são objetos com 99.9% de chance de serem Be's, sendo este o melhor gráfico para identificar candidatas a Be. Esquema de cor idêntico à Fig. 3.3.

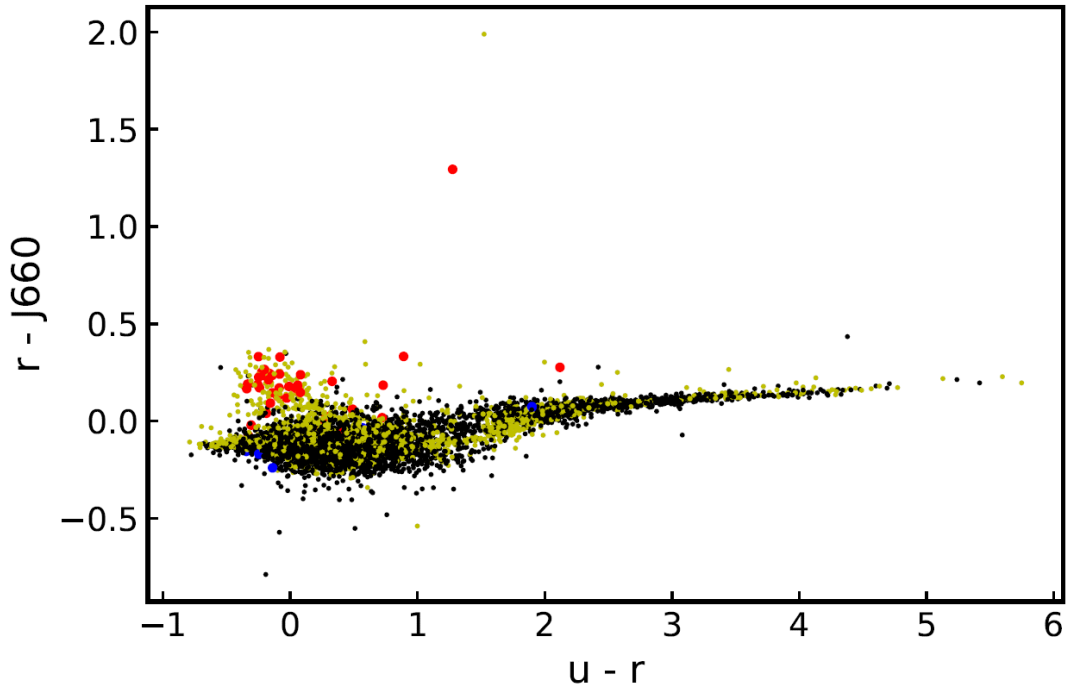

Figura 3.8: Diagrama cor-magnitude g *vs* (r-J660) de NGC 330 com a classificação realizada pelo modelo SVM pós calibração. Esquema de cor idêntico à Fig. 3.7.

Capítulo 4

Conclusões

O *machine learning* possui um potencial vasto e enormemente inexplorado e, desta maneira, toda a aplicabilidade de suas ferramentas está sendo pouco a pouco percebida. A astronomia possui inúmeros problemas em aberto e técnicas cada vez mais avançadas são utilizadas para ajudar a solucionar esses problemas.

A aplicação de classificadores de *machine learning* para encontrar estrelas Be pensada no presente estudo produziu resultados interessantes. A não utilização de *downsampling* tornou os modelos mais precisos, o que nos levanta questões sobre o desbalanceamento das classes nas bases de teste e o viés que isso gera nos preditores.

Os resultados dos modelos testados também em aglomerados sintéticos se mostraram mais corretos e poderosos, destacando-se *Support Vector Machines* que obteve resultados perfeitos. Para dados reais, a acurácia não é uma boa métrica de avaliação, ao contrário da precisão dos modelos, devido ao fato que não há certeza que todas as estrelas Be do aglomerado foram encontradas e corretamente rotuladas.

O modelo SVM se mostrou superior nesta métrica também para os dados reais. Com a utilização de uma calibração é possível definir uma probabilidade de uma estrela pertencer a uma classe e assim criar um valor de corte que restringirá ainda mais quais estrelas são candidatas a serem Be.

Além disso, a análise da importância das *features* pode fornecer informações relevantes para a classificação de Be's. Para os modelos SVM, os filtros i, J395, J660 e J861 se mostraram os mais relevantes para a precisão dos modelos.

Há ainda muitos passos a serem seguidos com estudos de classificação de grandes quantidades de estrelas. Fica claro o potencial destas ferramentas para problemas de classificação de estrelas Be. Isto pode ser aprimorado tanto para o estudo específico de separação de

estrelas Be quanto para qualquer objeto peculiar que apresente anomalias fotométricas em seu espectro.

Isto posto, nota-se que futuras melhorias podem ser feitas. Pode-se trabalhar melhor os dados do aglomerado real, levando em conta estrelas intrusas que não fazem parte do aglomerado, além de lidar com o avermelhamento destas estrelas. Também é necessário testar em diversos aglomerados e verificar se os resultados se mantêm coerentes. Ainda, técnicas mais modernas e avançadas como, por exemplo, redes neurais podem produzir resultados mais precisos no futuro.

Referências Bibliográficas

Baade D., Rivinius T., Pigulski A., Carciofi A. C., Martayan C., Moffat A., Wade G., Weiss W., Grunhut J., Handler G., et al., Short-term variability and mass loss in Be stars-I. BRITE satellite photometry of η and μ Centauri, Astronomy & Astrophysics, 2016, vol. 588, p. A56

Benitez N., et al., J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey, arXiv e-prints, 2014, p. arXiv:1403.5237

Bodensteiner J., Sana H., Mahy L., Patrick L. R., de Koter A., de Mink S. E., Evans C. J., Götberg Y., Langer N., Lennon D. J., Schneider F. R. N., Tramper F., The young massive SMC cluster NGC 330 seen by MUSE. I. Observations and stellar content, A&A, 2020, vol. 634, p. A51

Campesato O., Artificial intelligence, machine learning, and deep learning. Mercury Learning and Information, 2020

Carciofi A. C., Bjorkman J. E., Non-LTE Monte Carlo Radiative Transfer. I. The Thermal Properties of Keplerian Disks around Classical Be Stars, ApJ, 2006, vol. 639, p. 1081

Cardelli J. A., Clayton G. C., Mathis J. S., The Relationship between Infrared, Optical, and Ultraviolet Extinction, ApJ, 1989, vol. 345, p. 245

Castelli F., Kurucz R. L., New Grids of ATLAS9 Model Atmospheres. In Modelling of Stellar Atmospheres , vol. 210, 2003, p. A20

Dhar V., Data science and prediction, Communications of the ACM, 2013, vol. 56, p. 64

Feast M. W., The Cluster NGC 330 in the SMC (Paper II): Hα Emission in Main Sequence Stars, Monthly Notices of the Royal Astronomical Society, 1972, vol. 159, p. 113

Géron A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc.", 2022

Ghoreyshi M. R., Carciofi A. C., Rímulo L. R., Vieira R. G., Faes D. M., Baade D., Bjorkman J. E., Otero S., Rivinius T., The life cycles of Be viscous decretion discs: The case of ω CMa, MNRAS, 2018, vol. 479, p. 2214

Johnson H. L., Morgan W. W., Fundamental stellar photometry for standards of spectral type on the Revised System of the Yerkes Spectral Atlas., ApJ, 1953, vol. 117, p. 313

Keller S. C., Bessell M. S., Spectroscopy of Be stars in the Small Magellanic Cloud cluster NGC 330, A&A, 1998, vol. 340, p. 397

Labadie-Bartz J., Carciofi A. C., Henrique de Amorim T., Rubio A., Luiz Figueiredo A., Ticiani dos Santos P., Thomson-Paressant K., Classifying Be Star Variability With TESS. I. The Southern Ecliptic, AJ, 2022, vol. 163, p. 226

Lee U., Saio H., Osaki Y., Viscous excretion discs around Be stars, Monthly Notices of the Royal Astronomical Society, 1991, vol. 250, p. 432

Loveday J., The Sloan Digital Sky Survey : Status and Prospects.. In Dark Matter in Cosmology Quantam Measurements Experimental Gravitation , 1996, p. 215

Lynden-Bell D., Pringle J. E., The evolution of viscous discs and the origin of the nebular variables., MNRAS, 1974, vol. 168, p. 603

Martayan C., Floquet M., Hubert A.-M., Gutiérrez-Soto J., Fabregat J., Neiner C., Mekkas M., Be stars and binaries in the field of the SMC open cluster NGC 330 with VLT-FLAMES, Astronomy & Astrophysics, 2007, vol. 472, p. 577

McSwain M. V., Gies D. R., A Photometric Method to Search for Be Stars in Open Clusters, ApJ, 2005, vol. 622, p. 1052

Mendes de Oliveira C., et al., The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters, MNRAS, 2019, vol. 489, p. 241

Miles R., A light history of photometry: from Hipparchus to the Hubble Space Telescope, Journal of the British Astronomical Association, 2007, vol. 117, p. 172

Milone A. P., Marino A. F., Di Criscienzo M., D'Antona F., Bedin L. R., Da Costa G., Piotto G., Tailo M., Dotter A., Angeloni R., Anderson J., Jerjen H., Li C., Dupree A., Granata V., Lagioia E. P., Mackey A. D., Nardiello D., Vesperini E., Multiple stellar populations in Magellanic Cloud clusters - VI. A survey of multiple sequences and Be stars in young clusters, MNRAS, 2018, vol. 477, p. 2640

Mota BeAtlas: A grid of synthetic spectra for Be stars, IAG-USP, 2019, Tese de Doutorado

Müller A. C., Guido S., Introduction to machine learning with Python: a guide for data scientists. "O'Reilly Media, Inc.", 2016

Platt J., et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 1999, vol. 10, p. 61

Rivinius T., Carciofi A. C., Martayan C., Classical be stars, The Astronomy and Astrophysics Review, 2013, vol. 21, p. 1

Rubio A. C., *Eyes on Phact*: unraveling a Be star and its disk, Universidade de São Paulo, 2019, Dissertação de Mestrado, 162 p.

Saunders C., Stitson M. O., Weston J., Bottou L., Smola A., et al., Support vector machine-reference manual, 1998

Secchi A., Schreiben des herrn prof. secchi, directors der sternwarte des collegio romano, an den herausgeber, Astronomische Nachrichten, 1866, vol. 68, p. 63

Song H., Perello-Nieto M., Santos-Rodriguez R., Kull M., Flach P., et al., Classifier Calibration: How to assess and improve predicted class probabilities: a survey, arXiv preprint arXiv:2112.10327, 2021

Soszyński I., Udalski A., Szymański M. K., Kubiak M., Pietrzyński G., Wyrzykowski Ł., Ulaczyk K., Poleski R., Kozłowski S., Pietrukowicz P., Skowron J., The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XV. Long-Period Variables in the Galactic Bulge, Acta Astron., 2013, vol. 63, p. 21

Tinsley B. M., Evolution of the Stars and Gas in Galaxies, Fund. Cosmic Phys., 1980, vol. 5, p. 287

Vieira R. G., Carciofi A. C., The Viscous Disk Properties of 80 Be Stars. In The B[e] Phenomenon: Forty Years of Studies , vol. 508 of Astronomical Society of the Pacific Conference Series, 2017, p. 33

von Zeipel H., The radiative equilibrium of a rotating system of gaseous masses, MNRAS, 1924, vol. 84, p. 665