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

Isabela Souza de Almeida Garcia

Busca por galáxias anãs ultra compactas usando Machine Learning

São Paulo
2022

Isabela Souza de Almeida Garcia

Busca por galáxias anãs ultra compactas usando Machine Learning

Trabalho de Conclusão de Curso apresentado ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Astronomia.
Vertente: Pesquisa Básica.
Orientador: Prof. Laerte Sodré Júnior.

São Paulo

2022

Dedico este trabalho ao meu pai (in memorian) e à minha mãe, pelo apoio e amor incondicional durante minha jornada acadêmica e à Isabela de 11 anos, que tinha apenas uma paixão e um sonho.

Agradecimentos

À Deus, por guiar e iluminar meu caminho e me permitir estudar suas belíssimas criações no Universo;

À minha mãe, Lúcia, ao meu pai, Davison (*in memorian*), à minha avó, Erotides, à minha tia, Luciana e ao meu tio, Flávio, pelo suporte, apoio e amor incondicional que me permitiu seguir meu sonho e me dedicar à pesquisa;

Ao meu orientador, Laerte Sodré, pelos inúmeros ensinamentos, paciência, atenção e amizade que, desde o início da minha graduação em Astronomia, me guiou pela pesquisa científica. A todos os professores, por provocar desafios que fortaleceram o meu amor e dedicação à Astronomia, em especial, à professora Cláudia Mendes de Oliveira, pelas oportunidades e reconhecimento;

Aos meus amigos e colegas que me acompanharam nesta jornada, pelo convívio, troca de experiências e motivação. Ao Erik, Marcelo, Pablo, Vitor, Elismar, Juliana, Lucas e, mais recentemente, Carlos, Laura e Anna, meus amigos e colegas de pesquisa, que possibilitaram, durante nossas reuniões semanais, discussões e boas conversas;

Ao CNPq, pelo apoio financeiro, sob projetos n^o 2021-1048 e n^o 2022-1310;

Ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas e ao Instituto de Física pela oportunidade de conhecimento e aprendizado, com inúmeros desafios.

Esta tese/dissertação foi escrita em LaTeX com a classe IAGTESE, para teses e dissertações do IAG.

"*A paixão é a gênese do gênio.*"

Galileu Galilei

"*Ciência é competitiva, agressiva, exigente. Também é imaginativa, inspiradora, edificante.*"

Vera Rubin

Resumo

Galáxias anãs ultracompactas (UCDs, *ultracompact dwarf*) são objetos que pertencem à classe de galáxias anãs, sendo maiores e mais brilhantes que aglomerados globulares típicos e mais compactos que galáxias anãs de mesma luminosidade, com raio de meia luz de $10 < r_h < 100$ pc e luminosidade entre -13.0 ¡ Mg ¡ -10.0. O estudo dedica-se à busca e identificação destes objetos no aglomerado de Fornax utilizando algoritmos de *Machine Learning*, como *Random Forest* e SVM (*Support Vector Machine*). Retiramos uma amostra do terceiro e outra do quarto *data release* do S-PLUS, estas consistem em objetos com $PhotoFlagDet = 0$ e $M_{g\_petro} < -11$, para distinguir de aglomerado globulares, e as separamos em estrelas com FWHM_n < 1.3 *pixels* e galáxias através da classificação do *survey*. Criamos dois modelos, o primeiro possui galáxias com FWHM_n > 2.0 *pixels* e o segundo, galáxias sem restrição desta medida. O treinamento dos algoritmos usou 80% de cada amostra e verificamos, ao testá-los com os restantes 20%, que conseguiram distinguir estrelas e galáxias com acurácia maior que 97% em ambos os modelos. Consideramos candidatas a UCDs objetos classificados como galáxias, com FWHM_n < 2.1 *pixels* - consistente com UCDs conhecidas no aglomerado - e *redshift* fotométrico menor que 0.05. Logo, obtivemos como resultado deste trabalho 18 candidatas no DR3/SPLUS e seis no DR4/SPLUS. Estas são visualmente parecidas com as UCDs conhecidas e suas propriedades mostram que estão localizadas próximas ao centro do aglomerado ou ao raio do virial, são uma mistura de objetos vermelhos e azuis e possuem luminosidade de $-14.0 < M_{g\_petro} < -11.0$.

Abstract

This study is dedicated to the search and identification of UCDs (ultra-compact dwarf) galaxies, they belong to the dwarf galaxies' class, being larger and brighter than typical globular clusters and more compact than dwarf galaxies of the same luminosity, with half-light radii of $10 < r_h < 100$ pc and luminosities between $-13.0 < M_g < -10.0$, in the Fornax cluster using Machine Learning algorithms, such as Random Forest and SVM (Support Vector Machine). We retrieved a sample from the third and another from the fourth data release of S-PLUS, they consist of objects with $PhotoFlagDet = 0$ and $Mg_{petro} < -11$, to distinguish them from globular clusters, and we separated them on stars with FWHM_n < 1.3 pixels and galaxies through survey's classification, due to those algorithms classify based on known objects. We created two models, there are galaxies with FWHM_n > 2.0 pixels in the first one and there is no galaxy with restriction on this measure in the second one. The algorithms were trained using 80% of each sample and were evaluated using the remaining 20%. They were able to distinguish stars and galaxies with an accuracy greater than 97% in both models. We consider as UCD candidates the objects classified as galaxies with FMWH_n < 2.1 pixels - consistent with known UCDs in the cluster - and photometric redshift less than 0.05. Therefore, as a result of this work, we obtained 18 candidates in DR3/SPLUS and six in DR4/SPLUS. They are visually similar to known UCDs and their properties show that they are located close to the center of the cluster or the virial radius, they are a mixing box of red and blue objects and have a luminosity of $-14.0 < M_{g\_petro} < -11.0$.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A população de galáxias anãs é dominante no universo, todavia, são relativamente pouco estudadas devido a dificuldades observacionais. Elas apresentam uma grande diversidade de propriedades e são encontradas em ambientes também variados (Hodge, 1971; Simon, 2019). Dentre as várias morfologias existentes, estamos interessados na classe de galáxias anãs ultra compactas (UCDs, *ultra compact dwarfs*). Estes objetos se caracterizam por serem maiores e mais brilhantes que aglomerados globulares típicos, mas significantemente mais compactos que galáxias anãs de mesma luminosidade, com raio de meia luz de $10 < r_h < 100$ pc, luminosidade entre $-13.0 < M_g < -10.0$ e razão de massa dinâmica e estelar de $M_{dyn}/M_* > 1$ (Saifollahi et al., 2021). Em baixas luminosidades, são difíceis de distinguir de aglomerados globulares ou, visualmente, de estrelas. Seu processo de formação também é mal conhecido (Drinkwater et al., 2000; Mieske et al., 2002).

O *Southern Photometric Local Universe Survey*, S-PLUS; Mendes de Oliveira et al. (2019), com suas 12 bandas fotométricas, oferece uma excelente oportunidade para investigar este tipo de galáxia, devido à excelente qualidade da fotometria, à boa precisão de seus *redshifts* fotométricos, com erro típico de 0.015 (Lima et al., 2022), e à possibilidade de se obter parâmetros descrevendo as populações estelares desses objetos.

O objetivo deste trabalho consiste na busca e identificação das galáxias UCDs através de um modelo de classificação estrela/galáxias mais simplificando, onde considera apenas suas magnitudes aparentes, utilizando algoritmos de *Machine Learning*. A classificação estrela/galáxia/quasar do S-PLUS inclui como parâmetros a morfologia e as 12 magnitudes dos objetos (Nakazono et al., 2021) e, por isso, construímos novos modelos, uma vez que as UCDs podem ser confundidas visualmente com estrelas. Para essa finalidade, utilizamos os catálogos do terceiro e quarto *data releases* do S-PLUS para sua procura na região do

aglomerado de galáxias de Fornax. Nossa maior dificuldade é conseguir distingui-las de estrelas.

1.1 Galáxias UCDs

Galáxias UCDs são da classe das anãs que, para uma dada magnitude, são significativamente mais compactas (menores) que a maioria deste tipo, sendo um pouco maiores que os aglomerados globulares (Phillipps et al., 2001). Seu processo de formação ainda não é bem conhecido, existindo três principais hipóteses: podem ser aglomerados globulares supermassivos, núcleo remanescente de galáxias anãs que perderam as camadas mais externas (*stripped dwarf galaxies*) - principalmente galáxias elípticas anãs - ou um novo grupo de galáxias anãs compactas.

A primeira hipótese consiste nas UCDs serem resultado de fusões de dois ou mais aglomerados globulares, gerando um aglomerado supermassivo. Segundo Mieske et al. (2002), há uma sobreposição entre as distribuições de magnitudes das UCDs mais fracas e dos aglomerados globulares, corroborando com a ideia de existir uma transição entre estes objetos. Entretanto, para as mais brilhantes, como a UC03 de Fornax, não é possível chegar a mesma conclusão devido ao seu brilho e seu tamanho que não são típicos de um aglomerado globular.

Alguns aglomerados estelares jovens e massivos (YMCs, *Young Massive Cluster*) populam a mesma região do Plano Fundamental que as UCDs (Mieske et al., 2006). No entanto, para Gregg et al. (2003), estes dois tipos de objetos aparentam ser dinamicamente diferentes, com as UCDs tendo maior dispersão de velocidade interna e maior razão $< M/L >$ que a Via Láctea e o aglomerado globular M31, estando mais relacionadas a elípticas brilhantes que aglomerados globulares. Eles também verificaram que elas são, aproximadamente, duas magnitudes mais brilhantes que o maior aglomerado globular em NGC 1399 e que possuem linhas espectrais de absorção que indicam populações estelares velhas e relativamente pobres em metais. Mieske et al. (2006) concluiu que as metalicidades das UCDs de Fornax são um indicativo de que devem ter sido criadas em um evento de fusão cedo e violento.

Se as UCDs forem aglomerados globulares supermassivos, então não é esperado que possuam matéria escura. Hilker et al. (2007) concluiu que nenhum componente de matéria

escura é necessário para UCDs dentro de 1-3 raios de meia massa. Chilingarian et al. (2011) apontou que, para os 14 objetos para os quais estimaram as massas dinâmicas, 12 parecem não possuir matéria escura e, para dois, não há mais que 40% da fração de massa em matéria escura.

Estes objetos também podem ser núcleos remanescentes de galáxias elípticas anãs que sofreram efeito de maré, perdendo suas estruturas externas devido aos eventos violentos que acontecem em um aglomerado de galáxias, sendo esta a hipótese sobre *stripped dwarf galaxies*. Se o modelo de *threshing* de Bekki et al. (2001) estiver correto, as UCDs teriam perdido sua parte externa, cerca de 98% de sua luz, enquanto o núcleo quase não seria afetado. Desta forma, teriam começado sua vida com aproximadamente 4.25 magnitudes mais brilhantes, no regime de galáxias anãs nucleadas (Gregg et al., 2003). Para estes tipos de objetos, não se esperaria componentes de matéria escura (Hilker et al., 2007), pois o gás e o halo de matéria escura seriam removidos há muito tempo e não teriam oportunidade de auto enriquecimento e formação estelar desde seus *stripping* (Mieske et al., 2006).

Contudo, Mieske et al. (2006) aponta que as UCDs de Fornax são mais ricas em metais e mais vermelhas que os núcleos existentes de elípticas anãs sem indicação de uma diferença significativa de idade, uma explicação seria o evento de *tidal stripping* selecionar as galáxias elípticas anãs mais ricas em metais, contudo, ainda não explica completamente esta grande diferença. Portanto, eles também destacam que esta grande diferença de metalicidade não é consistente com o cenário de *stripping* como canal principal de formação destes objetos.

A análise da velocidade média destes objetos feita por Michielsen et al. (2004) no aglomerado de Fornax mostrou que esta componente em UCDs mais brilhantes é mais consistente com a de galáxias anãs, enquanto UCDs mais fracas não. Este resultado apoia a hipótese de que objetos compactos, como as UCDs brilhantes ($V < 20$ *mag*) se originaram, principalmente, em galáxias anãs. Enquanto os objetos fracos, ($V > 20$ *mag*), principalmente de aglomerados globulares.

No entanto, eles tambem destacam que se elas fossem núcleos de *stripped galaxies* ou fusão de super-aglomerados estelares, deveria existir mais exemplos de transição, o que apoia a hipótese das UCDs serem um novo grupo de galáxias compactas. Desta forma, é esperado uma componente considerável de matéria escura por serem formadas em halos pequenos de matéria escura. Hilker et al. (2007) perceberam que a massa bariônica das esferoidais anãs do Grupo Local é comparável com a das UCDs.

Afanasiev et al. (2018) encontraram um buraco negro de 3.5 milhões de massas solares na UC03 de Fornax, a UCD mais massiva e brilhante entre as cinco clássicas deste aglomerado, o que é consistente com o cenário de que esta galáxia foi formada por *tidal stripping* de uma progenitora massiva. A Figura 1.1 apresenta a imagem de uma galáxia UCD de Fornax e de uma galáxia anã mais extensa.

(a) Galáxia UC03 de Fornax.

(b) Galáxia anã NGC185.

Figura 1.1: Imagem de uma galáxia anã UCD (1.1a) e uma extensa (1.1b) obtida pelo *Legacy Survey*.

1.2 Aprendizagem de Máquina

Machine Learning é a ciência (ou a arte) de programar computadores para que possam aprender com os dados (Géron, 2019, p. 6), hoje em dia, está em todos os lugares: nos filtros de *spam* dos e-mails, sistemas de recomendações em sites de *streaming* como Netflix ou Youtube ou detecção de doenças em imagens de radiografias. A utilização de técnicas de *Machine Learning* está sendo cada vez mais utilizada na astronomia, ajudando em análises de *big data* em diversas áreas, como na classificação de estrela-galáxia. Para realizar um aprendizado supervisionado, é necessário dividir a amostra de dados em um conjunto de treinamento, assim cada algoritmo criará um modelo próprio, e um conjunto de teste que será usado para testar o algoritmo e calcular uma (ou mais) medidas de performance.

Machine Learning é ótimo em conseguir simplificar problemas complexos e possui excelentes ferramentas para classificação (Géron, 2019, p. 28), sendo esta a principal motivação para utilizá-lo no nosso problema de distinção de objetos compactos e extensos através de suas magnitudes em diferentes bandas. Existem diversos algoritmos para problemas de classificação, de modo que é necessário escolher alguns com a finalidade de realizar o nosso estudo e comparar a performance dos modelos. Em um primeiro momento, utilizamos a

base de dados de nossos estudos anteriores sobre UCDs (vide Seção 2.2.1), realizamos a comparação de oito algoritmos diferentes e escolhemos aqueles com as maiores acurácias. Comparamos os algoritmos *Naive Bayes* (acurácia de 72.22%), Árvore de Decisão (acurácia de 74.07%), *Random Forest* (acurácia de 79.63%), Regras (acurácia de 68.51%), *Majority Learning* (acurácia de 60%), KNN (*K-Nearest Neighbour* com acurácia de 77.78%), Regressão Logística (acurácia de 68.52%) e SVM (*Support Vector Machine* com acurácia de 81.48%). Logo, escolhemos os algoritmos *Random Forest* e SVM para os estudos sobre a região do aglomerado de galáxias de Fornax.

O algoritmo *Random Forest* utiliza várias árvores de decisão que escolhe de forma aleatória n atributos. Assim, realiza o *Ensemble Learning* (aprendizagem em conjunto). O resultado final provém do 'voto' majoritário das árvores.

No algoritmo SVM, a classificação é feita ao localizar o hiperplano, encontrado a partir da minimização das margens, que melhor diferencia as classes. Se esta fronteira não for linear, então podemos utilizar o *Kernel Trick* que transforma uma superfície ou variedade não linear em linear.

Analisar métricas dos modelos gerados pelos algoritmos é uma forma de verificar se o resultado é bom, estas são diferentes cálculos de acertos e erros. A acurácia é o número de acertos dividido pelo número total de objetos; ks (teste Kolmogorov-Smirnov) avalia se duas amostras possuem diferenças significantes uma da outra ao calcular a maior distância entre as distribuições das classes; a área sob a curva ROC (AUC, *Area Under the Curve*) mede a área sob uma curva formada pelo gráfico entre a taxa verdadeiros positivos e a taxa de falsos positivos; f1 *score* é a média harmônica entre a precisão — razão entre a quantidade de verdadeiros positivos e o total objetos classificados como positivos — e o *recall* - razão entre a quantidade de verdadeiros positivos e a quantidade total de positivos; gini normaliza AUC em que os melhores modelos possuem estes valores próximos de um; o log_loss é o calculo logaritmo da probabilidade do objeto pertencer a classe real, em que valores próximos a zero indicam melhor resultado.

Capítulo 2

Pré-processamento de Dados

O pré-processamento de dados é fundamental ao se trabalhar com algoritmos de inteligência artificial. Nesta etapa, estudamos e verificamos se a amostra condiz com o nosso objetivo final e eliminamos possíveis sujeiras, como dados faltantes. Desta forma, analisamos os objetos desde onde ela é retirada até a seleção dos dados.

Após este processamento, separamos a amostra em base de treinamento e de teste, em que 80% se encontra no primeiro conjunto e 20% no segundo. Destaca-se que, devido à importância da reprodutibilidade do estudo, usou-se uma semente pré-definida nesta etapa (semente: $random\_state = 0$), portanto, sempre serão selecionados os mesmos objetos.

2.1 Amostragem do S-PLUS

O projeto S-PLUS utiliza o telescópio T80-South, de 0.8 metros, localizado no Observatório Inter-americano Cerro Tololo no Chile e, quando o *survey* for finalizado, abrangerá, aproximadamente, 9300 $graus^2$ da esfera celeste em doze filtros fotométricos, que são uma combinação de cinco bandas largas (u, g, r, i e z) para restringir o espectro contínuo das fontes e sete estreitas ($J0378$, $J0395$, $J0410$, $J0430$, $J0515$, $J0660$, $J0861$) que coincidem respectivamente com os *features* de [OII], Ca H+K, Hδ, *G-band*, tripleto Mgb, $H\alpha$ e tripleto Ca em $z = 0$ (Mendes de Oliveira et al., 2019). A Figura 2.1 apresenta as curvas de transmissão total deste sistema fotométrico.

Estas características e a grande precisão de seus *redshifts* fotométricos nos oferecem uma grande oportunidade de investigar as galáxias UCDs no aglomerado de Fornax. Desta forma, utilizamos o terceiro *data release* para a primeira investigação e o quarto com a finalidade de complementar o estudo, uma vez que o novo *data release* oferece melhores

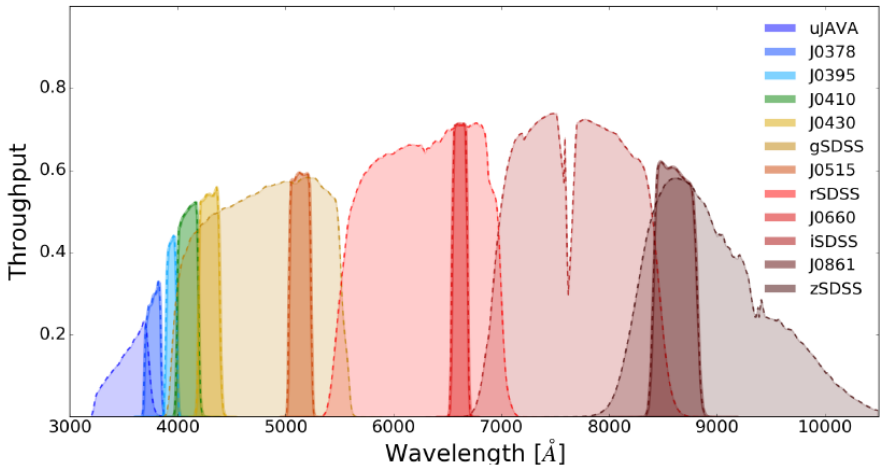


Figura 2.1: Sistema de doze filtros utilizados pelo S-PLUS; (Mendes de Oliveira et al., 2019).

fotometrias e aprimoramento da classificação estrela/galáxia/quasar.

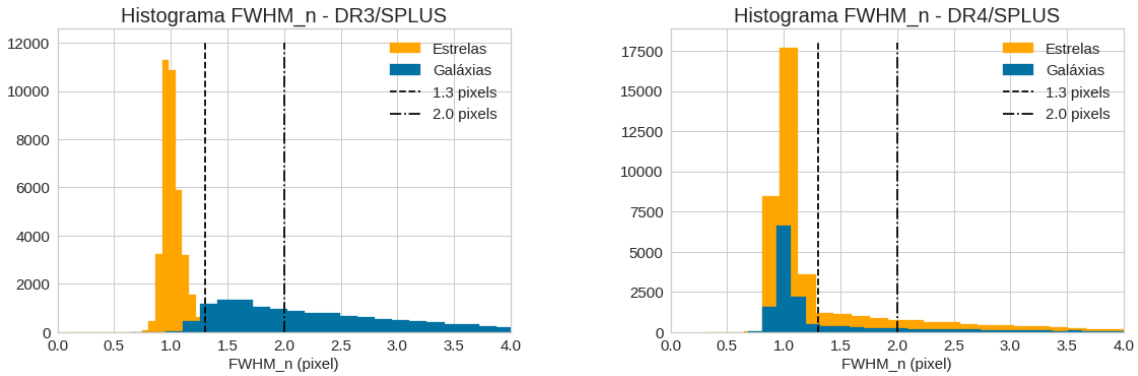
2.2 Seleção de amostra de treinamento e teste

A partir das amostras do S-PLUS, selecionamos apenas os objetos que possuem boa fotometria ($PhotoFlagDet = 0$) e com a magnitude absoluta $Mg_{petro} < -11$, desta forma evitamos os aglomerados globulares, que dominam em $Mg > -11$ (Mieske et al., 2006). Devido ao nosso interesse em utilizar apenas as medidas de magnitudes nos classificadores, descartamos todos os objetos com medida faltante em alguma banda na abertura *petro* (petrosiana) ou de abertura de três arco-segundos.

As magnitudes *petro* representam as aberturas totais que integram a maior parte da luz, sendo a estimativa da magnitude total das galáxias. A magnitude de abertura de três *arcsec* foi adicionada com a finalidade de ajudar na separação de objetos extensos e compactos, pois estas tendem a ter um brilho superficial central maior.

Devido aos algoritmos de *Machine Learning* utilizados serem supervisionados, de forma que é necessário uma amostra de objetos com a classe conhecida para o treinamento dos modelos, baseamos esta separação na probabilidade de cada objeto ser uma galáxia dada pelo S-PLUS, em que acima de 50% denotamos classe 1, das galáxias, e abaixo, classe 0, das estrelas. Comparamos uma medida de tamanho de ambas as classes, para melhor separá-las. A Figura 2.2 apresenta a distribuição de FWHM_n dentro da nossa amostra, em 2.2a é relativo ao DR3/SPLUS e 2.2b, ao DR4/SPLUS. Percebemos que a maioria das estrelas

possui esta medida em um valor próxima a um, mas as galáxias são mais distribuídas. Desse modo, para diferenciar de forma mais restritiva as classes, selecionamos estrelas com FWHM_n < 1.3 e galáxias com FWHM_n > 2.0.



(a) DR3/SPLUS. *(b)* DR4/SPLUS.

Figura 2.2: Histogramas de FWHM_n das galáxias e das estrelas para a região de Fornax dos catálogos de DR3/SPLUS e DR4/SPLUS.

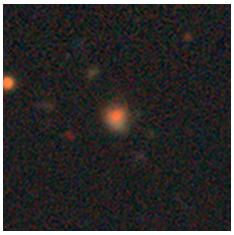
Antes de separar as bases de treinamento e teste, é necessário verificar o balanceamento das classes, pois uma grande diferença na quantidade de cada uma torna a medida de acurácia enviesada para a classe de maior número. Neste passo, verificamos que existem muito mais estrelas que galáxias na nossa amostra, uma proporção de 38.350 para 14.755 no DR3/SPLUS (72% para 28% da amostra) e 41.018 para 15.126 no DR4/SPLUS (73% para 27% da amostra). O balanceamento foi realizado apenas retirando a quantidade necessária de estrelas para as respectivas amostras e modelos, de forma que tenham a mesma quantidade das galáxias. Finalmente, dividimos a amostra em duas bases: 80% para o treinamento dos algoritmos e 20% para o teste e obtenção das métricas dos modelos.

2.2.1 Galáxias UCDs já conhecidas

Os estudos realizados durante as últimas duas décadas sobre as galáxias UCDs, introduzidos na Seção 1.1, resultaram na identificação de cinco UCDs que denominamos "clássicas"do aglomerado de galáxias de Fornax, vistas na Figura 2.3. Suas coordenadas e o valor de FWHM_n e das magnitudes aparente e absoluta na banda *g_petro* obtidas pelo DR3/SPLUS estão contidas na Tabela2.1. É interessante analisar a classe atribuída a cada uma delas pelos algoritmos, assim como, a probabilidade calculada.

Tabela 2.1 - Parâmetros das cinco UCDs clássicas do aglomerado de Fornax obtidas pelo DR3/S-PLUS.

Nome	RA	DEC	g_{petro}	Mg_{petro}	FWHM_n
UCD01	54.25	-35.63	21.16	-10.35	2.27
UCD02	54.53	-35.49	15.43	-16.08	23.71
UCD03	54.72	-35.56	19.12	-12.39	1.71
UCD04	54.91	-35.48	21.04	-10.47	2.03
UCD05	54.96	-35.06	19.32	-12.19	1.46



(a) UCD01.

(b) UCD02.

(c) UCD03.

(d) UCD04.

(e) UCD05.

Figura 2.3: Imagem das cinco galáxias UCDs clássicas do aglomerado de Fornax, obtida pelo *Legacy Survey*.

Devemos analisar estes parâmetros das UCDs clássicas do aglomerado de Fornax. As galáxias UCD01 e UCD04 possuem a magnitude absoluta mais fraca que o valor considerado na seleção, o que está relacionado com a questão em aberto destes objetos serem aglomerados globulares ou galáxias, apresentada na Seção 1.1. A galáxia UCD02 possui o valor de FWHM_n de 23.71, bem maior que o intervalo analisado nas outras galáxias da amostra. Ao observar sua imagem, Figura 2.3b, vemos que ela é relativamente grande comparada com as outras, sendo, também, a que possui a magnitude mais brilhante. É interessante analisar, por fim, a UCD03 apresentada na Figura 2.3c, pois esta se destaca por ser maior, mais brilhante e possuir um buraco negro em seu centro (Afanasiev et al., 2018).

Observa-se que o valor de FWHM_n é menor que dois *pixels* e é mais brilhante que um aglomerado globular típico, com $Mg\_petro = -12.39$. É visualmente mais avermelhada, maior e mais brilhante que as outras.

As UCDs são de nosso interesse há um ano em projetos de Iniciação Científica. Neste tempo, procuramos identificá-las no aglomerado de galáxias de Fornax pela técnica dos valores de FWHM_n, em que selecionamos todos objetos do DR2/SPLUS com $-14.5 < M_{g\_petro} < -11.0$, $Prob\_Gal > 0.5$ e com FWHM_n < 6 *pixels*, sendo este o valor comparado com estrelas, encontrando 134 candidatas. Neste projeto, nosso interesse se reserva em sua busca através de *Machine Learning*. Portanto, também é interessante saber como os algoritmos as classificaram e as probabilidades atribuídas a cada uma, fazendo uma comparação entre os resultados anteriores e os deste trabalho.

Este conjunto de 134 candidatas encontradas e as cinco UCDs clássicas formam a amostra, totalizando 139 galáxias de interesse. Devido aos algoritmos utilizarem 13 variáveis (uma sendo a abertura de três arco-segundos e as restantes serem as magnitudes medidas sob a abertura petrosiana) para a classificação, aquelas que não possuem alguma magnitude medida pelo S-PLUS foram descartadas. Logo, temos 82 galáxias de interesse, três são galáxias clássicas do aglomerado de Fornax, que são estudadas e analisadas com maior cautela na Seção 3.

Capítulo 3

Resultados e Análise

Construímos dois modelos para treinar os algoritmos de *Machine Learning*. A amostra 1 consiste em:

- estrelas: objetos com $Prob\_Gal < 0.5$ e FWHM_n < 1.3 *pixels*

- galáxias: objetos com $Prob\_Gal > 0.5$ e FWHM_n > 2.0 *pixels*.

Ela possui maior separação entre as classes e, assim, melhor diferenciação entre ambos os objetos. Contudo, valor pequeno de FWH_n não estariam na amostra e, devido ao nosso interesse em encontrar UCDs, decidimos também treiná-los com um modelo que inclui galáxias sem esta restrição. Desta forma, a amostra 2 consiste em:

- estrelas: objetos com $Prob\_Gal < 0.5$ e FWHM_n < 1.3 *pixels*

- galáxias: objetos com $Prob\_Gal > 0.5$ e todos valores de FWHM_n.

Logo, temos uma amostra de candidatas a partir de ambos os modelos e algoritmos. Destaca-se que eles possuem hiper-parâmetros importantes que influenciam na acurácia dos resultados, por essa razão, usamos um *framework* do Python chamado *optuna* para encontrar os melhores de cada modelo.

3.1 Amostra do DR3/SPLUS

A primeira análise foi realizada em uma amostra retirada do DR3/SPLUS na região do aglomerado de Fornax, onde selecionamos apenas estrelas com $FWHM\_n < 1.3$ *pixels*, obtendo 37.313 objetos. Para o primeiro modelo, as galáxias possuem a restrição desta medida em FWHM_n, totalizando 8.556 galáxias, logo, necessitou retirar 28.757 estrelas de forma que a amostra possua a mesma quantidade de ambas as classes. As bases de treinamento e de teste possuem 13.676 e 3.500 objetos, respectivamente. As Figuras das

matrizes de confusão e gráfico da relevância dos parâmetros de cada modelo encontram-se no Apêndice A.1.

O algoritmo *Random Forest* utilizou 490 árvores de decisão com profundidade máxima de 25 galhos, resultando em uma acurácia de 99.22%, com 11 falsos positivos e 16 falsos negativos, como visto na matriz de confusão, Figura A.1a. As métricas do modelo estão na Tabela 3.1, em que os valores de ks, AUC, f1 e gini estão próximas a um e log_loss próximo a zero, apresentando bom resultado. É interessante observar que a abertura de três arco-segundos é a variável mais importante para a classificação, como visto no histograma de relevância dos parâmetros, Figura A.1b. Isso possivelmente acontece porque, para uma dada magnitude, as UCDs têm brilho central maior que as estrelas

O algoritmo SVM utilizou o *kernel* polinomial, resultando em uma acurácia de 99.37% e sua matriz de confusão se encontra na Figura A.2, onde verificamos mais acertos na classificação, com 10 falsos positivos e 12 falsos negativos. As métricas apresentadas na Tabela 3.1 possuem valores melhores que os obtidos pelo *Random Forest* mas muito próximos, onde deduzimos que estes algoritmos conseguiram resultados parecidos.

As candidatas são todos os objetos classificados como galáxias neste treinamento, com FWHM_n < 2.1 *pixels*, para ser coerente com as UCDs clássicas, e com o valor do *redshift* fotométrico (zml, *redshift machinle learning*) menor que 0.03, por ser este 2σ da incerteza desta medida. Desta forma, ambos os algoritmos resultaram em um mesmo e único objeto como candidata a UCD.

Em seguida, também testamos este modelo para galáxias sem restrição de FWHM_n, com a finalidade de entender como os algoritmos classificariam as que possuem este valor pequeno. *Random Forest* conseguiu acurácia de 91.68% e SVM, de 91.43%, apresentando uma piora na capacidade de acertar a classe. As matrizes de confusão da Figura A.3 mostram maior quantidade de erro ao classificá-las, em que muitas foram entendidas como estrelas (classe 0). Nesta situação, encontramos nove candidatas pelos mesmos critérios.

O segundo modelo treinado consiste em classificar estrelas, com FWHM_n < 1.3 *pixels*, e todas as galáxias, ou seja, sem restrição desta medida. Excluímos 22.586 daquelas 37.313 estrelas de forma que cada classe possui 14.727 objetos. As bases de treinamento e de teste possuem 23.546 e 5.967 respectivamente.

O *Random Forest* utilizou 817 árvores de decisão com profundidade máxima de 29 galhos. Atingiu acurácia de 98.19% e suas métricas podem ser vistas na Tabela 3.1 em

que observamos ótimos resultados. Sua matriz de confusão se apresenta na Figura A.4a, mostrando que obteve 54 falsos positivos e 54 falsos negativos. A abertura de três arco-segundos se destaca novamente como o parâmetro mais importante, Figura A.4b.

O SVM utilizou o *kernel* rbf (função de base radial, *Radial Basis Function*) obtendo acurácia de 97.95%, suas métricas, na Tabela 3.1, são ótimas, contudo, são as piores entre os quatro resultados devido ao maior log_loss e menor valor das outras métricas. A matriz de confusão, presente na Figura A.5, mostra que o algoritmo obteve 65 falsos negativos e 58 falsos positivos.

Portanto, observamos que ao incluir galáxias de valor de FWHM_n pequeno, apesar desta medida não ser incluída entre os parâmetros dos algoritmos, estes tendem a ter maior dificuldade em diferenciá-las de estrelas. Contudo, as métricas, que são diferentes cálculos da relação entre acertos e erros, mostram que o segundo modelo também é excelente, conseguindo dez candidatas ao seguir o mesmo critério realizado no primeiro modelo.

Tabela 3.1 - Valores das métricas dos modelos para cada algoritmo para a amostra do DR3/SPLUS.

	Acurácia (%)	ks	log_loss	roc_auc	f1	Gini
modelo 1 - RF	99.22	0.986	0.045	0.999	0.992	0.999
modelo 1 - SVM	99.37	0.987	0.031	0.999	0.994	0.998
modelo 2 - RF	98.19	0.964	0.074	0.998	0.982	0.995
modelo 2 - SVM	97.94	0.961	0.075	0.996	0.979	0.992

Considerando os três resultados (dois obtidos pelo primeiro modelo e um através do segundo), a amostra de candidatas dentro do DR3/SPLUS possui 18 objetos, sobre os quais verificamos serem parecidos com as UCDs clássicas pela análise visual, Figura 3.1.

O mapa da Figura 3.2 apresenta a distribuição espacial das UCDs clássicas, que se encontram próximas ao centro do aglomerado, e das candidatas, destacando a cor (g-r)_petro, que estão distribuídas pela região. Estas se localizam próximas ao centro e ao raio do virial, ressaltando que alguns estão fora deste raio; podem ser galáxias *splash-back*, ou seja, estão entrando no aglomerado pela segunda vez.

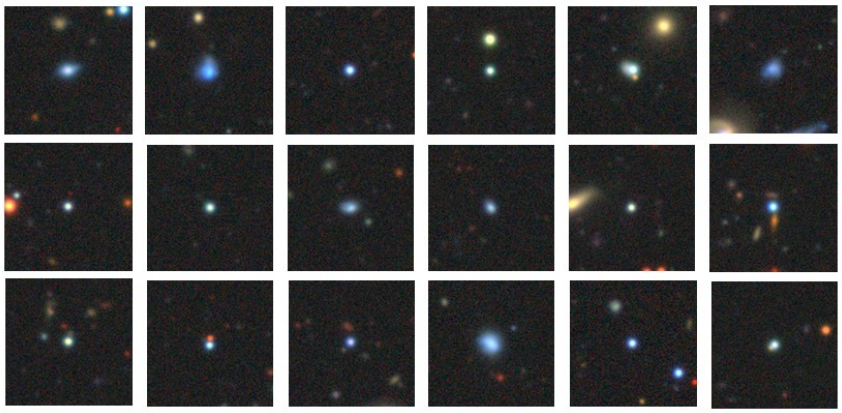


Figura 3.1: Imagens das 18 candidatas obtidas pelos dois modelos dentro da amostra do DR3/SPLUS, fonte: *Legacy Survey*.

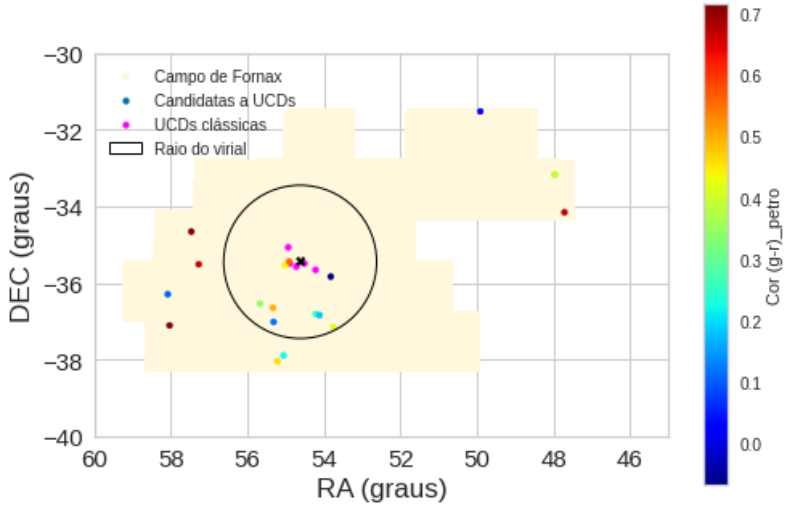


Figura 3.2: Distribuição espacial das UCDs clássicas, em magenta, e das candidatas obtidas, suas cores variam de acordo com a cor (g-r)_petro. O raio do virial e centro do aglomerado estão representados em preto. A área observada de Fornax está em amarelo claro.

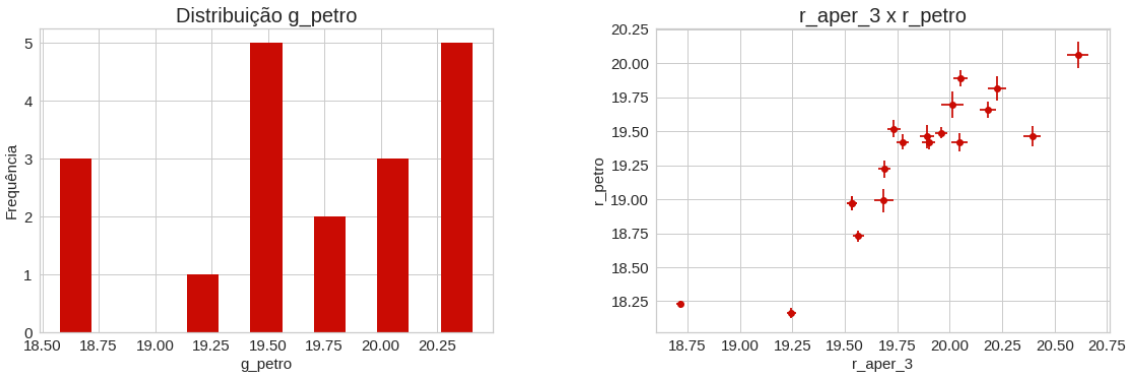
A magnitude g_petro varia de 18.50 a 20.50, como apresentado na Figura 3.3a, de forma que as candidatas possuem intervalo de magnitude absoluta nesta banda, dada pela Equação 3.1, de $-13.0 < M_{g\_petro} < -11$

$$M_{g\_petro} = g\_petro - DM \tag{3.1}$$

Onde DM é o módulo da distância que possui valor de 31.5 para Fornax.

A Figura 3.3b apresenta o gráfico da magnitude de abertura de três arco-segundos pela magnitude r_petro, mostrando uma tendência do número de objetos crescer com a magnitude. Os histogramas das cores (g-r)_petro e (r-J0660)_petro, Figura 3.4a e Figura

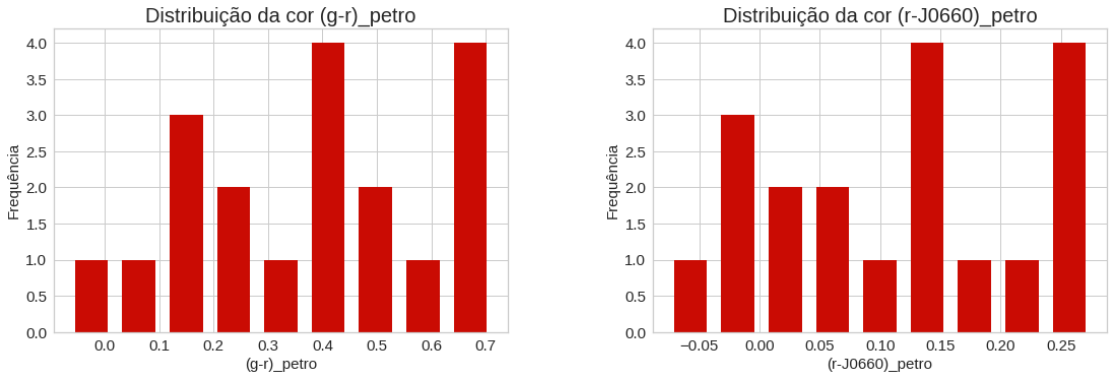
3.4b respectivamente, mostram que estão distribuídas dentro do intervalo de $0.0 < (g - r)\_petro < 0.7$ e $-0.05 < (r - J0600)\_petro < 0.250$.



(a) Histograma de g_petro.

(b) Gráfico de r_petro pela abertura de três arco-segundos.

Figura 3.3: 3.3a Histograma da magnitude g_petro e 3.3b Gráfico da magnitude r_petro em função da abertura de três arco-segundos.



(a) Histograma da cor (g-r)_petro g_petro.

(b) Histograma da cor (r-J0660)_petro g_petro.

Figura 3.4: Histograma da cor 3.4a (g-r)_petro e 3.4b (r-J0660)_petro das candidatas obtidas pelos modelos na amostra do DR3/SPLUS.

O diagrama cor-cor, Figura 3.5, nos mostra que as candidatas são uma mistura de objetos vermelhos e azuis. O mapa da Figura 3.2 revela que isto independe de sua localização dentro do aglomerado, indicando diversas origens para estes objetos, uma vez que esta característica relaciona-se, entre outros fatores, com sua localização dentro do aglomerado devido a interações violentas com outras galáxias. Logo, as UCDs próximas umas as outras, mas com taxa de formação estelar diferente, podem ter sido formadas por métodos distintos.

Finalmente, analisamos a distribuição do *redshift* fotométrico das candidatas, Figura 3.6. A maioria das candidatas possui os maiores valores desta medida, em que metade está no limite selecionado, $zml = 0.03$. Notamos, também, que uma possui este valor igual ao *redshift* espectroscópico de Fornax, $z = 0.005$. Por outro lado, é sabido que os *redshifts* fotométricos dessas galáxias obtidos pelo S-PLUS não são confiáveis pelo fato do *redshift* de Fornax ser menor que a metade do erro estimado dos *zmls* nesse intervalo.

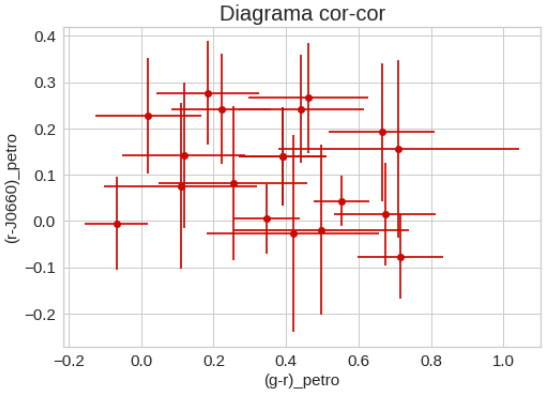


Figura 3.5: Diagrama cor-cor das candidatas obtidas pelo DR3/SPLUS.

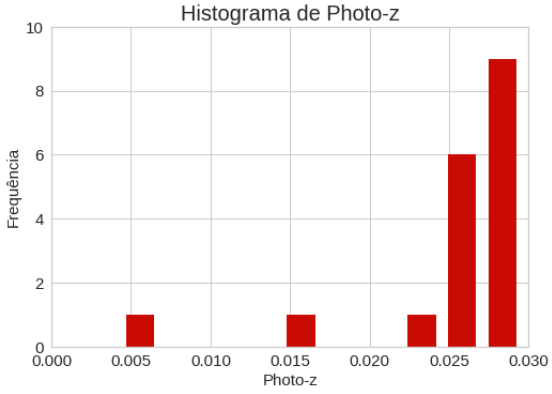


Figura 3.6: Distribuição do *redshift* fotométrico das candidatas obtidas pelo DR3/SPLUS.

Em Outubro de 2022, foi requerido tempo de observação no telescópio Gemini destas candidatas com a finalidade de analisar seus espectros, obter o *redshift* espectroscópico e visualizá-las melhor. As observações e a redução dos dados estão sendo realizadas até Dezembro de 2022, mas já foi possível notar que algumas possuem movimento próprio e, portanto, são estrelas.

3.2 Amostra do DR4/SPLUS

De forma análoga, a partir de uma amostra do DR4/SPLUS na região do aglomerado de Fornax, selecionamos apenas estrelas com $FWHM\_n < 1.3$ *pixels*, obtendo 40.164 objetos. Para o primeiro modelo, as galáxias possuem restrição em FWHM_n, totalizando 8.772 galáxias e, logo, necessitou retirar 31.392 estrelas de forma que a amostra possua a mesma quantidade de ambas as classes. As bases de treinamento e de teste possuem 14.035 e 3.509 objetos, respectivamente. As matrizes de confusão e histogramas de parâmetros mais importantes podem ser visto no Apêndice A.2.

O algoritmo *Random Forest* utilizou 690 árvores de decisão com profundidade máxima de 20 galhos, obteve acurácia de 99.06%, métricas excelentes, Tabela 3.2, e resultou em 16

falsos negativos e 17 falsos positivos, Figura A.6a. Semelhante aos outros modelos, o filtro de abertura de três arco-segundos é a variável mais importante, Figura A.6b.

O algoritmo SVM usou o *kernel* polinomial e resultou em uma acurácia de 99.34% com métricas parecidas com o obtido pelo *Random Forest*, Tabela 3.2, mas atingiu menos erros, com 18 falsos negativos e 5 falsos positivos, como visto na matriz de confusão, Figura A.7.

A seleção de candidatas é semelhante a amostra anterior: objetos classificados como galáxias pelos algoritmos, com FWHM_n < 2.1 *pixels* e $zml < 0.05$, em que resultou em duas candidatas. Acrescentamos, nesta amostra, os objetos falsos negativos, galáxias classificadas como estrelas, pois, neste caso, são objetos considerados estrelas pelo nosso algoritmos mas galáxias pela classificação do S-PLUS e, devido essa confusão entre os modelos, estas podem ser candidatas a UCDs. Contudo, o primeiro modelo não encontrou nenhum objeto dentro dos critérios necessários.

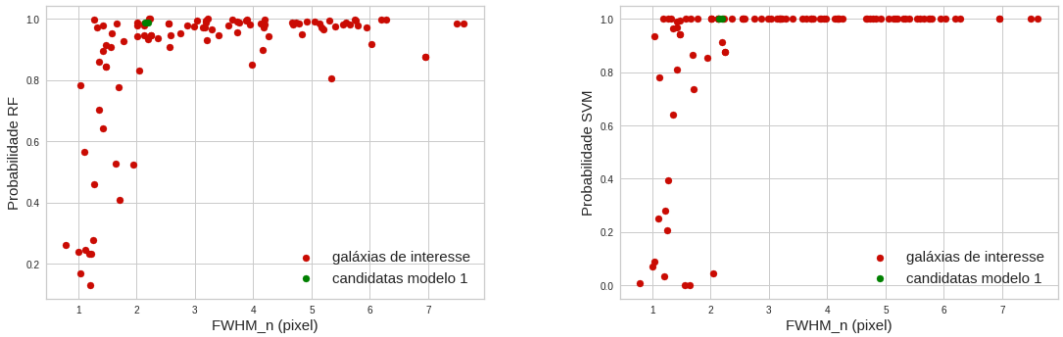
Aplicamos ambos os algoritmos na amostra das galáxias de interesse e nas candidatas obtidas anterioriormente, totalizando 100 objetos, com a finalidade de observar como estes foram classificados. Obtiveram acurácia de 91.0% para *Random Forest* e 88.0% para SVM.

O valor de FWHM_n se relaciona com as classificações apesar de não ser inserido como um parâmetro do modelo, como mencionado na Seção 3.1. A Figura 3.7 apresenta o gráfico desta medida em função da probabilidade de ser uma galáxia calculada para *Random Forest* e SVM, Figura 3.7a e Figura 3.7b em respectiva ordem. Observamos que a probabilidade se mantém alta para os maiores valores de FWHM_n, mas para estes valores pequenos, aproximadamente entre 1 e 2 *pixels*, a probabilidade está bem distribuída. Em relação às candidatas, possuem probabilidades altas mesmo com valor pequeno de FWHM_n.

O segundo modelo, que consiste em classificar estrelas e galáxias sem restrição desta medida, ou seja, todas da amostra, necessitou excluir 25.079 de 40.164 estrelas de forma que cada classe possui 15.085 objetos. As bases de treinamento e de teste possuem 24.136 e 6.034 respectivamente.

O algoritmo *Random Forest* utilizou 495 árvores de decisão com profundidade máxima de 18 galhos. Atingiu acurácia de 98.12% com métricas ótimas mas um pouco piores que o modelo anterior, Tabela 3.2. Sua matriz de confusão, Figura A.8a, mostra que obteve 56 falsos positivos e 57 falsos negativos. a magnitude de abertura de três arcos segundos novamente é a variável mais importante, FiguraA.8b.

O algoritmo SVM utilizou o *kernel* polinomial atingindo acurácia de 97.81% e ótimas



(a) Random Forest. *(b)* SVM.

Figura 3.7: Gráficos que relacionam a medida FWHM_n e a probabilidade de ser galáxia calculada pelo 3.7a *Random Forest* e 3.7b SVM para o modelo 1. Verde representa as candidatas deste modelo e vermelho a amostra das candidatas de interesse.

métricas mas não tão boas quanto os modelos anteriores, Tabela 3.2. A matriz de confusão, Figura A.9, mostra que obteve 79 falsos positivos e 53 falsos negativos.

Da mesma forma como observado na Seção 3.1, a não restrição das galáxias ocasiona maior confusão na classificação dos algoritmos e, assim, maiores erros. Contudo, estes modelos se mostram eficazes para a classificação estrela/galáxia uma vez que possuem acurácia de aproximadamente 98% e métricas excelentes.

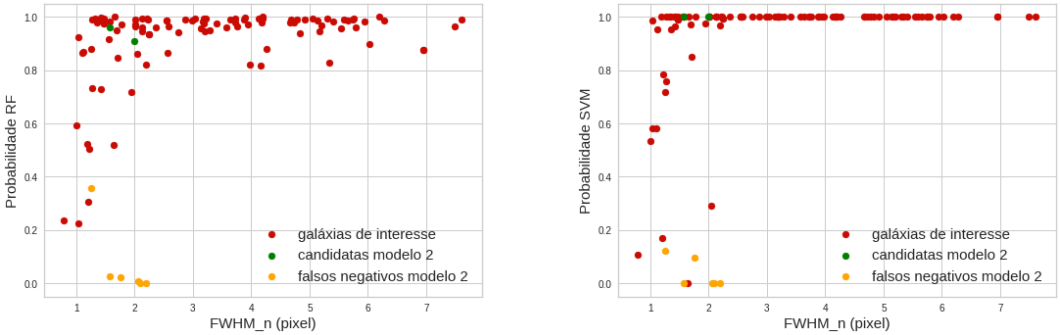
As candidatas foram obtidas de forma análoga ao primeiro modelo, resultando em duas candidatas classificadas como galáxias e seis candidatas através do falso negativo. A amostra das galáxias de interesse obteve acurácia de 88.0% para *Random Forest* e 85% para o SVM, apresentando maior dificuldade em classificá-las.

Tabela 3.2 - Valores das métricas dos modelos para cada algoritmo para a amostra do DR4/SPLUS.

	Acurácia (%)	ks	log_loss	roc_auc	f1	Gini
modelo 1 - RF	99.06	0.981	0.047	0.999	0.990	0.999
modelo 1 - SVM	99.34	0.987	0.025	0.999	0.993	0.999
modelo 2 - RF	98.13	0.962	0.071	0.998	0.981	0.996
modelo 2 - SVM	97.81	0.958	0.082	0.996	0.978	0.991

A relação entre a medida FWHM_n e a probabilidade de ser galáxias estão presentes na Figura 3.8a para o *Random Forest* e na Figura 3.8b para o SVM. Vemos relação similar à

encontrada no primeiro modelo entre estes parâmetros, destacando as altas probabilidades das candidatas em verde, consideradas galáxias, e das baixas probabilidades das candidatas falsas negativas em amarelo, tendo valor máximo menor que 40%.



(a) Random Forest. *(b)* SVM.

Figura 3.8: Gráficos que relacionam a medida FWHM_n e a probabilidade de ser galáxia calculada pelo 3.7a *Random Forest* e 3.7b SVM para o modelo 2. Verde representa as candidatas deste modelo, amarelo as candidatas através do falso negativo, vermelho a amostra das candidatas de interesse.

Portanto, a partir de uma amostra do DR4/SPLUS, obtivemos dez candidatas iniciais através destes modelos. Verificamos se estas possuem movimento próprio através do terceiro *data release* do gaia e, assim, descartamos quatro objetos, resultando em seis candidatas finais cujas imagens podem ser vistas na Figura 3.9. Estas se assemelham com galáxias compactas. Uma destas galáxias, ponta inferior direita da Figura 3.9, é candidata em ambas as amostras do DR3/SPLUS e DR4/SPLUS e apresenta alguma estrutura.

Figura 3.9: Imagens das seis candidatas obtidas pelos dois modelos dentro da amostra do DR4/SPLUS, fonte: *Legacy Survey.*

A distribuição espacial, Figura 3.10, mostra que três estão próximas ao raio do virial, sendo uma na parte de fora. A magnitude g_petro varia de 17.50 a 20.0, como apresentado na Figura 3.11a, assim as candidatas possuem magnitude absoluta dentro de intervalo de $-14.0 < M_{g\_petro} < -11.5$, dada pela Equação 3.1. A Figura 3.11b apresenta o gráfico da magnitude de abertura de três arco-segundos pela magnitude r_petro, os objetos estão dentro do intervalo de $18.0 < r\_aper\_3 < 20.25$ e $17.0 < r\_petro < 19.5$.

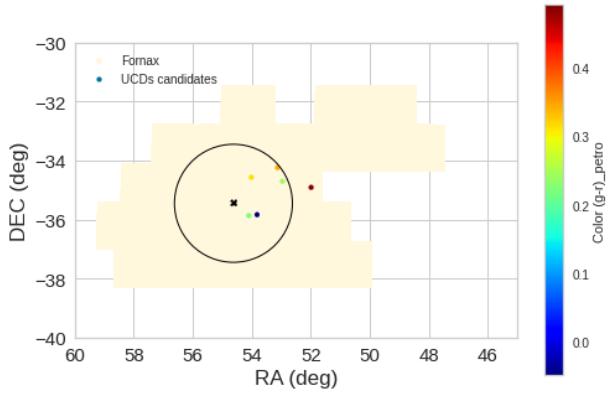
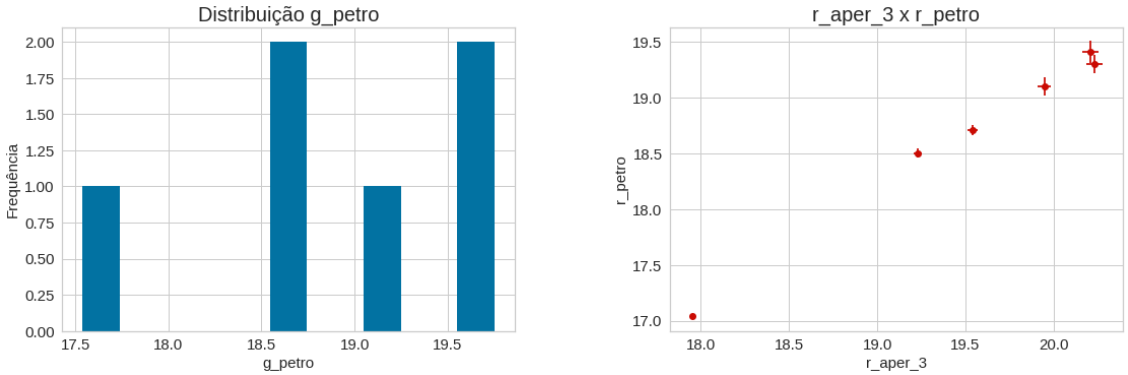


Figura 3.10: Distribuição espacial das candidatas obtidas, suas cores variam de acordo com a cor (g-r)_petro. O raio do virial e centro do aglomerado estão representados em preto. A área observada de Fornax está em amarelo claro.

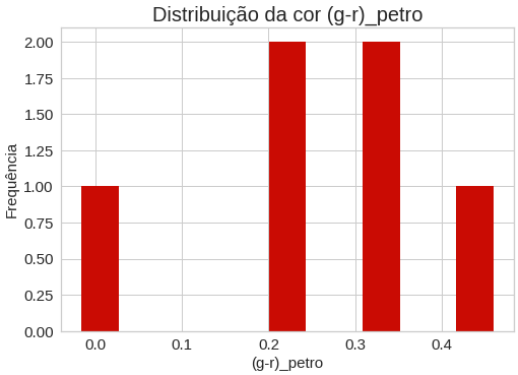


(a) Histograma de g_petro.

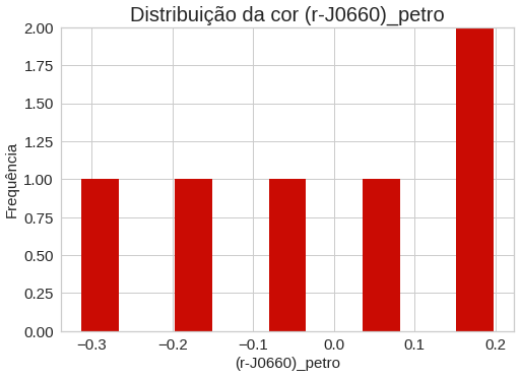
(b) Gráfico de r_petro pela abertura de três arco-segundos.

Figura 3.11: 3.11a Histograma da magnitude g_petro e 3.11b. Gráfico da magnitude r_petro em função da abertura de três arco-segundos.

Os histogramas das cores (g-r)_petro e (r-J0660)_petro, Figura 3.12a e Figura 3.12b respectivamente, mostram que estão distribuídas dentro do intervalo de $0.0 < (g-r)\_petro < 0.5$ e $-0.03 < (r - J0600)\_petro < 0.20$, muito parecido com as candidatas anteriores.



(a) Histograma da cor (g-r)_petro g_petro.



(b) Histograma da cor (r-J0660)_petro g_petro.

Figura 3.12: Histograma da cor 3.12a (g-r)_petro e 3.12b (r-J0660)_petro das candidatas obtidas pelos modelos na amostra do DR4/SPLUS.

O diagrama cor-cor, Figura B.4, nos mostra que, de forma semelhante às candidatas anteriores, as obtidas nesta amostra também são uma mistura de objetos vermelhos e azuis. O mapa da Figura 3.10 revela a distribuição de cores independe da localização dentro do aglomerado e, logo, podemos ter conclusões semelhantes às obtidas na Seção 3.1, condizentes com os diversos canais possíveis de formação destes objetos, introduzido na Seção 1.1. A análise da distribuição do *redshift* fotométrico das candidatas, Figura 3.14, revela que as candidatas estão distribuídas em $0.0325 < zml < 0.05$, próximas ao limite estabelecido de $zml = 0.05$ e maiores que o limite da Seção 3.1, de $zml = 0.03$.

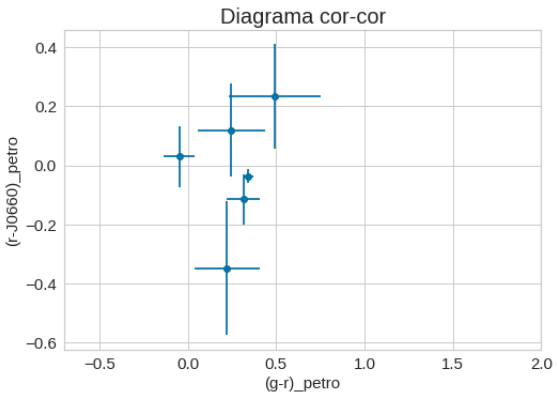


Figura 3.13: Diagrama cor-cor das candidatas obtidas pelo DR4/SPLUS.

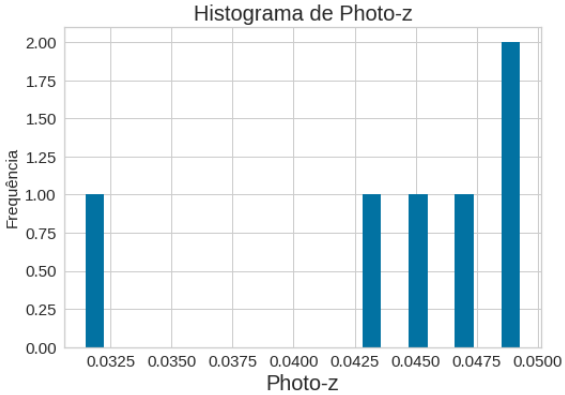


Figura 3.14: Distribuição do *redshift* fotométrico das candidatas obtidas pelo DR4/SPLUS.

Assim, para o DR4/SPLUS obtivemos seis candidatas a UCDs as quais, visualmente, se parecem com galáxias. Suas propriedades são muito parecidas com as candidatas resultantes do DR3/SPLUS e, uma, está presente em ambas amostras finais.

3.3 Candidatas Finais

Obtivemos em dois *data releases* 23 candidatas a UCDs, 18 através do DR3/SPLUS e seis pelo DR4/SPLUS, sendo uma em comum, suas imagens podem ser vistas na Figura 3.15. No Apêndice B, disponibilizamos um compêndio com os resultados apresentados nas Seções 3.1 e 3.2. A Tabela B.1 contém suas coordenadas, FWHM_n, zml e as magnitudes M_{g_petro}, g_petro, r_petro e de abertura de três arco-segundos. Os objetos marcados com asterisco estão sendo observados pelo Gemini e, com dois asteriscos, é a galáxias em comum. Observamos que a maioria dos objetos possui FWHM_n maior que um *pixel* e menor dois *pixels*. Em relação às magnitudes no filtro g, como também observado na Figura B.2a, a maioria está próxima do limite de $M_g = -11.0$, onde apenas duas são mais brilhantes que $M_g = -13.0$. Em relação às magnitudes r_petro e r_aper_3, também há concentração de objetos nas magnitudes mais fracas, como também observado na Figura B.2b. A análise dos *redshifts* fotométricos, também vista na Figura B.5, mostra que a maioria possui esta medida entre $0.02 < zml < 0.03$ e uma possui *redshift* negativo. A distribuição de cores, Figura B.3a e Figura B.3b, mostra que as candidatas estão concentradas em $-0.5 < (g - r)\_petro < 1.0$ e $-0.6 < (r - J660)\_petro < 0.4$ e a análise do mapa, Figura B.1, apresenta a distribuição espacial de todas as candidatas obtidas, onde observamos que se localizam próximas ao centro do aglomerado e ao raio do virial.

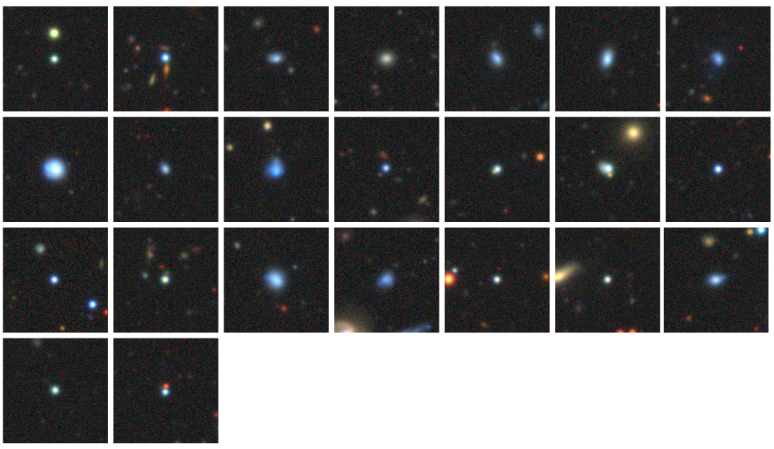


Figura 3.15: Imagens das candidatas obtidas em ambos os *data releases*, fonte: *Legacy Survey*.

Capítulo 4

Conclusões

Galáxias UCDs são uma classe de galáxias anãs. Elas são pouco estudadas devido a dificuldades observacionais e seu processo de formação ainda é pouco conhecido, podendo ser aglomerados globulares supermassivos, núcleos remanescente de galáxias elípticas anãs ou um novo grupo de galáxias. São visualmente parecidas com estrelas e, em baixas luminosidades, podem ser confundidas com aglomerados globulares.

Com a finalidade de identificá-las no aglomerado de Fornax, realizamos a classificação de estrelas e galáxias através de algoritmos de classificação de *Machine Learning*, como *Random Forest* e SVM, a partir de suas magnitudes aparentes petrosianas em 12 filtros e no filtro r de abertura em três *arco-segundos* do S-PLUS, utilizando seu terceiro e quarto *data release*, em que candidatas UCDs são os objetos classificados como galáxias, com FWHM_n < 2.1 *pixels* e $zml < 0.05$.

A partir dos catálogos do S-PLUS, selecionamos todos os objetos com boa fotometria ($PhotoFlagDet = 0$) e magnitude absoluta mais brilhante que a dominada por aglomerados globulares ($M_{g\_petro} < -11$). As classes foram separadas através da probabilidade de ser uma galáxias dada pelo S-PLUS, onde estrelas, classe 0, são os objetos com $Prob\_Gal < 0.5$ e galáxias, classe 1, com $Prob\_Gal > 0.5$.

Criamos dois modelos para cada *data release*. A base de treinamento do primeiro possui estrelas com $FWHM\_n < 1.3$ *pixels* e galáxias com $FWHM\_n > 2.0$ *pixels* para maior separação de objetos, atingindo acurácias acima de 99% em cada *data release*. Na amostra do DR3/SPLUS, também testamos o modelo com galáxias sem restrição desta medida e identificamos mais galáxias classificadas erradas que os outros resultados. A base de treinamento do segundo modelo consiste em estrelas com $FWHM\_n < 1.3$ *pixels* e galáxias sem esta restrição, ou seja, as classes não são tão bem separadas quanto no

primeiro modelo. A acurácia é um pouco menor, aproximadamente 98%, o que nos mostra que, apesar da medida de FWHM_n não ser um parâmetro para o aprendizado da máquina, ela afeta o resultado. Ambos os modelos apresentam resultados excelentes, com ótimas métricas e possuem como parâmetro mais importante para a classificação a magnitude r de abertura de três arco-segundos.

Obtivemos 18 candidatas dentro do DR3/SPLUS, as quais estão sendo observadas pelo telescópio Gemini, onde foi verificado que algumas possuem movimento próprio e, portanto, são estrelas. Através do DR4/SPLUS conseguimos seis candidatas cujo movimento próprio foi verificado no terceiro *data release* do gaia.

Realizamos sua análise visual, onde percebemos similaridade com as UCDs clássicas de Fornax. A distribuição espacial mostrou que se localizam próximas ao centro do aglomerado ou ao raio do virial, e algumas estão fora deste raio, podendo ser galáxias *splash-back* — galáxias que ao entrar no aglomerado, são rebatidas para fora do raio do virial e, em seguida, entram novamente. Através da distribuição de cores, vemos que são uma mistura de objetos vermelhos e azuis e, nos mapas, que sua localização não necessariamente interfere nas cores; assim, esta pode ser uma evidência de seus diversos canais de formação. Elas estão dentro do intervalo de magnitude $-14.0 < M_{g\_petro} < -11.0$.

Aplicamos os classificadores na amostra de galáxias de interesse, objetos obtidos em projetos anteriores apenas considerando o FWHM_n, as cinco UCDs clássicas e, para o DR4/SPLUS, as candidatas do DR3/SPLUS. Verificamos que a maioria foi classificada como galáxias com alta probabilidade e, as que foram identificadas como estrelas possuem esta medida distribuída, mas abaixo de 40%. O gráfico que relaciona o FWHM_n com a probabilidade calculada pelos algoritmos mostrou que, apesar desta medida não ser usada nos modelos, os classificadores apresentam distribuição de probabilidade maior em valores pequenos de FWHM_n e probabilidades próximas de 100% para o valor desta medida alta.

Assim, conclui-se que o objetivo deste trabalho foi cumprido com a obtenção de 23 candidatas a UCDs no aglomerado de Fornax, 18 através do DR3/SPLUS e seis através do DR4/SPLUS sendo uma em comum. Para etapas futuras, iremos analisar as observações feitas pelo Gemini e discutir suas propriedades e possíveis canais de formação em um artigo. Estes resultados poderão contribuir para os estudos de UCDs e para modelos de classificação estrela/galáxias mais simplificados, onde é utilizado apenas as magnitudes como parâmetros.

Referências Bibliográficas

Afanasiev A. V., Chilingarian I. V., Mieske S., Voggel K. T., Picotti A., Hilker M., Seth A., Neumayer N., Frank M., Romanowsky A. J., Hau G., Baumgardt H., Ahn C., Strader J., den Brok M., McDermid R., Spitler L., Brodie J., Walsh J. L., A 3.5 million Solar masses black hole in the centre of the ultracompact dwarf galaxy fornax UCD3, Monthly Notices of the Royal Astronomical Society, 2018, vol. 477, p. 4856

Bekki K., Couch W. J., Drinkwater M. J., Gregg M. D., A New Formation Model for M32: A Threshed Early-Type Spiral Galaxy?, The Astrophysical Journal, 2001, vol. 557, p. 39

Chilingarian I. V., Mieske S., Hilker M., Infante L., Dynamical versus stellar masses of ultracompact dwarf galaxies in the Fornax cluster72 , Monthly Notices of the Royal Astronomical Society, 2011, vol. 412, p. 1627

Drinkwater M. J., Jones J. B., Gregg M. D., Phillipps S., Compact Stellar Systems in the Fornax Cluster: Super-massive Star Clusters or Extremely Compact Dwarf Galaxies?, Publications of the Astronomical Society of Australia, 2000, vol. 17, p. 227

Gregg M., Drinkwater M., Hilker M., Phillipps S., Jones J. B., Ferguson H. C., , Astrophysics and Space Science, 2003, vol. 285, p. 113

Géron A., Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, 2019, 1109

Hilker M., Baumgardt H., Infante L., Drinkwater M., Evstigneeva E., Gregg M., Weighing Ultracompact Dwarf Galaxies in the Fornax Cluster, The Messenger, 2007, vol. 1, p. 49

Hodge P. W., Dwarf Galaxies, Annual Review of Astronomy and Astrophysics, 1971, vol. 9, p. 35

Lima E., Sodré L., Bom C., Teixeira G., Nakazono L., Buzzo M., Queiroz C., Herpich F., Castellon J. N., Dantas M., Dors O., de Souza R. T., Akras S., Jiménez-Teja Y., Kanaan A., Ribeiro T., Schoennell W., Photometric redshifts for the S-PLUS Survey: Is machine learning up to the task?, Astronomy and Computing, 2022, vol. 38, p. 100510

Mendes de Oliveira C., Ribeiro T., Schoenell W., Kanaan A., Overzier R. A., Molino A., Sampedro L., Coelho P., Barbosa C. E., Cortesi A., Costa-Duarte M. V., Herpich F. R., Hernandez-Jimenez J. A., Placco V. M., Xavier H. S., Abramo L. R., Saito R. K., Chies-Santos A. L., Ederoclite A., de Oliveira R. L., Gonç alves D. R., Akras S., Almeida L. A., Almeida-Fernandes F., Beers T. C., Bonatto C., Bonoli S., Cypriano E. S., Vinicius-Lima E., de Souza R. S., de Souza G. F., Ferrari F., Gonçalves T. S., Gonzalez A. H., Gutiérrez-Soto L. A., Hartmann E. A., Jaffe Y., Kerber L. O., Lima-Dias C., Lopes P. A. A., The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters, Monthly Notices of the Royal Astronomical Society, 2019, vol. 489, p. 241

Michielsen D., Rijcke S. D., Zeilinger W. W., Prugniel P., Dejonghe H., Roberts S., Evidence for a warm interstellar medium in Fornax dwarf elliptical galaxies - II. FCC032, FCC206 and FCCB729, Monthly Notices of the Royal Astronomical Society, 2004, vol. 353, p. 1293

Mieske S., Hilker M., Infante L., Ultra compact objects in the Fornax cluster of galaxies: Globular clusters or dwarf galaxies?, Astronomy &amp Astrophysics, 2002, vol. 383, p. 823

Mieske S., Hilker M., Infante L., Jordá n A., Spectroscopic Metallicities for Fornax Ultra-compact Dwarf Galaxies, Globular Clusters, and Nucleated Dwarf Elliptical Galaxies, The Astronomical Journal, 2006, vol. 131, p. 2442

Nakazono L., de Oliveira C. M., Hirata N. S. T., Jeram S., Queiroz C., Eikenberry S. S., Gonzalez A. H., Abramo R., Overzier R., Espadoto M., Martinazzo A., Sampedro L., Herpich F. R., Almeida-Fernandes F., Werle A., Barbosa C. E., Jr. L. S., Lima E. V.,

Buzzo M. L., Cortesi A., Menéndez-Delmestre K., Akras S., Alvarez-Candal A., Lopes A. R., Telles E., Schoenell W., Kanaan A., Ribeiro T., On the discovery of stars, quasars, and galaxies in the Southern Hemisphere with S-PLUS DR2, Monthly Notices of the Royal Astronomical Society, 2021, vol. 507, p. 5847

Phillipps S., Drinkwater M. J., Gregg M. D., Jones J. B., Ultracompact Dwarf Galaxies in the Fornax Cluster, The Astrophysical Journal, 2001, vol. 560, p. 201

Saifollahi T., Janz J., Peletier R. F., Cantiello M., Hilker M., Mieske S., Valentijn E. A., Venhola A., Kleijn G. V., Ultra-compact dwarfs beyond the centre of the Fornax galaxy cluster: hints of UCD formation in low-density environments, Monthly Notices of the Royal Astronomical Society, 2021, vol. 504, p. 3580

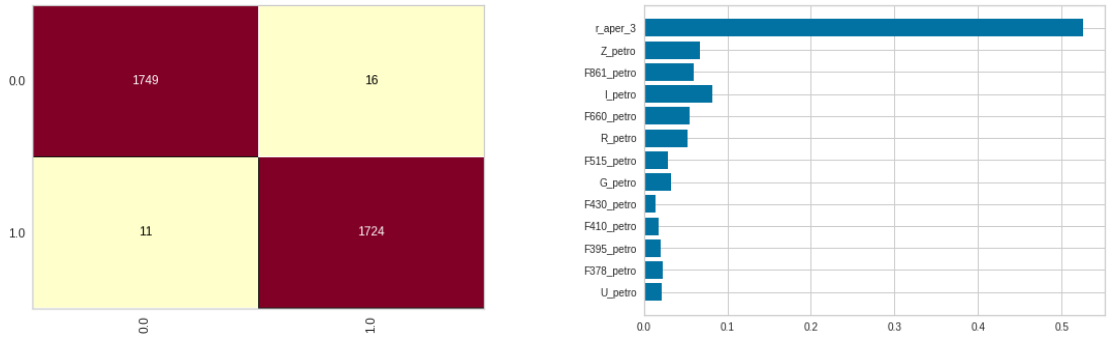
Simon J. D., The Faintest Dwarf Galaxies, Annual Review of Astronomy and Astrophysics, 2019, vol. 57, p. 375

Apêndice

Apêndice A

Matrizes de confusão e Histograma de relevância dos parâmetros

A.1 Obtidos com DR3/SPLUS



(a) Matriz de Confusão. *(b)* Gráfico dos melhores parâmetros.

Figura A.1: A.1a Matriz de Confusão obtido pelo *Random Forest* ao aplicar o primeiro modelo. A.1b Melhores parâmetros encontrados pelo algoritmo para a classificação - DR3/SPLUS.

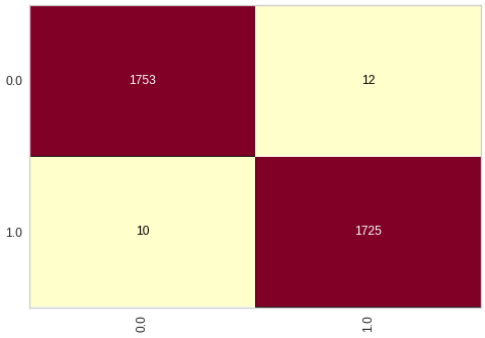
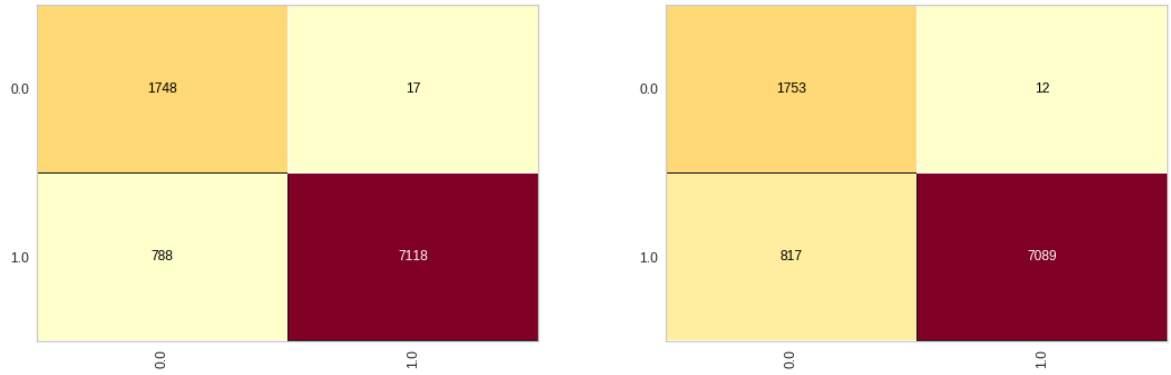
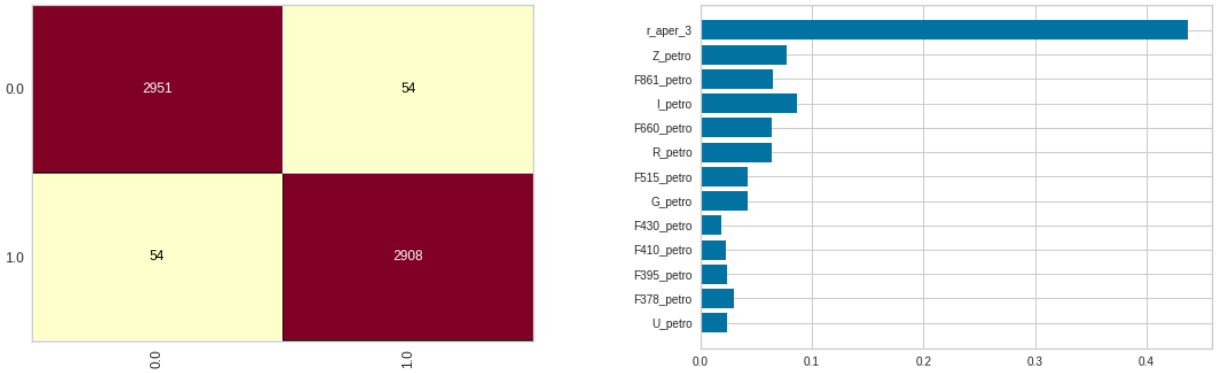


Figura A.2: Matriz de confusão obtida pelo algoritmo SVM ao treiná-lo com o primeiro modelo - DR3/SPLUS.



(a) Random Forest. *(b)* SVM.

Figura A.3: Matriz de confusão do algoritmo A.3a *Random Foreste* A.3b SVM ao testar o primeiro modelo em galáxias sem restrição de FWHM_n - DR3/SPLUS.



(a) Matriz de Confusão. *(b)* Gráfico dos melhores parâmetros.

Figura A.4: A.4a Matriz de Confusão obtido pelo *Random Forest* ao aplicar o segundo modelo. A.4b Melhores parâmetros encontrados pelo algoritmo para a classificação - DR3/SPLUS.

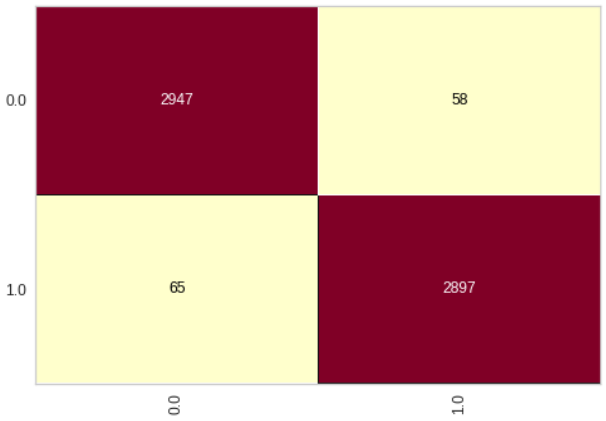
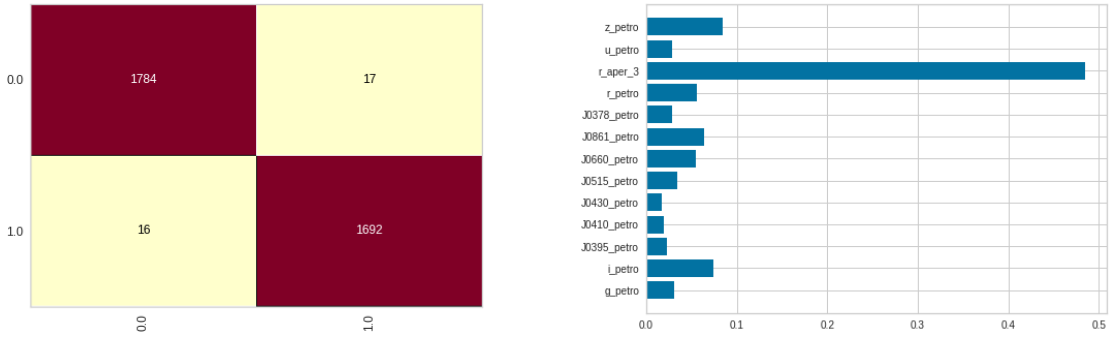


Figura A.5: Matriz de confusão obtida pelo algoritmo SVM ao treiná-lo com o segundo modelo - DR3/SPLUS.

A.2 Obtidos com DR4/SPLUS



(a) Matriz de Confusão. *(b)* Gráfico dos melhores parâmetros.

Figura A.6: A.6a Matriz de Confusão obtido pelo *Random Forest* ao aplicar o primeiro modelo. A.6b Melhores parâmetros encontrados pelo algoritmo para a classificação.

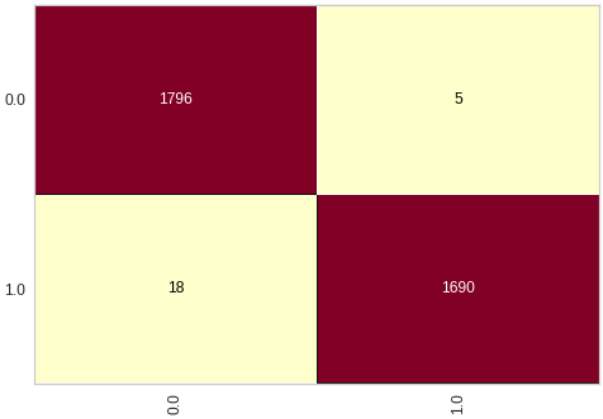
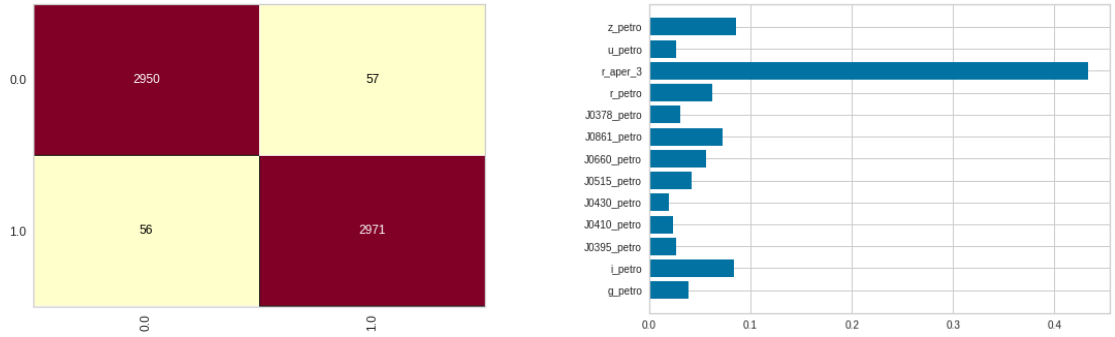


Figura A.7: Matriz de confusão obtida pelo algoritmo SVM ao treiná-lo com o primeiro modelo.



(a) Matriz de Confusão. *(b)* Gráfico dos melhores parâmetros.

Figura A.8: A.8a Matriz de Confusão obtido pelo *Random Forest* ao aplicar o segundo modelo. A.8b Melhores parâmetros encontrados pelo algoritmo para a classificação.

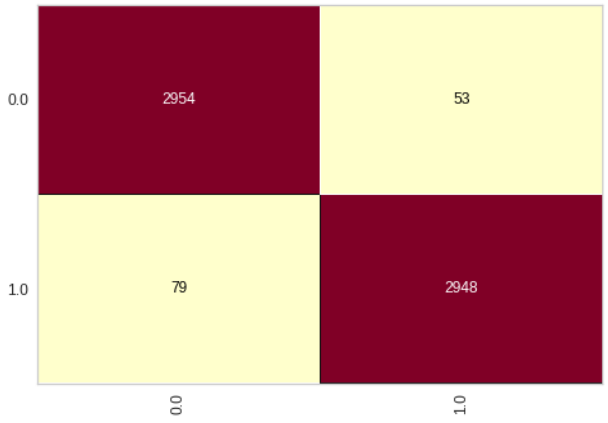


Figura A.9: Matriz de confusão obtida pelo algoritmo SVM ao treiná-lo com o segundo modelo.

Apêndice B

Candidatas finais - Gráficos de propriedades

Tabela B.1 - Parâmetros das candidatas obtidas pelo DR3/SPLUS e pelo DR4/SPLUS. O asterisco representa os objetos que estão sendo observados pelo Gemini; a galáxia com dois asteriscos foi obtida em ambos os *data releases*.

	RA	DEC	FWHM_n	M_{g_petro}	g_petro	r_petro	r_aper_3	zml
*	55.01	-35.53	1.23	-11.55	19.96	19.52	19.73	0.028
*	54.21	-36.81	1.25	-11.79	19.71	19.46	19.89	0.028
*	47.96	-33.17	1.27	-12.15	19.36	18.97	19.53	0.029
**	53.82	-35.84	2.07	-12.85	18.66	18.73	19.56	0.028
*	49.91	-31.52	1.71	-12.07	19.44	19.42	20.05	0.024
*	47.71	-34.16	1.22	-11.42	20.09	19.42	19.89	0.026
*	54.11	-36.84	1.28	-12.10	19.40	19.22	19.69	0.025
*	55.67	-36.54	1.77	-13.00	18.51	18.16	19.25	0.028
*	53.75	-37.15	1.59	-11.03	20.48	20.06	20.61	0.025
*	55.06	-37.89	1.42	11.63	19.88	19.66	20.18	0.028
*	57.47	-34.66	2.01	-11.10	20.40	19.69	20.01	-0.021
*	57.27	-35.51	1.03	-11.42	20.09	19.42	19.78	0.027
*	55.21	-38.05	0.81	-11.16	20.35	19.89	20.05	0.025
*	55.33	-36.65	1.09	-11.19	20.31	19.81	20.22	0.006
*	55.31	-37.02	2.08	-11.92	19.58	19.46	20.39	0.016
*	58.08	-36.29	1.68	-12.41	19.09	18.99	19.68	0.029
*	54.91	-35.43	1.24	-12.72	18.78	18.23	18.71	0.029
*	58.03	-37.11	1.05	-11.30	20.20	19.48	19.95	0.025
	53.12	-34.24	2.19	-14.12	17.38	17.04	17.94	0.048
	54.09	-35.86	2.13	-11.98	19.52	19.30	20.23	0.049
	54.01	-34.56	1.56	-12.69	18.81	18.50	19.23	0.043
	52.96	-34.70	1.99	-12.16	19.34	19.10	19.94	0.044
	51.98	-34.91	2.05	-11.60	19.90	19.40	20.20	0.047

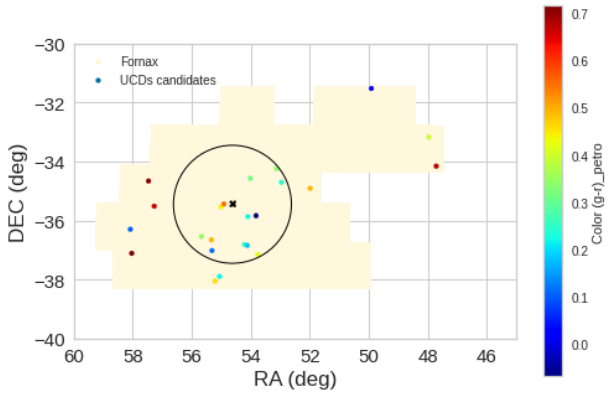
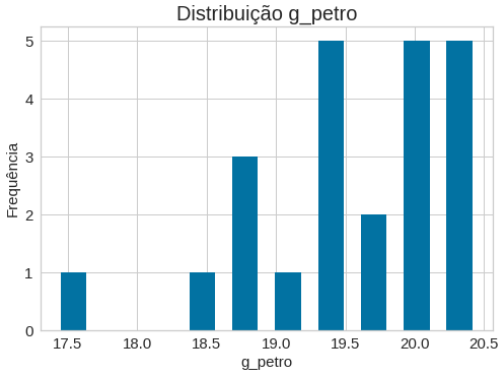
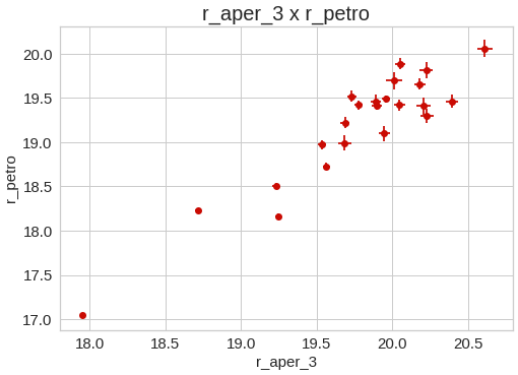


Figura B.1: Distribuição espacial das candidatas obtidas em ambos os *data releases*, suas cores variam de acordo com a cor (g-r)_petro. O raio do virial e centro do aglomerado estão representados em preto. A área observada de Fornax está em amarelo claro.

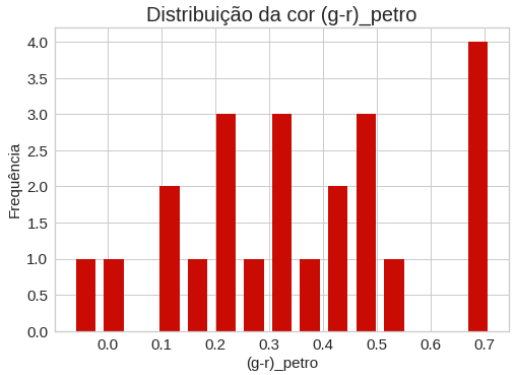


(a) Histograma de g_petro.

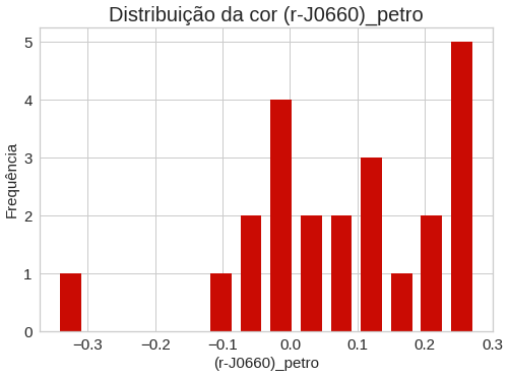


(b) Gráfico de r_petro pela abertura de três arco segundos.

Figura B.2: B.2a Histograma da magnitude g_petro e B.2b Gráfico da magnitude r_petro em função da abertura de três arco segundos.



(a) Histograma da cor (g-r)_petro g_petro.



(b) Histograma da cor (r-J0660)_petro g_petro.

Figura B.3: Histograma da cor B.3a (g-r)_petro e B.3b (r-J0660)_petro das candidatas obtidas em ambos os *data releases*

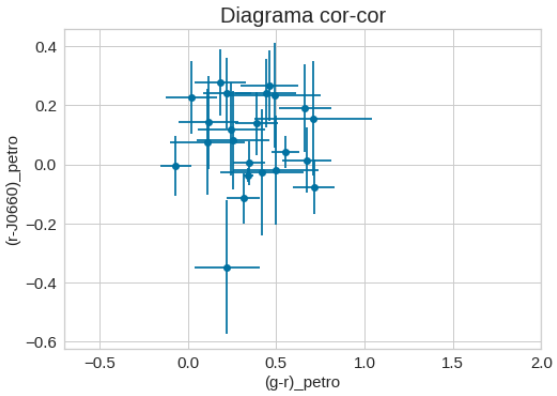


Figura B.4: Diagrama cor-cor das candidatas obtidas em ambos os *data releases*.

Figura B.5: Distribuição do *redshift* fotométrico das candidatas obtidas em ambos os *data releases*.

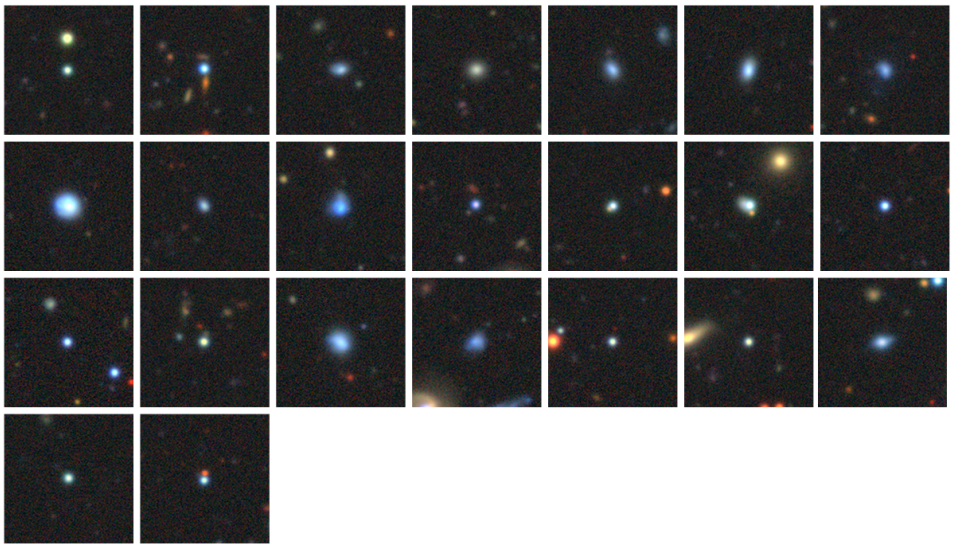


Figura B.6: Imagens das candidatas obtidas em ambos os *data releases*, fonte: *Legacy Survey*.