UNIVERSIDADE FEDERAL DO ESPIRITO SANTO
CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

PEDRO OTAVIO SOUZA BAQUI

SEPARAÇÃO DE ESTRELAS-GALÁXIAS USANDO ALGORITMOS DE MACHINE LEARNING APLICADOS AOS DADOS PRELIMINARES DO SURVEY MINIJPAS

VITÓRIA
2020

PEDRO OTAVIO SOUZA BAQUI

SEPARAÇÃO DE ESTRELAS-GALÁXIAS USANDO ALGORITMOS DE MACHINE LEARNING APLICADOS AOS DADOS PRELIMINARES DO SURVEY MINIJPAS

Tese apresentada ao Programa de Pós-Graduação em Física do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Doutor em Física, na área de concentração de Física Teòrica.

Orientador: Valério Marra
Coorientador: Luciano Casarini

Vitória
Abril de 2020

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Baqui, Pedro Otavio Souza, 1991-
B111 : Separação de estrelas-galáxias usando algoritmos de machine learning aplicados aos dados preliminares do survey MINIJPAS. / Pedro Otavio Souza Baqui. - 2020.
120 f. : il.

Orientador: Valerio Marra.
Coorientador: Luciano Casarini.
Tese (Doutorado em Física) - Universidade Federal do Espírito Santo, Centro de Ciências Exatas.

1. Classificação Estrelas/Galáxias. 2. Aprendizado de máquina. 3. Análise de dados. 4. Fotometria. I. Marra, Valerio. II. Casarini, Luciano. III. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. IV. Título.

CDU: 53

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

"Separação de estrelas-galáxias usando algoritmos de machine learning aplicados aos dados preliminares do survey J-PAS"

Pedro Otavio Souza Baqui

Tese submetida ao Programa de Pós-Graduação em Física da Universidade Federal do Espírito Santo, por webconferência, utilizando MConf, como requisito parcial para a obtenção do título de Doutor em Física.

Aprovada por:

_____ Prof. Dr. Luis Raul Weber Abramo (USP)	_____ Prof. Dr. Valerio Marra (Orientador - PPGFis/UFES)
_____ Prof. Dr. Miguel Boavista Quartin (UFRJ)	_____ Prof. Dr. Oliver Fabio Piattella (PPGFis/UFES)
_____ Prof. Dr. Júlio César Fabris (PPGFis/UFES)	_____ Prof. Dr. Davi Cabral Rodrigues (PPGFis/UFES)
_____ Prof. Dr. Luciano Casarini (Coorientador – UFSE)	

Vitória-ES, 08 de maio de 2020

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
JULIO CESAR FABRIS - SIAPE 297051
Departamento de Física - DF/CCE
Em 08/05/2020 às 18:38

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/21886?tipoArquivo=O

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
DAVI CABRAL RODRIGUES - SIAPE 1816732
Programa de Pós-Graduação em Física - PPGF/CCE
Em 08/05/2020 às 15:49

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/21846?tipoArquivo=O

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
OLIVER FABIO PIATTELLA - SIAPE 2847692
Departamento de Física - DF/CCE
Em 08/05/2020 às 15:07

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/21800?tipoArquivo=O

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
VALERIO MARRA - SIAPE 2179379
Departamento de Física - DF/CCE
Em 08/05/2020 às 13:46

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/21756?tipoArquivo=O

Dedico esse trabalho aos meus pais Luciana e Marcus, à minha irmã Mariana e a uma flor que passou pelo meu caminho chamada Jessica Buge (in memoriam).

Agradecimentos

Agradeço aos meus pais Luciana Souza e Marcus Baqui pela educação e disciplina que me deram. Pelos incentivos nos estudos. Pelas lições tomadas desde criança e que me fizeram tomar gosto pela leitura e ciência. Agradeço aos professores Valério Marra e Luciano Casarini pelo trabalho em conjunto, orientação e pela grande dedicação. Aos quais tornaram esse periodo de pesquisa um fase prazerosa. Agradeço a todos os meus amigos da pós-graduação aos quais tornaram o ambiente de trabalho divertido e de bastante aprendizagem ao mesmo tempo. Agradeço a todos que contribuiram para que essa tese pudesse ser concluída.

Agradeço à CAPES pelo financiamento do trabalho assim como o grupo Cosmo-UFES com as discussões no campo da astronomia/astrofísica os quais foram de fundamental importância. Também agradeço aos pesquisadores do survey J-PAS os quais trabalhei em parceria e que contribuíram de forma excepcional, em particular o *Data Validation Team*.

Não poderia também deixar de agradecer, a todos aqueles que lutaram para a contrução de uma sociedade mais justa, igualitária e que permitiram pessoas de menor poder aquisitivo estudar/pesquisar. O presente que temos foi contruído a partir do trabalho de muitos anônimos e personalidades. Presto aqui meus sinceros agradecimentos a todos esses.

*"**Vou aprender a ler pra ensinar aos meus camaradas.**"*
(Cantinga popular de Samba Chula do Reconcavo Baiano)

Resumo

Futuras levantamentos em astronomia/astrofísicas como o J-PAS, SDSS e LSST produzirão conjuntos de dados enormes chegando à uma taxa de 150 TB por dia. Portanto, novas ferramentas para processamento dessa quantidade de dados devem ser empregadas. De preferência que nos forneçam uma resposta quase que em tempo real, de forma eficiente e precisa. Cenário ideal para a aplicação de métodos de Machine Learning. Neste trabalho, analisamos dados do Pathfinder miniJPAS Survey, que observou $\sim 1deg^2$ sobre o campo AEGIS com 56 filtros de banda estreita e 4 filtros de banda larga *ugri*. Aqui, discutiremos a classificação de fontes observadas pelo miniJPAS em objetos compactos e estendidos, uma etapa necessária para os estudos científicos subsequentes. Assumimos em primeira aproximação estrelas como objetos compactos e galáxias como extendidos. Nosso objetivo é desenvolver um classificador de Machine Learning (ML) complementar às ferramentas tradicionais baseadas em outras modelagem. Em particular, nosso objetivo é construir um catálogo de valor agregado com nossas melhores classificações. Para treinar e testar nossos classificadores, cruzamos o conjunto de dados do miniJPAS com os dados SDSS e HSC-SSP, cuja classificação assumimos confiável dentro dos intervalos $15 < r < 21$ e $18.5 < r < 23.5$, respectivamente. Treinamos e testamos 6 algoritmos de ML diferentes nos dois catálogos com correspondência cruzada: K-Vizinhos mais Próximos (KNN), Árvores de Decisão (DT), Floresta Aleatória (RF), Redes Neurais Artificiais (ANN), Árvores Extremamente Randomizadas (ERT) e Ensemble Learning (Ensemble). Como entrada para os algoritmos ML, usamos 60 magnitudes associadas a cada banda fotométrica, com e sem os parâmetros morfológicos. Concluímos que, de acordo com a classificação SDSS, o algoritmo Ensemble apresenta melhor desempenho, obtendo $AUC = 0.9992$ (área sob a curva ROC) e $MSE = 0.009$ (erro quadrático médio). Ao se trabalhar com magnitudes mais fracas usando a classificação de HSC-SSP, o algoritmo Ensemble alcança o melhor desempenho, obtendo $AUC = 0.9744$ e $MSE = 0.0370$. Os últimos resultados são obtidos usando bandas fotométricas juntamente a parâmetros morfológicos. Os algoritmos de ML podem competir com classificadores tradicionais de estrela-galáxia, potencialmente superando o último em magnitudes mais fracas ($r \gtrsim 21$). Por fim contruímos um catálogo para a faixa $15 \leq r \leq 23.5$ utilizando máquinas treinadas a partir da fusão de rótulos entre os surveys SDSS e HSC-SSP.

Palavras-chave: Classificação Estrelas/Galaxias - Aprendizado de máquina - Análise de dados - Fotometria.

Abstract

Future astrophysical research such as JPAS will produce huge datasets never seen before, reaching a rate of 150 TB per day. Therefore, new tools for processing this amount of data must be employed. Preferably they will provide us with an almost real-time response in an efficient and accurate manner. Ideal scenario for the application of Machine Learning methods. In this work we analyzed data from the Pathfinder miniJ-PAS Survey, which observed $1 deg^2$ over the AEGIS field with 56 narrowband filters and 4 *ugri* broadband filters. Here, we will discuss the classification of miniJPAS sources into point and extended objects, a necessary step for subsequent scientific studies. Our goal is to develop an ML classifier complementary to traditional tools based on other models. In particular, our goal is to build a value-added catalog with our best classifications. To train and test our classifiers, we cross-check the miniJPAS data set with the SDSS and HSC-SSP data, whose classification we assume is reliable within the $15 < r < 21$ and $18.5 < r < 23.5$ ranges, respectively. We trained and tested 6 different ML algorithms in the two cross-referenced catalogs: K-neighbor (KNN), decision trees (DT), random forest (RF), artificial neural nets (RNA), extremely randomized trees (ERT) and classification ensemble (EC). As input for the ML algorithms, we use the magnitudes of the 60 filters, with and without morphological parameters. We concluded that, according to the SDSS classification, the EC algorithm presents better performance, obtaining $AUC = 0.9992$ (area under the ROC curve) and $MSE = 0.009$ (mean square error). By working with weaker magnitudes using the HSC-SSP rating, the EC achieves the best performance, obtaining $AUC = 0.9744$ and $MSE = 0.0370$. The latest results are obtained using photometric bands along with morphological parameters. ML algorithms can compete with traditional star-galaxy classifiers, potentially outperforming the latter in weaker magnitudes ($r \gtrsim 21$). Finally we built a catalog for the $15 \leq r \leq 23.5$ range using machines trained from the merger of labels between the SDSS and HSC-SSP surveys.

Keywords: Stars classification/Galaxies - Machine learning - Data analysis - Photometry.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Uma etapa importante na análise dos dados da estrutura em grande escala é a classificação de fontes luminosas em estrelas ou galáxias. Essa separação é imprescindível para muitas áreas desde a cosmologia até a astrofísica. Na prática, esses objetos são muito semelhantes em certos limites, e tem sido uma tarefa desafiadora desde o século XVIII com o trabalho pioneiro de Messier [4]. Diferentes modelos de classificação foram propostos na literatura, mas todos com seus prós e contras.

Um dos métodos amplamente utilizado nas pesquisas é a separação morfológica. Esses utilizam parâmetros relativos à estrutura do objeto em função da magnitude de uma banda fotométrica. Nesses métodos assumimos que as estrelas aparecem como fontes pontuais, enquanto as galáxias como fontes estendidas. Fato que tem sido mostrado consitente com as observações espectroscópicas [5] [6] [7]. À medida em que a caminhamos para magnitudes mais fracas essa diferença estrutural pontual e estendida diminui ao ponto de ambas serem confundidas, tornando-se um método pouco preciso. Para ilustrar essa dificuldade podemos observar a figura 1.1., que explora a questão com imagens do survey miniJPAS. Notemos que para magnitudes a partir de $rSDSS \sim 21$ começamos encontrar dificuldades para se separar esses objetos, região justamente as quais se encontra a maior parte dos dados do miniJPAS, como pode ser observado na figura 1.2. Esta figura nos dá um panorama de como os dados preliminares estão distribuídos com relação a magnitude da banda *rSDSS*. Os primeiros esforço em separação de estrelas-galáxias, utilizando métodos automatizados, se deram na década de 1970/80 após a digitalização de imagens. Dentre eles estão Macgillivray et al. (1976) [8] e Heydon-Dumbleton et al.(1989) [9].

Levantamentos fotométricos atuais e futuros como Dark Universe Survey (DES) [10], Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) [11], Sloan Digital Sky Survey (SDSS), [12] e o Large Synoptic Survey Telescope (LSST) [13] detectarão um grande número de objetos acima dessa faixa de difícil separação, i.e., $rSDSS \sim 21$. Portanto devemos pensar em alguma solução para amenizar este tipo de problema. Além do mais teremos a questão da análise da enorme quantidade de dados nunca antes produzida no que diz respeito a taxa e volume. Em números, a taxa chega à zeta byte/ano (10^{21}) como

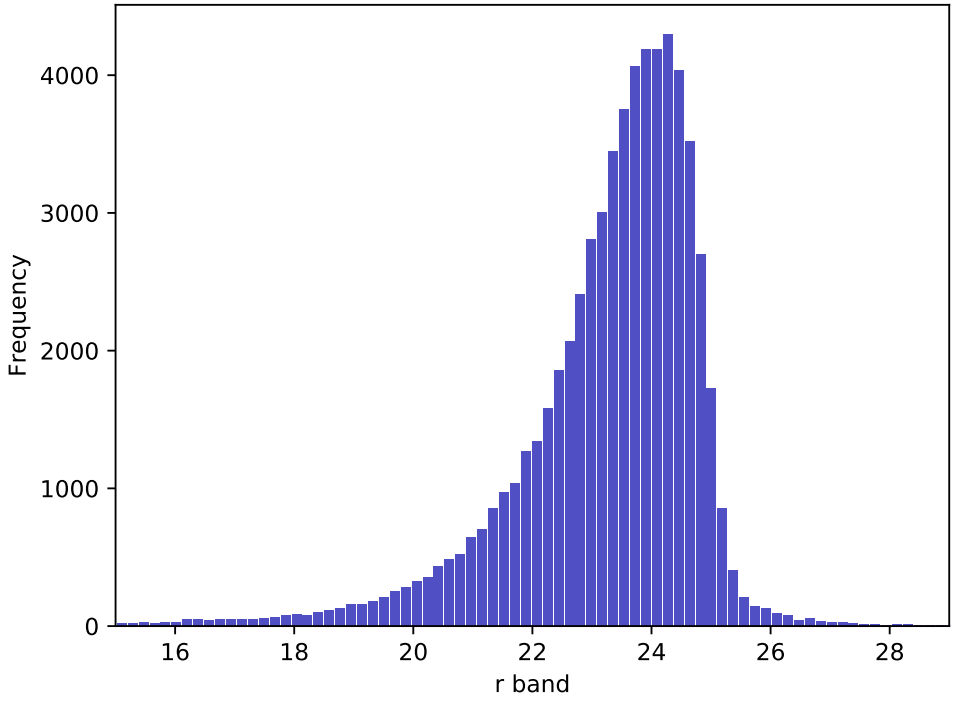

Figura 1.1. Distribuição dos dados do miniJPAS em função da magnitude na banda *rSDSS*

no survey Square Kilometre Array (SKA) [14], sendo imprescindível o desenvolvimento de novos métodos com respostas rápidas e com boa acurácia para o tratamento.

Recentemente tem-se empregado algoritmos de Machine Learning (ML) com o intuito de se contornar essas barreiras. Além de classificarem de forma rápida, podemos inserir várias características extras como morfológicas ou fotométricas para aumentar a acurácia do código separador. Essa classificação se faz com certa facilidade desde que os conjunto de dados de treinamento do algoritmo seja de boa qualidade. A introdução dos CCDs (Charge-Coupled Devices) nas observações nos proporcionou essa qualidade. Portanto, treinando o algoritmo em surveys, com "rótulos"seguros"oriundo de espectroscópia ou fotometria propriamente, podemos extender os limites das classificações de surveys fotométricos com menor resolução.

Nosso objetivo é analisar a maior quantidade de dados do miniJPAS e separá-los entre estrelas e galáxias com uma alta precisão. Para tal finalidades utilizaremos métodos de Machine Learning. Os algoritmos implementados neste trabalho são do tipo supervisionados, i.e, precisam de "rótulos"(objetos já classificados por algum outro survey com alta confiabilidade) para o processo de aprendizagem. Como entrada desses algoritmos, teremos magnitudes associadas a 60 bandas fotométricas juntamente a 4 parâmetros morfológicos. Compondo essas bandas, temos 4 bandas "largas"e 56 bandas "estreitas". Como saída, temos valores que variam entre 0 e 1 associado probabilidade de pertencer à classe de estrelas ou galáxias. Esses dados foram coletados a partir da observação de uma área de $\sim 1\ deg^2$ pro-

Input :	60 bandas fotométricas (Magnitudes) + 4 parâmetros morfológicos.
Output:	Probabilidade de pertencer a uma classe (Estrela ou Galáxia).

Tabela 1.1. Entrada e saída dos algoritmos de Machine Learning.

fundo no campo do All-Wavelength Extended Groth Strip International Survey (AEGIS) [15].

A análise foi dividida em duas etapas utilizando diferentes bins para a banda fotométrica *rSDSS*. Na primeira utilizamos rótulos do survey SDSS dentro da faixa $15 < r < 21$ e na segunda etapa utilizamos rótulos do Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) para a faixa $18.5 < r < 23.5$. Dentro das duas etapas ainda subdivimos o processo. Num primeiro instante utilizamos unicamente bandas fotométricas como entrada e num segundo momento bandas fotométricas juntamente aos parâmetros morfológicos. Nosso objetivo com essa subdivisão é analisar o quão importante a morfologia e a fotometria se fazem no processo de forma independentes. Além do mais, a classificação de Quasi-Stellar Object (QSO) como objetos extra-galácticos é um pouco melhor usando apenas bandas fotométricas [16].

Fizemos uma comparação de diferentes modelos de ML entre si. Também os comparamos à dois classificadores adotados pelo miniJPAS, são eles: *CLASS_STAR* fornecido pelo software (SExtractor) [17] e um classificador baseado em análise bayesian Stellar-Galaxy Loci Classifier (SGLC) [18]. No contexto de ML aplicamos aos dados os métodos de Random Forest (RF), Decision Trees (DT), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN) e um Ensemble Learning (Ensemble) baseado na união de três algoritmos.

Como resultado, mostramos que os parâmetros morfológicos aumentam consideravelmente a acurácia do separador quando combinados com as bandas fotométricas. Para magnitudes onde $15 < rSDSS < 21$, observamos pouca diferença entre a performace de SExctractor, SGLC e os algoritmos de ML. Entretanto, quando caminhamos para magnitudes mais fracas essa diferença aumenta consideravelmente. Para se medir a performace dos classificadores utilizamos as curvas ROC (Receiver Operating Characteristic) e Completeza x Pureza. Utilizando o algoritmo RF mostramos que os parâmetros morfológicos são mais importantes que as bandas fotométricas. Também traçamos as curvas Locus-Estelar para nos dar uma idéia do quão eficaz é o modelo.

Propostas de aplicações de ML à separação estrelas-galáxias tem sido realizada com sucesso a muitos surveys. Vasconcellos et al. 2011 [19] por exemplo utiliza diferentes métodos de árvores para se classificar as fontes do SDSS. Em Kim et al. [20] como outro exemplo, utiliza classificadores baseados em métodos que misturam métodos supervisionados e não supervisionados em ML à dados do CFHTLenS. Recentemente tem se introduzido Redes Neurais Convolucionais (CNN) aos quais se utilizam apenas imagens como em Kim et al 2016 [21] conseguindo obter uma $AUC > 0.99$ para dados do CFHTLenS e SDSS. Para mais aplicações em ML na classificação de estrelas e galáxias podemos consultar [16] [22] [23] [24].

Além de aplicações em separação de estrelas e galáxias, o ML tem sido aplicado amplamente do contexto da Cosmologia/Astrofísica como é o caso do classificação fotométrica de supernovas [25] [26], a análise de ondas gravitacionais [27], [28], a predição de redshift fotométrico [29] [30], morfologia de galáxias [31] [32]. Para mais aplicações dados observacionais em cosmologia/astrofísica consulte o livro escrito por Ivezic [33].

A estrutura desta tese está dividida como seguinte. No capítulo 2, falaremos brevemente sobre os modelos de classificadores aplicados atualmente na separação de estrelas e galáxias. No capítulo 3, descrevemos sobre o ML e seus diferentes tipos de algoritmos. No capítulo 4, descreveremos os dados do miniJPAS e como foram obtidos. No capítulo 5, mostraremos os resultados e no capítulo 6, finalizo com as conclusões e perpectivas futuras. Os resultados dessa tese podem ser encontrados no paper Baqui et al. (2020) [34].

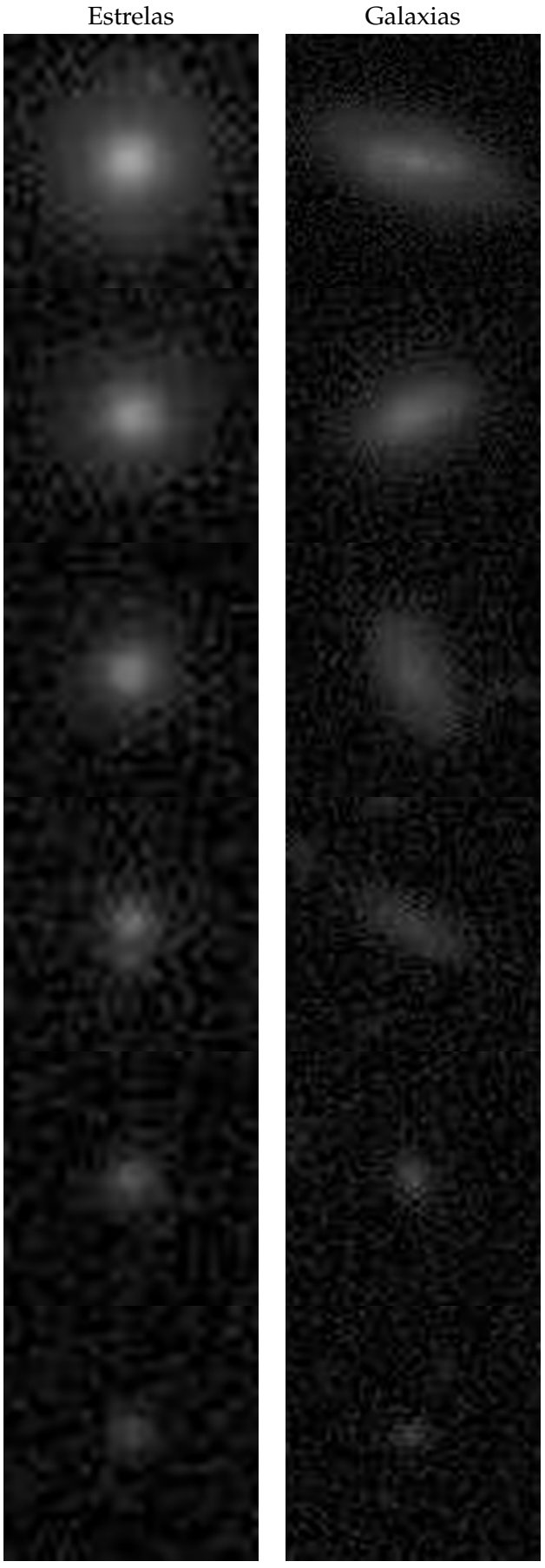

Figure 1.2: Conjunto de dados de estrelas (em cima) e galáxias (em baixo) para dados do J-PAS survey em diferentes magnitudes (rSDSS). À medida em que caminhamos para magnitudes mais fracas a dificuldade de se classificar fontes aumeta, justamente onde estão a maioria dos dados observados.

Capítulo 2

Tipos de Classificadores

Abordaremos brevemente aqui sobre os diferentes tipos de classificadores utilizados no processo de separação de fontes luminosas do survey miniJPAS. De maneira mais específica falaremos brevemente sobre as classificações do tipo Morfológicas, a classificação fornecidas por *CLASS_STAR* e uma classificação que utiliza Análise Bayesiana.

2.1 Classificadores Morfológicos

A classificação morfológica é um dos modelos mais simples aplicados para a separação estrelas e galáxias. O modelo consiste em se traçar um "corte duro" no espaço dos parâmetros das fontes luminosas. Podemos exemplificar o modelo a partir da figura 2.1 no qual aplicamos o corte. Nela calculamos a diferença entre duas formas distintas de se medir magnitudes: Mag_{PSF} e Mag_{cmodel}, em função de Mag_{cmodel} observadas pela banda r para fontes do survey HSC-SSP.

Podemos observar uma diferença entre dois conjuntos de objetos e uma linha vermelha os separam no eixo $y = 0.015$. A linha vermelha é o que chamamos de classificação morfológica: um simples corte que divide dois grandes grupos de objetos: extendidos e compactos. Objetos extendidos acima e compactos abaixo da linha vermelha.

A desvantagem desse modelo é que ele fornece uma classificação absoluta para um cenário onde a incerteza aumenta à medida em que caminhamos em direção a magnitudes mais fracas. Note na figura 2.1 que a partir de $Mag_{cmodel} \sim 24.5$ a sepação entre ambos conjuntos por esse modelo não é tão confiável.

2.2 CLASS_STAR

O *CLASS_STAR* é um classificador de objetos pertencente a um software com diversas funções chamado SExtractor (Source Extractor) [17]. O SExtractor foi desenvolvido para processamento de grande imagens (60 000 x 60 000 pixels) produzidas, por exemplo, por câmeras de CCD. Tem sido aplicadas amplamente em surveys fotométricos incluindo o miniJPAS. Além de detectar fontes e fazer um tratamento sobre as imagens, o SExtractor fornece

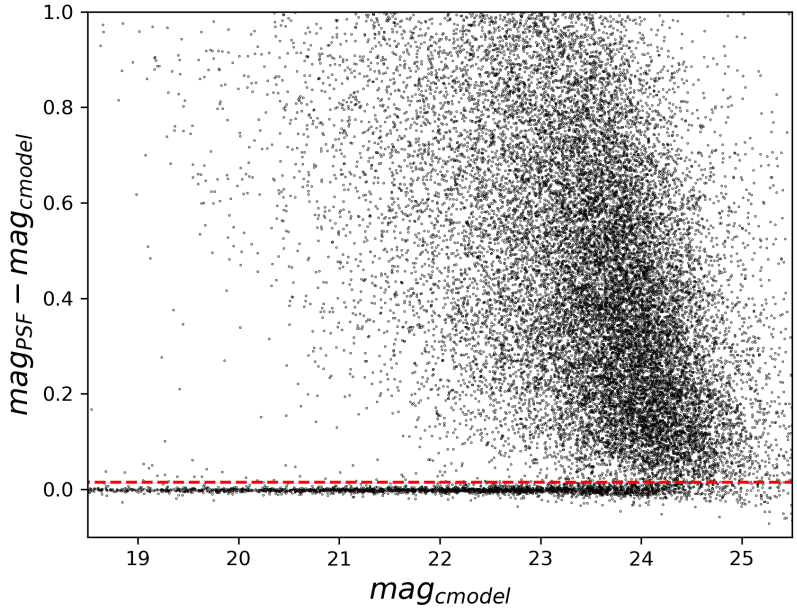

Figura 2.1. Diferença entre as magnitudes Mag_{PSF} e Mag_{cmodel} em função de Mag_{cmodel} observados pela banda fotométrica r_{cmodel} para dados do survey HSC-SSP. O corte em vermelho no espaço dos dividindo dois subgrupos de pontos representa uma classificação morfológica. Abaixo da linha temos objetos classificados como estrelas e acima como galáxias.

os fluxos e as magnitudes associados a cada banda fotométrica do survey, além de gerar uma classificação para cada objeto em estrela ou galáxia. O software possui dois algoritmos internos classificadores, o *CLASS_STAR* e *SPREAD_MODEL*. O método classificador aplicado sobre as imagens do miniJPAS escolhido pela colaboração foi *CLASS_STAR* e é construído a partir de Redes Neurais.

As Redes Neurais são algoritmos de Machine Learning compostos por elementos chamados *perceptron*. Inspirados em neurônios biológicos, esses elementos quando conectados conseguem extrair padrões de um conjunto de dados nos permitindo classificar objetos. As redes neurais são compostas por camadas de entrada, camadas escondidas e uma camada de saída. O processo de aprendizagem se dá através de um método chamado *back-propagation*.

A rede é treinada a partir de 600 imagens contendo estrelas e galáxias com 512x512 pixels geradas utilizando a banda azul B que varia em magnitude entre 10 e 27. O número de galáxias e estrelas na simulação é o mesmo. O teste é realizado com dados simulados e dados reais para diferentes surveys com diferentes profundidades. O resultado deste processo de aprendizagem em classificação é confiável para magnitudes até $r \sim 21$ [17]. Falaremos um pouco mais sobre as redes neurais mais à frente no tópico em que abordamos os métodos de ML.

2.3 Análise Bayesiana: Classificador Locus Estelar/Galaxia

O miniJPAS inclui o classificador bayesiano Stellar-Galaxy Loci Classifier (SGLC) ou Classificador Locus Estelar/Galaxia desenvolvido por López-Sanjuan et al. (2019) [18] para dados survey Javalambre Photometric Local Universe Survey (J-PLUS). O diagrama de concentração versus magnitude apresenta uma distribuição bimodal, correspondendo a objetos compactos e fontes extendidas. Esse método modela ambas distribuições para obter a probabilidade de cada fonte em ambas opções. O modelo com priors adequados é então usado para estimar a probabilidade bayesiana de uma fonte pertencer a classe das estrelas ou das galáxias. Também neste caso, espera-se que os quasares sejam classificados como "estrelas"uma vez que os objetos em sua essência são classificados em estendidos e compactos. Este método foi atualizado para os dados do miniJPAS, veja Bonoli et al. (2020) [2] para mais detalhes.

Capítulo 3

Machine Learning

Para melhor entendimento do trabalho daremos uma breve idéia do que é o Machine Learning e seus diferentes algoritmos. Em particular, focaremos nos tipos supervisionados no contexto de classificação. Apresentaremos as Redes Neurais assim como algoritmos baseados em árvores. Discutiremos como se encontrar as características mais importantes do sistema, utilizando esses últimos. Introduziremos algumas métricas que nos permitem avaliar o quão os modelos se ajustam aos dados. Em seguida apresentaremos alguns conceitos como *underfitting* and *overfitting*, os quais o algoritmo pode vir a sofrer. Comentaremos brevemente sobre a importância de sua introdução não só em na ciência, mas como também na indústria.

3.1 General Overview

O Machine Learning é um ramo da inteligência artificial definido como um processo automatizado, que consegue extrair padrões a partir de dados. Essa técnica habilita os computadores a aprenderem uma tarefa sem precisar serem explicitamente programados. Uma máquina, por exemplo, com esses algoritmos pode jogar melhor um jogo de damas do que quem o programou (Arthur Samuel 1959) [35]. O machine learning é empregado em uma variedade de tarefas de computação, onde a programação de algoritmos são difíceis ou inviáveis devido à complexidad dinâmica do sistema.

Estruturalmente, temos uma máquina a qual inserimos uma parte do conjunto de dados. O objetivo é ajustar os parâmetros internos do algoritmo. Após esse processo chamado de treinamento, inserimos o conjunto restante não vistos pela máquina com o intúito de se medir a acurácia do algoritmo, comparando valores preditos com os reais. Esse processo pode ser observado na figura 3.1.

Podemos subdividir o ML em métodos supervisionados e não supervisionados. Nos supervisionados, o algoritmo aprende com dados rotulados, i.e, com entradas e saídas do problema conhecidas. Entretanto esses rótulos nem sempre estarão ao nosso alcance. Para

[1]Figura disponível em: https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13

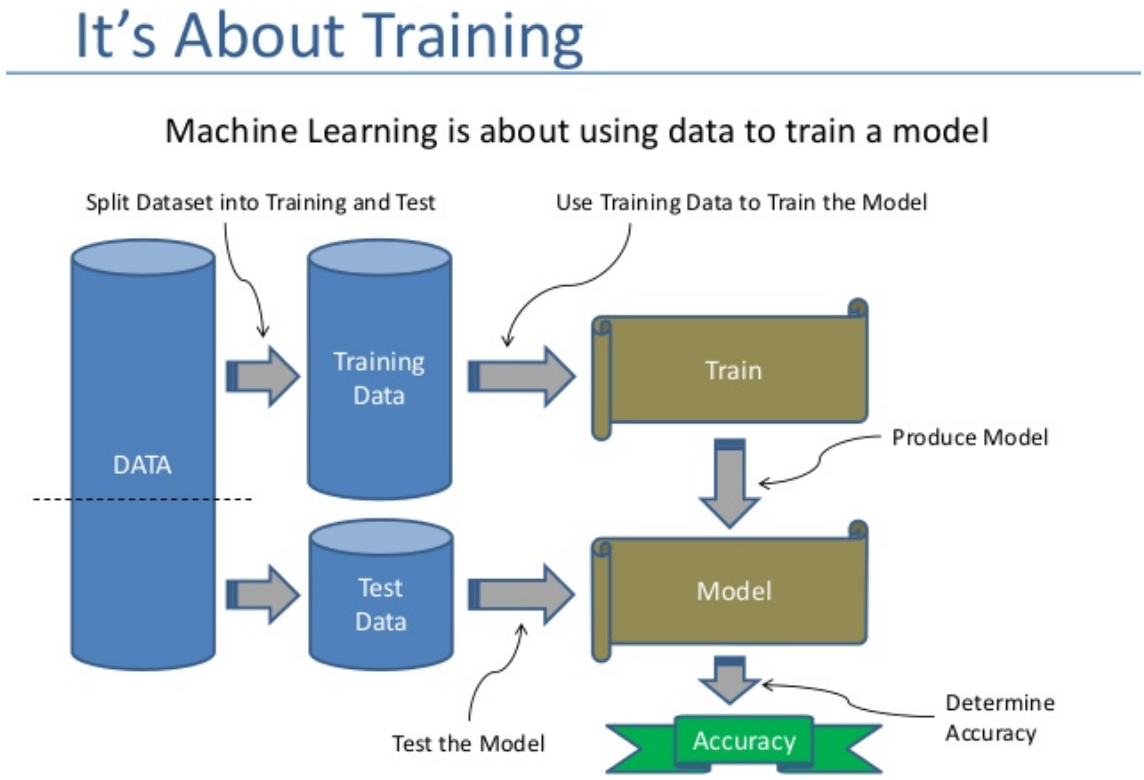

Figura 3.1. Processo de treinamento e teste dos algoritmos de machine learning para algoritmos supervisionados [1].

esses casos utilizamos os métodos não supervisionados, nos quais o algoritmo consegue agrupar por objetos semelhantes, i.e, aprender sem informação externa.

Nesta tese, nos concentramos em métodos binários de classificação supervisionada. Nesse caso, o modelo e seus parâmetros internos são implicitamente ajustados pelo "conjunto de treinamento". Seu desempenho é então testado com a parte restante do conjunto de dados, o "conjunto de testes". Especificamente, os parâmetros internos da função de predição $f : R^n \rightarrow Y$ são treinados através do conjunto de dados de treinamento $x_i \in R^n$ (n é a dimensionalidade do espaço das características, 'i' rotula os elementos do conjunto de treinamento) com classificações $y_i \in [0, 1]$, em que 1 significa galáxia e 0 estrela. Os classificadores são divididos em classificadores não probabilísticos e probabilísticos. O primeiro tipo de classificador gera a melhor classe, enquanto o segundo a probabilidade das classes (a melhor classe é considerada como a com maior probabilidade). Aqui, consideraremos apenas classificadores probabilísticos binários, de modo que $f : R^n \rightarrow [0, 1]$, ou seja, f fornece a probabilidade de que um objeto seja uma galáxia. a probabilidade de uma pertencer à classe das estrela é simplesmente $1 - f$. Um valor de f próximo a 1 significa que o objeto é provavelmente uma galáxia.

Existem vários métodos numéricos de se modelar essa função predição $f : R^n \rightarrow [0, 1]$ linares e não lineares. Neste trabalho investigamos seis métodos. São eles o K-Nearest Neigh-

bors (KNN), a Decision Tree (DT), Random Forest (RF), Extremely Randomized Trees (ERT), uma rede neural simples chamada Multilayer Perceptron (MPL) e um método utilizando Ensemble de Classificadores.

Geralmente separamos valores entre 80% ou 70% dos dados para treinar nossa máquina e obter parâmetros ótimos. O restante separamos para medir sua perfomace. Numericamente, implementamos esses algoritmos com o auxílio do pacote scikit-learning [2] escrito em python [36]. Para mais informações sobre aprendizagem supervisionada o leitor pode consultar [37] [38].

3.2 Algoritmos

A seguir descrevemos um pouco sobre os algoritmos utilizados no trabalho que modelam a função predição $f : R^n \rightarrow [0, 1]$. Todos os algoritmos que descreveremos a seguir possuem sua forma regressora e classificadora, entretanto neste trabalho apenas as formas classificadoras serão aplicadas

- K-Nearest Neighbours (KNN)
- Tree Methods (DT, RF and ERT)
- Neural Network
- Ensemble Method

3.2.1 K-Nearest Neighbors

O algoritmo KNN (K-Nearest-Neighbours) [39] [37] é um dos algoritmos mais simples em Machine Learning. Neste calcula-se a distância física entre o elemento o qual se quer predizer a classe (dados de teste), dos k vizinhos mais próximos pertencente aos dados de treino. A classe predita será calculada a partir do voto majoritário destes k vizinhos. A figura 3.2 ilustra o processo de classificação de um objeto no qual temos 3 votos para a classe "△", um voto para "+"e outro para "o". Pelo voto majoritário temos que a classe atribuída ao objeto "?"será "△".

Podemos esquematizar o algoritmo da seguinte forma:

- Escolhemos o número de k-vizinhos e uma métrica.
- Calculamos a distância entre os k-vizinhos e o os dados que queremos classificar.
- Atribuimos a classe utilizando o voto majoritário.

[2] http://scikit-learn.org
[3] Figura disponível em: https://github.com/rasbt/python-machine-learning-book-3rd-edition

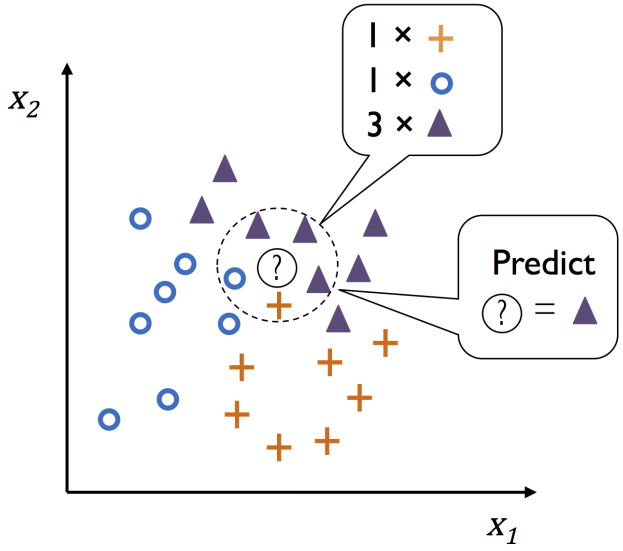

Figura 3.2. Representação do voto majoritário no processo de predição do algoritmo KNN onde '?' representa um elemento do conjunto de teste e os simbolos o, + e △ representando dados de treinamento.[3]

Nesta tese calculamos a distância entre os k vizinhos a métrica euclideana, entretando é possível se escolher outras como a de Minkowsky. Este método é bem rápido e seu custo computacional proporcional ao número de pontos do *training set*.

O modelo descrito utilizando o voto majoritário é discreto. Entretanto é possível extendê-lo ao caso contínuo $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ao qual associamos uma probabilidade a cada classe. Para isso calculamos uma média dos k vizinhos mais próximos associado a cada classe, ao invés de utilizar o valor mais comum.

$$f(x_q = c_j) = \frac{1}{k} \sum_{i=1}^{k} \delta(x_i = c_j) \tag{3.1}$$

Onde $f(x_q = c_j)$ será a probabilidade da fonte x_q pertencer a classe c_j e δ representa uma Delta de Dirac o qual assumirá valor 1 se o elemento x_i pertence à classe c_j e 0 caso contrário.

Podemos ainda melhorar nosso modelo inserindo pesos sob a contribuição de cada vizinho no processo de predição. Para o caso em que o peso é inversamente proporcional à distância quadrada, por exemplo, calculamos a função f como:

$$f(x_q = c_j) = \frac{\sum_{i=1}^{k} w_i \delta(x_i = c_j)}{\sum_{i=1}^{k} w_i} \quad com \quad w_i = \frac{1}{d(x_q, x_i)^2} \tag{3.2}$$

Suponhamos por exemplo que gostaríamos de predizer a classe de um objeto "?"e utilizamos os 5 vizinhos mais próximos para se fazer a predição, como na figura 3.2. Assumindo

que todos os vizinhos possuem o mesmo peso $w_i = 1$, a probabilidade desse objeto "?"pertencer as classes $\triangle$, "+"e "o", utilizando a equação 3.2, será de 3/5, 1/5 e 1/5 respectivamente .

3.2.2 Decision Trees

As Decision Trees [40] [37] são algoritmos que mapeiam possíveis resultados definidos a partir de escolhas. Começam com um nó que se ramifica em várias possibilidades dividindo o espaço das características gerando novos sub conjuntos. Novos nós são gerados a partir desses subconjuntos e assim por diante. Essa configuração de muitos ramos gera uma árvore. Essa estrutura nos permite avaliar ações que podem estar associadas a custo de produção/benefícios de uma empresa, por exemplo. Além do mais, esses mapas nos permitem conduzir discussões sobre quais decisões se tomar para uma determinada tarefa.

Na figura 3.3 temos um exemplo ilustrativo simples do processo de ramificações e decisões numa Decision Tree.

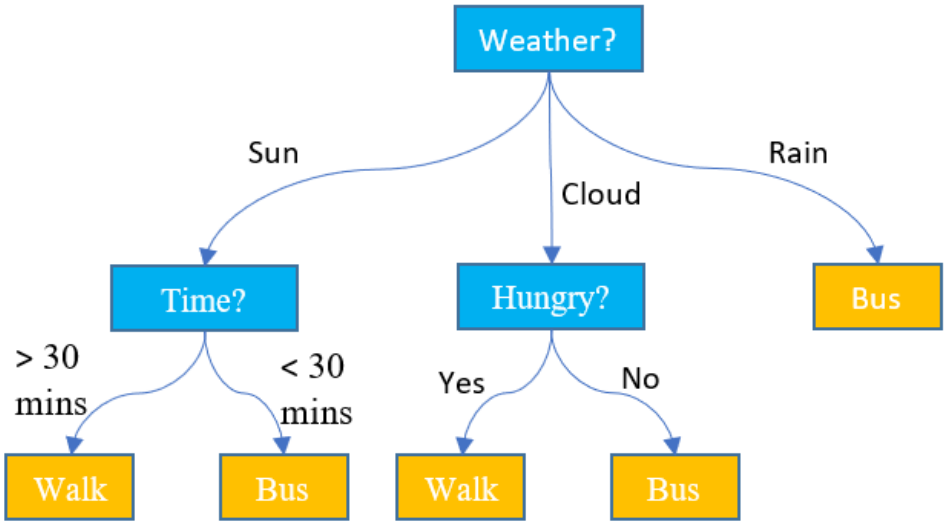

Figura 3.3. Representação do processo de decisão e ramificações de uma Decision Tree. O processo de ramificação nos permitem predizer sobre quais condições um passageiro pegará um onibus ou irá caminhando para o trabalho[4].

A figura 3.4 nos mostra um outro exemplo sobre a tomada de decisões para uma Decision Tree sobre o espaço composto por duas características X_1 e X_2. A cada nó que surge na árvore o sub espaço das características é subdividido. Isso ocorre até um determinado ponto definido pelo desenvolvedor.

[4]Figura disponível em: https://www.displayr.com/what-is-a-decision-tree/
[5]Figura retirada da referência [37].

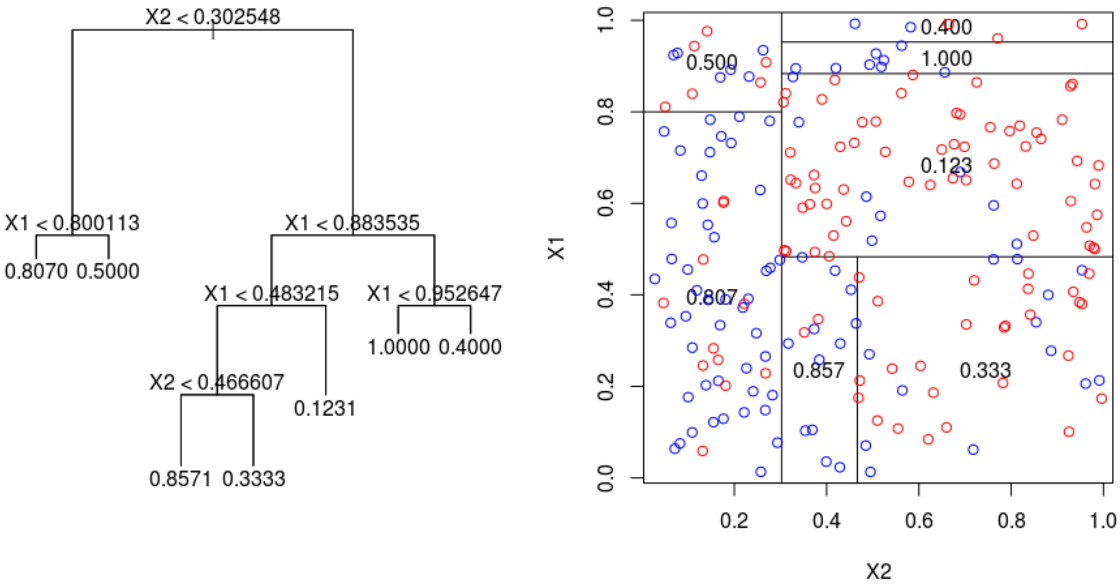

Figura 3.4. À esquerda temos a representação do processo de decisão e ramificações de uma Decision Tree. Após a maximização da função Ganho de Informação (IG) temos um espaço subdivido com probabilidades associadas à classe azul representado à direita [5].

Para a construção de uma Decision Tree devemos primeiramente definir uma função Ganho de Informação (IG):

$$IG(D_p, x_t) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \tag{3.3}$$

Onde D_p é o dataset pai e D_{left}, D_{right} datasets filhos. As grandezas N_{left} e N_{right} são os número de dados dos datasets filhos respectivamente e I uma função chamada impureza. Os conjuntos dos datasets pais e filhos, i.e, divisão do espaço das características em retângulos, são definidos a partir da escolha da característica e threshold x_t que maximizam a função IG. Note que a função Ganho de Informação nada mais é que a subtração das funções Impureza dos nós filhos da função Impureza do nó pai. E que quanto menor a Impureza dos filhos, maior o Ganho de Informação.

Para o processo de classificação existem diferentes funções impureza como a Entropia e Erro de Classificação. Abaixo como exemplo utilizamos a função Impureza Gini (I_G):

$$I_G(m) = 1 - \sum_{i=1}^{class} p(class|m)^2 \tag{3.4}$$

Onde $p(class|m)$ é a proporção dos dados pertencentes a classe para um particular nó m

$$p(class|m) = \frac{1}{N_m} \sum_{x_i \in D_m} \delta(y_i = class) \tag{3.5}$$

Novamente aqui temos uma Delta de Dirac que assumirá valor 1 caso y_i pertença à determinada *classe*. Após essas sequências de divisões no espaço das características, baseado na maximização da função IG, teremos um espaço subdivido com probabilidades associadas a cada classe. Os dados utilizados no processo de crescimento da árvore são pertencentes aos dados de treino. A probabilidade de um dado de teste, pertencer à uma classe, será justamente aquela associadas a região definida pelas coordenadas das características $\mathbb{R}^n$. Para o caso discreto o dados de teste assumirá o valor associado à classe majoritária.

A figura 3.4 mostra o de decisão e ramificações de uma Decision Tree. Após cada maximização da função Ganho de Informação (IG) temos a formação de dois subconjuntos filhos. À direita temos os espaços das características subdivididos para cada processo da maximização da função IG de maneira recursiva. Dentro de cada subconjunto dividido temos as probabilidades de um dado de teste pertencer à classe azul.

3.2.2.1 Importância de Características

Para o processo de ramificação dos algorítmos de Decision Trees escolhe-se a característica juntamente a um ponto de corte x_t (threshold) que maximizam a função IG. Algumas características aparecem mais vezes do que outras nesse processo de ramificação. Com essa frequência podemos medir o quão cada característica foi importante no processo de predição. Definimos a importância de cada característica $Imp(x)$ como:

$$Imp(x) = \sum_t \frac{N_p}{N_{tot}} IG(D_p, x_t) \tag{3.6}$$

Onde N_p é o número de dados do nó pai, N_{tot} é o número total de dados tratado no problema e $IG(D_p, x_t)$ é o valor da função ganho de informação onde a característica x subdividiu um nó. Portanto, quanto maior o número de vezes que a característica X subdivide a árvore, maior será sua importância. Notemos que as primeiras características que subdividem a árvore tenderão a ter maior importância devido ao fator N_p/N_{tot} que decresce conforme a árvore cresce.

3.2.3 Random Forest

O Random Forest (RF) [41] [37] é um algoritmo ensemble construído a partir de um conjunto de árvores (DT), nas quais cada uma se diferencia uma da outra ligeiramente. Esse conjunto de árvores gera uma floresta. Cada árvore pertencente ao ensemble gerará uma classificação particular e a predição do RF será a combinação dos diferentes outputs de cada árvore.

O fato de cada árvore gerar outputs diferentes se dá pela método aleatório os quais as features são encontradas no processo de maximização da função IG. Além do mais, utiliza-se

[6]Figura disponível em: https://medium.com/@williamkoehrsen/random-forest-simple-explanation\-377895a60d2d

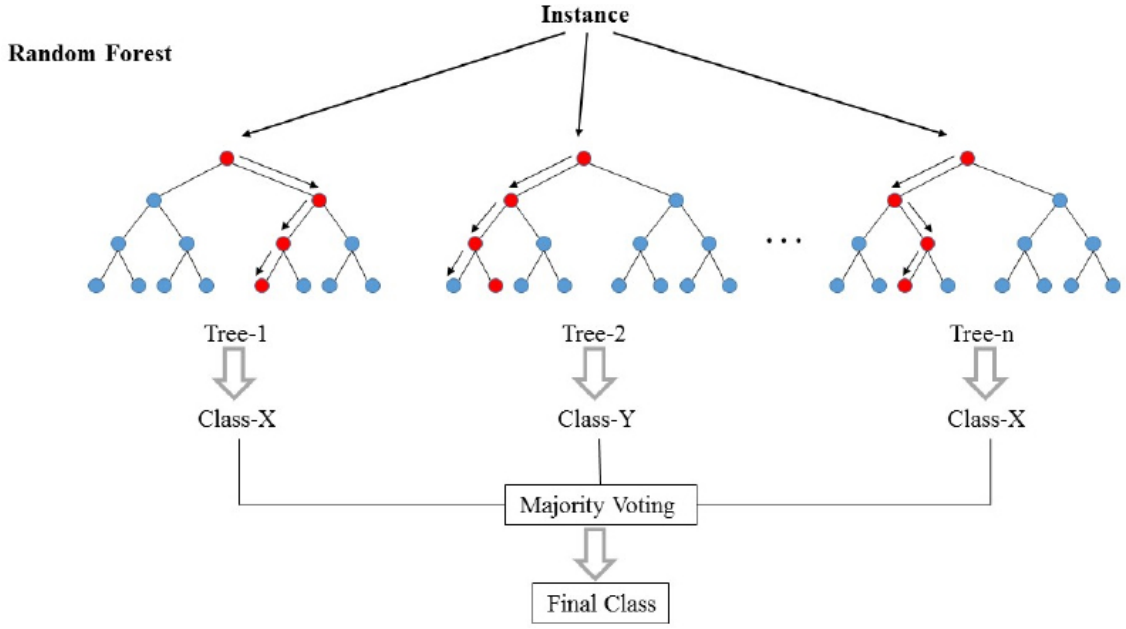

Figura 3.5. Estrutura do algoritmo Random Forest[6].

o método estatístico bootstrap nos quais construímos diferentes datasets a partir do original para diferentes árvores. Podemos esquematizar o algoritmo RF da seguinte forma:

- Escolher o número de árvores.
- Construir diferentes datasets para cada árvore (Bootstrap) e rodar as DT.
- Em cada nó de cada árvore escolher K features de forma aleatória de forma a maximizar a função IG.
- Combinar os resultados de todas árvore a fim gerar um output para a RF

Para o caso discreto o output é construído a partir do voto majoritário como no algoritmo KNN. Para o caso probabilístico calculamos o output da RF como a média das probabilidades de cada classe para cada árvore. Por exemplo:

Sejam C_1, C_2 e C_3 três diferentes árvores com suas respectivas classificações probabilísticas para a classe 1 e 2 de um particular dado $\mathbf{x}$, respectivamente:

$$C_1(\mathbf{x}) \rightarrow [0.9, 0.1], \quad C_2(\mathbf{x}) \rightarrow [0.8, 0.2], \quad C_3(\mathbf{x}) \rightarrow [0.4, 0.6] \tag{3.7}$$

Calculamos a função f em sua forma probabilística como:

$$f(\mathbf{x}|class) = \frac{1}{N}\sum_{j=1}^{m} p_j(i = class|\mathbf{x}) \tag{3.8}$$

Onde $p_j(class|\mathbf{x})$ é a probabilidade predita pela árvore j para a classe i e N o número de árvores. Assim $f(\mathbf{x}|class)$ assumirá os valores:

$$\sum_j p_j(i_0|\mathbf{x}) = (0.9 + 0.8 + 0.4)/3 = 0.7$$
$$\sum_j p_j(i_1|\mathbf{x}) = (0.1 + 0.2 + 0.6)/3 = 0.3 \tag{3.9}$$

Isto é, assumirá uma probabilidade de 0.7 de pertencer à classe 0 e 0.3 de pertencer à classe 1.

3.2.3.1 Feature Importance

O processo de definição da importância das características nos algoritmos de Random Forest é feita de maneira semelhante às Decicion Trees. Para cada árvore pertencente ao ensemble calculamos a importância da característica X e em seguida fazemos uma média sobre todas as árvores. Portanto, em forma matemática, a escrevemos como:

$$Imp(x) = \frac{1}{N_T} \sum_T \sum_t \frac{N_p}{N_{tot}} IG(D_p, x_t) \tag{3.10}$$

Onde N_T é o número de árvores pertencentes ao ensemble. Na figura 3.6 exemplificamos o output no cálculo da importância das características num processo de classificação de estrelas-galáxias-quasares realaizadas no apêndice B. Na ocasião utilizamos 12 bandas fotométricas do survey JPLUS assim como suas combinações e o parâmetro morfológico concentração c_r definido mais adiante no capítulo 4. Esse parâmetro nos diz respeitos sobre o caráter pontual ou extensivo da fonte. Os dados de rótulo para se treinar a máquina foram retirados do SDSS[7].

3.2.4 Extremely Randomized Trees

As Árvores Extremamente Aleatórias (ERT) [42] são um método de ensemble que utilizam (DT) e que são muito semelhantes aos algoritmos de Random Forest. Exitem apenas duas diferenças entre RF e ERT. A primeira é que ERT não utiliza bootstrap embora a implementação no pacote scikit-learn nos permita inseri-lo na análise. A segunda é que enquanto RF tenta encontrar o melhor threshold para uma característica, em ERT a divisão é feita de forma aleatória entre os valores da amostra em cada divisão. Isso faz com que os cortes não sejam ótimos. Podemos esquematizar o algoritmo ERT da seguinte forma:

- Escolher o número de árvores.
- De forma aleatória escolher o número K de característica X que maximizarão a função IG.

[7] Uma análise para separação de estrelas, galáxias e quasares para o survey J-PLUS pode ser encontrados no apêdice C.

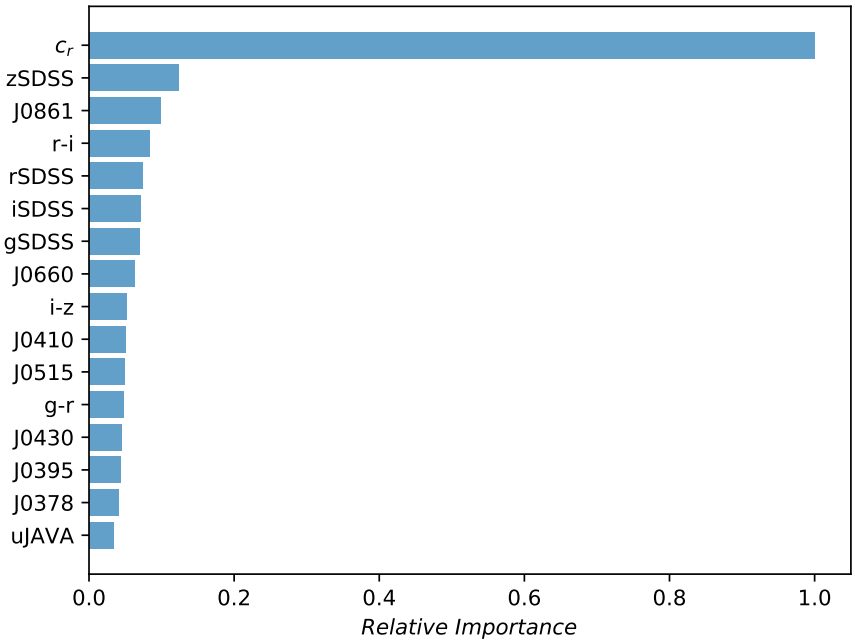

Figura 3.6. Importancia das features no processo de classificação do survey fotométrico JPLUS para $15 < rSDSS < 21$ utilizando como rótulos dados do SDSS. Utilizamos o algoritmo RF neste exemplo.

- Dentro dessa feature calcular seu mínimo e máximo, i.e, X_i^{min} e X_i^{max}.
- Selecionar um ponto de corte aleatório dentro do intervalo $[X_i^{min}, X_i^{max}]$.
- Retornar o corte no intervalo $X_i < X_c$.
- Seleciona o corte s_i tal que maximiza IG a analisando as K características.
- A árvore cresce até um ponto definido pelo desenvolvedor.

Devido a esse método aleatório de escolha de caraterísticas e thresholds, encontraremos diferentes valores como output para cada árvore. Os métodos de árvores em scikit-learning são bastante parecidos estruturalmente. Na tabela 3.1 uma tabela mostrando as diferenças no processo de predição de cada método.

3.2.5 Redes Neurais

As Redes Neurais Artificiais (ANN) consistem em uma técnica computacional que mimetiza o funcionamento do sistema nervoso, conseguindo reconhecer padrões a partir de um conjunto de dados. Devido ao seu sucesso nas predições, as redes neurais ganharam tanta notoriedade que hoje constituem um ramo à parte chamado Deep Learning (DL), dentro do ML. Em Deep Learning existem várias estruturas de algoritmos. O modelo que

x	**Decision Tree**	**Random Forest**	**Extremely Randomized Trees**
Número de árvores	1	Várias	Várias
Número de features consideradas ao se fazer o corte de decisão	Todas	Subconjunto Aleatório de Features	Subconjunto Aleatório de Features
Bootstrapping	Não Aplicável	Sim	Não
Método de corte	Melhor Corte	Melhor Corte	Corte Aleatório

Tabela 3.1. Diferença entre os métodos de árvores DT, RF e ERT.

iremos utilizar em nossas análises consiste em um modelo supervisionado simples chamado Multilayer Perceptron Layer (MPL), algoritmo composto por vários neurônios. Mas antes de apresentá-lo, explicaremos o Perceptron, um algoritmo composto por apenas um neurônio.

Perceptron

O Perceptron [43] é um algoritmo composto por apenas um neurônio. Sua simplicidade nos permite entender de maneira mais clara a matemática envolvida por trás das redes neurais, consistindo em uma excelente introdução. Comecemos defindo uma função ativação $\phi(z)$ que possui como argumento a combinação de inputs x_i (features) [8] e seus pesos associados w_i:

$$z = w_0 x_0 + w_1 x_1 + \ldots + w_n x_n \tag{3.11}$$

Como função ativação utilizamos uma simples função degrau nos quais assumirá valores:

$$\phi(z) = \begin{cases} 1 & \text{if } z \geq \theta \text{ (threshold)} \\ -1 & \text{caso contrário} \end{cases} \tag{3.12}$$

Onde w_i assume valores aleatórios inicialmente que são atualizados no processo de aprendizagem. O algoritmo funciona da seguinte forma:

- Inicializar os pesos de forma aleatória.
- Para cada dado de treinamento calcular:
  - Output $\hat{y} = \phi(z)$.

[8] As features são características de um determinado problema como a classificação de flores Iris, por exemplo, nos quais utilizamos a largura e comprimento da sépala da flor. Como essas informações conseguimos sub classifica-la e setosa, virginica e versicolor.

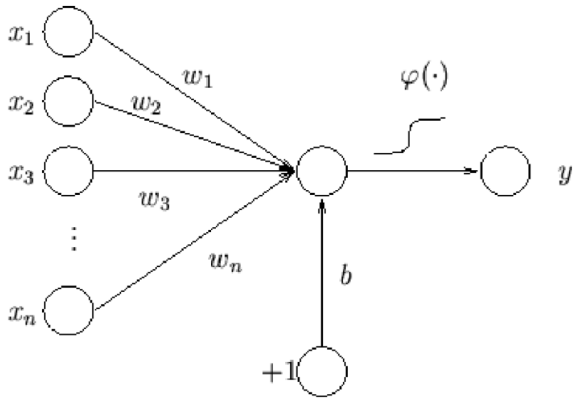

Figura 3.7. Estrutura de um perceptron com x_n entradas aos qual um bias b é adicionado e em seguida aplicamos uma função ativação $\phi(z)$ [1].

– Atualizar os pesos.

Os pesos associados a cada feature x_j são atualizados simultaneamente conforme a relação:

$$w_j = w_j + \Delta w_j \tag{3.13}$$

Onde Δw_j é calculado segundo a relação:

$$\Delta w_j = \eta(y^i - \hat{y}^i)x_j^i \tag{3.14}$$

onde η é a taxa de aprendizagem definido pelo programador que pode assumir valores entre $[0, 1]$, y^i e $\hat{y}^i$ os valores reais e preditos respectivamente para o dado de treino i. Note que quanto maior a quantidade de dados, maior o número de atualizações dos pesos Δw_j, portanto mais preciso tende a ser nosso algoritmo. O Perceptron é um algoritmo que consegue separar casos lineares. Para casos não lineares, devemos trabalhar com modelos mais complexos com mais de um neurônio organizados em camadas como o MPL.

Na figura 3.7 podemos observar a estrutura do perceptron. Somamos as informações $w_k x_k$ a um bias b que inseridas na função ativação gerará um output $\hat{y}$ para cada valor do dado de treino. Atualizamos os pesos Δw_j até um momento de convergência do algoritmo pré-definido.

Multilayer Perceptron Layer

O Multilayer Perceptron Layer (MPL) é composto por um conjunto de perceptrons organizado em diferentes camadas. São elas as camadas de entrada, onde se inserem as features do problema estudado; As camadas escondidas, onde ocorre o processo de aprendizagem; E a camada de saída onde temos a classificação do objeto. A informação na região das camadas escondidas é passada por cada perceptron por várias vezes até o momento da

convergência do algoritmo. Em um MPL podemos ter várias camadas contendo centenas de perceptrons. Na figura 3.8 podemos ver uma estrutura de rede neural simples com quatro camadas: uma de entrada, duas escondidas e uma de saída com dois perceptrons.

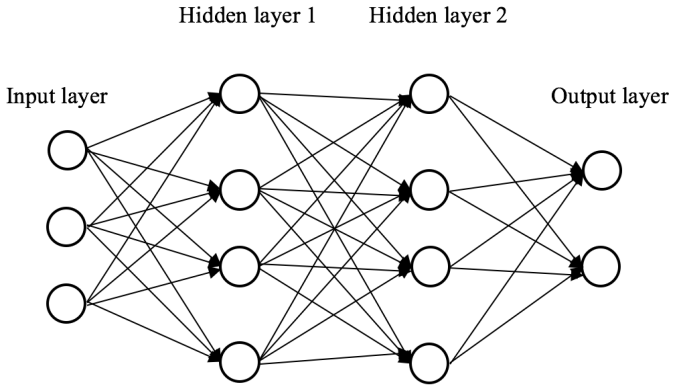

Figura 3.8. Rede Neural (MPL) composta por quatro camadas com duas escondidas e uma de saída.

A rede acima a qual trabalhamos é totalmente conectada. A informação que cada neurônio recebe vem da camada anterior. De forma que a propagação é feita de forma semalhante ao perceptron segundo a relação abaixo.
Para as camadas escondidas (hidden):

$$h_k^{(1)} = \phi^{(1)}\left(\sum_j w_{kj}^{(1)} x_j + b_k^{(1)}\right) \quad e \quad h_k^{(2)} = \phi^{(2)}\left(\sum_j w_{kj}^{(2)} h_j^{(1)} + b_k^{(2)}\right) \tag{3.15}$$

Onde $h_k^{(1)}$ e $h_k^{(2)}$ são as informações que chegam na primeira e segunda camadas escondidas respectivamente, associados ao k neurônios. Em ambas camadas aplica-se a função ativação $\phi^{(1)}$ e $\phi^{(2)}$ de maneira semelhante ao perceptron. A primeira camada do algoritmo receberá informações diretamente das features que serão multiplicadas pelos pesos $w_{kj}^{(1)}$ e que em seguida aplicada a função ativação. A segunda camada receberá informações processadas da primeira camada. Esta aplicará a função ativação sobre os outputs $h_k^{(2)}$ multiplicados pelos pesos $w_{kj}^{(2)}$. Em ambos processos adicionamos um viés para cada camada b_k^1 e b_k^2 respectivamente.

O processo de transmissão de informações de camada para camada são semelhantes. Para a camada de saída temos a idéia:

$$y_k = \phi^{(3)}\left(\sum_j w_{kj}^{(3)} h_j^{(2)} + b_k^{(3)}\right) \tag{3.16}$$

Onde aplicamos uma função ativação $\phi^{(3)}$ sobre os outputs da segunda camada multiplicados pelos pesos $w_{kj}^{(3)}$. Para esse processo também adicionamos um bias $b_k^{(3)}$. Esse processo de propagação de informação da camada de entrada até a camada de saída chamamos

feedforward.

Os pesos da rede são atualizados segundo a relação:

$$w_{kj}^{(l)} = w_{kj}^{(l)} - \eta \nabla J(w_{kj}^{(l)}) \tag{3.17}$$

Onde $\nabla J(w_{kj}^{(l)})$ é o gradiente da função custo que deve ser minimizada. Esse método é conhecido do *gradiente descendente* e o processo de minimação do erro no sentido da camada de saída até a camada de entrada é chamado *backpropagation*. Para mais informações sobre Redes Neurais consulte [37].

3.2.6 Métodos de Ensemble

Os métodos de ensemble tem como objetivo construir um meta-classificador a partir da união de outros algoritmos classificadores. Geralmente quando bem combinados esses classificadores conseguem fazer predições melhores do que utilizando únicamente um algoritmo. Seu método de combinação de classificadores é semelhante ao descrito pelo Random Forest. A diferença é que temos a liberdade de associar pesos a cada classificador. Inserindo esses pesos, nossa função predição f probabilítica será escrita como:

$$f(\mathbf{x}|class) = \sum_{j=1}^{m} w_j p_j(i = class|\mathbf{x}) \tag{3.18}$$

Onde $p_j(class|\mathbf{x})$ é a probabilidade predita pelo classificador C_j para a classe i. E w_j o peso associado a cada classificador. Por exemplo:

Sejam C_1, C_2 e C_3 três diferentes algoritmos com suas respectivas classificações probabilísticas para a classe 1 e 2:

$$C_1(\mathbf{x}) \rightarrow [0.9, 0.1], \quad C_2(\mathbf{x}) \rightarrow [0.8, 0.2], \quad C_3(\mathbf{x}) \rightarrow [0.4, 0.6] \tag{3.19}$$

e $\mathbf{w}$ um peso associado a cada classificador (arbitrariamente escolhido nesse caso):

$$\mathbf{w} = [0.2, 0.2, 0.6] \quad (weights) \tag{3.20}$$

A função predição assumirá o valor máximo das relações abaixo:

$$\begin{aligned} \sum_j p_j(i_0|\mathbf{x}) &= 0.2 \times 0.9 + 0.2 \times 0.8 + 0.6 \times 0.4 = 0.58 \\ \sum_j p_j(i_1|\mathbf{x}) &= 0.2 \times 0.1 + 0.2 \times 0.2 + 0.6 \times 0.6 = 0.42 \end{aligned} \tag{3.21}$$

Isto é, assumirá uma probabilidade de 0.58 de pertencer à classe 0 e 0.42 de pertencer à classe 1.

3.3 Performance

Para se medir o quanto os algoritmos descritos acima "aprendem"com os dados e saber se nossos modelos são enviesados, i.e. ter uma noção sobre underfitting/overfitting ou precisão/acurácia devemos introduzir algumas grandezas.

3.3.1 Métricas

Para se construir algumas métricas em classificadores probabilísticos devemos definir o conceito de matriz de confusão. Essa matriz nos permite visualizar a performace do algoritmo com relação aos acerto e erros das predições comparados aos dados reais. A matriz de confusão é construída a partir de quatro quantidades. Essas quantidades são definidas a partir da comparação do resultado dos valores preditos pelos algoritmos e valores rotulados (i.e, valores assumidos como verdadeiros) para um determinado *threshold*. Definimos as quatro quantidades como:

- TP (True Positive) : Objetos preditos como positivos e rotulados como verdadeiros.
- TN (True Negative) : Objetos preditos como negativos e rotulados como negativos.
- FP (False Positive) : Objetos preditos como positivos e rotulados como negativos.
- FN (False Negative) : Objetos preditos como negativos e rotulados como verdadeiros.

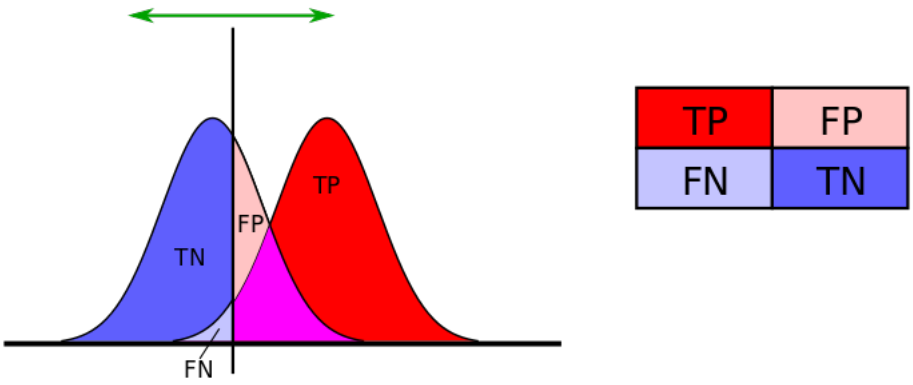

Figura 3.9. Posição da grandezas TP, TN, FP, FN na distribuição das duas classes de objetos estudados[9].

Em nosso trabalho tomamos os valores positivos como galáxias e assumindo os valores 1. Enquanto que as estrelas assumem os valores 0.

A figura 3.9 mostra a posição de cada definição de uma matrix de confusão assumida na distribuição de cada classe. A linha vertical que corta ambas distribuições representa o threshold, quantidade que nos permite separar as grandezas TN,TP,FN e FP. Na diagonal

[9]Figura disponível em: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

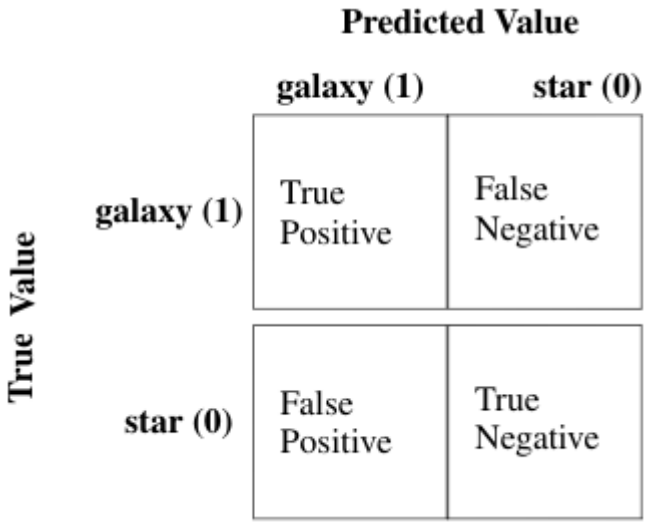

Figura 3.10. Matriz de confusão no qual assumimos valores 1 para galáxias enquanto que 0 para estrelas. A linha vertical que corta ambas distribuições representa o threshold.

principal temos as classificações corretas TN e TP. Enquanto que na diagonal secundária demos as classificações incorretas FP e FN. Os rótulos mencionados acima devem ser de origem segura. Para o caso, por exemplo, de algoritmos de ML construído para se classificar câncer como maligno ou benigno poderíamos utilizar resultados de biópicias. Para classificação de estrelas galáxias, como outro exemplo, poderíamos utilizar rótulos espectroscópicos. A figura 3.10 também representa a matriz de confusão. Nos permitindo comparar resultados preditos com "verdadeiros".

3.3.2 Característica de Operação do Receptor

Uma vez definida a matriz de confusão podemos introduzir a Curva Característica de Operação do Receptor ou simplesmente Curva ROC. Essa curva é uma representação gráfica para um classificador binário obtida a partir da Taxa de Verdadeiros Positivos (TPR) e Taxa de Falso Positivos (TFP) definidas como:

$$TPR = \frac{TP}{TP + FN} \quad \text{e} \quad FPR = \frac{FP}{FP + TN} \quad . \tag{3.22}$$

Para cada *threshold* teremos diferentes valores para as quantidades TP, TF, FN e FP e por consequência diferentes valores para TPR e TFP. A Cuva ROC é basicamente uma curva que representa a variação de TPR e TFP conforme variamos o *threshold* entre 0 e 1.

A partir da Curva ROC definimos outra quantidade chamada a AUC (área sobre a Curva ROC). Com a AUC podemos medir a performace do classificador assim como compará-los a outros métodos classificadores. A figura 3.11 representa duas curvas Curva ROC para o caso de classificadores binário. Note a variação das taxas TPR e TFP à medida em que variamos o threshold. Os valores do threshold variam entre zero e um e estão em

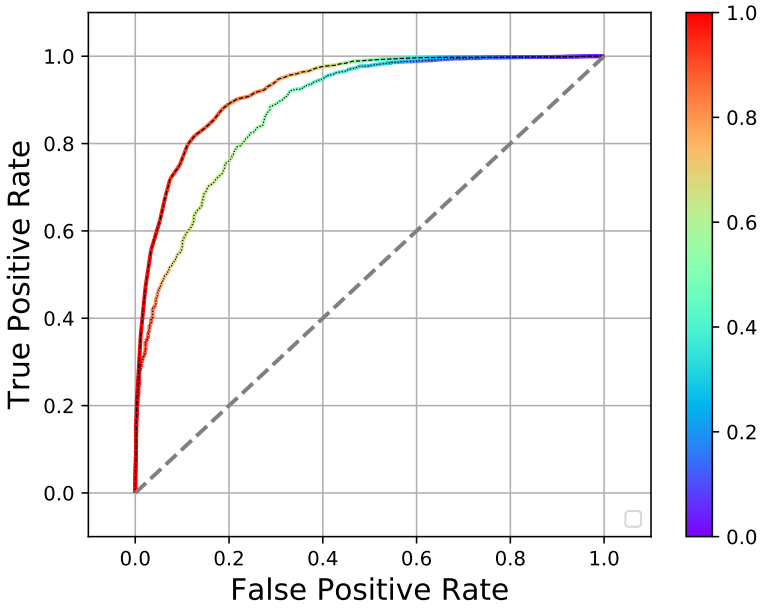

Figura 3.11. Cuva ROC para um classificador binário. As cores indicam os diferentes thresholds as quais foram calculadas as Taxa de Vedadeiro Positivo e Falso Positivo.

cores representados por uma barra à direta.

A AUC pode assumir valores entre 0 e 1. Um classificador perfeito possuirá valor igual à 1. Isso significa que possuirá a taxa de Falsos Negativos FN e Falso Positivo FP igual à zero. Podemos ter uma idéia sobre a qualidade da Curva ROC a partir da AUC da seguinte maneira:

- Entre 0.9 e 1 excelentes.
- Entre 0.8 e 0.9 boa.
- Entre 0.7 e 0.8 razoável.
- Entre 0.6 e 0.7 ruim.
- Entre 0.5 e 0.6 péssima.

Outros fatores também devem ser levados em consideração como por exemplo se o conjunto de dados é balanceados ou não.

3.3.3 Pureza and Completeza

Além da Curva ROC, podemos construir a partir da matriz de confusão outra curva que também nos fornecem uma perfomace de um classificador. Para isso definimos a Pureza e Completeza:

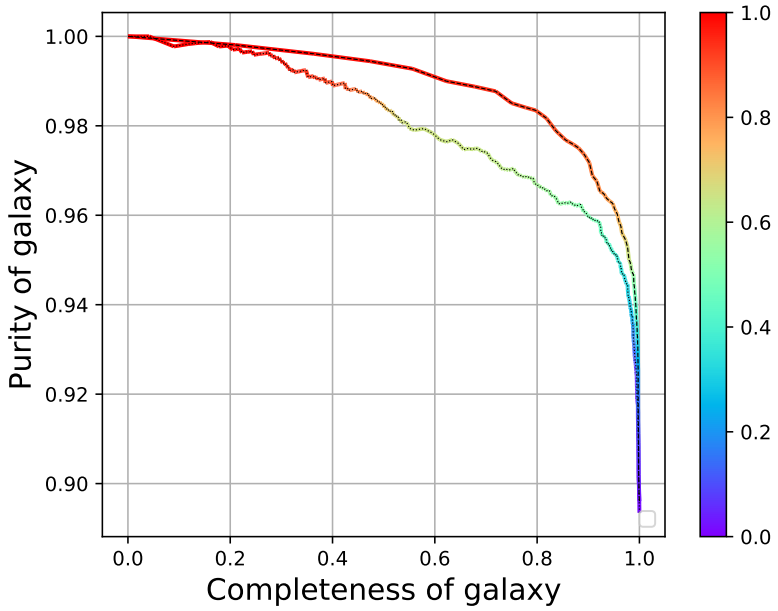

Figura 3.12. Completeza e Pureza para um classificador binário. As cores indicam os diferentes thresholds as quais foram calculadas as Completeza e Pureza.

$$Pureza = \frac{TP}{TP + FP} \quad \text{e} \quad Completeza = \frac{TP}{TP + FN} \tag{3.23}$$

Essas grandezas também são conhecidas na literatura como Precision e Recall respectivamente. Notemos que a Completeza possui a mesma defição que a taxa de verdadeiros positivos (TPR) definidas na Curva ROC. De maneira semelhante à Curva ROC podemos traçar uma curva que também varia conforme o threshold. Para medir a performance do algoritmo definimos a Precisão Média (AP):

$$AP = \sum_n (C_n - C_{n-1}) P_n \tag{3.24}$$

Onde C_n e P_n é a Completeza e a Pureza medidas no threshold n. Enquanto C_{n-1} é a Completeza medida no threshold $n-1$. Os pontos são ligados de maneira linear fomando uma área trapezodial que nos fornecerá a performace do classificador. A figura 3.12 representa duas Curvas de Completeza e Pureza para o caso de dois classificadores binário. Como no caso da Curva ROC, os thresholds variam entre zero e um e estão representados pela barra de cor à direita do gráfico. O valor máximo assumido pela AP também é igual à 1. Isso indica que os Falso Negativos (FN) e Verdadeiros Negativos (TN) assumem valor zero.

3.3.4 Erro Quadrático Médio (MSE)

Finalmente, pode-se medir o desempenho do algoritmo com o erro quadrático médio (MSE) definido como:

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (y_i - f(\mathbf{x}_i))^2 \tag{3.25}$$

onde y_i são as classificações do conjunto de testes e N_{test} é o tamanho do conjunto de testes. $MSE = 0$ caracteriza um desempenho perfeito.

3.4 Overfitting e Underfitting

Após o processo de treinamento dos algoritmos em ML podemos nos deparar com dois diferentes resultados no cálculo de sua performace. O primeiro é quando o modelo se ajusta muito bem aos dados de treino, mas no momento de extrapolar as predições aos dados de teste, possuem má performace. A esse caso em que há um sobreajuste aos dados de treino chamamos de overfitting ou sub-ajuste. O segundo caso é quando o algoritmo não se ajusta nem aos dados de treino. A esse caso chamamos de underfitting. Na figura 3.13 podemos observar os casos em que há overfitting/underfitting e ajuste ótimo.

Para avaliar a capacidade de generalização dos modelos de ML utilizamos o método de Validação Cruzada "K-Fold". Para esse método utilizamos os dados de treino separados em k partes iguais e mutuamente exclusivas. O modelo é treinado em $k-1$ partes e validado na restante. Essa parte restante é chamada validação. Repetimos esse processo k de forma cíclica. A performace final da técnica é calculada a partir da média de todas as acurácias medidas no loop. A figura 3.14 nos mostra o processo de divisão de dados durante o processo de validação cruzada para $k = 5$ folds. Observemos que os dados de teste e treino são separados inicialmente e em seguida separamos em 5 partes os dados de treino. Com quatro partes treinamos os algoritmos e com uma o validamos. Esse processo se repete por cinco vezes para diferentes conjunto de dados. Podemos avaliar os modelos da seguinte forma:

- Se a média do K-Fold possui uma acurácia excelente e o resultado de teste também, então o modelo não sofre overfitting.
- Se a média do K-Fold é excelente e os resultados do teste ruim, então o modelo sofre de overfitting.
- Se a média do K-Fold é ruim, então o modelo sofre de underfitting.

Além de nos permitir identificar modelos com overfitting e underfitting o K-fold nos permite calcular os melhores hiper-parâmetros para um algoritmo de ML em estudo. Os hiper parâmetros são grandezas inseridas à mão no algoritmo que influenciam diretamente

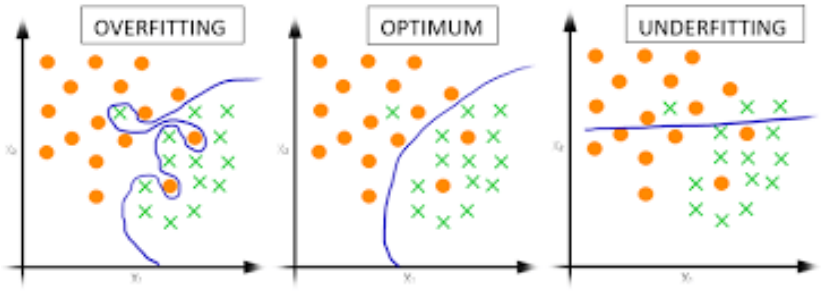

Figura 3.13. Exemplo de classificadores quando há um underfitting, overfitting e um ótimo ajuste sobre os dados[10].

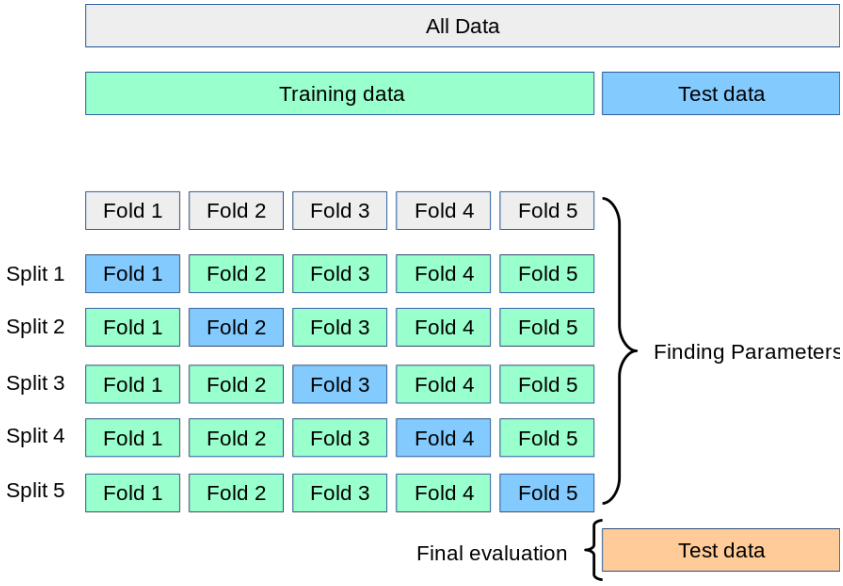

Figura 3.14. Representação de um processo de K-fold Cross Validation para $k = 5$.[11]

sua performace. Como exemplo podemos citar o número de árvores numa RF ou o número de vizinhos em um método KNN. Na prática aplicamos o K-fold utilizando diferentes hiperparametros e selecionamos o melhor.

3.5 Por quê Machine Learning ?

A quantidade de dados que temos nos deparado atualmente na indústria, nos grandes levantamentos astronômicos ou mesmo na Web nos leva a pensar em desenvolvimento de meios com rápida e eficaz análise. O volume total nos levantamentos astronômicos chegam à Zettabyte com uma velocidade Terabytes por dia [44]. Na tabela 3.2 esquematizamos alguns dos principais surveys em andamento e construção com os volumes, velocidade alcançados e os tipos de dados que serão obtidos. Numa escala global, existem 2,5 quintilhões (10^{18})

[10]Figura disponível em: https://machinelearningmedium.com/2017/09/08/overfitting-and-\regularization/

[11]Figura disponível em: https://scikit-learn.org/stable/modules/cross_validation.html

Sky Survey	Volume	Velocity	Variety
SDSS	50 TB	200 GB por dia	imagens, catalogos, redshift
GAIA	100 TB	40 GB por dia	mais que 100 parâmetros
Pan-STARRS	5 PB	5 TB por dia	imagens, catalogos
LSST	60 PB	10 TB por dia	imagens, catalogos
SKA	3 ZB	150 TB por dia	imagens, catalogos, redshift

Tabela 3.2. Descrição da quantidade de informação de levantamentos astronômicos atuais e futuros.

bytes de dados criados diariamente em nosso ritmo atual [45]. A cada minuto por dia por exemplo:

- Usuários do Snapchat compartilham 527.760 fotos.
- 456,000 tweets são enviados no Twitter.
- Usuários do Instagram postam 46.740 fotos.
- Spotify adiciona 13 novas músicas.
- Passageiros do Uber fazem 45,788 viagens.

No ramo empresarial/comercial a extração de informações sobre o perfil do consumidor ou a conhecimento sobre a dinâmica do sistema torna-se decisivo num ramo competitivo. O poder preditivo desses algoritmos tem chamado bastante atenção dessas empresas. Tendências em música [46], predição de doenças via imagens [47], filtro de spam [48] e detecção de fraude de cartão de crédito [49], estão entre algumas da inúmeras aplicações de ML ao processamento dessas informações.

Capítulo 4

Dataset: O survey miniJPAS

O J-PAS é um survey fotométrico que observará 8500 deg^2 do céu por meio da técnica de quasi-espectroscopia. Utilizando 56 filtros de banda estreita e 3 filtros *ugr* de banda larga, produzirá um pseudo-espectro com resolução média de ($R \sim 60$) [1]. Seu telescópio, o JST/T250, possui um espelho primário de 2.5 m e está equipado com uma câmera de 1.2 Gigapixel. Possui um campo de visão grande com 4.2 deg^2 sendo observados por 14 CCDs. Está localizado na cordilheira "Sierra de Javalambre" (Espanha), a uma altitude de 2000 metros. Uma região especialmente escura com a muito boa visão. O J-PAS se enquadra entre pesquisas fotométricas e espectroscópicas, combinando proveitosamente as vantagens da primeira (velocidade e baixo custo) com as do último (espectro). Em particular, graças ao excelente desempenho do redshift fotométrico (photo-z), será possível estudar com precisão o universo em grande escala usando os catálogos de galáxias e quasares.

O JPAS observará $\sim 9 \times 10^7$ Luminous Red Galaxies (LRG), Emission-Line Galaxy (ELG) e milhões de QSO com precisão para redshift fotométrico de $0.003(1 + z)$. Além de inferir a massa de 7×10^5 aglomerados de galáxias. A alta precisão do photo-z é devido a presença dos 56 filtros associados à bandas curtas no processo de observação das fontes, juntamente à 3 bandas largas. O que faz do survey único no campo da fotométria quasi-espectroscópica. Os surveys os quais se utiliza apenas bandas largas por sua vez alcançam uma precisão típica de $dz/(1 + z) \geq 3\%$ [11].

Entre maio e setembro de 2018, o seu telescópio JST/T250 foi equipado com a câmera Pathfinder, usada para testar seu desempenho e executar as primeiras operações científicas juntamente ao conjunto de 60 filtros. Essa câmera possui um CCD $9k \times 9k$, com um campo de visão de 0.3 deg^2 e tamanho de pixel de 0.225 arcos segundos. A figura 4.1 nos mostra o telescópio JST/T250 acoplado à câmera Pathfinder utilizado nas observações aos quais geraram o catálogo do miniJPAS. O catálogo principal do miniJPAS contém 64 293 objetos na banda de detecção *r*, com fotometria forçada em todos os outros filtros (Bonoli et al. (2020) [2]).

[1] A resolução do comprimento de onda R_λ é definido como $R_\lambda = \lambda/\Delta\lambda$. Para o sistema de filtros do J-PAS temos uma média de $R_\lambda \sim 60$.

	Telescope		Camera				
	Size	FoV	# CCD	CCD Format	# of pixels	Resolution	Filters
LSST	8.4 m	9.6 deg^2	189	4096 x 4096	3.2 Gpixels	0.2 "/pix	u,g,r,i,z,y
PanStarrs	1.8 m	6.7 deg^2	60	4600 x 4600	1.3 G pixels	0.26 "/pix	g,r,i,z,y
JPCam	2.5 m	4.9 deg^2	14	9231 x 9216	1.2 Gpixels	0.23 "/pix	56NB + 3BB
HyperSuprimeCam	8.2 m	1.8 deg^2	112	2048 x 4096	940 Mpixels	0.18 "/pix	r,i,z,y
VIS (EUCLID)	1.2 m	0.5 deg^2	36	4096 x 4096	520 M pixels	0.1 "/pix	R,I,Z
DECam	4 m	3 deg^2	62	2048 x 4096	500 M pixels	0.27 "/pix	g,r,i,z,y
MegaCam	3.6 m	1 deg^2	32	2048 x 4096	340 M pixels	0.19 "/pix	u,g,r,i,z
OmegaCam	2.6 m	1 deg^2	32	2048 x 4096	340 M pixels	0.21 "/pix	u,g,r,i,z
JPAS-Path Finder	2.5 m	0.45 deg^2	1	10580 x 10560	110 M pixels	0.23 "/pix	g,r,i +NBs
T80Cam	0.8 m	2.1 deg^2	1	10580 x 10560	110 M pixels	0.5 "/pix	u,g,r,i,z + 7NB
SuprimeCam	8.2 m	0.25 deg^2	10	2048 x 4096	80 M pixels	0.2 "/pix	g,r,i,z,y

Tabela 4.1. Comparação de diferentes câmeras e seus respectivos telescópios utilizadas em colaboração atuais e futuras.

A tabela 4.1 nos permite comparar diferentes câmeras e telescópios para diferentes colaborações. Em negrito temos as câmeras JCam e JPAS-Path Finder que estão associados à colaboração J-PAS, miniJPAS e o T80Cam associado ao J-PLUS um survey local composto por 7 bandas curtas mais 5 bandas largas [50].

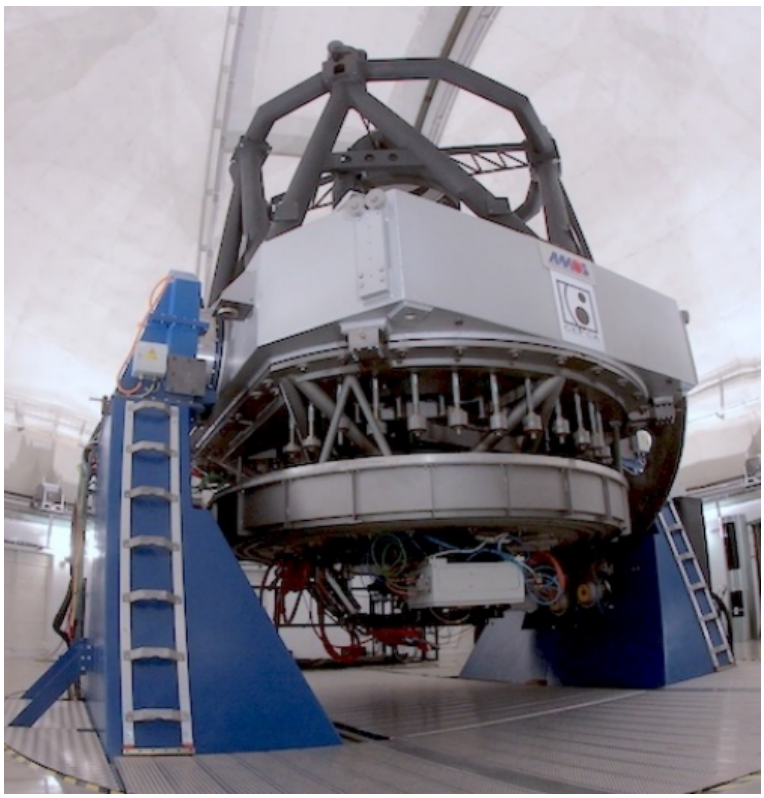

Figura 4.1. Imagem do telescópio JST/T250 utilizado para a pesquisa miniJPAS, com o JPAS-Câmera Pathfinder.

Neste trabalho utilizamos o miniJPAS dataset, um conjunto de observações já realizadas com área de 1 deg^2 profundos observadas no campo do AEGIS. Além do AEGIS, outros

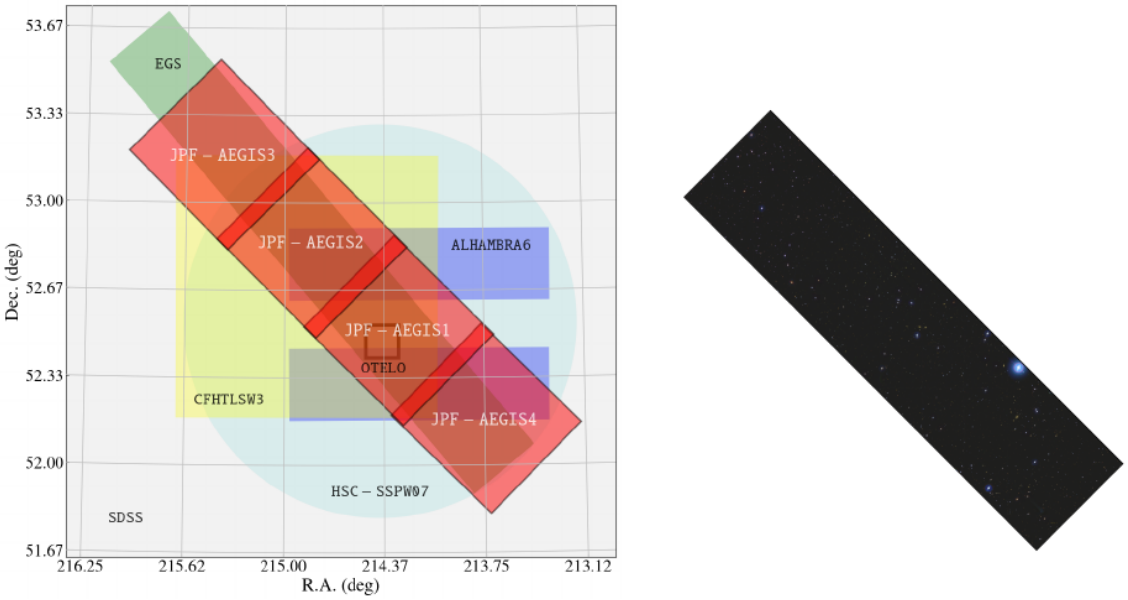

Figura 4.2. Footprint do miniJPAS com as telhas destacadas em vermelho. A áreas cobertas por outros surveys também são mostradas.

surveys já fizeram medidas sobre essa região. Na figura 4.2 podemos observar a sobreposição da área do miniJPAS sobre diferentes áreas cobertas por outros surveys. São justamente essas sobreposições que nos permitem treinar nossos algoritmos de ML, o qual estendemos as classificações para outras regiões.

Como já mencionado, o miniJPAS fez observações com 60 bandas fotométricas. A figura 4.3 representa a transmissão[2] em função do comprimento de onda para esse sistema de filtros, composto por 56 bandas curtas (narrows) e as 4 bandas largas (broad) uJPAS, gSDSS, rSDSS, iSDSS.[3] O J-PAS propriamente excluirá a banda *i*, fazendo medidas com 59 filtros.

4.1 Fluxo e Magnitudes

A fotometria é a medição da quantidade de luz oriunda de um objeto. No passado a principal forma de se medir essa grandeza era o olho humano. Depois da idade média vieram as lunetas/telescópios e as observações de Galileu Galilei no século XVII. No final do século XIX surgem as fotografias astronômicas as quais tem-se desenvolvido até os dias de hoje [51]. Diferentes dispositivos eletrônicos sensíveis à radiação eletromagnéticas permitiram o desenvolvimento da astronomia moderna como as câmeras de CCD (Charge-Coupled Device). O J-PAS é um survey fotométrico que utiliza deste tipo de câmera acoplada.

Para quantificar o fluxo de cada fonte, contamos o número de fótons do objeto dentro de uma área definida sobre as imagens observadas pelas CCDs. Esses fluxos podem ser

[2] A transmissão é definida como a fração do fluxo incidente que é transmitida através de um filtro.

[3] A última banda a direita, de cor violeta, representa a banda larga z pertencente ao JPLUS um survey fotométrico local com 12 filtros de banda curta e larga.

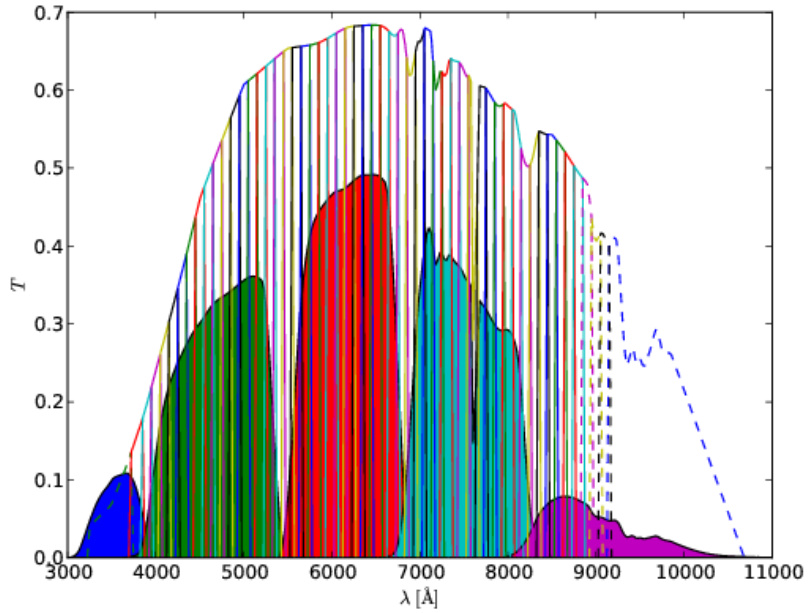

Figura 4.3. Curvas de transmissão das bandas fotométricas. As linha finas representam as curvas associadas a bandas curtas enquanto que as áreas coloridas representam as bandas grossas.

calculados de diferente formas conforme a definição da área. Na figura 4.5 podemos ver de forma qualitativa algumas das diferentes definições.

Na figura 4.4 podemos observar diferentes medidas realizadas pelo survey miniJPAS. Em colorido temos os fluxos associados as 60 bandas fotométricas. Em cinza tempos os fluxos espectroscópicos da fonte observada pelo SDSS juntamente à foto de cada fonte composto por estrelas, galáxias e quasares.

Uma outra forma de se medir essa quantidade de fótons que chegam nas CCDs é utilizando magnitude aparente m associadadas a uma banda fotométrica:

$$m = -2.5\, log_{10}\, F + const \tag{4.1}$$

Onde F é o fluxo associado a banda fotométrica utilizada na medição. Esse fluxo é calculado a partir de uma área sobre a imagem a qual pode ser definida de diferentes formas. Diferentes áreas definidas darão origem a diferentes magnitudes/fluxos medidos. Na figura 4.5 podemos observar diferentes definições. Em nosso caso utilizamos o *FLUX_AUTO* para a equação 4.1 que são calculadas a partir da área de uma elipse. Esses cálulos são realizados pelo software SExtractor. Esses fluxos nos fornecerão as *MAG_AUTO* [4] [52] [51], os quais utilizamos como entrada em nossos algoritmos. Cada filtro irá gerar uma *MAG_AUTO*; 60 no total.

[4] https://sextractor.readthedocs.io/en/latest/Param.html

Definições de Magnitudes

A seguir definiremos alguns tipos de magnitudes calculados pelo software SExtractor e que são amplamentes utilizados em diversas literaturas. Essas definições estão inseridas dentro do processo chamado *segmentação* que consiste na separação de diferentes regiões de imagem. Uma detecção é definida a partir de um threshold definido acima do fundo. Uma detecção é formada por um conjunto de pixel conectados e que estão acima de threshold é definido. A partir dessas detecções podemos definir os diferentes tipos de magnitudes.

MAGNITUDE ISO: Como dito anteriormente, as detecções são definidas a partir de um threshold. As detecções acima desse threshold constituem a área isophotal. Definimos MAGNITUDES ou FLUXOS ISO a partir da contagem desses pixel subtraído do background.

MAGNITUDE ISOCORR: As MAGNITUDES ISOCOR, i.e, ISO CORRIGIDAS são uma forma grossa de se obter os fluxos ou magnitudes perdidos nas regiões de "asas"assumindo um perfil simétrico gaussiano.

MAGNITUDE APERTURE: Estima o fluxo ou magnitude acima do fundo com uma abertura circular fixa.

MAGNITUDE AUTO: As magnitudes AUTO foram construídas com intúito de obter uma estimativa total das magnitudes pelo menos para galáxias uma vez que é definida a partir de uma abertura elíptica. os semi-eixos da elipse são definidos a partir dos momentos de segunda ordem da distribuição de luz do objeto.

A figura 4.4 nos da uma interpretação geométrica de cada definição de fluxo-magnitudes. Os quadrados em cinza representçao as detecções acima de um threshold enquanto que os circulos representam a área o qual será feita a contagem para o cálculo do fluxo ou magnitude. Para mais informações sobre as diferentes formas de se definir magnitudes e fluxos consulte [52] [3].

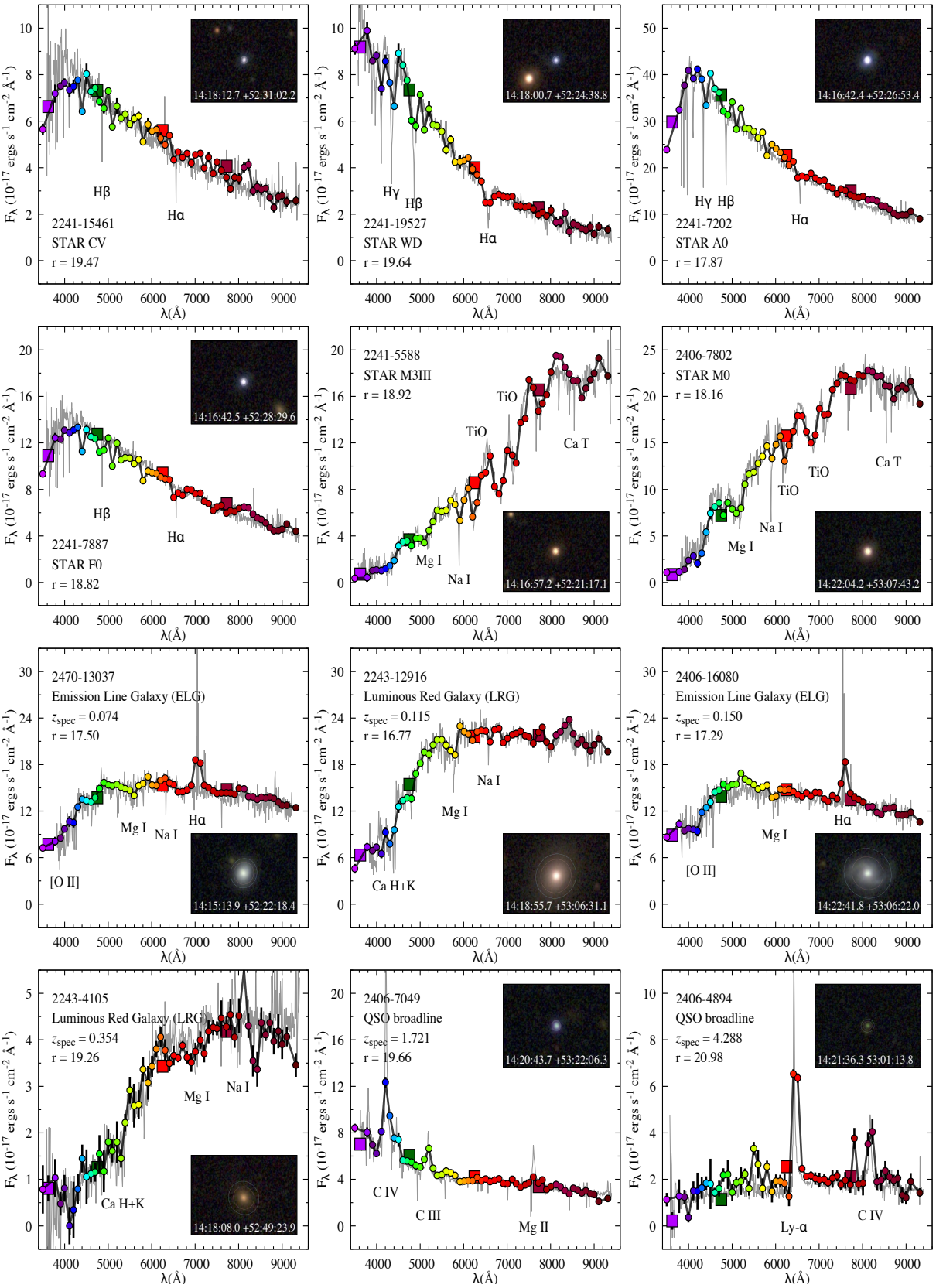

Figura 4.4. Fotometria das diferentes classes de estrelas, galáxias e quasares no campo miniJPAS (pontos coloridos) comparados com os espectros do SDSS. Figura retirada do paper miniJPAS PathFinder (Bonoli(2020) [2])

4.2 Parâmetros Morfológicos

Nas imagens captadas pelos telecópios, estrelas aparecem como objetos pontuais enquanto galáxias como fontes extendidas. Entretando à medida que observamos a magnitudes mais fracas, a diferença entre fontes pontuais e extendidas vai diminuindo se tornando indistinguíveis em um determinado momento. Existem modelos que conseguem diferenciar fontes extendidas e pontuais utilizando apenas parâmetros morfológicos. Na figura 4.6 (à direita), podemos observar um exemplo no survey HSC-SSP onde o classificador é gerado a partir da subtração de diferentes definições de magnitudes. Em nosso trabalho utilizamos quatro parâmetros morfológicos juntamente à 60 valores de magnitudes associadas a cada filtro. Os parâmetros morfológicos utilizados foram:

- A concentração $c_r = r_{1.5''} - r_{3.0''}$, onde $r_{1.5''}$ e $r_{3.0''}$ representam magnitudes observadas dentro da abertura circular fixa com 1.5"e 3.0"respectivamente, utilizando a banda *rSDSS*.
- O alongamento A/B, que consiste na divisão entre RMS da distribuição da luz ao longo das direções máxima e mínima de dispersão [52].
- Largura a meia altura (FWHM) assumindo um núcleo gaussiano que nos fornece o diâmetro do disco que contém metade do fluxo de objetos
- Pico de brilho da superfície acima do fundo dividido, pela magnitude com abertura fixa de 3.0" na a banda r, $\mu_{max}/r_{3.0''}$

O parâmetro μ_{max} é como uma fotometria medida em 1 pixel. Todos os parâmetros acima nos permitem ter uma noção sobre a formas das fontes. Na figura 4.9 podemos

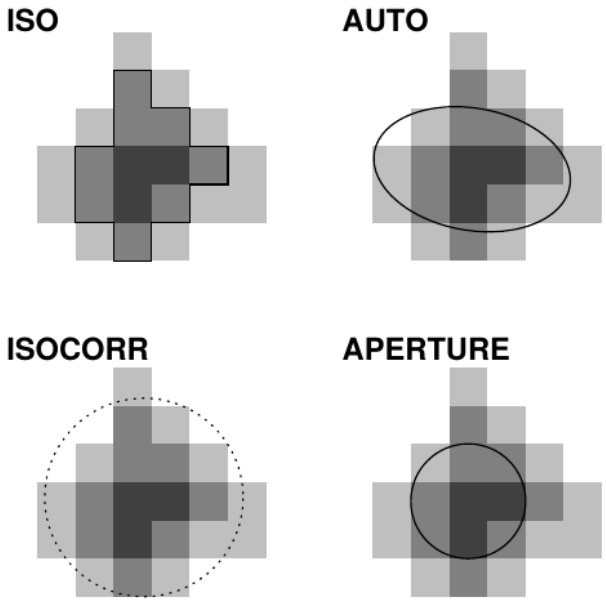

Figura 4.5. Diferentes áreas definidas, associadas aos seus respectivos fluxos [3].

observar a distribuição desses parâmetros para objetos classificados como estrelas e galáxias pelos surveys SDSS e HSC-SSP. À direita temos objetos classificados segundo o survey SDSS e à esquerda pelo survey HSC-SSP, onde as galáxias assumem a cor vermelha e as estrelas a cor azul. Note que o poder de separação desses parâmetros para magnitudes mais fortes é alto, enquanto que para magnitudes mais fracas necessitamos mais informações das fontes. Isso motiva a utilização das bandas fotométricas como informação adicional em nossos algoritmos.

Na figura 4.6 (à esquerda) temos a representação do parâmetro concentração c_r em função da banda *rSDSS* para os dados do miniJPAS. O corte em vermelho foi traçado para destacar a presença de duas distribuições marcantes até uma magnitude de ~ 21.

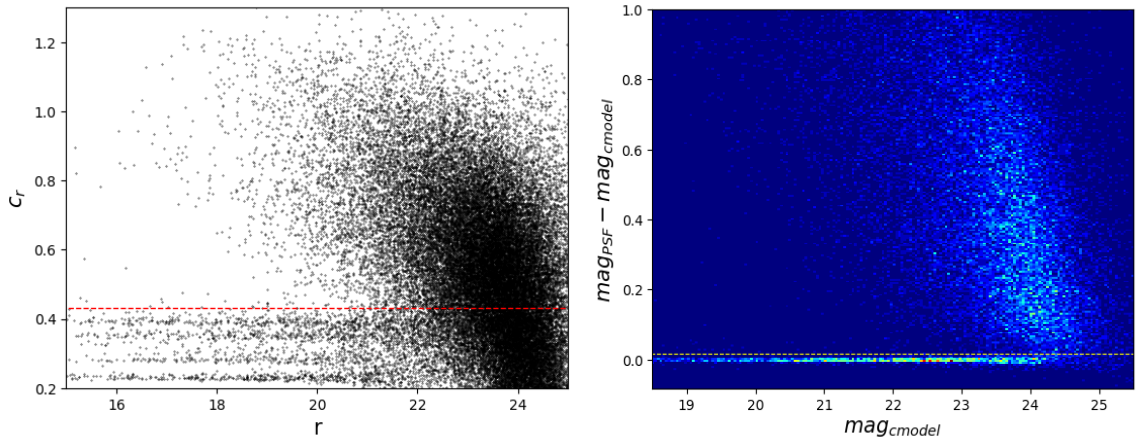

Figura 4.6. À esquerda: Representação do parâmetro concentração c_r em função da banda r. À direita: Dados do HSC-SUBARU e em função do seu classificador observada pela banda r_{cmodel}. O corte em amarelo separa fontes em estrelas (abaixo) e galáxias (acima).

Esse limite é característico de todos os survey photométricos. Na figura 4.6 (à direita), por exemplo, podemos observar uma diferença entre objetos extendidos e pontuais até uma magnitude de $\sim$ 24. Os levantamentos mais modernos observarão fontes mais distante $r > 21$, portanto novas técnicas de separação à magnitudes fracas devem ser pensadas. O ML surge com essa proposta; Extender classificações à regiões limitadas do survey.

4.3 Cross-Match

Para se treinar os algoritmos de ML utilizamos classificações oriundas de dois surveys de forma separadas e independentes classificações as quais assumimos ser confiável. Para magnitudes mais fortes (intensas), utilizamos rótulos do cross-match entre dados de miniJPAS e SDSS. Para magnitudes mais fracas (menos intensas) utilizamos o cross-match entre miniJPAS e HSC-SSP. Para ambos os casos aplicamos cortes de qualidade aos dados do miniJPAS inserindo as restrições $flag = 0$ e $mask = 0$. Isso é feito para se evitar dados com

algum tipo de problema na observação, nos permitindo aumentar o poder preditivo de nossos algoritmos. Os dados extraídos do miniJPAS foram do catálogo *minijpas.MagABDualObj*. Esse catálogo contém fotometria de todos os objetos detectados utilizando a banda *rSDSS*. As magnitudes estão no sistema AB e não há objetos duplicados em regiões de sobreposição das telhas. Quando isso ocorre descarta-se a pior medida[5].

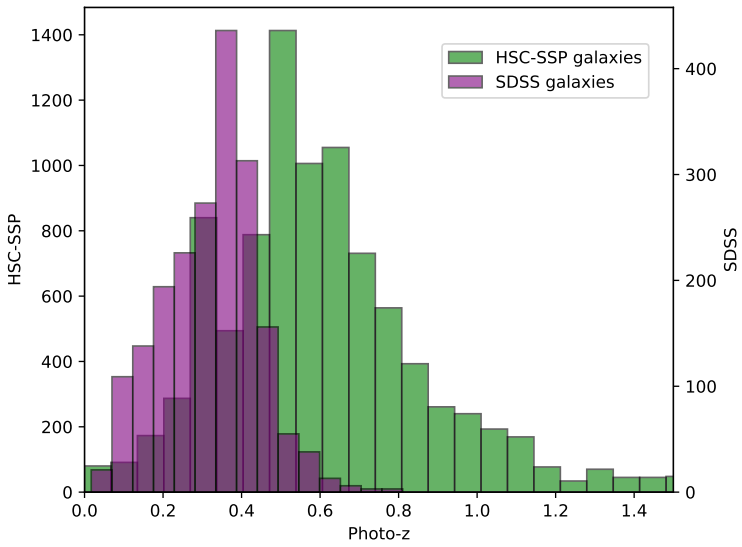

Figura 4.7. Distribuição dos redshifts fotométricos para os cross-match entre miniJPAS e os surveys SDSS e HSC-SSP. Em verde temos os dados de HSC e em violeta os dados de SDSS em menor número.

Na figura 4.7 podemos observar a distribuição do redshift das galáxias utilizados em nossas análises. Em violeta temos os dados cruzados com o SDSS e em verde com HSC-SSP. Os redshifts utilizados para contruir o histograma são photométricos e oriundos dos respectivos surveys. Em particular, os redshifts fotométricos oriundos do survey HSC-SPP foram os *photo-z mode* retirado do catálogo *photoz_mizuki* aplicado ao campo WIDE [6]. Os photo-z's oriundo do survey SDSS, z_{ph}, podem ser encontrados no catálogo *minijpas.xmatch_sdss_dr*12 do miniJPAS.

Na figura 4.8 podemos observar a distribuição de galáxias e estrelas em função da magnitude na bandas r os quais utilizamos para treinar e testar nossos algoritmos. À esquerda temos os dados do miniJPAS cruzados com o SDSS, enquanto que à direita com os dados do HSC-SSP. As galáxias assumem cores vermelhas enquanto estrelas as cores azuis. Note que o número de galáxias é bem maior que o de estrelas para o crosmatch com HSC-SPP. Em SDSS o número de objetos para cada classe é mais equilibrada. Isso se deve à faixas de magnitudes estudadas. Para magnitudes mais fortes $15 < r < 21$ observaremos um universo mais local, encontrando muito mais estrelas. Para magnitudes mais fracas $18.5 < r < 23.5$ observaremos objetos mais distante encontrando muito mais galáxias em nosso dataset.

[5] A representação das sobreposição das telhas pode ser observado na figura 4.2

[6] Para mais informação consulte: https://hsc.mtk.nao.ac.jp/ssp/science/photometric-redshfits/

4.3.1 The Sloan Digital Sky Survey (SDSS)

A primeira análise foi feita para o faixa $15 < r < 21$ utilizados rótulos fotométricos do SDSS. O SDSS é um dos surveys mais influentes na história da astronomia. Suas atividades começaram abril de 2000 e observou quase 1/3 da esfera celeste imageando e utilizando espectroscopia. É conduzido por Observatório de Apache Point [12] [53]. Possui um telescópio com espelho primário de 2.5 m situado sendo situado no Novo México, EUA. Observando fontes utilizando 5 bandas fotométricas: u, g, r, i, z. O SDSS pode ser dividido em quatro grandes fases:

- SDSS - I/II : pesquisa anterior (2000 - 2008)
  LEGACY SUPERNOVA SEGUE-1
- SDSS - III : pesquisa anterior (2008 - 2014)
  APOGEE BOSS MARVELS SEGUE-2
- SDSS - IV : pesquisa atual (2014 - 2020)
  APOGEE-2 eBOSS MaNGA

O **LEGACY** survey observou nas bandas *ugriz* milhares de fontes sobre um campo com mais de 7.500 deg^2. Os dados resultantes apoiaram estudos que variam de asteróides e estrelas próximas à estrutura em larga escala do universo. O **SUPERNOVA** survey observou imagens do Stripe 82 para encontrar restos de explosões de estrela, juntamente com acompanhamento espectroscópico e fotométrico para obter redshift e curvas de luz. A pesquisa descobriu quase 500 supernovas do tipo Ia com desvio para o vermelho de cerca de 0.4. O **SEGUE** (Sloan Exploration of Galactic Understanding and Evolution) observou o conteúdo estelar da Via Láctea, a fim de criar um mapa tridimensional detalhado da galáxia. O SEGUE obteve 3500 deg^2 observados e obteve espectros de 240.000 estrelas no disco e no esferóide. O **APOGEE** (Apache Point Observatory Galactic Evolution) é um survey espectroscópico que começou em 2011 e observará cerca de 100.000 estrelas, principalmente no disco da Via Láctea, para obter composições químicas e cinemáticas detalhadas. O **BOSS** (Baryon Oscillation Spectroscopic Survey) é um survey espectroscópico que obteve o redshift de 1,5 milhão de galáxias e espectros de 150.000 quasares, para medir o sinal de oscilação do bárion na função de correlação como uma sonda geométrica da cosmologia. O **eBOSS** (Extended Baryon Oscillation Spectroscopic Survey) concentra seus esforços na observação de galáxias e, em particular, quasares, em várias distâncias. Consiste numa extensão do survey BOSS. **MARVELS** (Multi-object APO Radial Velocity Exoplanet Large-area Survey) é um survey espectroscópico que observa repetidamente estrelas brilhantes para detectar as variações de velocidade radial causadas por planetas em órbita. **MaNGA** (Mapping Nearby Galaxies at APO) permite medições espectrais ($\sim$ 10.000) galáxias próximas. O MaNGA fornecerá a velocidade estelar e dispersão de velocidade, idade média estelar e histórico de formação de estrelas, metalicidade estelar, taxa de abundância de elementos, densidade de superfícial de

massa estelar, velocidade de gás ionizado, metalicidade de gás ionizado, taxa de formação de estrelas e extinção de poeira para uma amostra estatisticamente poderosa.

Para o cross-match entre miniJPAS e SDSS (DR12) encontramos 3 699 fontes com 1 988 galaxias e 1 711 estrelas. O cross-match foi feito pela colaboração miniJPAS e está disponível na UPAD no catálogo *minijpas.xmatch_sdss_dr12*.

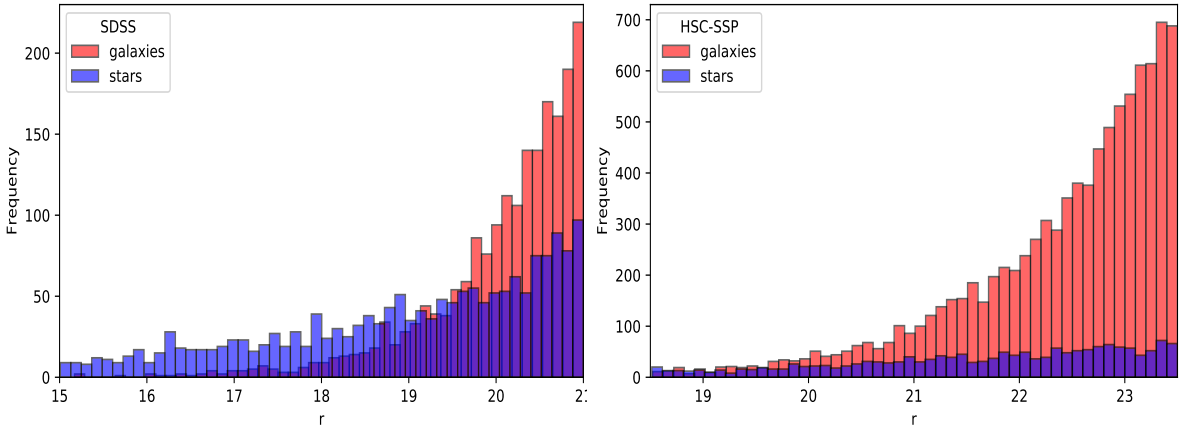

Figura 4.8. Distribuição dos dados SDSS e HSC-SSP cruzados com o miniJPAS. Em azul temos estrelas e em vermelho temos galáxias. Devido aos limites de magnitudes observados, temos distribuições diferentes para cada caso. Em magnitudes mais fracas temos menos estrelas do que galáxias enquanto que para magnitudes mais fortes temos um número de estrelas e galáxias aproximadamente iguais.

4.3.2 The Hyper Suprime-Cam Subaru Strategic (HSC-SSP)

A segunda análise foi feita para a faixa $18.5 < rSDSS < 23.5$ utilizados rótulos fotométricos do SDSS. Escolhemos o limite de 23.5 por que o survey até essa magnitude irá observar com uma completeza de $\sim$ 100%. O HSC-SSP é um survey fotométrico dirigido pela Observatório Nacional Astronômico do Japão que cobre 1400 deg^2 divididos em 3 processos de observação: Wide, Deep e UltraDeep. A parte utilizada para o cross-match com miniJPAS foi a parte Wide que cobre 1400 deg^2 observados em 5 bandas largas *grizy* até magnitude $\sim$ 26 [54] [55]. Para isso utilizamos o Public Data Release (PDR2) e encontramos após o cross-match 11 083 fontes, com 9 392 galaxias e 1 691 estrelas. Os campos Deep e Ultradeep cobrem 27 deg^2 e 3.5 deg^2 respectivamente. O campo Deep observa utilizando fontes utilizando as bandas *grizy*+4*NBs* até a magnitude $r \sim 27$. O campo UltraDeep por sua vez com as bandas *grizy* + 4*NBs* até uma profundidade magnitude de $r \sim 28$.

Para construção do catálogo, baixamos os dados do survey HSC-SSP a partir do site `hsc-release.mtk.nao.ac.jp/doc/`. Em seguida baixamos os dados do miniJPAS e fizemos um Cross-Match com o programa TOPCAT. Ambas consultas podem ser encontradas no apêndice C. Aplicamos corte de qualidade nos dados de HSC-SUBARU impondo a restrição *isprimary* = *True* e *r_inputcount_valeu* >= 4. A primeira condição nos permite excluir dados

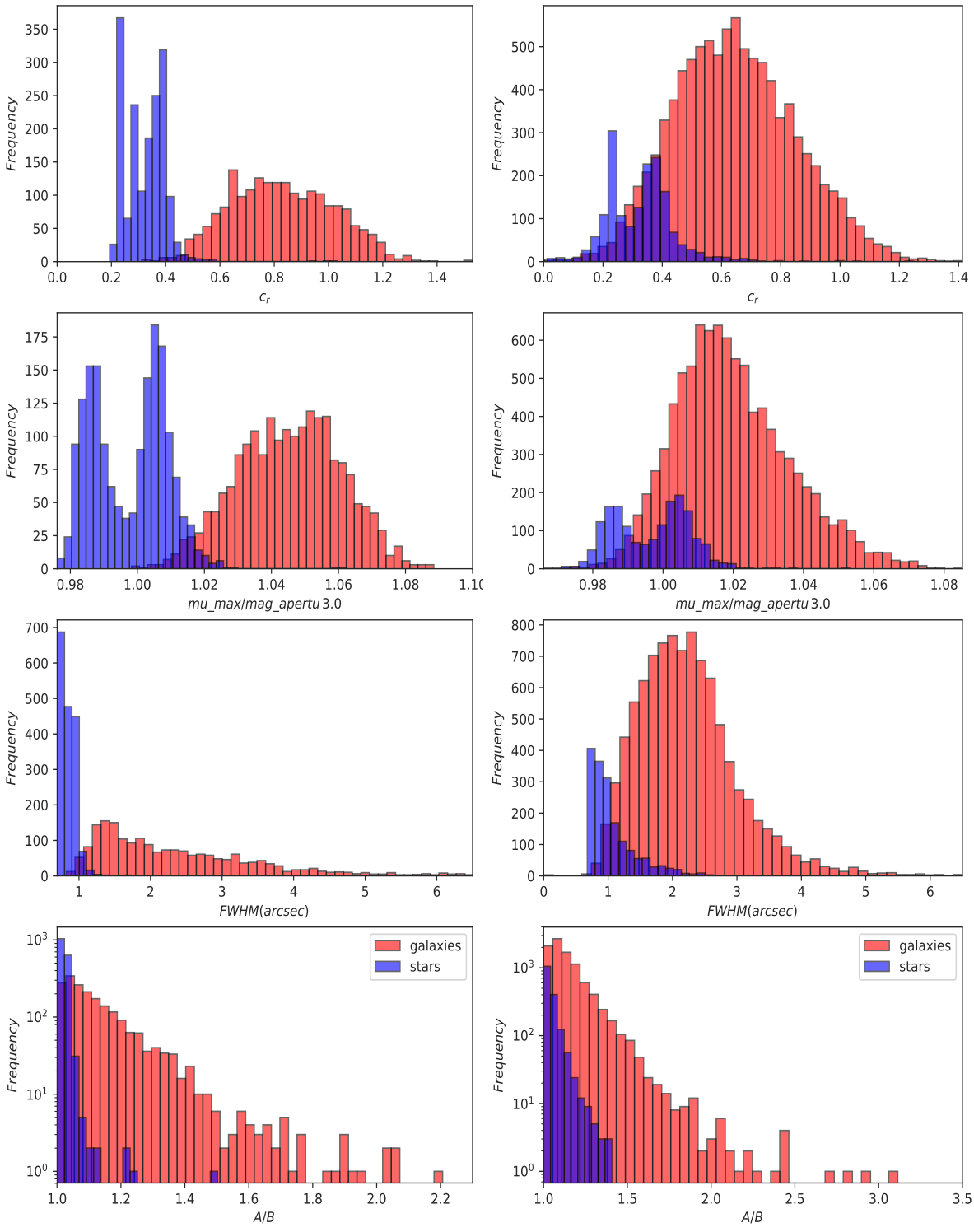

Figura 4.9. Distribuição dos parâmetros morfológicos para fontes obtidas através do cross-match entre miniJPAS e SDSS (à esquerda) e HSC-SSP (à direita). Em vermelho temos as galáxias e em azul estrelas, todas fontes classificadas pelos seus respectivos surveis cruzados. Note que esses parâmetros são muito expressivos a magnitudes mais fracas.

duplicados em regiões de overlap do survey enquanto que a segunda nos permite selecionar fontes observadas no mínimo por 4 bandas fotométricas.

O poder de extensão de classificação de nossos algoritmos, para regiões de magnitudes mais fracas do miniJPAS, se dá por conseguirmos treinar nossa máquina com rótulos do HSC-SSP. Rótulos obtidos a partir de um telescópio com espelho primário com 8.2 metros. Quanto maior o espelho, maior a captação de luz, i.e, teremos informações com maior qualidade. A colaboração HSC-SSP conseguiu observar um universo profundo, mas por outro lado, tem-se dificuldades em observar objetos mais próximos devido a alta intensidade da luz e saturação das CCDs. O limite de saturação para a banda *r* no campo Wide, utilizado no trabalho, se dá para $17.4^{+0.7}_{-0.4}$ mag. A profunidade alcançada é de $26.2^{+0.2}_{-0.3}$ *mag* e o tempo de exposição é varia entre 10^{+2}_{-5} minutos. Portanto quando é seguro tomar o limite o inferior como 18.5 sobre a banda *r* para o caso do em que utilizamos os rótulos deste dataset.

4.3.3 Pré-Processamento

Para rodar o algoritmo realizamos um pré processamento com os dados. Os padronizamos, i.e, centralizamos os valores das features na média 0 com desvio padrão 1 utilizando a equação:

$$x^i = \frac{x^i - \mu_x}{\sigma_x} \tag{4.2}$$

Tanto os dados de teste como os dados de treino.

Nas tabelas oriundas do catálogo *minijpas.MagABDualObj* podemos encontrar valores com magnitudes igual 99 para algumas bandas fotométricas. Esse valor indica uma má observação. Isso é bastante comum em aprendizagem de máquina no contexto de empresa e pesquisa. Existem diversas formas de se tratar esse problema. Em nosso trabalho os deixamos como estão, uma vez que os dados teste também possuem essa característica. Quanto mais semelhantes as distribuições dos dados de teste e treino maior tende a ser a acurácia de nosso modelo.

Capítulo 5

Resultados

Neste capítulo apresentamos os resultados das análises realizadas com os dados preliminares do miniJPAS. Comparamos diferentes modelos de Machine Learning entre si e com resultados oriundos dos classificadores SExctractor e SGLC; classificadores disponíveis no catálogo do miniJPAS. Separamos os resultados em duas faixas de magnitude. A primeira, foi realizada para o caso em que $15 < rSDSS < 21$ no qual treinamos nossos algoritmos sobre os dados do miniJPAS com rótulos do survey SDSS. A segunda etapa foi para $18.5 < rSDSS < 23.5$, no qual treinamos nossas máquinas com dados do miniJPAS rotulados pelo HSC-SSP.

Para medir e comparar a performace dos modelos utilizamos as Curvas ROC e Completeza x Pureza, assim como a área sobre a Curva ROC (AUC) e a Precisão Média (AP) para cada caso. As curvas de Completeza e Pureza foram feitas tanto para galáxias como para estrelas. Também fizemos outras análises como diagramas *cor* x *cor* para avaliar o nível de contaminação das classificações através do Locus Stellar. Aplicamos o método K-Fold Cross-Validation para avaliar se nossos modelos sofrem de Overffiting ou Underffiting e para escolha dos melhores parâmetros. Os algoritmos de ML utilizados nas análises foram o KNN, NN, DT, RF, ERT e um Ensemble Learning. Também calculamos a distribuição das probabilidades das galáxias e estrelas para cada caso. Isso nos permite observar de maneira qualitativa o quão bem os algoritmos classificam as fontes em estudo.

5.1 Usando rótulos SDSS

Comparamos aqui diferentes algoritmos de ML com os classificadores oriundos de *CLASS_STAR* e de SGLC disponíveis no catálogo do miniJPAS[1]. A análise com rótulos de SDSS foram feitas para o intervalo $15 < rSDSS < 21$ e foi subdivida em duas parte. Uma na qual utilizamos apenas bandas fotométricas como entrada dos algoritmos e uma segunda parte o qual utilizamos bandas fotométricas juntamente à parâmetros morfológicos. O objetivo dessa análise em forma dividida é observar o quão importante se faz as grandezas

[1] http://www.j-pas.org/datareleases

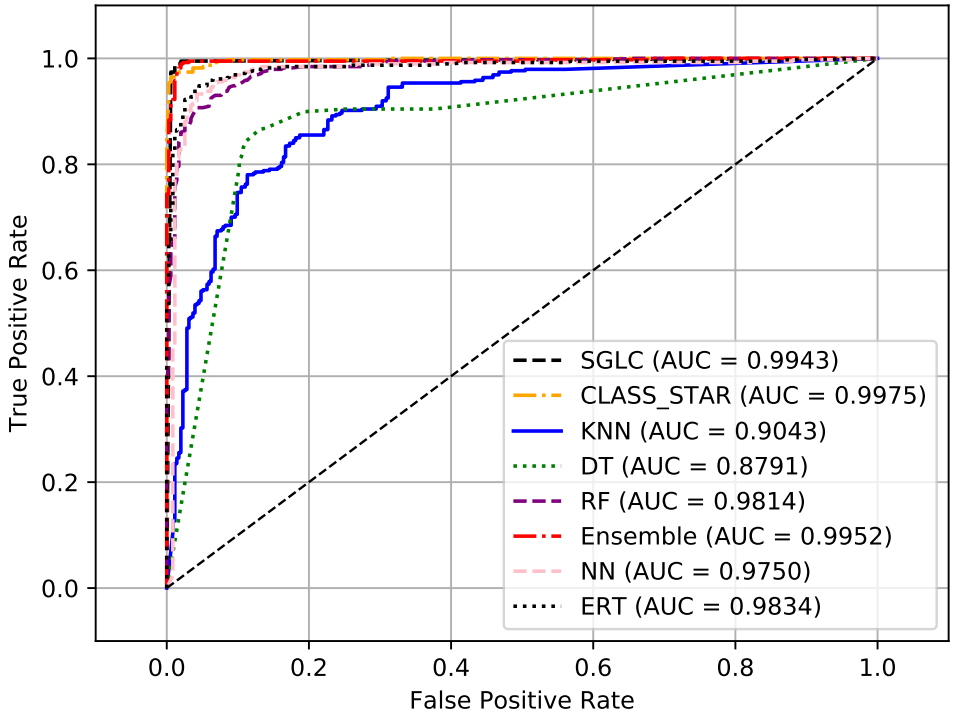

Figura 5.1. Curvas ROC para dados do miniJPAS com rótulos do SDSS utilizando apenas bandas fotometricas. A análise foi realizada para $15 < rSDSS < 21$.

fotométricas separadamente das morfológicas. Além do mais, análises as quais utilizam apenas bandas fotométricas tendem a classificar quasares como objetos extragaláticos melhor do que o caso em que se utiliza bandas fotométricas juntamente à parâmetros morfológicos [16].

Na figura 5.1, podemos observar as Curvas ROC geradas a partir de algoritmos os quais utilizamos apenas bandas fotométricas. Encontramos uma excelente AUC como comportamento geral. Possuindo o ERT o melhor desempenho com uma $AUC = 0.9984$. Embora ML possua um excelente performace, os classificadores *CLASS_STAR* e SGLC os superam. Isso não é de se surpreender, uma vez que estes utilizam de parâmetros morfológicos em seu processo de classificação. O algoritmo Ensemble Learning possui uma performace maior que ERT com uma $AUC = 0.9952$. Mas este por sua vez combina os classificadores SGLC juntamente à NN e ao RF, i.e, leva em consideração parâmetros morfológicos em um momento de suas predições. Dentre os métodos de árvores, o DT é o que possui menor desempenho.

Na figura 5.2 inserimos juntamente às bandas fotometricas, parâmetros morfológicos. Podemos observar uma melhora de desempenho considerável devido a esses parâmetros adicionais. De uma maneira mais geral, a performace dos algoritmos os quais utilizamos ML se aproximam bastante dos resultados gerados pelos métodos SGLC e *CLASS_STAR*. Em alguns casos SGLG e *CLASS_STAR* possuem performace menor que ML. O algoritmo com maior destaque é o ERT com $AUC = 0.9984$ superando a todos.

Uma outra forma de se medir a performace de um algoritmo é calculando a Completeza

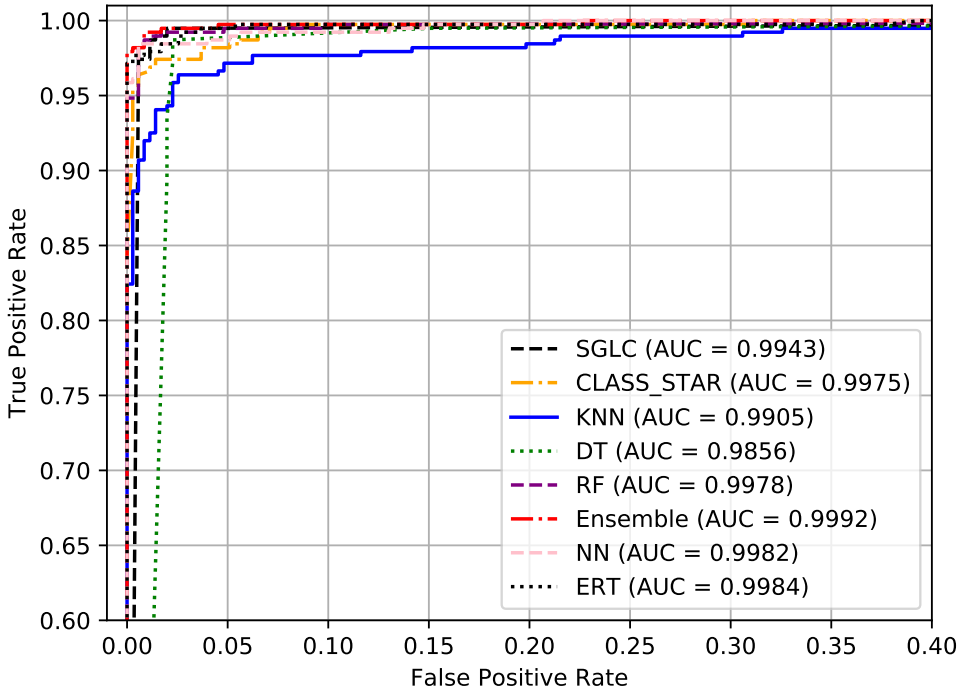

Figura 5.2. Curvas ROC para dados do miniJPAS com rótulos do SDSS utilizando bandas fotométricas juntamente à parâmetros morfológicos. A análise foi realizada para $15 < rSDSS < 21$.

em função da Pureza. Método também eficaz para o análise de dados não-balanceados. A figura 5.3 mostra essas grandezas calculadas para galáxias no caso em que utilizamos unicamente bandas fotométricas como entrada do algoritmo. De maneira semelhante à curva ROC, os métodos SGLC e *CLASS_STAR* superam os algoritmos que utilizam apenas bandas fotométricas. Dentre os que construímos, o ERT possui melhor performace com uma $AP^{\mathrm{gal}} = 0.9871$. O DT possui a menor performace entre os algoritmos de árvores. Ensemble learning consegue superar a análise SGLC.

Na figura 5.4, inserimos juntamente as bandas fotométricas, os parâmetros morfológicos. Como era de se esperar, há uma melhoria considerável na performace dos algoritmos. Alguns deles compostos puramente por ML superam SGLC. A maior performace possui o algoritmo Ensemble com $AP = 0.9993$ seguida por ERT e RF superando SExctractor e SGLC. Os algoritmos de menores performace são KNN e DT.

Na tabela 5.1 fizemos um resumo das performaces dos algoritmos para as Curvas ROC e Completeza x Pureza. Os índices $F + M$ informam que a análise foi realizada inserindo bandas fotométricas juntamente a parâmetros morfológicos. O índice F indica que apenas bandas fotométricas foram inseridas na análise. O índice *gal* em AP indica que estamos trabalhando com a Completeza e Pureza das galáxias. Como podemos observar de maneira mais compacta, é grande a diferença quando adicionamos morfologia nas análises comparadas ao caso em que utilizamos apenas bandas fotométricas. As melhores performaces estão em negrito. O interessante é notar que sempre conseguimos superar o classificador

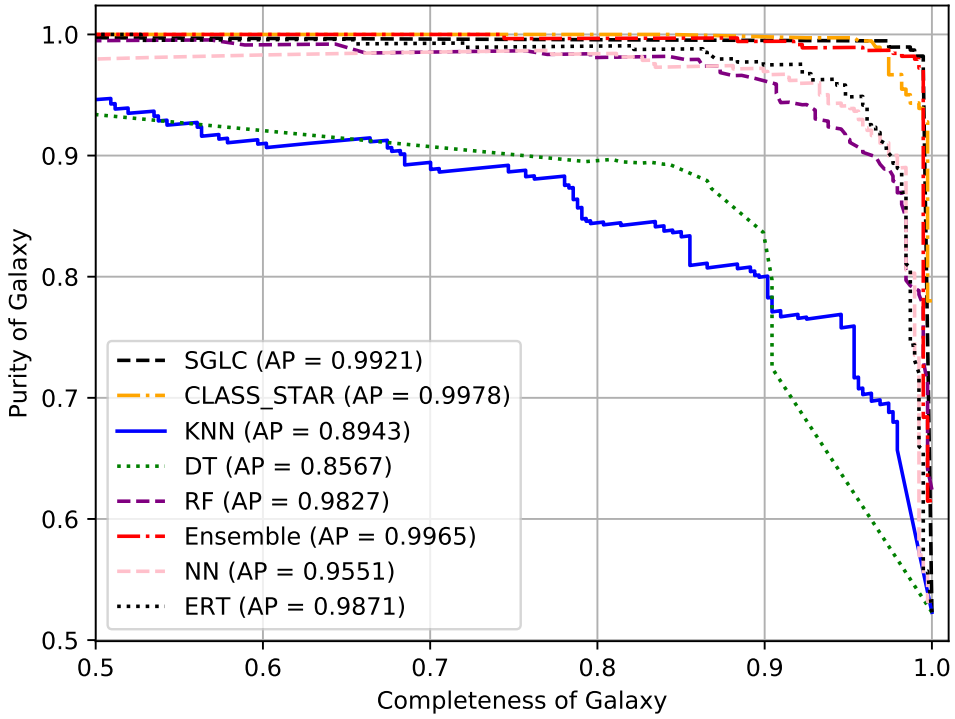

Figura 5.3. Curvas de Completeza e Pureza para galáxias utilizando dados do miniJPAS com rótulos do SDSS. A análise foi realizada para $15 < rSDSS < 21$ utilizando apenas bandas fotométricas.

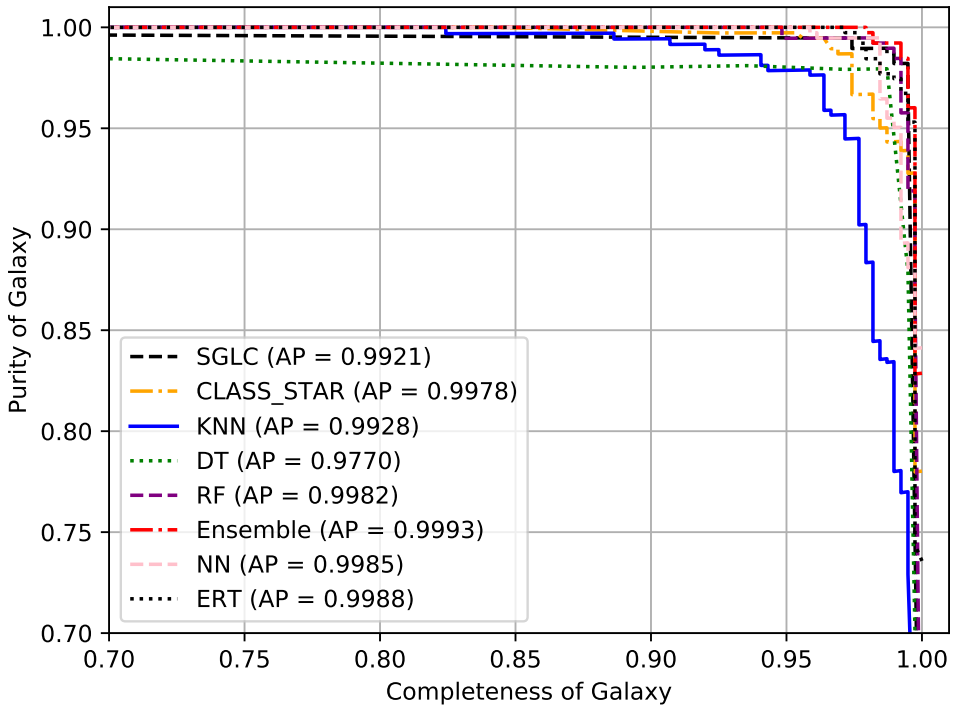

Figura 5.4. Curvas de Completeza e Pureza para galáxias utilizando dados do miniJPAS com rótulos do SDSS. A análise foi realizada para $15 < rSDSS < 21$ utilizando bandas fotométricas juntamente à parâmetros morfológicos.

miniJPAS-SDSS	AUC_{M+P}	AUC_P	AP^{gal}_{M+P}	AP^{gal}_P	MSE_{M+P}	MSE_P
SGLC	0.9943	–	0.9921	–	0.0122	–
CLASS_STAR	0.9975	–	0.9978	–	0.0395	–
KNN	0.9905	0.9043	0.9928	0.8944	0.0365	0.1375
DT	0.9856	0.8791	0.9770	0.8567	0.0189	0.1252
RF	0.9978	0.9814	0.9982	0.9827	**0.0097**	0.0563
Ensemble	0.9992	0.9952	0.9993	0.9965	0.0090	0.0232
ANN	0.9982	0.9750	0.9985	0.9551	0.0211	**0.0484**
ERT	**0.9984**	**0.9834**	**0.9988**	**0.9871**	0.0115	0.0523

Tabela 5.1. Desempenho dos classificadores para o catálogo miniJPAS comparado com o catálogo SDSS para a faixa $15 < rSDSS < 21$. O melhor desempenho é marcado em negrito. O índice F representa a análise que utiliza apenas bandas fotométricas, enquanto $M+F$ representa a análise que utiliza bandas fotométricas juntamente com parâmetros morfológicos. As melhores performaces estão destacadas em negrito.

CLASS_STAR e SGLC em alguma métrica. Isso nos mostra o quão competitivo os modelos de ML são.

Analisamos também as curvas de Completeza e a Pureza para as estrelas. Utilizando apenas bandas fotométricas, observamos na figura 5.5, que RF possui o melhor desempenho com $AP^{star} = 0.9796$. KNN e DT possuem o pior desempenho enquanto NN, ERT e RF possuem um bom desempenho. Quando inserimos parâmetros morfológicos podemos notar na figura 5.6 que todos os algoritmos possuem uma excelente performace com AP^{star} acima de 0.98.

Os gráficos acima são comuns dentro do processo de classificação de fontes em estrelas e galáxias. Podemos observar que possuem comportamento semelhantes ao das análises realizadas para os dados do DES Y1 [22], do COSMOS em [24] e para o survey PAU [23].

Após o processo de classificação, cada fonte pertencente aos dados de teste possuirão uma probabilidade de pertencer à uma classe: estrelas ou galáxias. Nos gráficos 5.7 e 5.8 plotamos essa distribuição de probabilidades. Em vermelho temos objetos classificados como galáxias e em azul estrelas, ambas rotulados pelo SDSS. Os gráficos foram esboçados para cinco métodos. São eles o *CLASS_STAR*, SGLC, ERT, Ensemble e o DT. Inserimos a essa análise os algoritmo ERT e Ensemble por que são os que mais se sobressaem sobre os outros. SGLC e *CLASS_STAR* devem ser inseridos por uma questão de comparação, já que nosso objetivo é de alguma forma superar sua performace. O algoritmo DT foi escolhido por possuir a menor performace dentre todos. O objetivo é mostrar que em alguns casos, esse último algoritmo ainda é competitivo com os modelos fornecidos no catálogo do miniJPAS.

Podemos ter uma visão qualitativa da performace dos algoritmos a partir deste histograma. Observemos que quanto mais separadas as ditribuições e quanto menor as interceções melhor será nosso algoritmo. Isto é, quanto menor a distribuição de galáxias abaixo de $p_{cut} < 0.5$ e estrelas acima de $p_{cut} > 0.5$ melhor tende a ser a performance do algoritmo.

De maneira geral, observamos que a distribuição da probabilidade das galáxias para *CLASS_STAR* está muito mais concentrado do que a distribuição da probabilidade para as

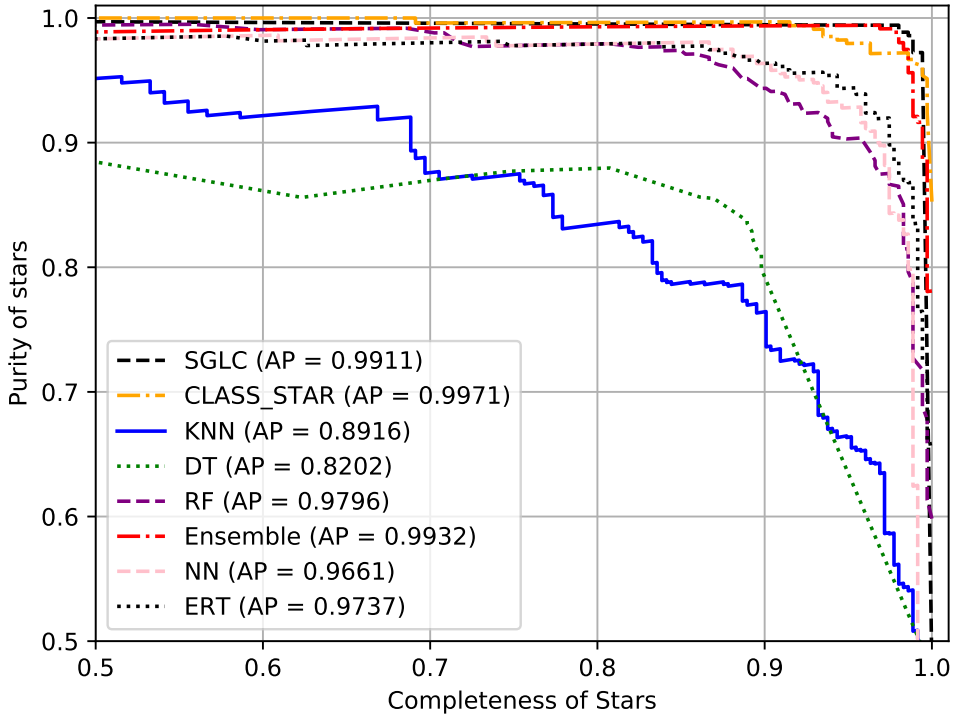

Figura 5.5. Curvas de completeza e pureza para estrelas usando dados miniJPAS com rótulos do SDSS. A análise foi realizada para $15 < rSDSS < 21$ usando apenas bandas fotométricas.

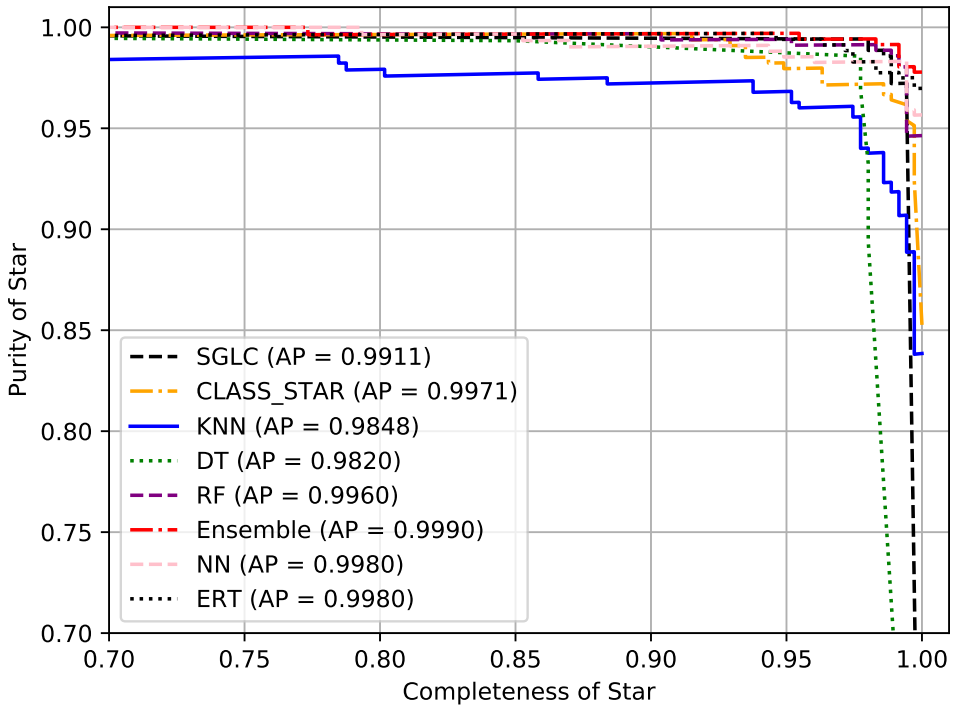

Figura 5.6. Curvas de integridade e pureza para estrelas usando dados miniJPAS com rótulos do SDSS. A análise foi realizada para $15 < rSDSS < 21$ usando bandas fotométricas juntamente com parâmetros morfológicos.

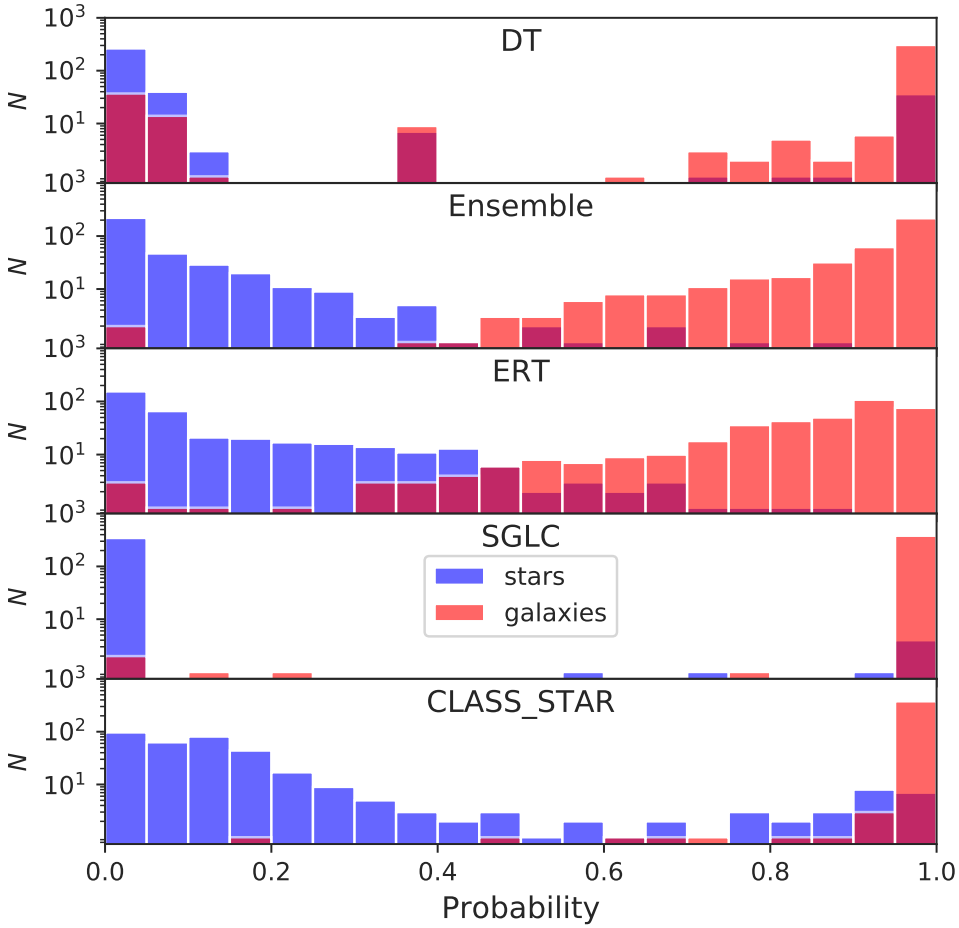

Figura 5.7. Histograma da prababilidade de uma fonte pertencer à classe das galáxias para diferentes métodos classificadores. Em azul temos estrelas e em vermelho temos galáxias do miniJPAS rotuladas pelo survey SDSS. Utilizamos apenas bandas fotométricas para a faixa $15 < rSDSS < 21$. Em roxo temos a superposição das duas distribuições.

estrelas. Isso nos leva a conclusão que SExctractor possui uma leve tendência em classificar galáxias melhor do que estrelas.

Podemos observar nitidamente um melhor comportamento das distribuições das fontes quando inserimos os parâmetros morfológicos. Na figura 5.7 podemos ver que enquanto utilizamos apenas bandas fotométricas, temos distribuições mais suaves havendo mais erros de classificação. Quando inserimos morfologia, a distribuição torna-se mais concentrada, havendo muito menos erros de classificação como pode ser observado na figura 5.8. Resultados semelhantes podem ser encontrados para o estudos dos dados do survey CFHTLenS [20].

Uma outra forma de analisar de forma qualitativa a performace de nosso algoritmos é traçando um gráfico do tipo *cor* x *cor* para os objetos classificados como estrelas. Nas figuras 5.9 e 5.10 plotamos os objetos classificados como estrela, tomando um *threshold* = 0.5, i.e, $p_{cut} < 0.5$ para as cores $g - r$ em função de $r - i$. A linha azul representa a interpolação dos dados do miniJPAS rotulados por SDSS como estrelas, para um polinômio de quinto grau. As marcações com diferentes cores representam as médias de cada classificador para

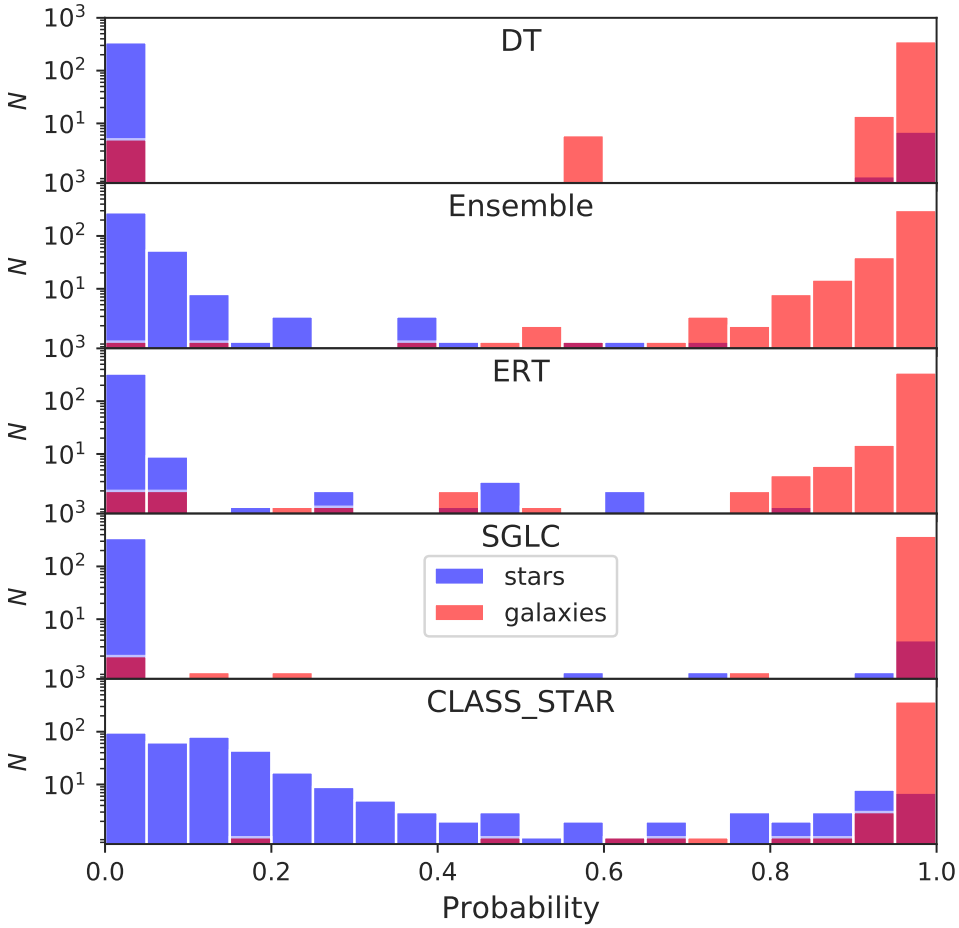

Figura 5.8. Histograma da prababilidade de uma fonte pertencer a classe das galáxias para diferentes métodos classificadores. Em azul temos estrelas e em vermelho temos galáxias do miniJPAS rotuladas pelo survey SDSS. Utilizamos parâmetros morfológicos juntamente à bandas fotométricas para a faixa $15 < rSDSS < 21$. Em roxo temos a superposição das duas distribuições.

diferentes bins. Na figura 5.9 utilizamos apenas bandas fotométricas nos algoritmos de ML. Podemos observar que há uma dispersão muito leve das médias em alguns bins para os classificadores em comparação a curva de interpolação de quinto grau. Isso indica que os resultados para os modelos estudados e a classificação realizada por SDSS estão em muito bom acordo.

Por outro lado, na figura 5.10 no qual inserimos parâmetros morfológicos na análise, essas médias estão sobrepostas umas as outras muito bem ajustadas à interpolação. Isso indica que os resultados dos algoritmos estão em excelente acordo com os rótulos do SDSS. Note a existência de um comportamento comum para as estrelas de cada classificador. Esse padrão é conhecido como Locus Estelar e é bastante conhecido no estudo de estrelas. Resultados semelhantes podem ser encontrados para o RF aplicado aos dados do survey SPLUS [16] e para diferentes classificadores de ML a aplicados aos dados do survey DES [22].

Como mencionado no capítulo 3, dentro do processo de treinamento dos algoritmos

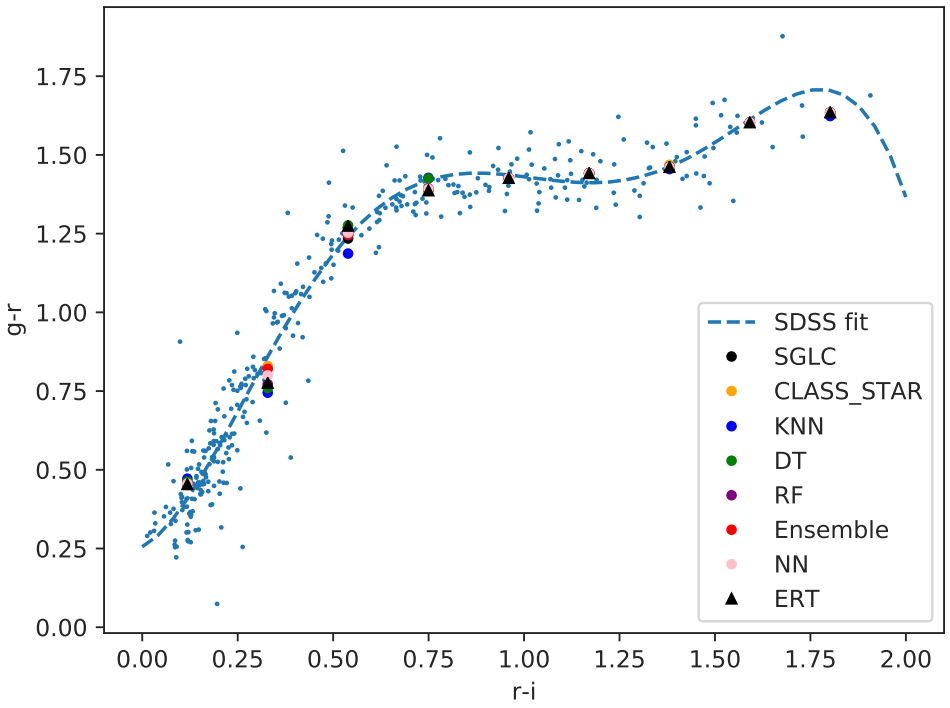

Figura 5.9. Locus estelar para objetos com $p_{cut} < 0.5$. A análise foi feita utilizando apenas bandas fotométricas para 15<rSDSS<21.

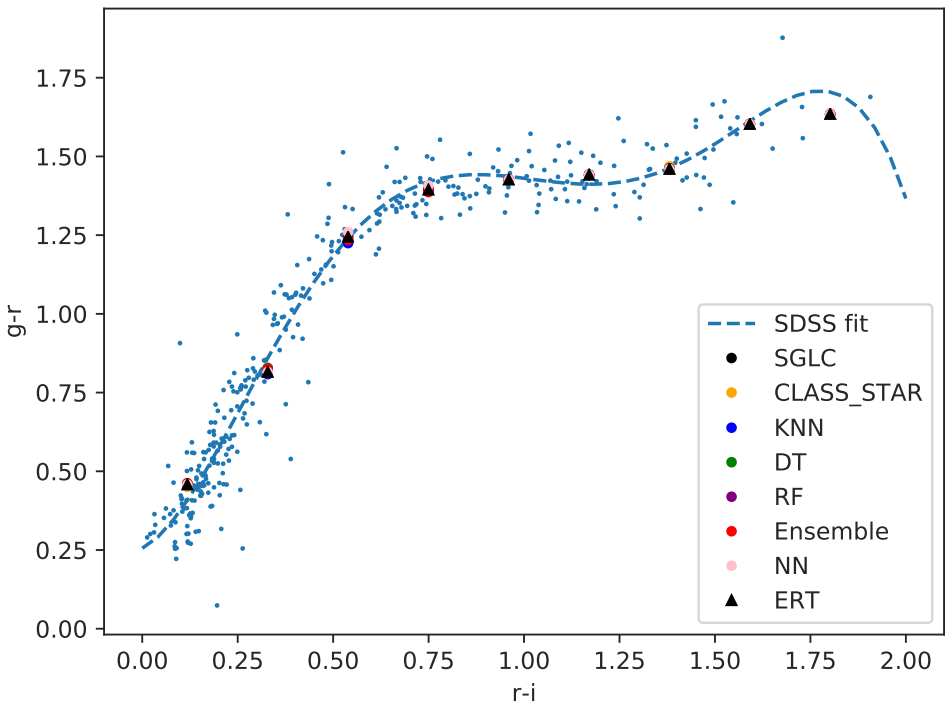

Figura 5.10. Locus estelar para objetos com $p_{cut} < 0.5$. A análise foi feita utilizando bandas fotométricas juntamente aos parâmetros morfológicos para a faixa $15 < rSDSS < 21$.

de árvores, é possivel se medir de forma analítica o quão cada característica foi importante. Na tabela no apêndice A na tabela A.1 podemos observar as características mais importantes utilizando-se apenas bandas fotométricas e bandas fotométricas juntamente à morfologia. O método utilizado para avaliação da importância foi o RF e os hiperparâmetros os mesmos utilizados em análises anteriores.

Como era de se esperar, os parâmetros morfológicos desempenham um papél mais importante quando comparado as bandas fotométricas. Como feature mais importante e de forma quase que exclusiva temos o parâmetro concentração c_r. Para o caso em que utilizamos apenas bandas fotométricas, as caracteristicas que assumem maior importância são a J0390, em seguida de J0460, iSDSS e J0810.

Os dados de ambas tabelas foram normalizados pela característica mais importante em seus respectivos casos. Em particular, podemos observar uma importância muito grande do parâmetro concentração c_r para o caso em que utilizamos morfologia juntamente à fotometria. Essa configuração de importância se deve a escolha dos hiperparâmetros que maximizam a performace do algoritmo. Resultados de testes extras com outras configurações de hiperparâmetros estudadas, mostraram uma maior importância para as outras features, tanto morfológicas como fotométricas. Entretanto, como comportamento geral, sempre atribuiram uma maior importância aos parâmetros morfológicos do que as bandas fotomométricas.

Para obter uma visão física das regiões do espectro que mais importam para classificação, mostramos na figura 5.11 a importância relativa dos filtros em função do comprimento de onda dos filtros, juntamente com o espectro médio das estrelas e galáxias. Podemos observar a existencia de regiões sistematicamente mais importantes que outras e que há correlação entre as regiões mais importantes e as características médias dos espectros.

Os algoritmos de ML possuem internamente vários hiper-parâmetros que variam de modelo para modelo. Esses hiperparâmetros influenciam diretamente na performace dos algoritmos. Diferentes hiperparâmetros foram prospostos neste trabalho para cada algoritmo, entretanto apresentamos apenas os que tiveram os melhores desempenhos. A lista de hiperparâmetros para cada algoritmo de ML está no apêndice A. Para se escolher o conjunto hiperparâmetros que geraram as melhores perfomances, utilizamos o método de validação cruzada K-Fold. Como métrica, utilizamos a AUC associada as curvas ROC e para esse processo de validação cruzada utilizamos $k = 10$. Cada conjunto de hiperparâmetros o qual aplicamos o método K-Fold irá gerar 10 curvas ROC aos quais calcularemos a AUC de forma individual. O resultado final após a aplicação do método será a média da AUC das 10 curvas. A configuração de hiperarâmetros finais utilizados nos algoritmos dessa tese foram justamente aquelas os quais tiveram maior média.

Muitas das vezes costuma-se olhar essas diferentes curvas ROC a fim de observar se há overfitting ou underfitting, comparando se as performaces oriundas dos dados de treino com as performaces dos dados de teste.

As tabelas em 5.2 nos permitem observar a performace dos dados de teste e dos dados de treino para o caso em que utilizamos dados do SDSS. Nelas analisamos as AUCs

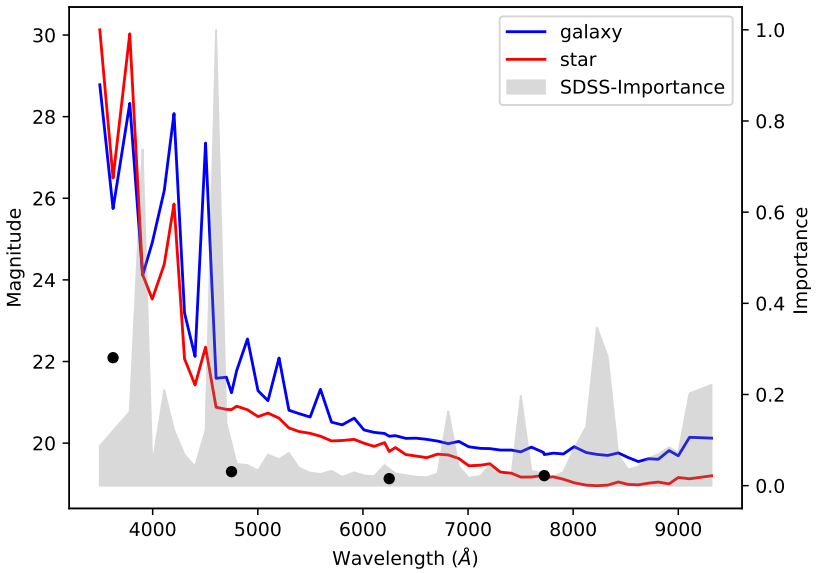

Figura 5.11. A área sombreada representa a importância relativa dos filtros das bandas estreitas em função do comprimento de onda dos filtros para as análises que usam apenas informações fotométricas. A análise foi feita com rótulos do SDSS. A importância das 4 bandas de banda larga filtros é mostrado usando círculos pretos. As linhas vermelha e azul mostram o espectro fotométrico médio de estrelas e galáxias, respectivamente.

e as AP assim como seus desvio-padrão (STD) respectivamente. A partir desse resultado conseguimos analisar o nível de overfitting ou underfitting em nossos algoritmos. Nesta tabela, na parte superior, utilizamos apenas bandas fotométricas como entrada do algoritmo. Na parte inferior utilizamos bandas fotométricas juntamente a parâmetros morfológicos. Como podemos observar, esses modelos quase não sofrem de overfitting ou underfitting.

Classificador	Média K-Fold (AUC ± STD)	Teste (AUC)	Média K-Fold (AP ± STD)	Teste (AP)
KNN	0.8995 ± 0.0176	0.904	0.9024 ± 0.0187	0.894
DT	0.9022 ± 0.0217	0.879	0.8837 ± 0.0264	0.857
RF	0.9784 ± 0.0055	0.9814	0.9782 ± 0.0075	0.9827
NN	0.9759 ± 0.0064	0.9750	0.9701 ± 0.0117	0.955
ERT	0.9829 ± 0.0035	0.9834	0.9835 ± 0.0051	0.9871
Classificador	Média K-Fold (AUC ± STD)	Teste (AUC)	Média K-Fold (AP ± STD)	Teste (AP)
KNN	0.9928 ± 0.0045	0.9905	0.9924 ± 0.0060	0.9928
DT	0.9891 ± 0.0068	0.9856	0.9857 ± 0.0089	0.9770
RF	0.9980 ± 0.0023	0.9978	0.9980 ± 0.0029	0.9982
NN	0.9971 ± 0.0028	0.9982	0.9952 ± 0.0072	0.9985
ERT	0.9986 ± 0.0019	0.9984	0.9983 ± 0.0029	0.9988

Tabela 5.2. Performace para os dados de teste e treino dos algoritmos de ML para o caso em que utilizamos apenas bandas fotométricas (acima) e o caso em que utilizamos bandas fotometricas juntamente à parâmetros morfológicos (abaixo). Nesta análise utilizamos rótulos do survey SDSS.

5.2 Usando rótulos HSC-SUBARU

A classificação de fontes luminosas em estrelas e galáxias para magnitudes na faixa $15 < rSDSS < 21$ é uma questão conhecida e relativamente bem resolvida. Diferentes modelos classificadores com performaces muito boas são conhecidos na literatura. Como é o exemplo temos o classificador SGLC e de *CLASS_STAR* adotados pelo survey miniJPAS. Entretando o cenário muda quando caminhamos para magnitudes mais fracas. Como a quantidade de luz que chega é menor, teremos menos informações chegando aos telescópio. Por consequência as performances dos algoritmos diminuem ao ponto de nos levarem a procurar por soluções alternativas. O ML surge com esse intuito. Treinando os algoritmos com rótulos de surveys que conseguem captar mais luz, conseguimos expandir o limite de classificação de fontes para surveys com menor precisão.

Nessa seção comparamos os diferentes modelos de ML com os modelos gerados por SGLC e *CLASS_STAR*. Os resultados foram realizados para a faixa $18.5 < rSDSS < 23.5$ utilizando dados do miniJPAS com rótulos de HSC-SSP. Duas análises foram realizadas de maneira independentes, uma inserindo apenas bandas fotométricas e outra utilizando bandas fotométricas juntamente à parâmetros morfológicos.

Na figura 5.12 observamos a Curva ROC para os classificadores em estudos no qual utilizamos apenas bandas fotométricas. Observamos uma performace razoável para os algorimos de ML com uma AUC acima de 0.8. O melhor algoritmo é o ERT com $AUC = 0.8944$ e os com menor performace são os DT e KNN respectivamente. Os algoritmos os quais contruímos possuem desempenho bem menor que SGLC e SExctractor como no tópico anterior. Isso se deve ao fato de que estes não possuem informações morfológicas enquanto SGLC e SExctractor as possuem. Interessante é o resultado de Ensemble Learning que nesse caso combina modelos com informação morfológica e sem, o qual consegue superar *CLASS_STAR*. Esse algoritmo poderia surgir como um classificador intermediário para um modelo sensível a quasares classificados como estrelas.

Na figura 5.13 traçamos a Curva ROC inserindo bandas fotométricas juntamente à parâmetros morfológicos. Observamos uma melhora considerável na performace dos algoritmos que agora possuem uma $AUC > 0.94$. As melhores performaces são desempenhadas pelo Ensemble Learning com $AUC = 0.9744$ seguido por ERT e SGLC. O menor desempenho é realizado por DT e KNN, mas ainda sim constituem um bom modelo. É de se esperar que SExctractor comece a ser superado uma vez que sua boa performace limita-se a magnitudes mais fortes ($r > 21$). O interessante é notar que a combinação de métodos de ML e SGLC geram uma performace maior que apenas um desses algoritmos sozinhos.

Como segunda análise de performace, calculamos as curvas de Completeza e Pureza para galáxias e estrelas. Como o dataset é não-balanceado, essa análise se faz fundamental. Na figura 5.14 calculamos as curvas de Completeza e Pureza para galáxias inserindo apenas bandas fotométricas na análise. Observamos que todos algoritmos possuem performace excelente. O melhor algoritmo para esse caso foi o RF com $AUC = 0.9708$. Os menores desempenhos foram realizados por KNN e DT. Os métodos de floresta são os melhores com

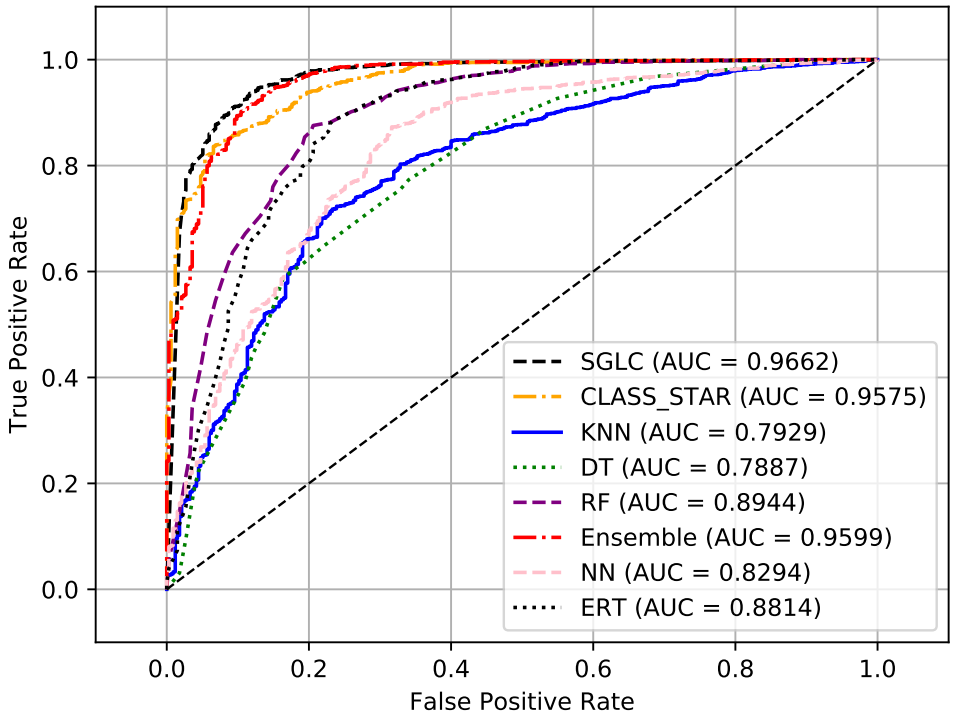

Figura 5.12. Curvas ROC para dados do miniJPAS com rótulos do HSC-SSP utilizando apenas bandas fotométricas. A análise foi realizada para $18.5 < rSDSS < 23.5$.

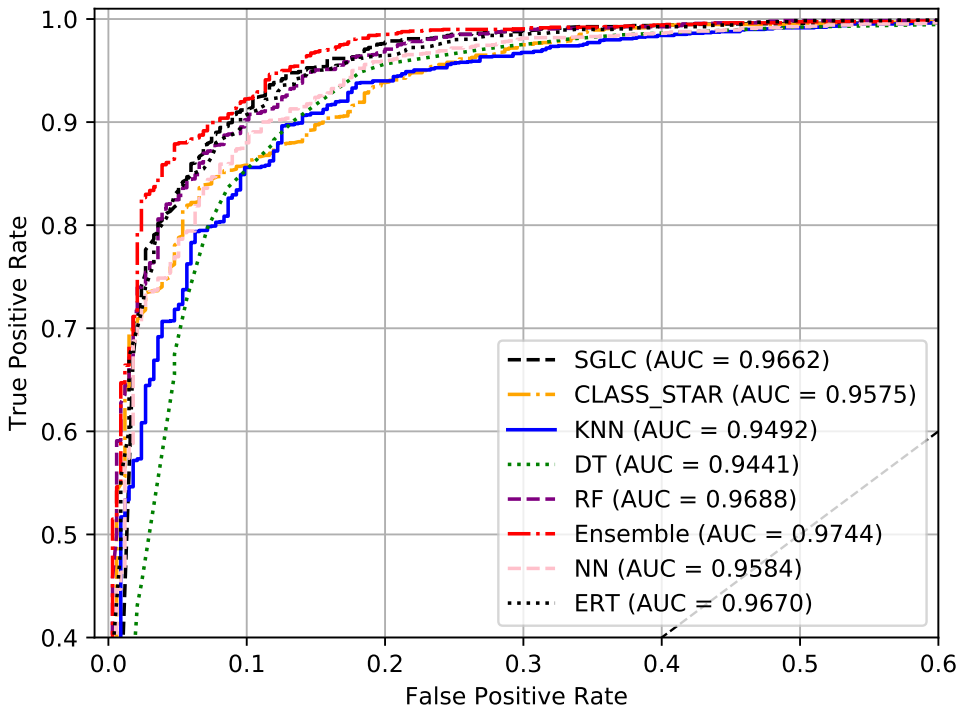

Figura 5.13. Curvas ROC para dados do miniJPAS com rótulos do HSC-SSP utilizando bandas fotométricas juntamente à parâmetros morfologicos. A análise foi realizada para $18.5 < rSDSS < 23.5$.

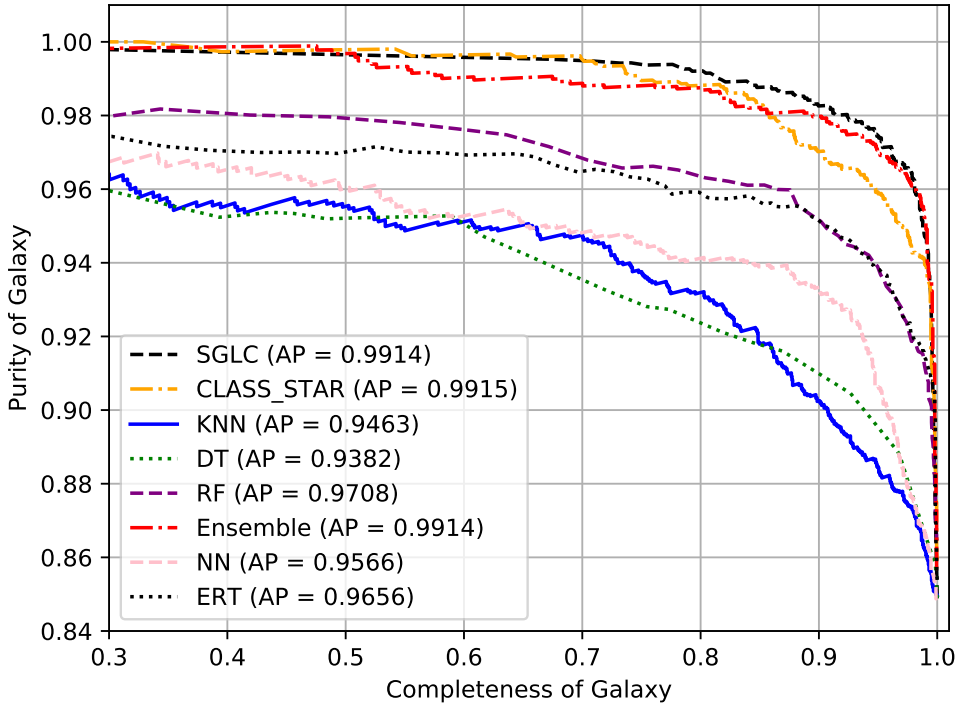

Figura 5.14. Curvas de Completeza e Pureza para galáxias utilizando dados do miniJPAS com rótulos do HSC-SSP . A análise foi realizada para $18.5 < rSDSS < 23.5$ utilizando bandas fotométricas.

$AP > 0.96$ dentre os algoritmos contruídos.

Na figura 5.15 calculamos as curvas de Completeza e Pureza para galáxias inserindo bandas fotométricas juntamente à parâmetros morfológicos na análise. Observamos uma melhora considerável nos resultados e que a performace excelente no geral. A maior performace é realizada por Ensemble com $AP = 0.9945$ seguida por RF e ERT. A menor performace é realizada por DT com $AP = 9796$.

Na figura 5.16 calculamos as curvas de Completeza e Pureza para estrelas. Inserimos unicamente bandas fotométricas. Observamos um desempenho baixo para os algoritmos no geral. A maior performace é de análise de ERT com $AP = 0.7365$ seguida por RF. Os piores desempenhos foram realizados por *DT* e *KNN* com $AP < 0.47$.

Na figura 5.17 calculamos as curvas de Completeza e Pureza para estrelas inserindo parâmetros morfológicos juntamente à bandas fotométricas na análise. A performace aumenta consideravelmente. Observamos que Ensemble possui o melhor desempenho com $AP = 0.9072$ seguido por RF e ERT. Notamos que para esse caso os algoritmos de ML superam os classificadores padrões. A menor performace é realizada por DT com $AP = 0.8180$ seguido por *KNN*.

Embora as *AUCs* associadas as Curvas ROC possuam um valor menor para magnitudes mais fracas, as AP^{gal} associadas as Curvas de Completeza e Pureza são excelentes. Isso ocorre tanto no caso em que utilizamos rótulos de HSC-SSP quanto no caso SDSS. Mesmo quando utilizamos apenas bandas fotométricas. Isso se deve ao fato de possuirmos muitos

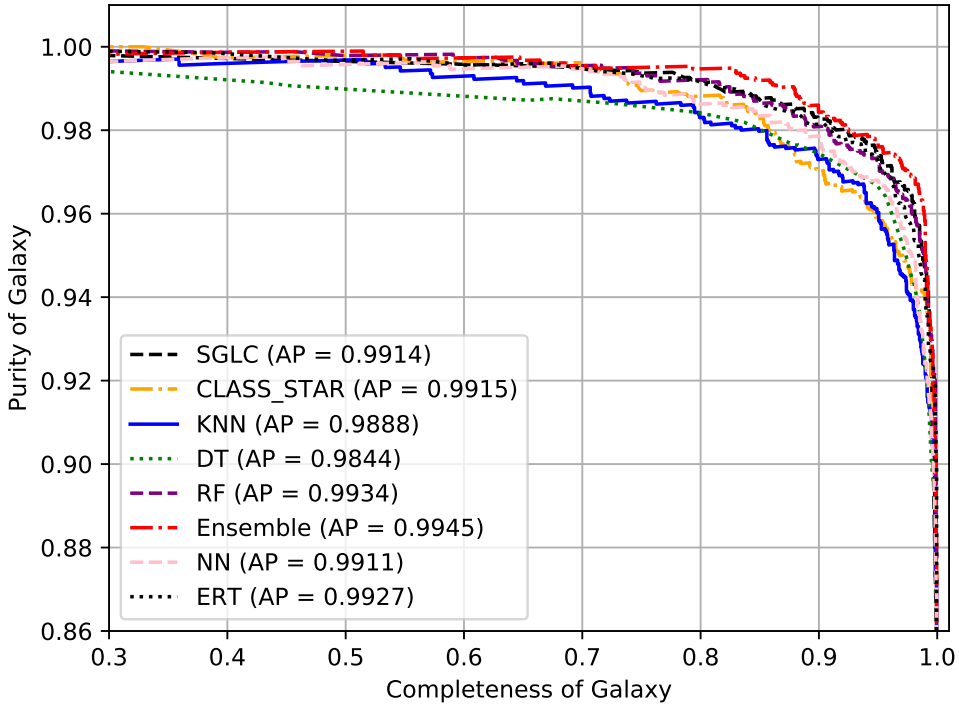

Figura 5.15. Curvas de Completeza e Pureza para galáxias utilizando dados do miniJPAS com rótulos do HSC-SSP . A análise foi realizada para $18.5 < rSDSS < 23.5$ utilizando bandas fotométricas juntamente a parâmetros morfológicos.

mais galáxias no dataset de teste, ie., apenas 15.3% de estrelas. Portanto, mesmo que haja contaminação de estrelas em regiões de distribuição das galáxias elas estão em menor número.

Por outro lado, as curvas de Pureza e Completeza para estrelas possuem valores menores para AP^{star}. Nos classificadores KNN, DT e NN temos uma performace ruim para o caso em que utilizamos apenas bandas fotométricas. Isso se deve ao fato de termos objetos classificados erroneamente como estrelas. Quando traçamos as mesmas curvas utilizando parâmetros morfológico juntamente a bandas fotométricas, a situação melhora, gerando performaces boas com $AP^{star} > 0.81$.

Na tabela 5.3 escrevemos os resultados das performaces oriundas das curvas ROC e Completeza x Pureza para galáxias. Notemos que a performace cai quando avançamos para magnitudes mais fracas. Isso é de se esperar já que o fluxo é menor, i.e, temos menos informação chegando aos telescópios. Embora a performace diminua, os resultados são muito bons. Observemos que o método de Ensemble possui a melhor performace quando utilizamos bandas fotométricas juntamente à parâmetros morfológicos e que DT possui a menor. Por outro lado, RF possui a melhor quando utilizamos apenas bandas fotométricas com DT e KNN possuindo as menores performaces.

Na figura 5.18 observamos a distribuição das probabilidades oriundas dos algoritmos para fontes rotuladas como estrelas ou galáxias pelos survey HSC-SSP. Para esse caso inse-

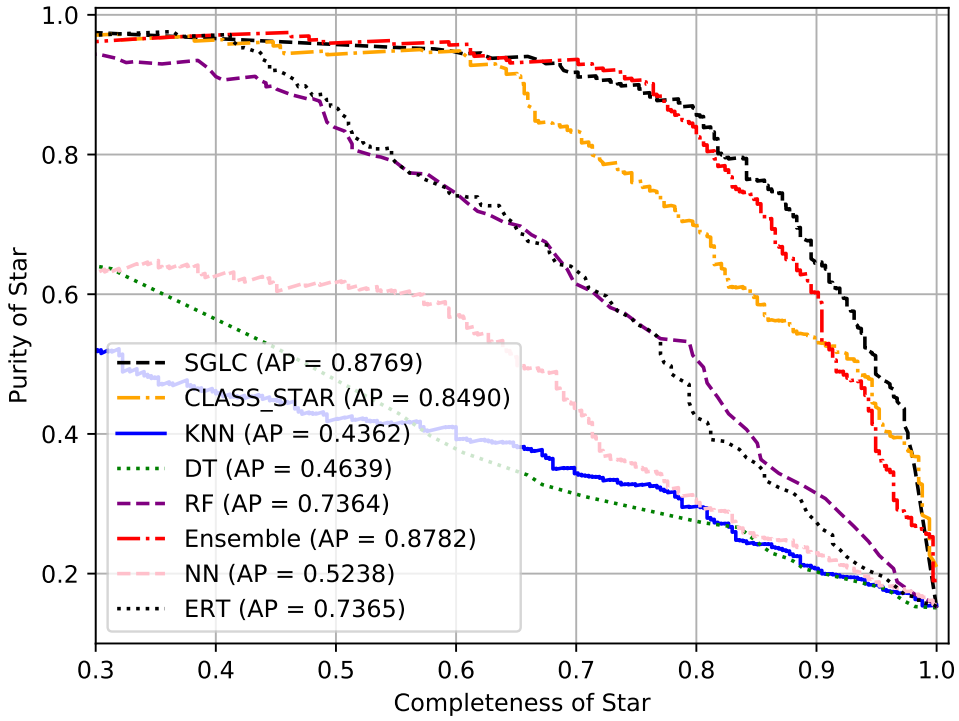

Figura 5.16. Curvas de Completeza e Pureza para estrelas utilizando dados do miniJPAS com rótulos do HSC-SSP. A análise foi realizada para $18.5 < rSDSS < 23.5$ utilizando apenas bandas fotométricas.

miniJPAS-HSC-SSP	AUC_{M+P}	AUC_P	AP^{gal}_{M+P}	AP^{gal}_P	MSE_{M+P}	MSE_P
SGLC	0.9662	–	0.9914	–	0.0439	–
CLASS_STAR	0.9575	–	0.9915	–	0.0533	–
KNN	0.9492	0.7929	0.9888	0.9463	0.0566	0.1071
DT	0.9441	0.7887	0.9844	0.9382	0.0719	0.1705
RF	**0.9688**	**0.8944**	**0.9934**	**0.9708**	**0.0419**	**0.0714**
Ensemble	0.9744	0.9599	0.9945	0.9914	0.0370	0.0501
ANN	0.9584	0.8294	0.9911	0.9566	0.0487	0.1002
ERT	0.9670	0.8814	0.9927	0.9656	0.0430	0.0715

Tabela 5.3. Desempenho dos classificadores para o catálogo miniJPAS comparado com o catálogo HSC-SSP para a faixa $18.5 < rSDSS < 23.5$. O melhor desempenho é marcado em negrito. F representa a análise que utiliza apenas bandas fotométricas, enquanto $M+P$ representa a análise que utiliza faixas fotométricas juntamente com parâmetros morfológicos. Em negrito temos a melhor performace desempenhada.

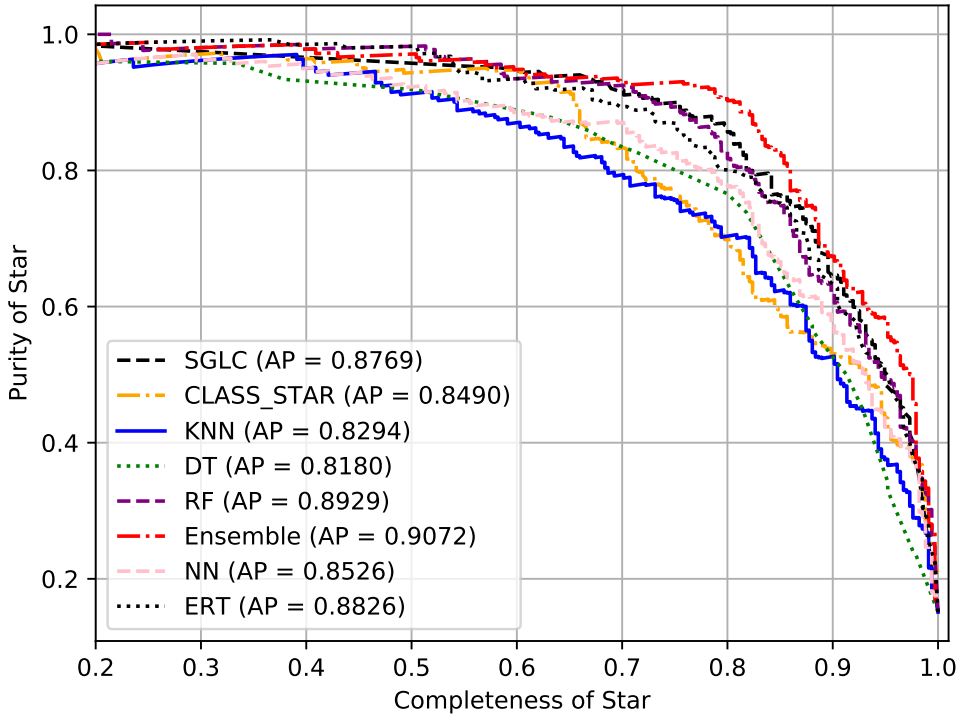

Figura 5.17. Curvas de Completeza e Pureza para estrelas utilizando dados do miniJPAS com rótulos do HSC-SSP. A análise foi realizada para $18.5 < rSDSS < 23.5$ utilizando bandas fotométricas juntamente à parâmetros morfologicos.

rimos na análise apenas bandas fotométricas. Na figura 5.19 inserimos bandas fotométricas juntamente à parâmetros morfológicos. O número de objetos classificados erroneamente ganha destaque visualmente apenas porque os eixos estão em escala logarítmica, mas na realidade são pequenos em ambos os casos.

Comparando a análise realizada utilizando rótulos de HSC-SSP com os resultados em que utilizamos SDSS, o número de galáxias é muito maior do que o de estrelas nos gráficos. Isso se deve a natureza do dataset que possui apenas 15.3% de estrelas. Diferentemente do dataset o qual utilizamos rótulos do SDSS que possui 46.3% de estrelas.

Como análise complementar também traçamos o Locus Estelar para os dados com probabilidade $p_{cut} < 0.5$. Comparando com os resultados obtidos para a faixa $15 < rSDSS < 21$, observemos que a dispersão é maior para magnitudes mais fracas nas figuras 5.20 e 5.21. Sobretudo quando utilizamos apenas bandas fotométricas na análise. A curva interpolada também é de quinto grau, mas oriunda dos dados o qual utilizamos rótulos do HSC-SSP. Quando há apenas bandas fotométricas na análise, observamos uma dispersão maior para KNN, DT e NN resultados que está de acordo com as performaces da tabela 5.3. Quando inserimos parâmetros morfológicos à análise, observamos que a dispersão diminui. Por fim, Ensemble, RF e ERT se ajustam melhor as curvas.

Na tabela A.2 no apêndice A calculamos as características mais importantes no processo de predição de treinamento dos algoritmos de RF. Os hiperparâmetros utilizados foram

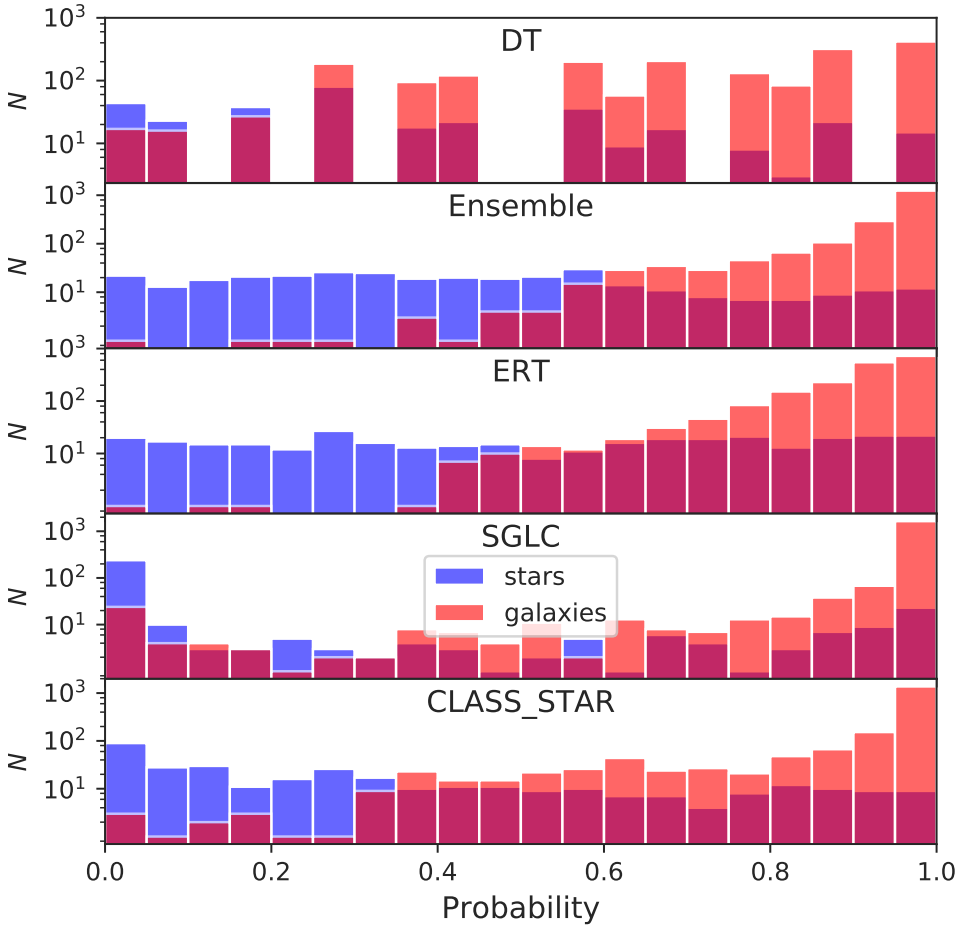

Figura 5.18. Histograma da prababilidade de uma fonte pertencer à classe das galáxias para diferentes métodos classificadores. Em azul temos estrelas e em vermelho temos galáxias do miniJPAS rotuladas pelo survey HSC-SSP. Utilizamos apenas bandas fotométricas para a faixa $18.5 < rSDSS < 23.5$. Em roxo temos a superposição das duas distribuições.

os mesmos que geraram os resultados da tabela 5.3. Observemos que as bandas J0390, J0460, iSDSS e J0810 assumem os valores mais importantes, quando utilizamos apenas bandas fotométricas na análise. Enquanto que quando utilizamos parâmetros morfológicos juntamente, observamos c_r, FWHM, $\mu_{max}/r_{3.0''}$ assumindo as primeiras posições. O parâmetro alongamento A/B perde importância à medida que caminhamos para magnitudes mais fracas como era de se esperar. O fluxo diminui, ao ponto de confundirmos fontes pontuais de extendidas. O interessante é notar que as bandas J0740 e J0390 foram mais importantes que o parâmetro alongamento A/B. Diferentemente dos resultados onde utilizamos rótulos do SDSS, na análise morfológica as bandas fotométricas e os pâmetros morfológicos, no geral, possuem uma maior importância quando comparadas à c_r.

Para obter uma visão física das regiões do espectro que mais importam para classificação, mostramos na figura 5.22 a importância relativa dos filtros em função do comprimento de onda dos filtros, juntamente com o espectro médio de fotos e galáxias.

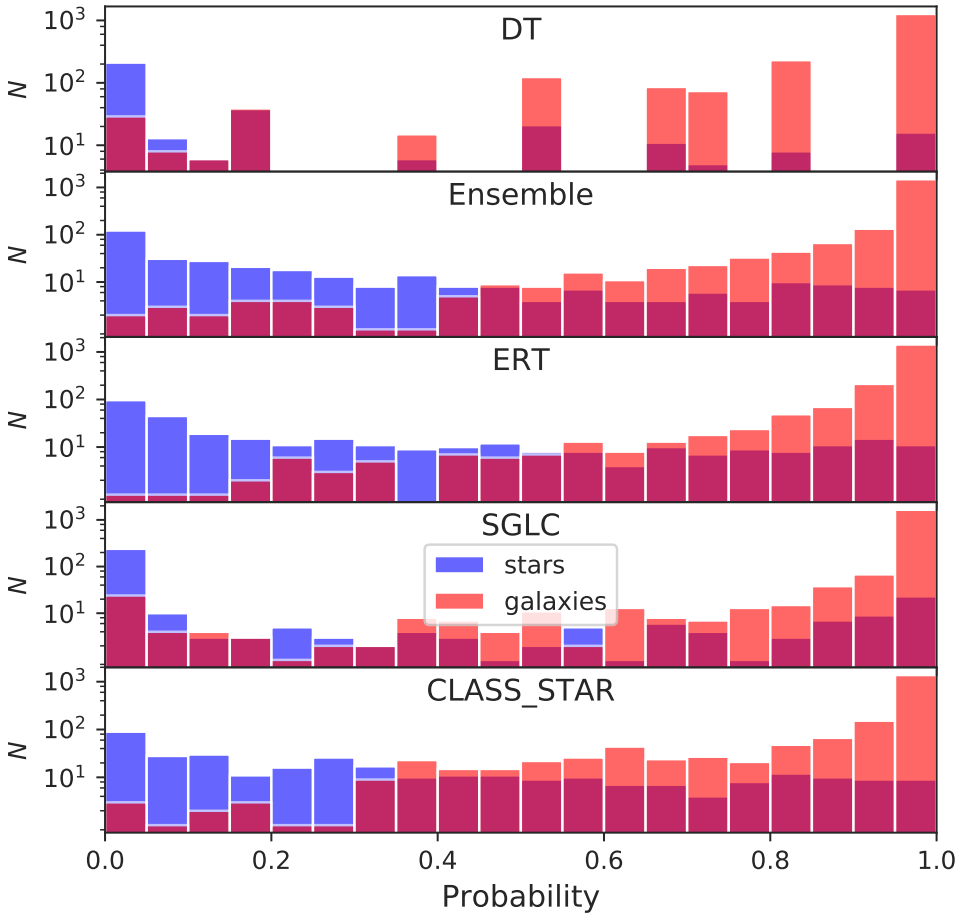

Figura 5.19. Histograma da prababilidade de uma fonte pertencer à classe das galáxias para diferentes métodos classificadores. Em azul temos estrelas e em vermelho temos galáxias do miniJPAS rotuladas pelo survey HSC-SSP. Utilizamos parâmetros morfológicos juntamente à bandas fotométricas para a faixa $18.5 < rSDSS < 23.5$. Em roxo temos a superposição das duas distribuições.

Assim como para a análise em que utilizamos rótulos do SDSS, aplicamos o método K-Fold para magnitudes mais fracas. Também utilizamos $k = 10$. e os hiperparâmetros com as melhores performaces podem ser encontrados no apêndice B.

Também construímos a mesma tabela para as performaces dos algoritmos treinados com rótulos do HSC-SSP. Nas tabelas 5.4, na parte superior utilizamos apenas bandas fotométricas enquanto que na parte inferior utilizamos bandas fotométricas juntamente à parâmetros morfológicos. Podemos também observar que os modelos sofrem quase nada de overfitting e underfitting.

5.3 O Campo AEGIS01

O survey miniJPAS é composto por 4 campos (mJP-AEGIS1, mJP-AEGIS2, mJP-AEGIS3 e mJP-AEGIS4), cada um dos campos de visão de 0.25 deg^2 (para detalhes, ver Bonoli et al.

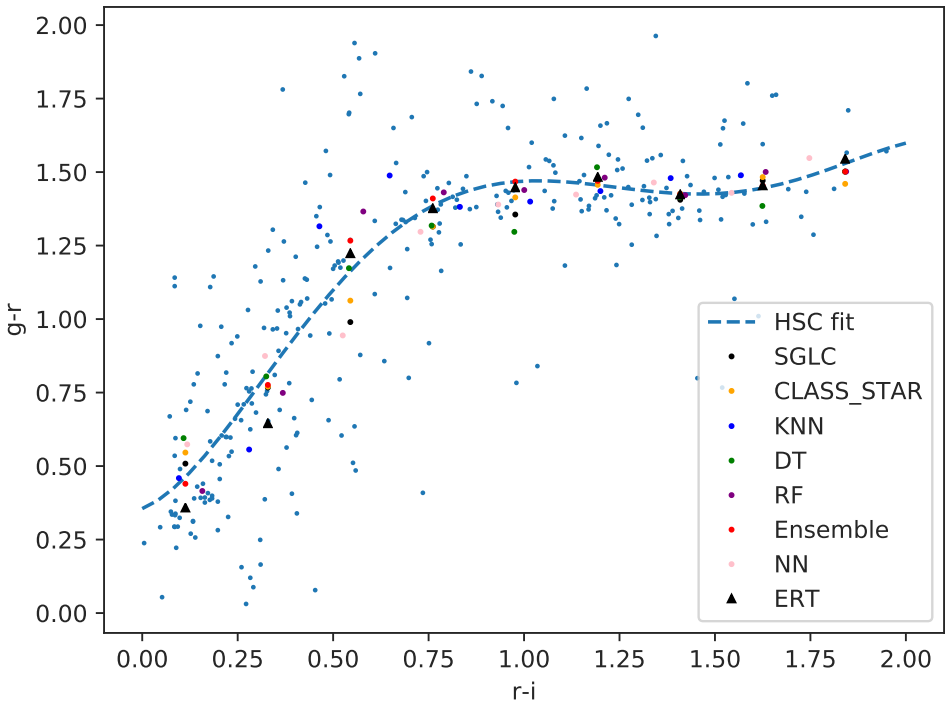

Figura 5.20. Locus estelar para objetos classificados como estrelas. A análise foi feita utilizando apenas bandas fotométricas para a faixa 18.5 < *rSDSS* < 23.5.

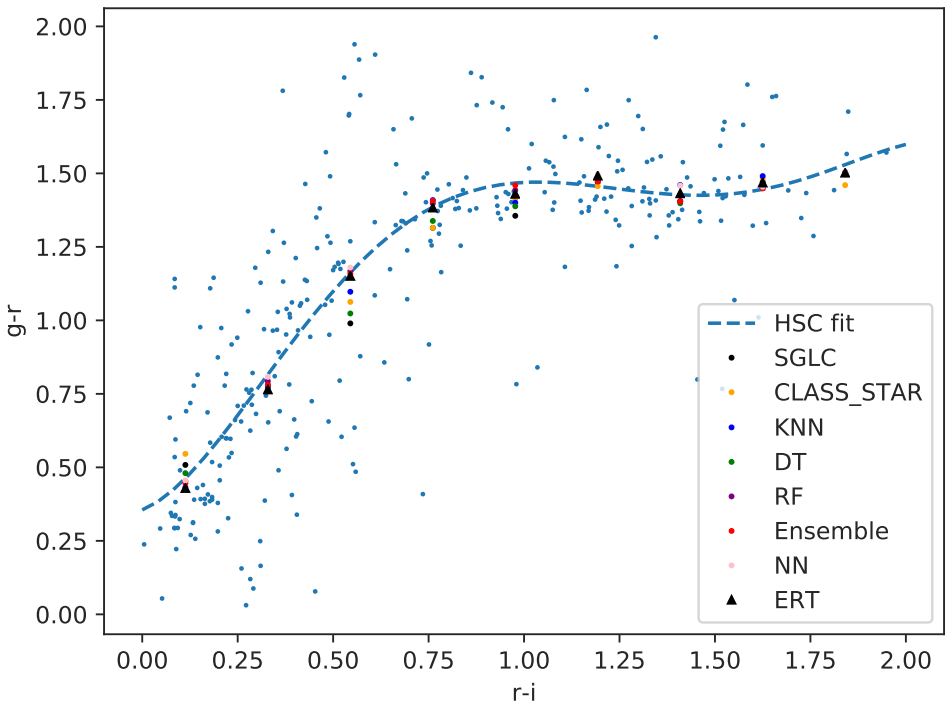

Figura 5.21. Locus estelar para objetos classificados como estrelas. A análise foi feita utilizando bandas fotométricas juntamente à parâmetros morfológicos para a faixa 18.5 < *rSDSS* < 23.5.

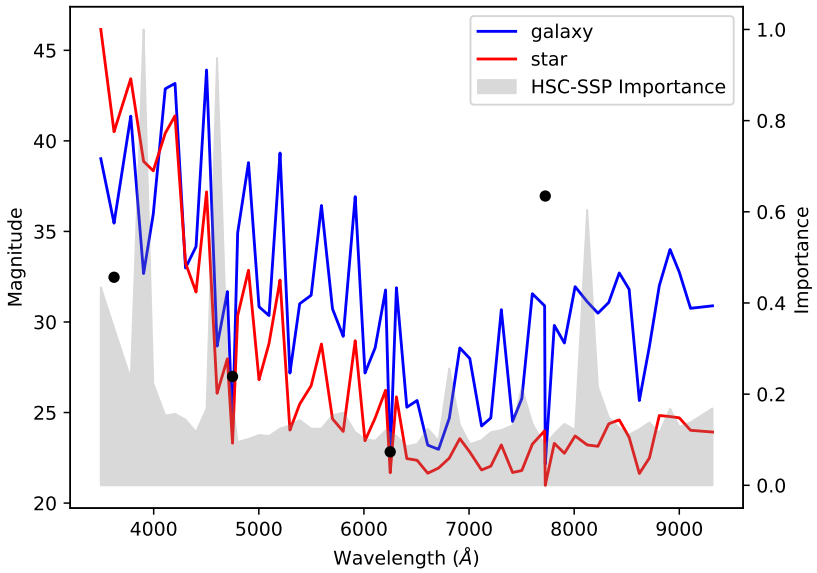

Figura 5.22. A área sombreada representa a importância relativa dos filtros das bandas estreitas em função do comprimento de onda dos filtros para as análises que usam apenas informações fotométricas. A análise foi feita com rótulos do HSC-SSP. A importância das 4 bandas de banda larga filtros é mostrado usando círculos pretos. As linhas vermelha e azul mostram o espectro fotométrico médio de estrelas e galáxias, respectivamente.

Classificador	Média K-Fold (AUC ± STD)	Teste (AUC)	Média K-Fold (AP ± STD)	Teste (AP)
KNN	0.8071 ± 0.0236	0.7929	0.9489 ± 0.0082	0.9463
DT	0.8044 ± 0.0236	0.7887	0.9443 ± 0.0091	0.9382
RF	0.9152 ± 0.0158	0.8944	0.9777 ± 0.0052	0.9708
NN	0.8164 ± 0.0277	0.8294	0.9481 ± 0.0099	0.9566
ERT	0.9081 ± 0.0170	0.8814	0.9740 ± 0.0066	0.9656
Classificador	Média K-Fold (AUC ± STD)	Teste (AUC)	Média K-Fold (AP ± STD)	Teste (AP)
KNN	0.9547 ± 0.0091	0.9492	0.9900 ± 0.0025	0.9888
DT	0.9558 ± 0.0075	0.9441	0.9498 ± 0.0060	0.9844
RF	0.9706 ± 0.0054	0.9688	0.9934 ± 0.0018	0.9934
NN	0.9622 ± 0.0061	0.9584	0.9911 ± 0.0025	0.9911
ERT	0.9720 ± 0.0059	0.9670	0.9941 ± 0.0013	0.9927

Tabela 5.4. Performace para os dados de teste e treino dos algoritmos de ML para o caso em que utilizamos apenas bandas fotometricas (acima) e o caso em que utilizamos bandas fotometricas juntamente à parâmetros morfológicos (abaixo). Nesta análise utilizamos rótulos do survey HSC-SSP

2020). O mJP-AEGIS1 possui 20 016 objetos e possui um r-PSF de banda semelhante ao mJP-AEGIS3 ($\sim 0.7''$) e melhor do que mJP-AEGIS2 e mJP-AEGIS4 ($\sim 0,8''$). Sendo assim, repetimos para o campo mJP-AEGIS1 a análise relativa ao HSC-SSP. Não consideramos a análise relativa a SDSS, já que os resultados já são excelentes e o número de dados seria muito pequeno.

Nas figuras 5.23 e 5.24 temos as Curvas ROC e Completeza x Pureza para o campo do AEGIS1. Podemos ver que a qualidade das predições aumentam com relação as análises realizadas na seção 5.2.

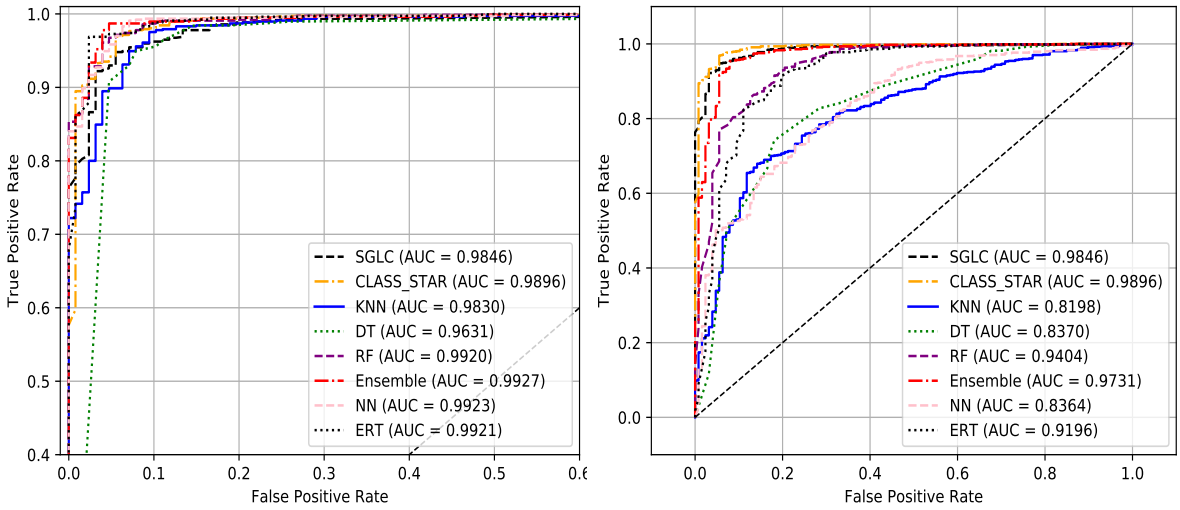

Figura 5.23. Curvas ROC obtidas a partir da aplicação dos algoritmos no campo mJP-AEGIS1 em comparação com o catálogo HSC-SSP no intervalo de magnitude $18.5 < r < 23.5$. Na figura à esquerda utilizamos bandas fotométricas juntamente à parâmetros morfológicos, enquanto à direita utilizamos apenas bandas fotométricas. Para comparação, é mostrada também a classificação por *CLASS_STAR* e SGLC que sempre utilizam parâmetros morfológicos.

Para a melhor performace, por exemplo, em ERT temos uma $AUC = 0.9670$ utilizando morfologia na seção 5.2 enquanto que utilizando dados do campo AEGIS1 encontramos $AUC = 0.9921$.

Na figura 5.25 podemos observar a distribuição dos parâmetros morfológicos para o campo do AEGIS01. Notemos que comparando com a distribuição dos dados da figura 4.9 observamos uma diferença nas distribuições. Enquanto que para AEGIS01 temos um pico apenas, utilizando todas telhas observamos dois picos nas distribuições das estrelas. Isso se deve ao fato de termos diferentes PSF aplicadas à diferentes telhas.

5.4 Construção de Catálogo

O objetivo final deste trabalho é lançar um catálogo de valor agregado com nossa melhor classificação alternativa. Na seção anterior, estudamos a classificação estrela/galáxia nas faixas de magnitude parcialmente sobrepostas $15 < r < 21$ e $18.5 < r < 23.5$. Aqui,

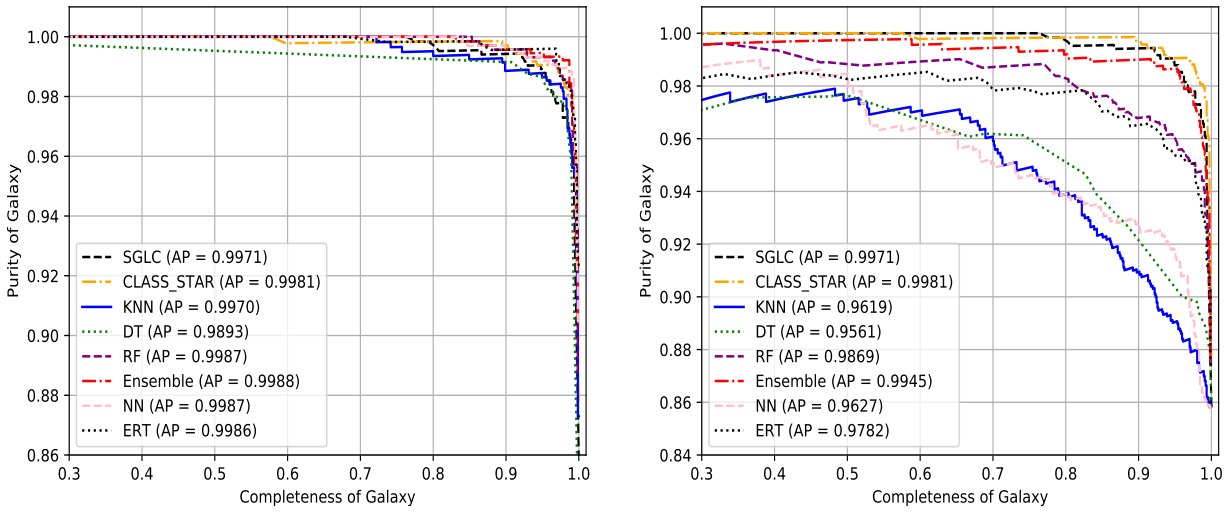

Figura 5.24. Curvas de Completeza e Pureza obtidas a partir da aplicação dos algoritmos no campo mJP-AEGIS1 em comparação com o catálogo HSC-SSP no intervalo de magnitude $18.5 < r < 23.5$. Na figura à esquerda utilizamos bandas fotométricas juntamente à parâmetros morfológicos, enquanto à direita utilizamos apenas bandas fotométricas. Para comparação, é mostrada também a classificação por *CLASS_STAR* e SGLC que sempre utilizam parâmetros morfológicos.

para ter uma dependência uniforme do p_{cut}, desejamos produzir um catálogo que é obtido usando um único classificador.

Para combinar os dois conjuntos de treinamento, é útil ver como as classificações do SDSS e do HSC-SSP concordam no intervalo comum $18.5 \leq r \leq 21$. A interseção dos dois catálogos apresenta 1 427 objetos; a matriz de confusão entre SDSS e HSC-SSP cruzadas com miniJPAS é mostrada na figura 5.26. Podemos ver que a concordância é boa: 0.35% dos objetos são classificados como estrelas pelo SDSS, mas como galáxias pelo HSC-SSP. Inversamente, 2.45% dos objetos são classificados como galáxias pelo SDSS, mas como estrelas pelo HSC-SSP. Para entender melhor a pequena discordância entre as duas classificações, mostramos na figura 5.27 as classificações conflitantes em função da magnitude r. O número de classificações diferentes aumenta próximo ao limite fraco do catálogo do SDSS. Dado que o HSC-SSP é uma pesquisa mais profunda, pode-se concluir que o SDSS está classificando algumas galáxias como estrelas. Portanto, adotamos $r = 20$ como a magnitude máxima para o conjunto de treinamento do SDSS. Além de estar longe do limite fraco do SDSS, $r = 20$ também está longe do limite de saturação do HSC-SSP. Este catálogo abrange a magnitude de $18.5 \leq r \leq 23.5$ e apresenta um total de 12 372 fontes, 9 794 galáxias e 2 578 estrelas.

Em seguida, treinamos e testamos todos os modelos deste catálogo. Usando apenas informações fotométricas, o melhor classificador é o RF, que atinge $AUC = 0.941$, próximo ao desempenho do SGLC que utiliza informações morfológicas. Usando informações fotométricas e morfológicas, o melhor classificador é o ERT, que com $AUC = 0.981$ supera o SGLC. A figura 5.28 mostra a curva ROC e a curva de Completeza x Pureza para galáxias para os classificadores SGLC, ERT (utilizando bandas fotométricas junto à morfologia) e RF

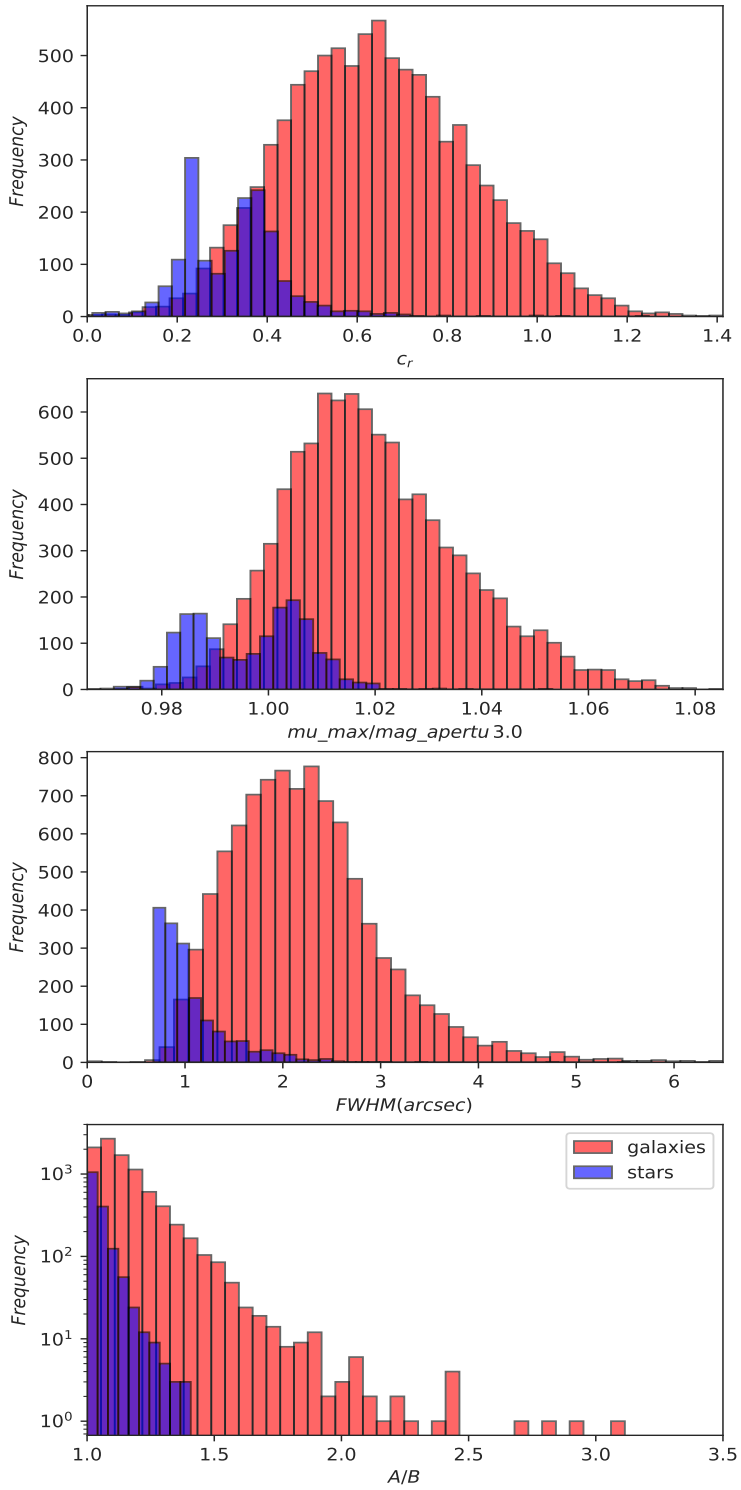

Figura 5.25. Distribuição dos parâmetros morfológicos para os dados do campo AEGIS01.

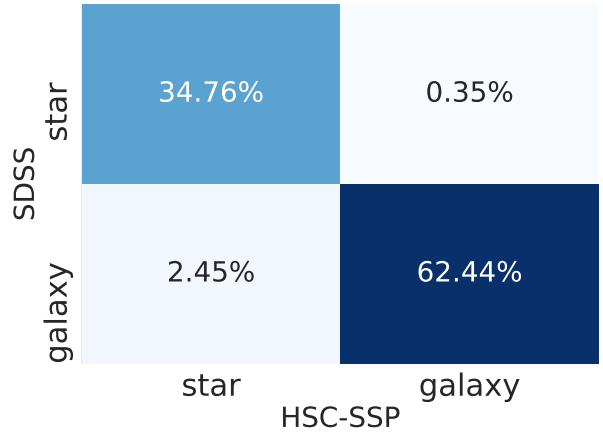

Figura 5.26. Matriz de confusão entre dados cruzados do miniJPAS como SDSS e HSC-SSP na área comum intervalo $18.5 \leq r \leq 21$.

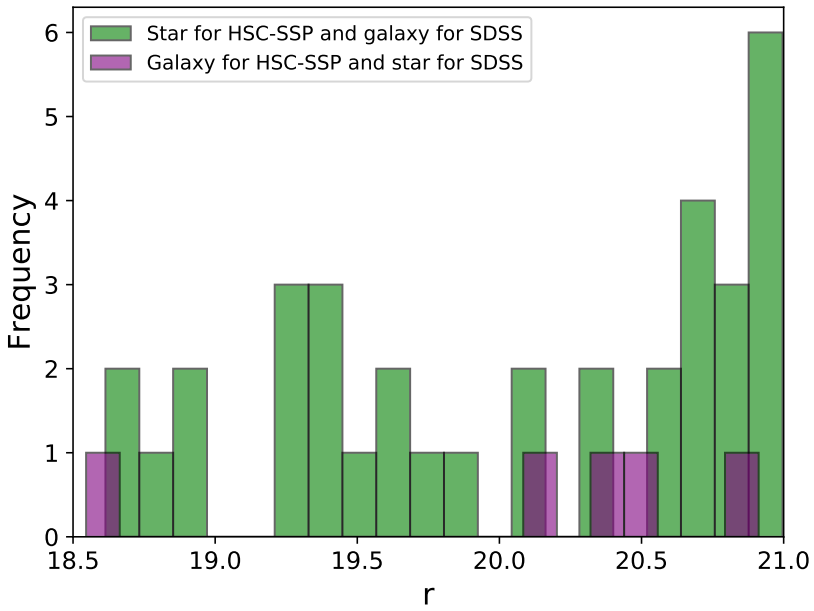

Figura 5.27. Classificações conflitantes pelo SDSS e HSC-SSP como uma função da magnitude r.

(utilizando apenas fotometria), com a adição do limite de probabilidade p_{cut} via código de cores. Esses gráficos destinam-se a ajudar a escolher o limite de probabilidade que melhor atenda às necessidades de integridade e pureza. Essas plotagens foram feitas com o código disponível em `github.com/PedroBaqui/minijpas-astroclass`.

Para construir nosso catálogo, aplicamos nossos dois melhores classificadores (RF sem morfologia e ERT com morfologia) às fontes 29 551 miniJPAS na faixa de magnitude $15 \leq r \leq 23.5$. É importante observar que, dada a integridade do miniJPAS [2], fontes fora desse intervalo de magnitude têm menos probabilidade de entrar em estudos científicos. O catálogo está disponível em `http://archive.cefca.es/catalogues/minijpas-pdr201912/`.

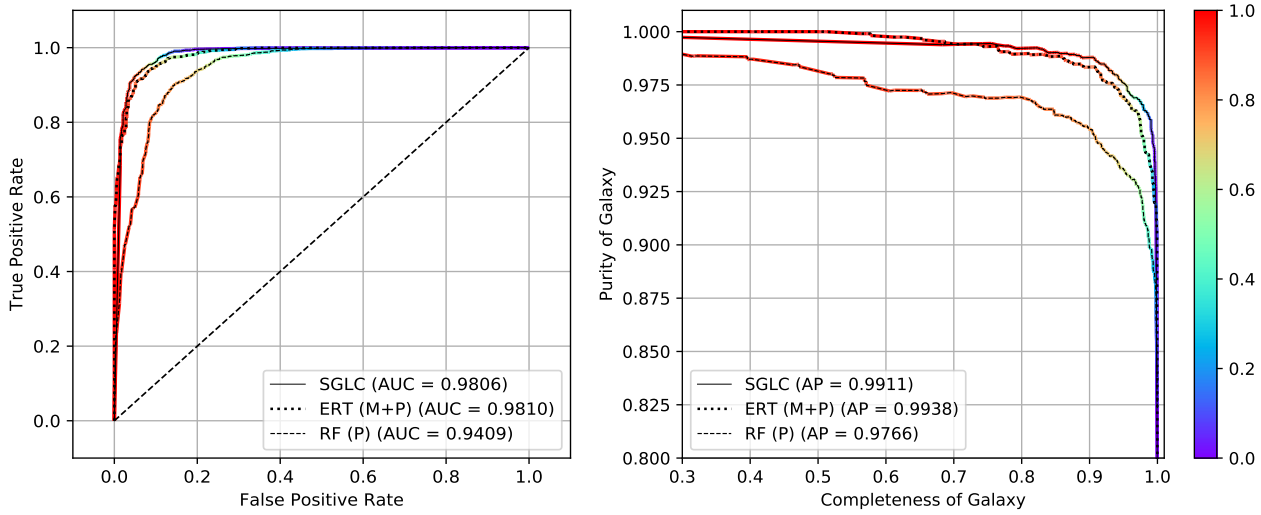

Figura 5.28. Curva ROC à esquerda e Completeza x Pureza para galáxias à direita para RF (sem morfologia), ERT (com morfologia) e SGLC. A análise foi realizada para fontes na faixa de magnitude $15 \leq r \leq 23.5$. As cores indicam o limite de probabilidade p_{cut}.

Na figura 5.29 podemos ver as performaces dos classificações geradas para todos os algoritmos treinados $15 \leq r \leq 23.5$. Notemos que temos uma melhora na performace quando comparamos com a análise as quais utilizamos unicamente rótulos do HSC-SSP. Veja na tabela 5.3 que além de conseguirmos uma melhor performace, conseguimos extender o limite do poder de classificação dos algoritmos de ML. Isso se da pelo fato de conseguirmos excelentes rótulos e com quantidade expressiva oriundas do cross-match miniJPAS e SDSS.

5.5 Cruzamentos com Outros Surveys

Nas análises realizadas até o momento assumimos os rótulos dos surveys SDSS e HSC-SSP como verdadeiros, os quais treinamos nossos algoritmos. Nesta seção daremos uma idéia do quão confiável são esses rótulos fotométricos quando comparados aos rótulos de outros surveys.

Quando cruzamos os dados do survey miniJPAS Pathfinder com outros surveys espectroscópicos observamos poucos dados em comum. Isso nos impossibilita de utiliza-los como treino na análise. Sendo assim, utilizamos rótulos fotométricos para treinar nossas máquinas. Para conferir o nível de confiança destes, os comparamos aos rótulos de outros surveys. Em particular comparamos os rótulos fotométricos do SDSS e ALHAMBRA para o bins $15 \leq r \leq 21$, enquanto que para magnitudes mais fracas comparamos rótulos fotométricos de HSC-SSP como rótulos espectroscópicos do Deep2 para o bins $18.5 \leq r \leq 23.5$.

Na figura 5.30 à esquerda podemos observar a discordância entre os surveys fotométricos SDSS e ALHAMBRA para a faixa $15 \leq r \leq 21$. Em verde temos fontes classificadas como estrelas para ALHMABRA e galáxias para SDSS, enquanto que em roxo temos objetos classificados como galáxias para ALHAMBRA e estrelas para SDSS. A direita temos a distri-

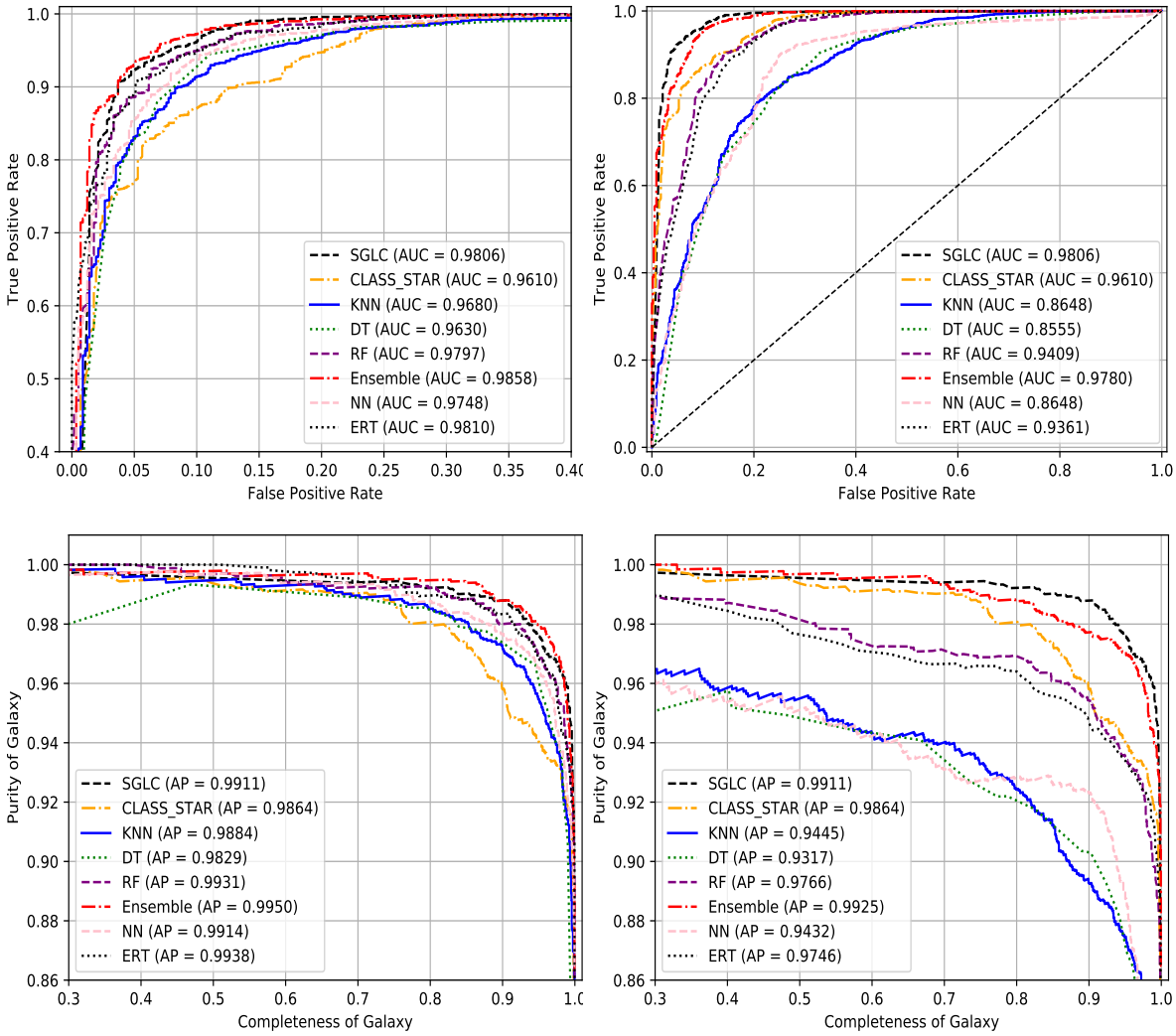

Figura 5.29. Acima: Curva ROC para as classificações geradas para os algoritmos treinados $15 \leq r \leq 23.5$. Abaixo: Curvas de Completeza e Pureza para as classificações geradas para os algoritmos treinados $15 \leq r \leq 23.5$. À esquerda temos resultados de máquinas treinadas utilizando parâmetros morfológicos e bandas fotométricas, enquanto que à direita apenas bandas fotoméricas.

buição dos dados para estrelas e galáxias classificadas por ALHAMBRA e que são comuns a ambos levantamentos. Podemos notar que o número de dados conflitantes é muito pequeno.

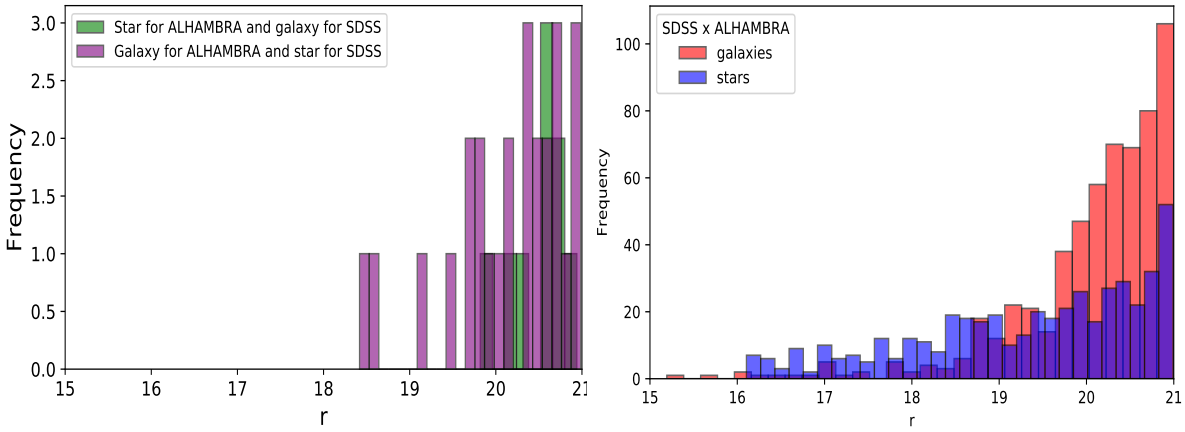

Figura 5.30. Direita: Comparação entre o catálogo SDSS usado neste pacote por e ALHAMBRA. Esquerda: Discordância entre SDSS e ALHAMBRA em função da magnitude r.

Em seguida fizemos uma análise semelhante comparando os rótulos fotométricos de HSC-SSP com os espectroscópicos de Deep2. Na figura 5.31 podemos observar os valores conflitantes entre ambos surveys. Em verde temos fontes classificadas como estrelas para HSC-SSP e galáxias para Deep2. A direita temos a distribuição de galáxias e estrelas classificadas segundo o Deep2 para os dados em comuns em ambos surveys.

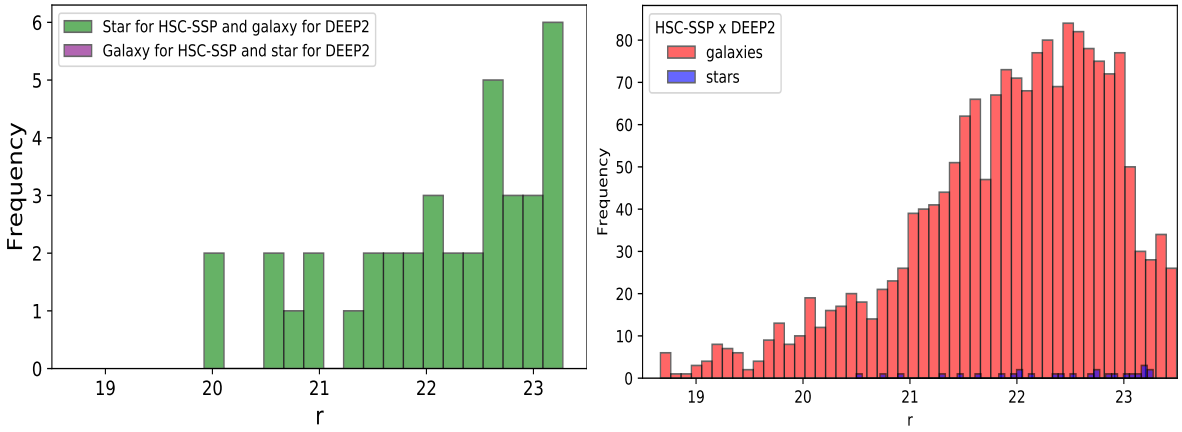

Figura 5.31. Esquerda: Comparação entre o catálogo HSC-SSP usado neste pacote por e DEEP2. Direita: Desacordo entre HSC-SSP e DEEP2 em função da magnitude r.

Essas análises nos permitem avaliar de certa forma o nível de confiabilidade dos rótulos fotométricos utilizadas para treinar nossas máquinas. Embora o cruzamento entre os surveys em estudo não englobe todos os dados, nos dão uma idéia de um comportamento para certos limites.

Capítulo 6

Conclusões

Neste trabalho aplicamos diferentes métodos de aprendizado de máquina para a classificação de fontes da pesquisa miniJPAS. O objetivo foi criar modelos competitivos e complementares aos existentes na literatura e oferecer à comunidade astronômica um catálogo de valor agregado com uma classificação alternativa. Como consideramos os algoritmos de ML supervisionados, trabalhamos com objetos do miniJPAS que são comuns ao SDSS (~ 3700 objetos) e ao HSC-SSP (~ 11100 objetos), cuja classificação assumimos ser confiável dentro dos intervalos de magnitude $15 < r < 21$ e $18.5 < r < 23.5$, respectivamente. As análises para cada survey foram realizadas de forma independentes. Utilizamos como entrada as magnitudes associadas aos 60 filtros, juntamente com 4 parâmetros morfológicos. A saída dos algoritmos é probabilística. Testamos e treinamos os algoritmos KNN, DT, RF, NN, ERT e um Ensemble Learning.

Nossos resultados mostram que o ML é capaz de classificar objetos em estrelas e galáxias sem o uso de parâmetros morfológicos, isso torna os classificadores de ML bastante valiosos os quais nos permitem classificar melhor quasares como objetos extragalácticos. A inclusão de parâmetros morfológicos melhora os resultados no geral a ponto dos métodos de ML superarem o *CLASS_STAR* e o SGLC (classificadores padrão no miniJPAS). O fato de conseguirmos agregar diferentes parâmetros morfológicos juntamente a bandas fotométricas no processo de classificação é algo que diferencia os modelos de ML dos existentes na literatura. Quanto mais parâmetros, maior a quantidade de informações, portanto, melhor tende a sera performance dos modelos de ML comparado a outros.

Utilizamos o algoritmo de RF para avaliar a importância das características. Ao usar parâmetros morfológicos, a concentração c_r é de longe a característica mais importante. Quando utilizamos apenas informações fotométricas, as características com importância relativa maior que 0.5 são J0460 e J0390. De maneira semelhante, nas classificações em SDSS e HSC-SSP temos como características mais importantes as observações associadas aos filtros J0390, J0460, iSDSS, J0810. Mostramos que o ML pode fornecer informações significativas sobre as regiões do espectro mais importantes para a classificação.

Após validar nossos métodos, aplicamos nossos melhores classificadores, com e sem

morfologia, ao conjunto completo de dados. Esta classificação está disponível como um catálogo de valor agregado em http://archive.cefca.es/catalogues. Nosso catálogo valida os resultados do classificador SGLC e produz uma classificação independente que é útil para testar a robustez das análises científicas subsequentes. Em particular, nossa classificação usa as informações fotométricas completas, com e sem morfologia, o que é importante para galáxias e estrelas fracas cuja morfologia são semelhantes.

6.1 Perspectivas Futuras

Pretendemos aplicar os Algoritmos de ML assim como os de Ensemble aos dados que serão coletados durante as observações do survey J-PAS. Tudo isso de forma complementar às classificações fornecidas pelo método SGLC e pela classificação fornecida por *CLASS_STAR*.

Temos aplicados esses métodos de classificação aos dados do survey Javalambre Photometric Local Universe Survey (J-PLUS), um survey fotométrico que nos permite observar um universo local com 12 filtros. Em particular, temos interesse em separar estrelas, galáxias e quasares. A idéia de separação de objetos em estrelas, galáxias e quasares também tem sido aplicada aos dados do miniJPAS. Utilizando catálogos sintéticos construídos por Carolina Queiroz e Raul Abramo, nós temos treinados nossos algoritmos a fim de se fazer estimativas sobre o número de quasares, assim como estatísticas secundárias sobre esses objetos.

Também temos idéias de aplicar redes neurais convolucionais (CNN) ao contexto de separação de fontes. Utilizando imagens dos surveys esperamos encontrar resultados competitivos com os modelos construídos nessa tese. De maneira mais específica, utilizaremos 60 imagens para cada fonte, uma associada a cada filtro, como entrada do algoritmo. A implementação de CNN no processo de classificação de objetos é algo que vem sendo empregado bastante em ciência no geral. E Recentemente tem ganhado popularidade no ramo da astrofísica.

Um projeto paralelo e quem vem sendo desenvolvido em parceria com o grupo do Data Validation (DAVA) é a aplicação de algoritmos de aprendizagem de máquina à predição de redshift fotométrico. O objetivo é treinar as máquinas com rótulos espectroscópicos de surveys cruzados com os dados do miniJPAS. Entretanto, há poucos dados em comum observados pelos surveys miniJPAS e BOSS, DEEP2. Portanto, temos poucos dados para treinar nossas máquinas. Como alternativa, temos trabalhado com catálogos sintéticos construídos por Raul Abramo e Carolina Queiroz. Chegamos a encontrar uma precisão de $\sigma_{NMAD} \sim 10^{-3}$ testando-se para dados reais. Um resumo dos resultados pode ser encontrado no anexo A.

Um projeto com a mesma essência é a aplicação desses códigos aos útimos dois anos de coleta de dados do Dark Energy Survey (DES). A idéia é trabalhar em parceria com o grupo do Laboratório Interinstitucional de e-Astronomia (LIneA). Em relação à quantidade de bandas fotométricas, temos um número muito reduzido quando comparadas ao número das existentes no miniJPAS.

Referências Bibliográficas

[1] Hematinezhad, M., Gholizadeh, M. H., Ramezaniyan, M., Shafiee, S. & Zahedi, A. G. Predicting the success of nations in asian games using neural network. *Sport Scientific & Practical Aspects* **8** (2011).

[2] Boloni, S. & colaboration, J. The pathfinder-minijpas survey: Paving the way for j-pas. *Astronomy & Astrophysics* .

[3] Holwerda, B. W. Source extractor for dummies v5. *arXiv preprint astro-ph/0512139* (2005).

[4] Messier, C. Catalogue des nébuleuses et des amas d'étoiles (catalog of nebulae and star clusters). *Connoissance des Temps ou des Mouvements Célestes, for 1784, p. 227-267* 227–267 (1781).

[5] Newman, J. A. *et al.* The deep2 galaxy redshift survey: design, observations, data reduction, and redshifts. *The Astrophysical Journal Supplement Series* **208**, 5 (2013).

[6] Dawson, K. S. *et al.* The baryon oscillation spectroscopic survey of sdss-iii. *The Astronomical Journal* **145**, 10 (2012).

[7] Fevre, O. L., Crampton, D., Lilly, S. J., Hammer, F. & Tresse, L. The canada-france redshift survey ii: Spectroscopic program; data for the 0000-00 and 1000+ 25 fields. *arXiv preprint astro-ph/9507011* (1995).

[8] MacGillivray, H. *et al.* A method for the automatic separation of the images of galaxies and stars from measurements made with the cosmos machine. *Monthly Notices of the Royal Astronomical Society* **176**, 265–274 (1976).

[9] Heydon-Dumbleton, N., Collins, C. & MacGillivray, H. The edinburgh/durham southern galaxy catalogue–ii. image classification and galaxy number counts. *Monthly Notices of the Royal Astronomical Society* **238**, 379–406 (1989).

[10] Collaboration, D. E. S. *et al.* The dark energy survey. *arXiv preprint astro-ph/0510346* (2005).

[11] Benitez, N. *et al.* J-pas: the javalambre-physics of the accelerated universe astrophysical survey. *arXiv preprint arXiv:1403.5237* (2014).

[12] York, D. G. *et al.* The sloan digital sky survey: Technical summary. *The Astronomical Journal* **120**, 1579 (2000).

[13] Tyson, J. A. Large synoptic survey telescope: overview. In *Survey and Other Telescope Technologies and Discoveries*, vol. 4836, 10–20 (International Society for Optics and Photonics, 2002).

[14] Dewdney, P. E., Hall, P. J., Schilizzi, R. T. & Lazio, T. J. L. The square kilometre array. *Proceedings of the IEEE* **97**, 1482–1496 (2009).

[15] Davis, M. *et al.* The all-wavelength extended groth strip international survey (aegis) data sets. *The Astrophysical Journal Letters* **660**, L1 (2007).

[16] Costa-Duarte, M. *et al.* The s-plus: a star/galaxy classification based on a machine learning approach. *arXiv preprint arXiv:1909.08626* (2019).

[17] Bertin, E. & Arnouts, S. Sextractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* **117**, 393–404 (1996).

[18] López-Sanjuan, C. *et al.* J-plus: Morphological star/galaxy classification by pdf analysis. *Astronomy & Astrophysics* **622**, A177 (2019).

[19] Vasconcellos, E. *et al.* Decision tree classifiers for star/galaxy separation. *The Astronomical Journal* **141**, 189 (2011).

[20] Kim, E. J., Brunner, R. J. & Carrasco Kind, M. A hybrid ensemble learning approach to star–galaxy classification. *Monthly Notices of the Royal Astronomical Society* **453**, 507–521 (2015).

[21] Kim, E. J. & Brunner, R. J. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society* stw2672 (2016).

[22] Sevilla-Noarbe, I. *et al.* Star–galaxy classification in the dark energy survey y1 data set. *Monthly Notices of the Royal Astronomical Society* **481**, 5451–5469 (2018).

[23] Cabayol, L. *et al.* The pau survey: star–galaxy classification with multi narrow-band data. *Monthly Notices of the Royal Astronomical Society* **483**, 529–539 (2018).

[24] Fadely, R., Hogg, D. W. & Willman, B. Star-galaxy classification in multi-band optical imaging. *The Astrophysical Journal* **760**, 15 (2012).

[25] Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O. & Winter, M. K. Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series* **225**, 31 (2016).

[26] Charnock, T. & Moss, A. supernovae: Photometric classification of supernovae. *Astrophysics Source Code Library* (2017).

[27] Biswas, R. *et al.* Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data. *Physical Review D* **88**, 062003 (2013).

[28] Carrillo, M., González, J., Gracia-Linares, M. & Guzmán, F. Time series analysis of gravitational wave signals using neural networks. In *Journal of Physics: Conference Series*, vol. 654, 012001 (IOP Publishing, 2015).

[29] Bilicki, M. *et al.* Photometric redshifts for the kilo-degree survey-machine-learning analysis with artificial neural networks. *Astronomy & Astrophysics* **616**, A69 (2018).

[30] Cavuoti, S. *et al.* Machine-learning-based photometric redshifts for galaxies of the eso kilo-degree survey data release 2. *Monthly Notices of the Royal Astronomical Society* **452**, 3100–3105 (2015).

[31] Gauci, A., Adami, K. Z. & Abela, J. Machine learning for galaxy morphology classification. *arXiv preprint arXiv:1005.0390* (2010).

[32] Banerji, M. *et al.* Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society* **406**, 342–353 (2010).

[33] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T. & Gray, A. *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data* (Princeton University Press, 2019).

[34] P.O.Baqui, L. C. e. a., V. Marra. Star-galaxy classification using machine learning in minijpas. *Astronomy & Astrophysics* .

[35] Samuel, A. 1959. some studies oin machine learning using the game of checkers. *IBM Journal of Researchand Development* **3**, 211–229.

[36] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).

[37] Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).

[38] Mitchell, T. M. Machine learning (1997).

[39] Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**, 175–185 (1992).

[40] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and regression trees. belmont, ca: Wadsworth. *International Group* **432**, 151–166 (1984).

[41] Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

[42] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine learning* **63**, 3–42 (2006).

[43] Rosenblatt, F. *The perceptron, a perceiving and recognizing automaton Project Para* (Cornell Aeronautical Laboratory, 1957).

[44] Garofalo, M., Botta, A. & Ventre, G. Astrophysics and big data: Challenges, methods, and tools. *Proceedings of the International Astronomical Union* **12**, 345–348 (2016).

[45] Marr, B. how much data do we create every day? the mind-blowing stats everyone should read. *Forbes, May 21st, www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/, accessed Dec 12th* (2019).

[46] Interiano, M. *et al.* Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science* **5**, 171274 (2018).

[47] Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).

[48] Guzella, T. S. & Caminhas, W. M. A review of machine learning approaches to spam filtering. *Expert Systems with Applications* **36**, 10206–10222 (2009).

[49] Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. Credit card fraud detection using bayesian and neural networks. In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, 261–270 (2002).

[50] Cenarro, A. *et al.* J-plus: the javalambre photometric local universe survey. *Astronomy & Astrophysics* **622**, A176 (2019).

[51] de Souza Oliveira Filho, K. & Saraiva, M. d. F. O. Astronomia e astrofısica. *Livraria da Fısica, Rio Grande do Sul,* (2014).

[52] Bertin, E. Sextractor documentation .

[53] Gunn, J. *et al.* The sloan digital sky survey photometric camera. *The Astronomical Journal* **116**, 3040 (1998).

[54] Bosch, J. *et al.* The hyper suprime-cam software pipeline. *Publications of the Astronomical Society of Japan* **70**, S5 (2018).

[55] Aihara, H. *et al.* Second data release of the hyper suprime-cam subaru strategic program. *Publications of the Astronomical Society of Japan* **71**, 114 (2019).

[56] Sbisà, F., Baqui, P. O., Miranda, T., Jorás, S. E. & Piattella, O. F. Neutron star masses in r2-gravity. *Physics of the Dark Universe* **27**, 100411 (2020).

Apêndice A

Análises Complementares

Nesta seção inserimos algumas análises complementares que nos permitem ter maior entendimento sobre os classificadores estudados. Inserimos tabelas com a importância das características e os hiperparâmetros utilizados nos códigos.

A.1 Feature Importances

Na tabela A.1 observamos as importância das características no processo de treinamente dos algoritmos de RF, tanto para o caso fotométrico como para o caso em que utilizamos fotometria juntamente à morfologia. Para essas tabelas utilizamos rótulos do survey SDSS. Os dados estão normalizados e representam os valores encontrados no gráfico 5.11. À esquerda temos a tabela com os resultados no qual utilizamos morfologia e à direita temos os resultados utilizando unicamente bandas fotométricas.

Característica	Importância
c_r	1.000000
A/B	0.005035
FWHM	0.003761
J0490	0.002929
$\mu_{max}/r_{3''}$	0.002396
J0470	0.001285
J0460	0.001285
J0640	0.000882
J0740	0.000865
J0820	0.000834
J0850	0.000765
J0410	0.000656
J0700	0.000586
J0730	0.000579
J0710	0.000527

Característica	Importância
J0390	1.000000
J0460	0.937650
iSDSS	0.634614
J0810	0.604139
uJPAS	0.456416
uJAVA	0.434258
J0470	0.266658
J0680	0.256502
gSDSS	0.238639
J0378	0.233176
J0400	0.223251
J0820	0.215911
J0750	0.214282
J1007	0.168681
J0890	0.168170

J0550	0.000499	J0450	0.168141
J0610	0.000471	J0580	0.159174
J0890	0.000456	J0420	0.156766
J0480	0.000417	J0570	0.153626
J0840	0.000414	J0410	0.153081
uJPAS	0.000344	J0830	0.148749
J0870	0.000342	J0430	0.144174
J0650	0.000336	J0540	0.143323
J0780	0.000327	J0910	0.138072
J0790	0.000303	J0870	0.137995
J0880	0.000286	J0760	0.135758
J0390	0.000279	J0790	0.134750
J0560	0.000276	J0690	0.133001
J0660	0.000274	J0530	0.131367
J0450	0.000272	J0740	0.131338
J0830	0.000268	J0900	0.128271
J0400	0.000263	J0840	0.125705
J0800	0.000254	J0520	0.124458
J0540	0.000253	J0560	0.124363
rSDSS	0.000252	J0550	0.124244
J0430	0.000235	J0660	0.124003
J0600	0.000234	J0860	0.122901
J0720	0.000223	J0800	0.121352
J1007	0.000211	J0730	0.121187
J0580	0.000207	J0620	0.120243
J0420	0.000203	J0590	0.116794
J0810	0.000202	J0440	0.116651
J0440	0.000193	J0720	0.116382
J0630	0.000191	J0880	0.114459
J0530	0.000184	J0780	0.112966
J0520	0.000178	J0850	0.111254
J0770	0.000170	J0500	0.110276
J0620	0.000170	J0630	0.108235
J0510	0.000163	J0510	0.108176
J0760	0.000163	J0490	0.101685
J0860	0.000159	J0600	0.101472
J0590	0.000147	J0710	0.098890
J0500	0.000146	J0670	0.097720
uJAVA	0.000143	J0610	0.097628
iSDSS	0.000143	J0480	0.094775
J0910	0.000131	J0700	0.090393

J0750	0.000122	J0650	0.090292
gSDSS	0.000117	J0770	0.087435
J0378	0.000099	J0640	0.085737
J0900	0.000094	rSDSS	0.073460
J0670	0.000076		
J0680	0.000071		
J0690	0.000056		
J0570	0.000048		

Tabela A.1. Tabela com o valor das importâncias das características utilizadas no processo de predição para os algoritmos de RF. As análises realizadas aqui foram para os casos em que utilizamos os rótulos do survey SDSS. À esquerda: Análise utilizando bandas fotométricas junto à parâmetros morfológicos. À direita: Análise utilizando apenas bandas fotometricas.

Na tabela A.2 podemos verificar a importância de cada característica para o caso em que utilizamos dados do survey HSC-SSP. A tabela à direita representa o valor de cada importância no gráfico 5.22 no qual utilizamos apenas bandas fotométricas como entrada enquanto. À esquerda temos os valores associados as importâncias o qual utilizamos bandas fotométricas juntamente à morfologia.

Característica	Importância	Característica	Importância
c_r	1.000000	J0390	1.000000
FWHM	0.223514	J0460	0.937650
$\mu_{max}/r_{3''}$	0.114960	iSDSS	0.634614
J0740	0.078108	J0810	0.604139
J0390	0.040912	uJPAS	0.456416
A/B	0.027084	uJAVA	0.434258
uJAVA	0.021331	J0470	0.266658
J0680	0.018030	J0680	0.256502
gSDSS	0.016422	gSDSS	0.238639
J0580	0.015705	J0378	0.233176
J0870	0.014923	J0400	0.223251
uJPAS	0.012753	J0820	0.215911
J0650	0.011579	J0750	0.214282
J0400	0.010792	J1007	0.168681
J0620	0.010357	J0890	0.168170
J0460	0.010249	J0450	0.168141
J0800	0.010172	J0580	0.159174
J0750	0.009736	J0420	0.156766
J0730	0.009623	J0570	0.153626
J0860	0.009613	J0410	0.153081

J0510	0.009546	J0830	0.148749
J0490	0.009437	J0430	0.144174
J0470	0.009356	J0540	0.143323
J0790	0.009334	J0910	0.138072
J0378	0.009319	J0870	0.137995
J0630	0.009221	J0760	0.135758
J0590	0.009202	J0790	0.134750
J0830	0.009009	J0690	0.133001
J0530	0.008909	J0530	0.131367
J0770	0.008812	J0740	0.131338
J0820	0.008636	J0900	0.128271
J0440	0.008516	J0840	0.125705
J0910	0.008459	J0520	0.124458
iSDSS	0.008451	J0560	0.124363
J1007	0.008395	J0550	0.124244
J0900	0.008341	J0660	0.124003
J0690	0.008239	J0860	0.122901
J0810	0.008141	J0800	0.121352
J0710	0.008105	J0730	0.121187
J0560	0.008096	J0620	0.120243
J0500	0.007942	J0590	0.116794
rSDSS	0.007889	J0440	0.116651
J0610	0.007809	J0720	0.116382
J0880	0.007764	J0880	0.114459
J0550	0.007751	J0780	0.112966
J0850	0.007713	J0850	0.111254
J0410	0.007687	J0500	0.110276
J0840	0.007563	J0630	0.108235
J0890	0.007510	J0510	0.108176
J0720	0.007477	J0490	0.101685
J0660	0.007215	J0600	0.101472
J0760	0.006984	J0710	0.098890
J0450	0.006976	J0670	0.097720
J0420	0.006882	J0610	0.097628
J0520	0.006809	J0480	0.094775
J0570	0.006800	J0700	0.090393
J0430	0.006573	J0650	0.090292
J0670	0.006409	J0770	0.087435
J0700	0.005794	J0640	0.085737
J0540	0.005792	rSDSS	0.073460
J0480	0.005706		

J0640	0.005612
J0780	0.005264
J0600	0.005242

Tabela A.2. Tabela com o valor das importâncias das características utilizadas no processo de predição para os algoritmos de RF. As análises realizadas aqui foram para os casos em que utilizamos os rótulos do survey HSC-SSP. À esquerda: Análise utilizando bandas fotométricas junto à morfologia. À direita: Análise utilizando apenas bandas fotométricas.

A.2 Overfitting/Underfitting

Abaixo seguem a grade com os hiperparâmetros inseridos nos algoritmos KNN, DT, NN, RF, ERT utilizados nesse trabalho.

```
KNN

param_dist = {"n_neighbors":[10,50,100],
"weights":['uniform', 'distance'],
"n_jobs":[1],
"random_state":[5]
}


Decision Trees

param_dist = { "max_features":[None],# max_features=n_features.
"criterion":['gini','entropy'],
"max_depth":[None,5,10,20],
"class_weight":['balanced'],
"random_state":[5]
             }


# Random Forest

param_dist = {"n_estimators":[100],
"max_features":[None],# max_features=n_features.
"max_depth":[None,5,10,20],
"criterion":['gini','entropy'],
"bootstrap":[True],
```

```
"class_weight":['balanced_subsample'],
"random_state":[5],
"n_jobs":[-1]
            }

# Neural Network

param_dist = { "hidden_layer_sizes":[200,200],
   "activation":['relu','logistic'],
   "solver":['adam', 'sgd'],
   "learning_rate":['constant', 'invscaling', 'adaptive'],
    "random_state":[5]
}


# Extremely Randomized Trees

param_dist = {"n_estimators":[200],
"max_features":[None],# max_features=n_features.
"criterion":['gini','entropy'],
"max_depth":[None,5,10,20],
"bootstrap":[False],
"class_weight":['balanced_subsample'],
"random_state":[5],
"n_jobs":[-1]
            }
```

Diferentes hiperparâmetros geram diferentes performaces. Segue abaixo os hiperparâmetros que resultaram a melhor AUC média oriunda da aplicação do método K-Fold.

```
Análise resultante utilizando apenas bandas fotométricas para dados do miniJPAS com rótul

KNN

AUC = 0.8995 (STD = 0.0176)          AP = 0.9024 (STD = 0.0187)
{'n_jobs': 1, 'n_neighbors': 10, 'weights': 'distance'}

DT

AUC = 0.9022 (STD = 0.0217)          AP = 0.8837 (STD = 0.0264)
{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 10, 'max_features':
```

None, 'random_state': 5}

RF

AUC = 0.9784 (STD = 0.0055) AP = 0.9782 (STD = 0.0075)
{'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': None, 'max_features': None, 'n_estimators': 100, 'n_jobs': -1,
'random_state': 5}

NN

AUC = 0.9759 (STD = 0.0064) AP = 0.9701 (STD = 0.0117)
{'activation': 'relu', 'hidden_layer_sizes': 200, 'learning_rate': 'constant',
'random_state': 5, 'solver': 'adam'}

ERT

AUC = 0.9829 (STD = 0.0035) AP = 0.9835 (STD = 0.0051)
{'bootstrap': False, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': 20, 'max_features': None, 'n_estimators': 200, 'n_jobs': -1,
'random_state': 5}

Análise resultante utilizando bandas fotométricas junto à parâmetros morfológicos para da

KNN

AUC = 0.9928 (STD = 0.0047) AP = 0.9924 (STD = 0.0060)

{'n_jobs': 1, 'n_neighbors': 50, 'weights': 'distance'}

DT

AUC = 0.9891 (STD = 0.0068) AP = 0.9857 (STD = 0.0090)
{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 5,
'max_features': None, 'random_state': 5}

RF

AUC = 0.9981 (STD = 0.0023) AP = 0.9980 (STD = 0.0029)
{'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'gini',

'max_depth': 5, 'max_features': None, 'n_estimators': 100, 'n_jobs': -1,
'random_state': 5}

NN

AUC = 0.9971 (STD = 0.0028) AP = 0.9952 (STD = 0.0072)
{'activation': 'logistic', 'hidden_layer_sizes': 200, 'learning_rate': 'constant',
'random_state': 5, 'solver': 'sgd'}

ERT

AUC = 0.9986 (STD = 0.0019) AP = 0.9983 (STD = 0.0029)
{'bootstrap': False, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': 10, 'max_features': None, 'n_estimators': 200, 'n_jobs': -1,
'random_state': 5}

Análise resultante utilizando apenas bandas fotométricas para dados do miniJPAS com rótul

KNN

AUC = 0.8071 (STD = 0.0236) AP = 0.9489 (STD = 0.0082)
{'n_jobs': 1, 'n_neighbors': 50, 'weights': 'distance'}

DT

AUC = 0.8044 (STD = 0.0236) AP = 0.9443 (STD = 0.0091)
{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 5, 'max_features':
None, 'random_state': 5}

RF

AUC = 0.9152 (STD = 0.0158) AP = 0.9777 (STD = 0.0052)
{'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': None, 'max_features': None, 'n_estimators': 100, 'n_jobs': -1,
'random_state': 5}

ANN

AUC = 0.8164 (STD = 0.0277) AP = 0.9481 (STD = 0.0099)

{'activation': 'relu', 'hidden_layer_sizes': 200, 'learning_rate': 'constant',
'random_state': 5, 'solver': 'adam'}

ERT

AUC = 0.9081 (STD = 0.0170) AP = 0.9740 (STD = 0.0066)
{'bootstrap': False, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': None, 'max_features': None, 'n_estimators': 200, 'n_jobs': -1,
'random_state': 5}

Análise resultante utilizando bandas fotométricas mais parâmetros morfológicos para dados

KNN

AUC = 0.9547 (STD = 0.0091) AP = 0.9900 (STD = 0.0025)
{'n_jobs': 1, 'n_neighbors': 100, 'weights': 'distance'}

DT

AUC = 0.9558 (STD = 0.0075) AP = 0.9498 (STD = 0.0060)
{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 5, 'max_features':
None, 'random_state': 5}

RF

AUC = 0.9706 (STD = 0.0054) AP = 0.9934 (STD = 0.0018)
{'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'entropy',
'max_depth': 10, 'max_features': None, 'n_estimators': 100, 'n_jobs': -1,
'random_state': 5}

NN

AUC = 0.9622 (STD = 0.0061) AP = 0.9911 (STD = 0.0025)
{'activation': 'relu', 'hidden_layer_sizes': 200, 'learning_rate': 'constant',
'random_state': 5, 'solver': 'sgd'}

ERT

```
AUC = 0.9720 (STD = 0.0059)           AP = 0.9941 (STD = 0.0013)
{'bootstrap': False, 'class_weight': 'balanced_subsample', 'criterion': 'gini',
 'max_depth': None, 'max_features': None, 'n_estimators': 200, 'n_jobs': -1,
 'random_state': 5}
```

Apêndice B

JPLUS Survey

Além da classificação de estrelas e galáxias para os dados do miniJPAS, começamos um trabalho com o DAVA team na classificação de fontes do miniJPAS em Estrelas, Galáxias e QSO. Mas para testar a robusts e acurácia do código, o aplicamos antes aos dados do Javalambre Photometric Local Universe Survey (JPLUS). O JPLUS é um survey fotométrico localizado em Javalambre, Espanha que está observando o céu com 12 filtros de bandas largas, intermediárias e estreitas. Seu telescópio, o T80Cam possui um diâmetro de 0.83 m com filtros que cobrem o faixa óptica de 3 500 a 10 000 Å. O J-PLUS possui uma profundidade AB de 21.25 *mag* por banda. Espera-se observar 35 milhões de estrelas e 24 milhões de galáxias até o fim das observações [50].

Para treinar nossa máquina utilizamos rótulos espectroscópico do SDSS DR12 obtidos a partir do catálogo disponível no site [1]. Esse catálogo foi construído a partir do Cross-Match entre J-PLUS e SDSS. A consulta do catálogo em SQL pode ser encontrada no apêndice C. Encontramos para essa consulta 127 714 galáxias, 45 442 estrelas e 27 672 quasares as quais dividimos em dados de teste e treino com 20% e 80% respectivamente. O método utilizado para a multiclassificação foi o *One vs All* no qual o algoritmos separa a classe A de *não-A* por exemplo. Isso quer dizer que a multiclassificação se transforma em um conjunto de classificações binárias.

Para separação dos objetos utilizamos o algoritmo Random Forest. Como entrada para o algoritmo utilizamos as masgnitudes associas as 12 bandas fotométricas juntamente ao parâmetro morfológico c_r e as combinações entre as cores $g - r$, $r - i$ e $i - z$.

Observemos nas figuras B.1, B.2 que encontramos excelentes resultados para as classificações. Analisando a Curva ROC obtemos $AUC = 0.9763$ para galáxias, $AUC = 0.9635$ Estrelas e $AUC = 0.9527$. Por outro lado, ao analisar a Curva de Completeza e Puridade obtemos $AP = 0.9827$ para galáxias, $AP = 0.9158$ para estrelas e $AP = 0.7897$ para QSOs.

Neste trabalho em particular, no processo de classificação *One vs All*, as curvas azuis representam objetos que foram classificados como QSO e não QSO, podendo esses objetos pertencer à classe das estrelas ou galáxias. A curva vermelha são para objetos classificados

[1] http://archive.cefca.es/catalogues/jplus-dr1

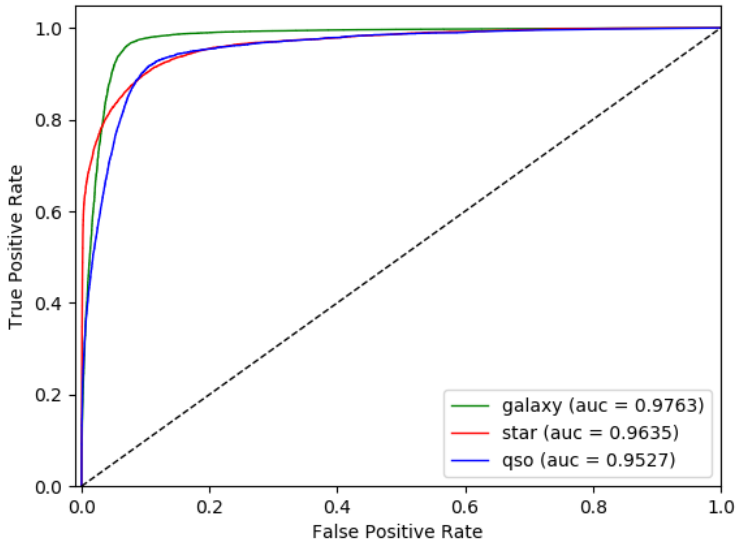

Figura B.1. Curvas ROC para fontes do J-PLUS utilizando o método *OnevsAll*. Utilizamos como entrada bandas fotométricas juntamente ao parâmetro concentração c_r.

como estrelas e não estrelas e a verde para objetos classificados como galáxias e não galáxias.

Este trabalho ainda está em construção. Como próximos passos iremos aplicar outros algoritmos além de RF assim como outras análises estatísticas. Também estamos extendendo essas análises aos dados do MiniJ-PAS, treinando as máquinas com dados sintéticos contruídos pelo professor Dr. Raul Abramo e Carolina Queiroz.

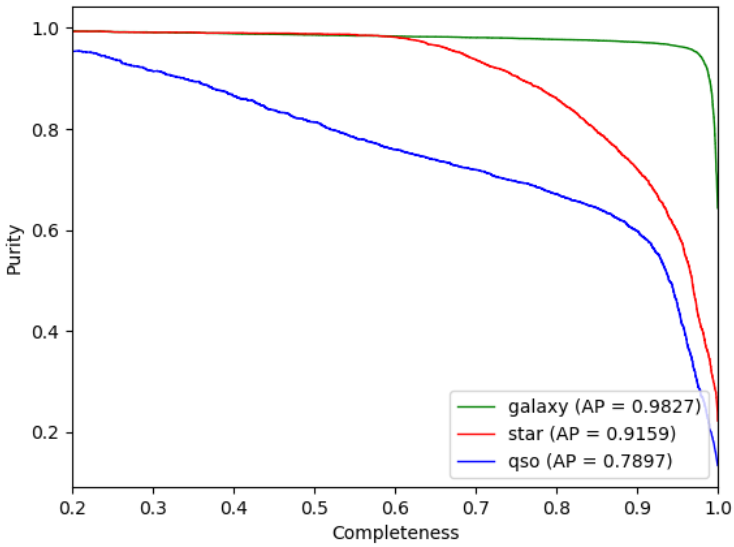

Figura B.2. Curvas de Completeza e Puridade para fontes do J-PLUS utilizando o método *OnevsAll*. Utilizamos como entrada bandas fotométricas juntamente ao parâmetro concentração c_r.

Apêndice C

Surveys Query

C.1 J-PAS Query

C.1.1 Cross-Match MiniJ-PAS x SDSS

Aqui disponibilizamos a consulta aplicada ao banco de dados do MiniJ-PAS (2019/10) para as análise em que treinamos os algoritmos de ML com rótulos do SDSS. Essa consulta é válida tanto para o caso em que utilizamos apenas fotometria como input, bem como fotometria juntamente à morfologia. O banco de dados pode ser encontrado em:

http://archive.cefca.es/catalogues/minijpas-idr201910

e o código em SQL, abaixo:

```
SELECT

t1.ALPHA_J2000, t1.DELTA_J2000, t1.CLASS_STAR as class_Sextractor,
t2.class as class_SDSS, t3.total_prob_star as  pdf,
t1.MAG_AUTO[minijpas::uJAVA] as uJAVA, t1.MAG_AUTO[minijpas::J0378] as J0378,
t1.MAG_AUTO[minijpas::J0390] as  J0390,  t1.MAG_AUTO[minijpas::J0400] as J0400,
t1.MAG_AUTO[minijpas::J0410] as J0410, t1.MAG_AUTO[minijpas::J0420] as J0420,
t1.MAG_AUTO[minijpas::J0430] as J0430, t1.MAG_AUTO[minijpas::J0440] as J0440,
t1.MAG_AUTO[minijpas::J0450] as J0450, t1.MAG_AUTO[minijpas::J0460] as  J0460,
t1.MAG_AUTO[minijpas::J0470] as J0470, t1.MAG_AUTO[minijpas::J0480] as J0480,
t1.MAG_AUTO[minijpas::J0490] as J0490,  t1.MAG_AUTO[minijpas::J0500] as J0500,
t1.MAG_AUTO[minijpas::J0510] as J0510, t1.MAG_AUTO[minijpas::J0520] as J0520,
t1.MAG_AUTO[minijpas::J0530] as J0530,  t1.MAG_AUTO[minijpas::J0540] as J0540,
t1.MAG_AUTO[minijpas::J0550] as J0550, t1.MAG_AUTO[minijpas::J0560] as J0560,
t1.MAG_AUTO[minijpas::J0570] as J0570, t1.MAG_AUTO[minijpas::J0580] as J0580,
t1.MAG_AUTO[minijpas::J0590] as J0590, t1.MAG_AUTO[minijpas::J0600] as J0600,
t1.MAG_AUTO[minijpas::J0610] as J0610, t1.MAG_AUTO[minijpas::J0620] as J0620,
```

```
t1.MAG_AUTO[minijpas::J0630] as J0630, t1.MAG_AUTO[minijpas::J0640] as J0640,
t1.MAG_AUTO[minijpas::J0650] as J0650, t1.MAG_AUTO[minijpas::J0660] as J0660,
t1.MAG_AUTO[minijpas::J0670] as J0670, t1.MAG_AUTO[minijpas::J0680] as J0680,
t1.MAG_AUTO[minijpas::J0690] as J0690, t1.MAG_AUTO[minijpas::J0700] as J0700,
t1.MAG_AUTO[minijpas::J0710] as J0710, t1.MAG_AUTO[minijpas::J0720] as J0720,
t1.MAG_AUTO[minijpas::J0730] as J0730, t1.MAG_AUTO[minijpas::J0740] as J0740,
t1.MAG_AUTO[minijpas::J0750] as J0750, t1.MAG_AUTO[minijpas::J0760] as J0760,
t1.MAG_AUTO[minijpas::J0770] as J0770, t1.MAG_AUTO[minijpas::J0780] as J0780,
t1.MAG_AUTO[minijpas::J0790] as J0790, t1.MAG_AUTO[minijpas::J0800] as J0800,
t1.MAG_AUTO[minijpas::J0810] as J0810, t1.MAG_AUTO[minijpas::J0820] as J0820,
t1.MAG_AUTO[minijpas::J0830] as J0830, t1.MAG_AUTO[minijpas::J0840] as J0840,
t1.MAG_AUTO[minijpas::J0850] as J0850, t1.MAG_AUTO[minijpas::J0860] as J0860,
t1.MAG_AUTO[minijpas::J0870] as J0870, t1.MAG_AUTO[minijpas::J0880] as J0880,
t1.MAG_AUTO[minijpas::J0890] as J0890, t1.MAG_AUTO[minijpas::J0900] as J0900,
t1.MAG_AUTO[minijpas::J0910] as J0910, t1.MAG_AUTO[minijpas::J1007] as J1007,
t1.MAG_AUTO[minijpas::uJPAS] as  uJPAS, t1.MAG_AUTO[minijpas::gSDSS] as  gSDSS,
t1.MAG_AUTO[minijpas::rSDSS] as  rSDSS, t1.MAG_AUTO[minijpas::iSDSS] as  iSDSS,
t1.MAG_APER_1_5[minijpas::rSDSS] - t1.MAG_APER_3_0[minijpas::rSDSS] as c_r,
t1.MU_MAX[minijpas::rSDSS] / t1.MAG_APER_3_0[minijpas::rSDSS] as mu_max_mag_apertu,
t1.FWHM_WORLD as fwhm, t1.A_WORLD / t1.B_WORLD as alb


FROM

minijpas.MagABDualObj t1

JOIN

minijpas.xmatch_sdss_dr12 t2

ON

t1.tile_id = t2.tile_id   AND  t1.NUMBER=t2.NUMBER


JOIN


minijpas.StarGalClass t3

ON
```

```
t1.tile_id = t3.tile_id   AND  t1.NUMBER=t3.NUMBER

WHERE


t1.flags[minijpas::rSDSS]=0 AND t1.mask_flags[minijpas::rSDSS]=0
```

C.1.2 Cross-Match MiniJ-PAS x HSC-SSP

Consulta aplicada ao banco de dados do MiniJ-PAS (2019/10) ao qual utilizaremos para se fazer o cross-match com os dados do HSC-SSP. Segue o código em SQL, abaixo:

```
SELECT

t1.ALPHA_J2000, t1.DELTA_J2000,
t1.CLASS_STAR as class_Sextractor, t3.total_prob_star as  pdf,
t1.MAG_AUTO[minijpas::uJAVA] as uJAVA, t1.MAG_AUTO[minijpas::J0378] as J0378,
t1.MAG_AUTO[minijpas::J0390] as J0390, t1.MAG_AUTO[minijpas::J0400] as J0400,
t1.MAG_AUTO[minijpas::J0410] as J0410, t1.MAG_AUTO[minijpas::J0420] as J0420,
t1.MAG_AUTO[minijpas::J0430] as J0430, t1.MAG_AUTO[minijpas::J0440] as J0440,
t1.MAG_AUTO[minijpas::J0450] as J0450, t1.MAG_AUTO[minijpas::J0460] as  J0460,
t1.MAG_AUTO[minijpas::J0470] as J0470, t1.MAG_AUTO[minijpas::J0480] as J0480,
t1.MAG_AUTO[minijpas::J0490] as J0490, t1.MAG_AUTO[minijpas::J0500] as J0500,
t1.MAG_AUTO[minijpas::J0510] as J0510, t1.MAG_AUTO[minijpas::J0520] as J0520,
t1.MAG_AUTO[minijpas::J0530] as J0530, t1.MAG_AUTO[minijpas::J0540] as J0540,
t1.MAG_AUTO[minijpas::J0550] as J0550, t1.MAG_AUTO[minijpas::J0560] as J0560,
t1.MAG_AUTO[minijpas::J0570] as J0570, t1.MAG_AUTO[minijpas::J0580] as J0580,
t1.MAG_AUTO[minijpas::J0590] as J0590, t1.MAG_AUTO[minijpas::J0600] as J0600,
t1.MAG_AUTO[minijpas::J0610] as J0610, t1.MAG_AUTO[minijpas::J0620] as J0620,
t1.MAG_AUTO[minijpas::J0630] as J0630, t1.MAG_AUTO[minijpas::J0640] as J0640,
t1.MAG_AUTO[minijpas::J0650] as J0650, t1.MAG_AUTO[minijpas::J0660] as J0660,
t1.MAG_AUTO[minijpas::J0670] as J0670, t1.MAG_AUTO[minijpas::J0680] as J0680,
t1.MAG_AUTO[minijpas::J0690] as J0690, t1.MAG_AUTO[minijpas::J0700] as J0700,
t1.MAG_AUTO[minijpas::J0710] as J0710, t1.MAG_AUTO[minijpas::J0720] as J0720,
```

```
t1.MAG_AUTO[minijpas::J0730] as J0730, t1.MAG_AUTO[minijpas::J0740] as J0740,
t1.MAG_AUTO[minijpas::J0750] as J0750, t1.MAG_AUTO[minijpas::J0760] as J0760,
t1.MAG_AUTO[minijpas::J0770] as J0770, t1.MAG_AUTO[minijpas::J0780] as J0780,
t1.MAG_AUTO[minijpas::J0790] as J0790, t1.MAG_AUTO[minijpas::J0800] as J0800,
t1.MAG_AUTO[minijpas::J0810] as J0810, t1.MAG_AUTO[minijpas::J0820] as J0820,
t1.MAG_AUTO[minijpas::J0830] as J0830, t1.MAG_AUTO[minijpas::J0840] as J0840,
t1.MAG_AUTO[minijpas::J0850] as J0850, t1.MAG_AUTO[minijpas::J0860] as J0860,
t1.MAG_AUTO[minijpas::J0870] as J0870, t1.MAG_AUTO[minijpas::J0880] as J0880,
t1.MAG_AUTO[minijpas::J0890] as J0890, t1.MAG_AUTO[minijpas::J0900] as J0900,
t1.MAG_AUTO[minijpas::J0910] as J0910, t1.MAG_AUTO[minijpas::J1007] as J1007,
t1.MAG_AUTO[minijpas::uJPAS] as  uJPAS, t1.MAG_AUTO[minijpas::gSDSS] as  gSDSS,
t1.MAG_AUTO[minijpas::rSDSS] as  rSDSS, t1.MAG_AUTO[minijpas::iSDSS] as  iSDSS,
t1.MAG_APER_1_5[minijpas::rSDSS] - t1.MAG_APER_3_0[minijpas::rSDSS] as c_r,
t1.MU_MAX[minijpas::rSDSS] / t1.MAG_APER_3_0[minijpas::rSDSS] as mu_max_mag_apertu,
t1.FWHM_WORLD as fwhm, t1.A_WORLD / t1.B_WORLD as alb


FROM

minijpas.MagABDualObj t1


JOIN

minijpas.StarGalClass t3

ON

t1.tile_id = t3.tile_id   AND  t1.NUMBER=t3.NUMBER


WHERE

t1.flags[minijpas::rSDSS]=0 AND t1.mask_flags[minijpas::rSDSS]=0
```

C.2 HSC-SSP Query

Aqui disponibilizamos a consulta aplicada ao banco de dados do HSC-SSP (DR2). De maneira análoga ao caso SDSS, os rótulos obtidos aqui são aplicados as análises em ML utilizando fotometria e fotometria juntamente à morfologia. O banco de dados pode ser encontrado em:

https://hsc-release.mtk.nao.ac.jp/datasearch/

e o código em SQL, abaixo:

```
SELECT
        t1.ra,
        t1.DEC,
        t1.r_extendedness_value,
        t2.r_psfflux_mag - t1.r_cmodel_mag as morphology,
        t1.r_cmodel_mag


FROM
pdr2_wide.forced as t1

JOIN

pdr2_wide.forced2 as t2

ON

t1.object_id=t2.object_id

 WHERE
       t1.isprimary= True
       AND t1.r_inputcount_value>=4
       AND t1.dec BETWEEN 51.13 AND 53.55
       AND t1.ra BETWEEN 213.14 AND 216.00
```

C.3 TOP CAT

Para se treinar os algoritmos com os rótulos do HSC-SSP fizemos o cross-match deste com os dados do MiniJ-PAS utilizando o programa TOPCAT. O processo foi realizado a partir da identificação das coordenadas declinação (DEC) e ascenção reta (RAC) com tolerância máxima de 1″ utilizando a função "SKY". Na figura C.1 podemos observar a configuração do parâmetros utilizados:

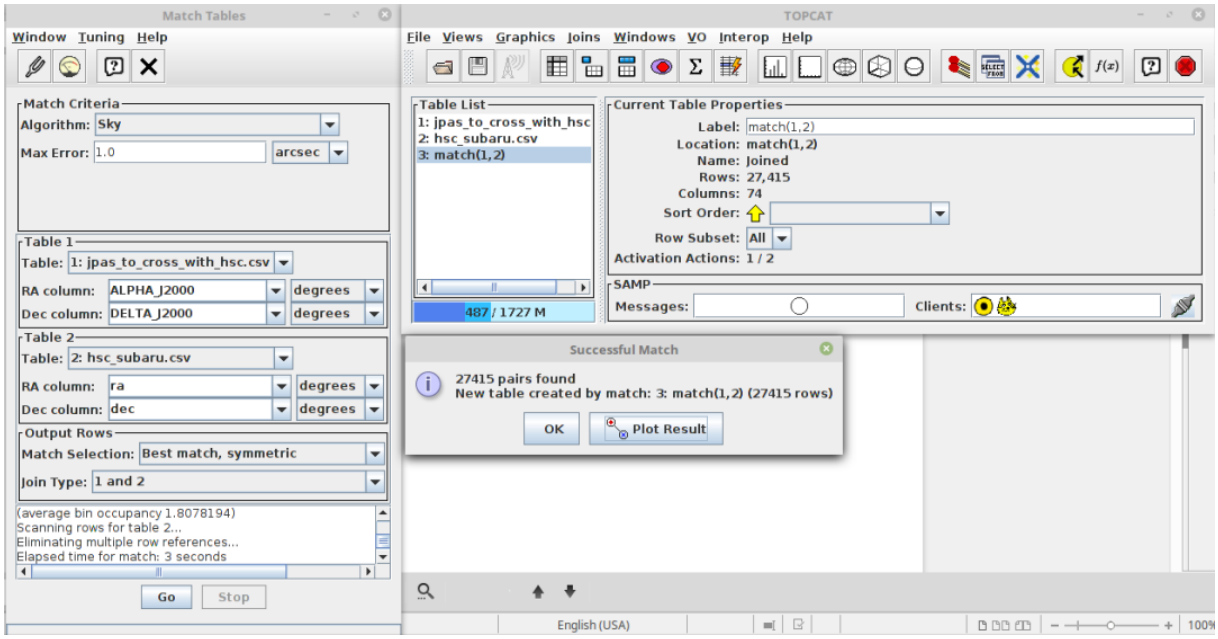

Figura C.1. Configuração dos parâmetros no programa TOPCAP ao se fazer os cross-match entre os dados do MiniJ-PAS e do HSC-SUBARU.

C.4 Query JPLUS

Para separação de fontes em estrelas/galaxias e quasares realizadas no apêndice B, utilizamos dados do J-PLUS com rótulos espectrocópicos do SDSS. Abaixo está a consulta realizada no banco de dados em SQL.

```
SELECT

t1.ALPHA_J2000,t1.DELTA_J2000, t1.MAG_AUTO[jplus::rSDSS] as rSDSS,
t1.MAG_AUTO[jplus::gSDSS] as gSDSS, t1.MAG_AUTO[jplus::iSDSS] as iSDSS,
t1.MAG_AUTO[jplus::zSDSS] as zSDSS, t1.MAG_AUTO[jplus::uJAVA] as uJAVA,
t1.MAG_AUTO[jplus::J0378] as J0378, t1.MAG_AUTO[jplus::J0395] as J0395,
t1.MAG_AUTO[jplus::J0410] as J0410, t1.MAG_AUTO[jplus::J0430] as J0430,
t1.MAG_AUTO[jplus::J0515] as J0515, t1.MAG_AUTO[jplus::J0660] as J0660,
t1.MAG_AUTO[jplus::J0861] as J0861,
t1.MAG_AUTO[jplus::gSDSS] - t1.MAG_AUTO[jplus::rSDSS] as g_r,
t1.MAG_AUTO[jplus::rSDSS] - t1.MAG_AUTO[jplus::iSDSS] as r_i,
t1.MAG_AUTO[jplus::iSDSS] - t1.MAG_AUTO[jplus::zSDSS] as i_z,
t1.MAG_APER_1_5[jplus::rSDSS] - t1.MAG_APER_3_0[jplus::rSDSS] as c_r,
t2.spCl

FROM jplus.MagABDualObj t1

JOIN
```

```
jplus.xmatch_sdss_dr12 t2

ON

t1.tile_id = t2.tile_id   AND  t1.NUMBER=t2.NUMBER

WHERE

t1.flags[jplus::rSDSS]=0 AND t1.mask_flags[jplus::rSDSS]=0

AND t2.spCl IS NOT NULL
```

Anexo A

Outros Trabalhos

Além de trabalhar sobre classificação de fontes em galáxias/estrelas, trabalhamos em outras linhas de pesquisa aplicando Machine Learning ao contexto de astrofísica/cosmologia. Foram elas :

- Photo-Z.
- Machine Learning e Simulações Cosmológicas.

Além disso tive oportunidade de continuar um trabalho sobre análise de diferentes definições de massa no contexto de estrela de nêutrons.

A.1 Photo-Z

Temos desenvolvido com o DAVA e LSS team da colaboração miniJPAS/J-PAS um trabalho de predição de redshift fotométrico aplicados à galáxias vermelhas luminosas (LRG). Para treinar nossos algoritmos utilizamos dados sintéticos. Esses dados sintéticos foram contruídos a partir do código desenvolvido por Carolina Queiroz e Raul Abraamo. Nele utilizamos fluxos espectropicos do BOSS DR12 das LRG para convoluir com os filtros de transmissão do miniJPAS. Utilizamos o catálogo $galaxy\_DR12v5\_CMASSLOWZTOT\_North.fits.gz$ composto por informações espectrocópicas e fotométricas das LRGs do catálogo CMASS e LOWZ presentes no hemisfério norte. Foram retirados 105 849 dados do catálogo para treinar a máquina de ML. Na figura A.1 observamos sua distribuição. Podemos notar que temos mais dados de CMASS do que propriamente dados de LOWZ.

Uma vez que o catálogo foi construído a partir da convolução dos filtros, treinamos nossa máquina com os fluxos associados as 60 bandas fotométricas. O Algoritmo utilizado na predição de photo-z foi o RF e utilizamos o caso regressivo. Na figura A.2 podemos observar as predições para o redshitf fotométrico em função de seus valores espectroscópicos, treinando e testando com catalogos mock. Treinamos apenas com os dados do fluxo e sem os erros associados a cada medida. Uma forma de se calcular a performace do algoritmo em regressivo é analisar o Coeficiente de Person R^2 [36] definido como:

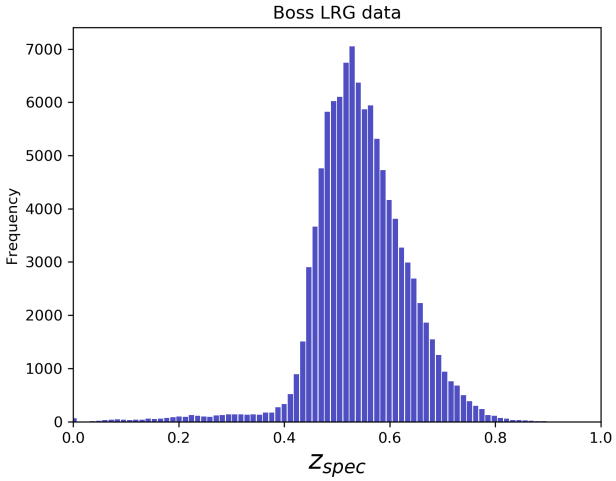

Figura A.1. Distribuição dos dados retidos do cátalogo CMASS + LOWZ norte em função do redshift.

$$R^2 = 1 - \frac{\sum(y^i_{test} - y^i_{predicted})^2}{\sum(y^i_{test} - y_{mean,test})^2} \tag{A.1}$$

Quanto mais próximo à 1 melhor sua performace. Outra forma é o cálculo do σ_{nmad} definido por:

$$\sigma_{nmad} = 1.4826 * median(\mid \Delta z \mid - median(\Delta z)) \tag{A.2}$$

Onde

$$\Delta z = \sum \frac{y^i_{pred} - y^i_{true}}{1 + y^i_{true}} \tag{A.3}$$

O σ_{nmad} foi calculado para diferentes bins. Notemos na figura A.2 que à medida que caminhamos para objetos mais distantes, σ_{nmad} para dados mock aumentam. Obtivemos resultados exelentes para dados de test mock com $R^2 = 0.93$ e $\sigma_{nmad} = 0.0027$.

Uma outra análise realizada foi a utilização de máquinas treinadas em dados sintéticos e testada em dados reais. Para o caso, testamos nossa máquina em 105 LGRs do miniJPAS que foram obtidas a partir do cross-match entre miniJPAS e SDSS surveys. Como resultado encontramos $R^2 = 0.708$ e $\sigma_{nmad} = 0.0109$. Na figura A.3 superior, observamos os redshifts photométricos em função dos espectroscópicos, abaixo observamos o comportamento de σ_{nmad} para diferentes bins. Em ambas análises treinamos em dados sintéticos e testamos em dados reais do miniJPAS.

Existem ainda várias tarefas a ser realizadas nesse trabalho como inserir mais dados

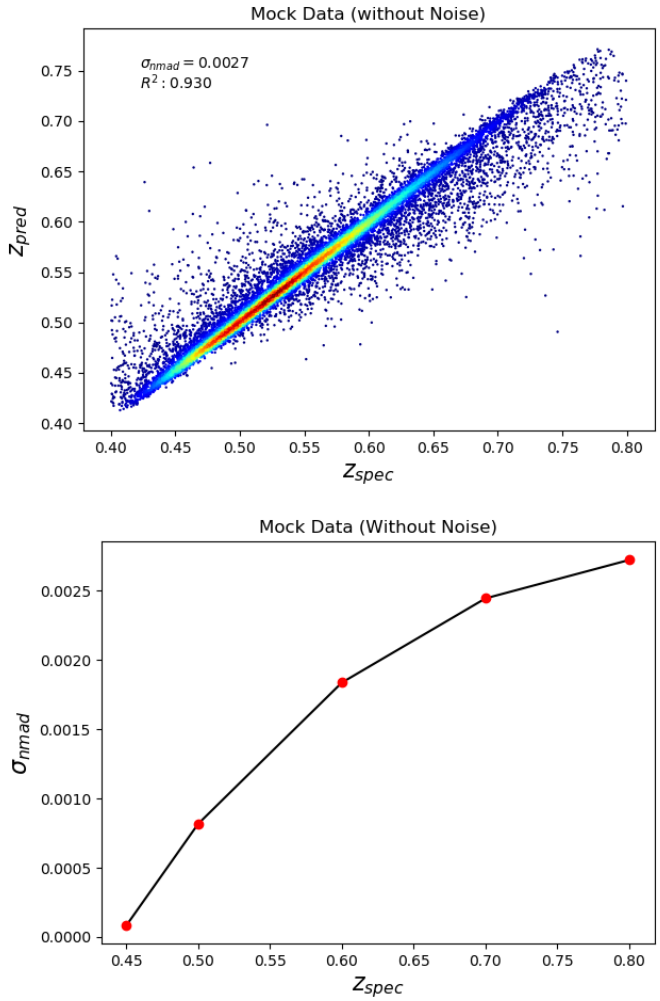

Figura A.2. Em cima: Redshift fotométrico predito em função do espectrocópico treinados e testados em catálogos mock. ; Em baixo: σ_{nmad} calculado para dados de teste mock. Em ambos os casos não utilizamos os erros associados a aos fluxo.

de LOWZ na análise, aplicar diferentes métodos de ML assim como extender a análise para QSO.

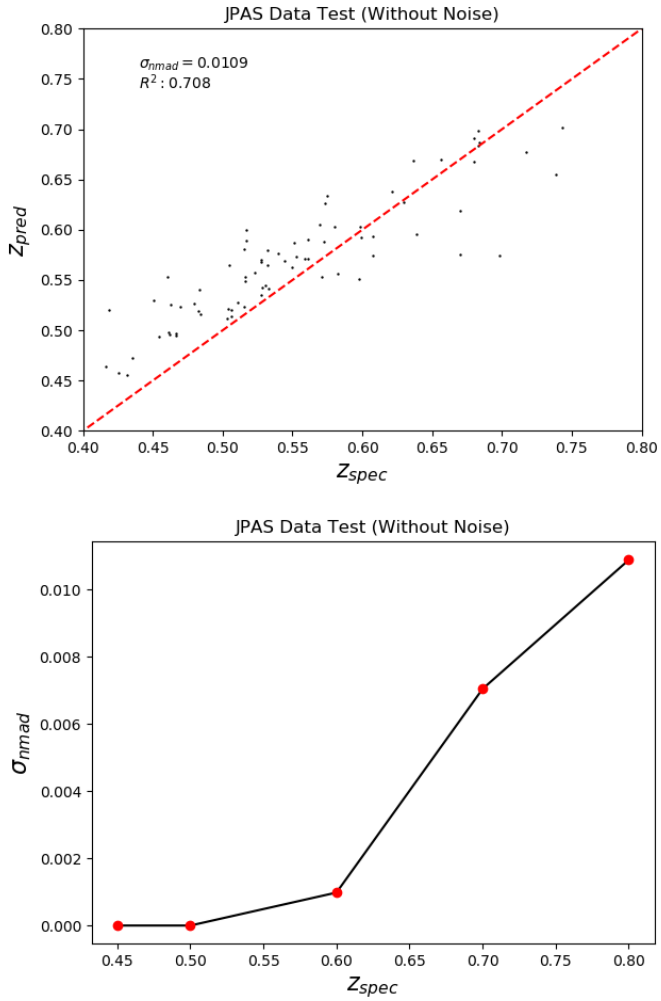

Figura A.3. Em cima: Redshift fotométrico predito em função do espectrocópico treinados em catálogos mock e testados em dados do miniJPAS. ; Em baixo: σ_{nmad} calculado para dados do miniJPAS. Em ambos os casos não utilizamos os erros associados a aos fluxo.

A.2 Machine Learning e Simulações Cosmológicas

Apresentamos um método de aprendizado de máquina para reconstruir as propriedades de matéria escura (por exemplo, raio de meia massa, massa crítica e raio crítico), a partir de quantidades bariônicas como (luminosidade de raios-x, velocidade máxima de rotação, a razão entre massas do bojo/estelares e $u - r$, r, bandas fotométricas), que derivam tipicamente de grandes surveys pesquisas como SDSS, GAMA e, no futuro, LSST. Este estudo preliminar mostra que as técnicas de aprendizado de máquina são uma ferramenta promissora para inferir o conteúdo de DM de grandes amostras de sistemas gravitacionais, como galáxias e aglomerados de galáxias, com um número limitado de quantidades observáveis. Essa técnica abrirá novas alternativas para a análise dinâmica direta, para usar o conteúdo de matéria escura da galáxia para validar as cosmologias, além das simulações usadas para treinar a abordagem de aprendizado de máquina escolhida. Mostramos que essa abordagem

de ML pode fornecer previsões precisas das propriedades do halo do DM, e mostramos que, usando apenas a velocidade máxima de rotação como variável de entrada, também obtemos resultados de alta precisão, indicando que ela representa a característica mais relevante para nossos estudos. Concluímos mostrando que o aprendizado de máquina pode fornecer ferramentas promissoras para caracterizar a massa do halo em torno dos sistemas gravitacionais e desempenhará um papel importante na cosmologia observacional.

O paper está em preparação e sendo construído em colaboração com Professor Dr. Luciano Casarini e Professor Dr. Nicola Napolitano.

A.3 Massa de Estrelas Neutrons em gravidade R^2

Neste trabalho, abordamos a questão da existência de diferentes definições de massa gravitacional na gravidade R^2. Apresentamos várias definições de massa gravitacional e discutimos as relações formais entre elas. Consideramos, então, o caso concreto de uma estrela de nêutrons estática e esfericamente simétrica e resolvemos numericamente as equações de movimento para vários valores do parâmetro livre do modelo. Comparamos as características das relações *Massa x Raio* obtidas para cada definição de massa gravitacional e comentamos sua dependência do parâmetro livre. Argumentamos então que a R^2 é um modelo valioso para discutir a existência de definições desiguais de massa gravitacional em uma teoria genérica da gravidade modificada e apresentar alguns comentários sobre o caso geral.

Este paper, o qual tive oportunidade de contribuir, foi publicado na revista Physics of the Dark Universe. Para mais informações consulte [56]. Embora deslocado do assunto de inteligencia artificial, esse paper é a continuação de um trabalho iniciado no mestrado sob a supervisão do Professor Dr. Oliver Piatella.