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

Vitor Martins Cernic

Síntese de populações estelares de galáxias com técnicas de aprendizado de máquina

São Paulo
2023

Vitor Martins Cernic

Síntese de populações estelares de galáxias com técnicas de aprendizado de máquina

Trabalho de Conclusão de Curso apresentado ao Instituto de Astronomia, Geofisica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Mestre em Astronomia.

Vertente: Astrofísica Extragalática
Orientador: Prof. Dr. Laerte Sodré Júnior

São Paulo
2023

Essa tese é dedicada ao meu pai, que apesar de não poder me ver concluindo essa jornada sempre esteve ao meu lado nos momentos que precisei. Sinto falta do seu abraço.

Agradecimentos

Ao meu pai, que me apoiou desde o início da minha carreira;

À minha mãe e minha família, que me deu forças para continuar em frente;

À Milena, meu porto seguro, que continua ao meu lado depois de tantos anos. Não sei onde estaria hoje sem você;

Ao meu orientador Laerte, por ser quem eu mais pude contar com suporte dentro da Universidade;

Aos amigos Tajan e Fabrícia, por me ajudarem nos momentos difíceis da Universidade. Vocês merecem o melhor;

Aos meus amigos de infância, em especial ao Gabriel Hardt, que nunca me deixaram na mão e sempre me apoiaram;

Aos colegas que conheci através do Magic, em especial a comunidade de juízes, que me ajudaram a amadurecer e me mostraram um lado da vida que nunca seria capaz de buscar;

À CAPES, pelo apoio financeiro sob o projeto 88887.604781/2021-00;

E à todos aqueles cujos nomes estarão pra sempre guardados em mim.

"*No fim das contas é tudo bolinha girando em volta de bolinha, né?*"

Ivan Cernic, meu pai

Resumo

Esse projeto tem como objetivo encontrar os parâmetros de populações estelares de galáxias (massa estelar, extinção de poeira, idades e metalicidades médias) a partir da fotometria do levantamento S-PLUS. Criamos um conjunto de treinamento a partir de uma aplicação do STARLIGHT que combinou os espectros do SDSS com a fotometria do GALEX, melhorando as sínteses espectrais. Após um pré-processamento dos dados que consistiu na adição das linhas espectrais, cálculo da fotometria, calibração com as magnitudes do S-PLUS e imputação de dados faltantes, criamos um conjunto de treinamento com 137,734 galáxias, cada uma com sua fotometria no sistema do S-PLUS e seus respectivos parâmetros de populações estelares fornecidos pelo STARLIGHT. Comparamos 5 algoritmos de regressão diferentes entre Regressão Linear, K-Nearest Neighbours, Random Forest, XGBoost e Redes Neurais. Acabamos escolhendo uma rede neural de *Deep Learning* para as estimativas, consistindo de uma camada de entrada com 13 neurônios (12 magnitudes do S-PLUS + 1 valor de *redshift*) seguido por 4 camadas ReLU com 256 neurônios cada e uma última camada final com 1 neurônio correspondente ao parâmetro de população estelar. Utilizando um método de *Data Augmentation* fomos capazes de gerar uma estimativa de erro para cada predição.

Estimamos 6 parâmetros de populações estelares: massa estelar; absorção por poeira no visível; metalicidade média ponderada em fluxo e em massa; idade média ponderada em fluxo e em massa. Também estimamos as larguras equivalentes de 4 linhas de emissão: Hα, Hβ, [OIII]5007Åe [NII]6583Å. As estimativas são bem satisfatórias e condizem com métodos clássicos de SED *fitting*. Mostramos que o algoritmo é robusto em *redshift* até 0.1 mesmo para estimativas das larguras equivalentes. Realizamos uma aplicação a 182 galáxias do aglomerado de Fornax, obtendo suas relações de idade e metalicidade. Por

fim, recriamos o diagrama BPT através da estimativa das larguras equivalentes das linhas de emissão e vimos que a rede consegue predizer em qual região do diagrama a galáxia pertence.

Abstract

This project aims to obtain the stellar population's parameters of galaxies (stellar mass, dust attenuation, mean ages and metallicities) from the S-PLUS survey photometry. We've created a training set from a STARLIGHT application that combined SDSS spectra with GALEX photometry, improving the spectral synthesis. After preprocessing the data, in which we added the emission lines, calculated the photometry, calibrated with S-PLUS data and dealt with missing data, we've created a training set with 137,734 galaxies, each one with their respective S-PLUS-*like* photometry and their stellar population parameters obtained by STARLIGHT. We compared 5 different regression algorithms, including Linear Regression, K-Neares Neighbours, Random Forest, XGBoost and Neural Networks. We ended up choosing a Deep Learning Neural Network for the estimations, which consisted of an entry layer with 13 neurons (12 S-PLUS magnitudes + 1 redshift value) followed by 4 ReLU activation layers with 256 neurons each and a final layer with a single neuron corresponding to the stellar population parameter. Using a Data Augmentation method we were capable of obtaining an error estimation for each prediction.

We've estimated 6 stellar population parameters: stellar mass; dust attenuation; mean ages weighted by flux and mass; mean metallicity weighted by flux and mass. We've also estimated the equivalent width of 4 emission lines: Hα, Hβ, [OIII]5007Åand [NII]6583Å. The estimations were satisfactory and are in line with classical SED fitting methods. The algorithm is robust up to a 0.1 redshift even for the emission lines estimation. We've applied our models to a sample of 182 galaxies from the Fornax cluster, obtaining their color age relationship. Finally, we've created the BPT diagram through the estimation of equivalent widths of the spectral lines and saw that the network can predict in which region of the diagram a galaxy belongs just from their photometry.

Lista de Figuras

1.1 Uma esquematização geral de como funciona um algoritmo de *Machine Learning*. 26

2.1 Resultados da análise de Werle et al. (2018) para três objetos diferentes, um em cada linha, mostrando o fluxo $F_\lambda/F_{\lambda 0}$ por comprimento de onda λ. $F_{\lambda 0}$ é o valor do fluxo no comprimento de onda de normalização, 5634Å. Em azul está o espectro sintetizado combinando-se a espectroscopia óptica do SDSS com a fotometria do Galex. Em vermelho está a extrapolação para o ultravioleta do ajuste obtido usando-se apenas o espectro óptico. Os fluxos das bandas do Galex são representados por círculos azuis. Os círculos laranja marcam o fluxo esperado no UV quando não se considera a fotometria do GALEX na síntese. Créditos: Werle et al. (2018) 30

2.2 Na figura de cima temos um espectro sem linhas de emissão (somente contribuição estelar) da amostra de Werle et al. (2018). Na figura de baixo vemos o espectro com as linhas de emissão adicionadas (contribuição estelar mais nebular) do ajuste do DOBBY. 31

2.3 Gráfico de fluxo em função do comprimento de onda (*wavelength*) de um espectro. Em azul temos um espectro direto da amostra de Werle et al. (2018) em um redshift no referencial de repouso ($z = 0$) enquanto em vermelho temos esse mesmo espectro em $z = 0.15$ por meio da conta da Equação 2.1. . 31

2.4 Distribuição de *redshift* da amostra de Werle et al. (2018). 33

2.5 Os 12 filtros do S-PLUS em gráfico de comprimento de onda λ por eficiência $E(\lambda)$. Cada filtro é colorido de acordo com a legenda da direita. Créditos: Mendes de Oliveira et al. (2019). 34

2.6 Cálculos de fluxo F_λ em função do comprimento de onda λ. As linhas vermelhas representam o espectro enquanto os pontos azuis representam a fotometria calculada para cada filtro. A posição das abscissas dos pontos é dada pelo λ_{eff}. Na figura de cima, o espectro está em redshift de repouso ($z = 0$) enquanto na figura de baixo temos o mesmo espectro em $z = 0.15$. 34

2.7 Gráficos de calibração para os filtros do S-PLUS. O eixo X são as magnitudes sintéticas calculadas, enquanto o eixo Y são as magnitudes reais vindas diretamente do S-PLUS. As linhas de contorno cinzas mostram a densidade dos dados antes da calibração, enquanto os pontos azuis e as linhas de contorno azuis mostram os dados após aplicar a relação descrita dentro de cada plot. A reta pontilhada representa a linha $X = Y$. 36

2.8 Porcentagem de dados faltantes por filtro (esquerda) e porcentagem de dados faltantes por por objeto (direita). 37

3.1 Representação visual da melhor rede neural obtida. Ela é composta por uma camada de entrada com 13 neurônios (12 filtros do S-PLUS + 1 valor de redshift) e uma de saída composta por 1 único neurônio (um parâmetro de população estelar). Entre elas existem 4 camadas de ativação ReLU escondidas com 256 neurônios cada. 45

3.2 Gráfico da raiz do erro médio quadrático (RMSE) do conjunto de validação para cada uma das 1000 épocas. A área sombreada representa os valores máximos e mínimos de um K-Fold *Cross-Validation* com $K = 5$, enquanto a linha representa a média entre os *folds*. 46

3.3 Representação visual do método *Test-Time Data Augmentation* para saber a chance de existir um ganso na imagem. Um dado inicial passa por uma transformação diversas vezes, e um algoritmo é utilizado para realizar predições pontuais. Por fim, combinamos as predições pontuais para obter uma distribuição de probabilidades. 47

3.4 Histograma de 10,000 estimativas de massa estelar para a galáxia S-PLUS 0073-0011797 utilizando o método TTA. O eixo-X corresponde ao valor da massa predito. O desvio padrão da distribuição está descrito no título enquanto a linha tracejada corresponde à estimativa sem erros observacionais. 48

4.1 Estimativa dos parâmetros de populações estelares. Em cada gráfico, o eixo-X representa o valor obtido espectroscopicamente pelo STARLIGHT, enquanto o eixo-Y representa a estimativa da rede neural. A linha vermelha mostra a identidade $x = y$. 51
4.2 Estimativa das larguras equivalentes das linhas de emissão. Em cada gráfico, o eixo-X representa o valor obtido espectroscopicamente pelo STARLIGHT, enquanto o eixo-Y representa a estimativa da rede neural. A linha vermelha mostra a identidade $x = y$. 52
4.3 Na figura da esquerda, vemos a predição da massa estelar para o conjunto de teste, com os valores pintados dependendo do seu *redshift*. Na figura da direita vemos o erro (*bias*) de cada uma dessas galáxias em relação ao seu *redshift*. 54
4.4 Histogramas de erros para predição da massa estelar utilizando o *redshift* fotométrico (azul) e o *redshift* espectroscópico (laranja) como parâmetro de entrada. 55
4.5 Relação idade *versus* cor G-R para um conjunto de 182 galáxias do aglomerado de Fornax. 57
4.6 Igual à Figura 4.3 porém estimando a largura equivalente de Hα. 57
4.7 Diagramas BPT utilizando dados "reais"provenientes do STARLIGHT (esquerda) e com dados preditos pela rede neural através das magnitudes do S-PLUS (direita). Os pontos são coloridos a partir da posição do eixo-X no diagrama da esquerda. 58
4.8 Valores SHAP obtidos a partir da rede treinada para obter a massa estelar. Quanto maior o valor SHAP de uma *feature*, maior o impacto dela no resultado final da predição. Na esquerda vemos os valores SHAP, enquanto os histogramas da direita nos mostram a média do módulos dos valores SHAP. 59
4.9 Igual a Figura 4.8 mas para a estimativa da largura equivalente de $H\alpha$. . . 60

Lista de Tabelas

1.1 Nome do filtro, comprimento de onda efetivo (ou central) λ_{eff} e intervalo de comprimento de onda $\Delta\lambda$ para cada um dos 12 filtros do S-PLUS. A coluna "Comment" traz informações adicionais sobre a relação dos filtros com outros *surveys* ou quais linhas espectrais as bandas estreitas podem detectar em baixos *redshifts*. Créditos: Mendes de Oliveira et al. (2019) . . 24

2.1 MAE e RMSE do conjunto de teste quando substituímos as os dados faltantes por cada valor. MICE corresponde ao algoritmo de imputação. Utilizamos a idade média ponderada em fluxo para a comparação. 38

3.1 Resultados dos modelos para o conjunto de teste. 44

4.1 Parâmetro estimado e erros do conjunto de teste para este trabalho e o de Thainá-Batista et al. (2023) para galáxias com S/N = 100. Os valores de erro de cada medida correspondem ao σ_{NMAD} do *bias* das predições. 53

Sumário

1. Introdução 21

1.1 Motivação 21

1.2 Espectros e parâmetros de galáxias 22

1.3 Levantamentos fotométricos 23

1.4 Uma visão geral sobre aprendizado de máquina 25

2. Dados 27

2.1 A síntese espectral de Ariel Werle 27

2.2 Adição de linhas espectrais em emissão 29

2.3 Espectro em diferentes *redshifts* 29

2.4 Cálculo da fotometria de cada espectro 32

2.5 Calibração da fotometria com o S-PLUS 35

2.6 Erros e dados faltantes 35

2.7 Os parâmetros das populações estelares 39

3. Algoritmo 41

3.1 A escolha do modelo 42

3.2 A rede neural 44

3.3 Estimativa de erros 45

4. Análises 49

4.1 Estimativa dos parâmetros 49

4.2 Dependência de *redshift* 53

4.3 Aplicação em aglomerados próximos: Fornax 56

4.4 Diagrama BPT 56
4.5 Estudo de importância das *features* 58

5. Conclusões 61

Referências 63

Capítulo 1

Introdução

1.1 Motivação

Com a globalização e o avanço da tecnologia nos últimos anos, o tempo onde a humanidade era capaz de estudar todo o conteúdo produzido em uma única área do conhecimento há muito se fora. A área de *data science* cresce como nenhuma outra, e instituições ao redor do mundo buscam maneiras eficientes de tratar seus grandes conjuntos de dados, incluindo a Astronomia (Baron, 2019). Vivemos hoje no período do *Big Data*, uma época onde o avanço tecnológico e científico da sociedade estão atrelados à eficiência dos modelos que analisam as informações de forma massiva, ao invés de estudar os dados separadamente. Mas não se deve deixar enganar pensando que esse saturamento de informações é algo ruim para a humanidade. Grandes conjuntos de dados, a partir de análises estatísticas robustas, revelam informações sobre eventos que a mente humana nunca foi capaz compreender completamente, e relações que antes pareciam abstratas começam a tomar formas mais concretas. Com *data science* conseguimos explorar as relações desde tarefas que parecem mundanas, como diferenciar computacionalmente um gato de um cachorro, até tarefas que pareciam nunca antes possíveis, como obter o *redshift* dada uma fotometria.

É inevitável que a Astronomia, como a maioria das outras áreas da ciência, seja afetada pelo fenômeno da avalanche de dados. Conforme o avanço na tecnologia de telescópios aumenta, o volume de dados gerado cresce de maneira significativa, fazendo com que a quantidade de informações que eram geradas por levantamentos inteiros (*surveys*) há uma década são geradas hoje em uma noite. Esse gigantesco fluxo atual é discutido por Kremer et al. (2017), onde se nota que a tendência para o futuro é que esse volume só aumente, devido aos levantamentos em andamento ou aos esperados para um futuro

próximo. Com isso, é necessário a criação de novas ferramentas para que esses dados consigam ser compreendidos pelos astrônomos, já que esses levantamentos são as bases para abordar as grandes perguntas da astronomia atual: "do que é feito o Universo?"; "o que é a misteriosa matéria escura?"; "como a Via Láctea se formou?"; entre muitas outras.

1.2 Espectros e parâmetros de galáxias

Analisar a distribuição espectral de uma galáxia é uma das principais formas de obter informações sobre suas propriedades. Os diferentes processos físicos que ocorrem dentro de cada galáxia deixam marcas em seus espectros, e isso pode ser um indicativo de como entender as propriedades daquela galáxia (Walcher et al., 2010). Para extrair tais informações, são necessários modelos que conectem as propriedades físicas com os espectros. *Spectral Energy Distribution Fitting*, ou SED *fitting*, é o método tradicional para estudar a energia espectral das galáxias.

A luz das galáxias se origina da luz das estrelas, seja diretamente ou reprocessada pelo gás e poeira do meio interestelar. Isso só não vale para galáxias com núcleo ativo, onde um disco de acreção em torno de um buraco negro supermassivo central pode dominar sua emissão (Marshall et al., 2022). Podemos modelar a luz dessas estrelas simplificando a emissão de uma galáxia como uma combinação de diversas populações estelares simples, onde as estrelas de uma população são todas da mesma idade e metalicidade. Esse método é utilizado pela maior parte dos modelos de populações estelares e já está estabelecido há muitos anos (Charlot e Bruzual 1991; Chiosi et al. 1988), porém se mostram limitados pelo nosso conhecimento sobre evolução estelar.

Através da modelagem das populações estelares junto com o meio interestelar é possível obter as propriedades das galáxias. Existem diversos códigos que foram criados, e cada um deles tem suas peculiaridades. Como o espectro pode ser representado por uma soma dos espectros individuais das populações estelares, se desconsiderarmos efeitos como poeira nos resta simplesmente um problema linear que pode ser solucionado a partir de uma inversão de matriz. Esse é o processo ao qual a maioria dos códigos funcionam, como MOPED (Heavens et al., 2000), STECKMAP (Ocvirk et al., 2006) e FADO (Gomes e Papaderos, 2017). Esses códigos mostram bons resultados quando aplicados à espectro de galáxias, porém obter o espectro em primeiro lugar é um privilégio, e não a norma. Aplicar esses códigos

para fotometrias é algo imprescindível, porém é algo que ainda está evoluindo. Códigos como o Prospector (Johnson et al., 2021) são muito reconhecidos pela sua performance, mas acabam sendo limitados pela escolha dos priores e tempo computacional.

Neste projeto iniciamos a criação de uma ferramenta baseada em aprendizado de máquina (*machine learning*, ML) para extrair alguns parâmetros de populações estelares como metalicidade e idade média a partir da fotometria de um dos levantamentos fotométricos mais importantes atualmente, o *Southern Photometrical Local Universe Survey* (S-PLUS, Mendes de Oliveira et al. 2019).

1.3 Levantamentos fotométricos

Os procedimentos que iremos desenvolver serão aplicados inicialmente aos objetos do Universo local ($z < 0.1$) observados pelo levantamento S-PLUS, porém podem ser generalizado para outros *surveys*. Diferentemente dos levantamentos fotométricos mais comuns de 5 filtros, como o *Sloan Digital Sky Survey* (SDSS, Kollmeier et al. (2019)), o S-PLUS é um *survey* fotométrico com 12 filtros (7 bandas estreitas e as 5 bandas largas *u, g, r, i, z*) descritos na Tabela 1.1. Enquanto as bandas largas contêm informações sobre o contínuo de um espectro, as bandas estreitas representam melhor, em baixos *redshifts*, regiões de linhas de emissão e absorção. Essas informações adicionais sobre o espectro devem ajudar o algoritmo a fazer predições melhores.

O projeto S-PLUS está em andamento em Cerro Tololo, no Chile, onde o telescópio T80-S está instalado. O projeto começou a coletar dados em janeiro de 2017, e em 18/12/2017 foi anunciado o *First Internal Data Release of Stripe82 fields*, consistindo de 88 apontamentos que cobrem grande parte da Stripe 82, previamente observada pelo SDSS e por outros *surveys*, o que permite aferir a qualidade das observações além de fazer ciência. Atualmente, o S-PLUS está no seu quinto *Data Realease* interno.

O levantamento J-PLUS, acrônimo de *Javalambre Photometric Local Universe Survey* (Cenarro et al., 2019), está em andamento desde o final de 2015 com o telescópio T80-N e é equivalente ao S-PLUS mas para o hemisfério Norte. Em 29 de setembro de 2017 o *Centro de Estudios de Física del Cosmos de Aragón* (CEFCA) anunciou o acesso aberto ao *Early Data Release* (EDR) deste *survey*, que consiste em 18 apontamentos em diferentes regiões do céu, cobrindo 36 graus quadrados. Os catálogos disponibilizados juntos com as imagens

Filter name	λ_{eff} [Å]	$\Delta\lambda$ [Å]	Comment
uJava	3574	330	Javalambre u
J0378	3771	151	[O II]
J0395	3941	103	Ca H+K
J0410	4094	201	Hδ
J0430	4292	200	G-band
gSDSS	4756	1536	SDSS-like g
J0515	5133	207	Mgb Triplet
rSDSS	6260	1462	SDSS-like r
J0660	6614	147	Hα
iSDSS	7692	1504	SDSS-like i
J0861	8611	408	Ca Triplet
zSDSS	8783	1072	SDSS-like z

Tabela 1.1 - Nome do filtro, comprimento de onda efetivo (ou central) λ_{eff} e intervalo de comprimento de onda $\Delta\lambda$ para cada um dos 12 filtros do S-PLUS. A coluna "Comment" traz informações adicionais sobre a relação dos filtros com outros *surveys* ou quais linhas espectrais as bandas estreitas podem detectar em baixos *redshifts*. Créditos: Mendes de Oliveira et al. (2019)

contêm mais de 400 mil objetos astronômicos observados nas 12 bandas do projeto, sendo 150 mil estrelas e 101 mil galáxias até a magnitude $r < 21$. Diversos artigos com resultados desse *survey* já foram publicados.

O J-PAS, *Javalambre Physics of the Accelerating Universe Astrophysical Survey* (J-PAS, Benitez et al. 2014) é uma colaboração hispano-brasileira que deve conduzir um levantamento fotométrico de aproximadamente 8500 graus quadrados do céu com um sistema de 59 filtros (54 estreitos, com aprox. 110 deg A de largura cada, e 5 de banda larga), a partir de um observatório dedicado, o *Javalambre Astrophysical Observatory* (JAO), em Teruel, Espanha, também operado pelo CEFCA. O JAO conta com dois telescópios, um de 80cm (o T80-N) para calibração, e um de 2.5m (o T250) que conduzirá o levantamento propriamente dito. Este último está equipado com uma câmera CCD de 1.2 Gpixel, com um campo de visão de 3 graus quadrados - a segunda maior do mundo em número de pixels, atrás apenas da usada no projeto Pan-Starrs (Kaiser et al., 2010).

1.4 Uma visão geral sobre aprendizado de máquina

Machine Learning e Inteligência Artificial são ferramentas estatísticas que fazem parte de nosso dia-a-dia. Elas se baseiam em encontrar relações complexas e são aplicadas em diversos assuntos: ferramentas de pesquisa (Radford et al., 2019), análises financeiras de crédito (Chow, 2017), reconhecimento facial (Balaban, 2015), autonomia de veículos (Kuutti et al., 2020), entre outros.

Apesar de ter um mundo imenso de aplicações, os modelos de aprendizado de máquina sempre partem de um mesmo princípio e funcionam de formas similares. Existe um conjunto de dados de treinamento que, como o nome diz, treina um modelo, isto é, serve para determinar seus parâmetros. Esse modelo pode variar muito, desde métodos clássicos (como regressão linear ou árvores de decisão) até algoritmos de *Deep Learning* (DL) baseados em redes neurais. Após o treinamento do modelo, podemos aplicá-lo a novos dados que não fizeram parte do conjunto de treinamento. Uma esquematização geral de ML pode ser visto na Figura 1.1.

Existem alguns tipos diferentes de apredizado, com uma diferença especial entre supervisionado e não-supervisionado. Aprendizado supervisionado está relacionado a problemas de classificação e regressão, onde conhecemos tanto os valores de entrada (*inputs*, *features*) quanto os valores de saída (*outputs*), e alguns exemplos são como diferenciar imagens de cães e gatos (Parkhi et al., 2012) ou estimar *redshifts* fotométricos (Benítez, 2000). Apredizado não-supervisionado se refere a problemas onde não existe um resultado esperado bem definido, como agrupamento de objetos (Barbosa et al., 2022) ou redução de dimensionalidade com autoencoders (Pat et al., 2022).

Durante os últimos anos o desenvolvimento de técnicas de aprendizado de máquina dentro da Astronomia creceu como nunca antes. Seguindo a análise de Meher e Panda (2021), temos uma compilação de exemplos de aplicações: habitabilidade de exoplanetas (Jagtap et al. 2021; Jakka 2023); descoberta e classificação de anomalias (Etsebeth et al. 2023; Böhm et al. 2023); busca de ondas gravitationais (Meijer et al. 2023; Nousi et al. 2023); estimativa de *redshift* fotométrico (Jones et al. 2023; Ait-Ouahmed et al. 2023).

Em nosso caso queremos achar a relação entre fotometria e parâmetros das populações estelares. Os dados (conjunto) de treinamento devem conter, de um lado, fotometrias, e do outro lado, propriedades de populações estelares. Nessa tese discutimos a criação do

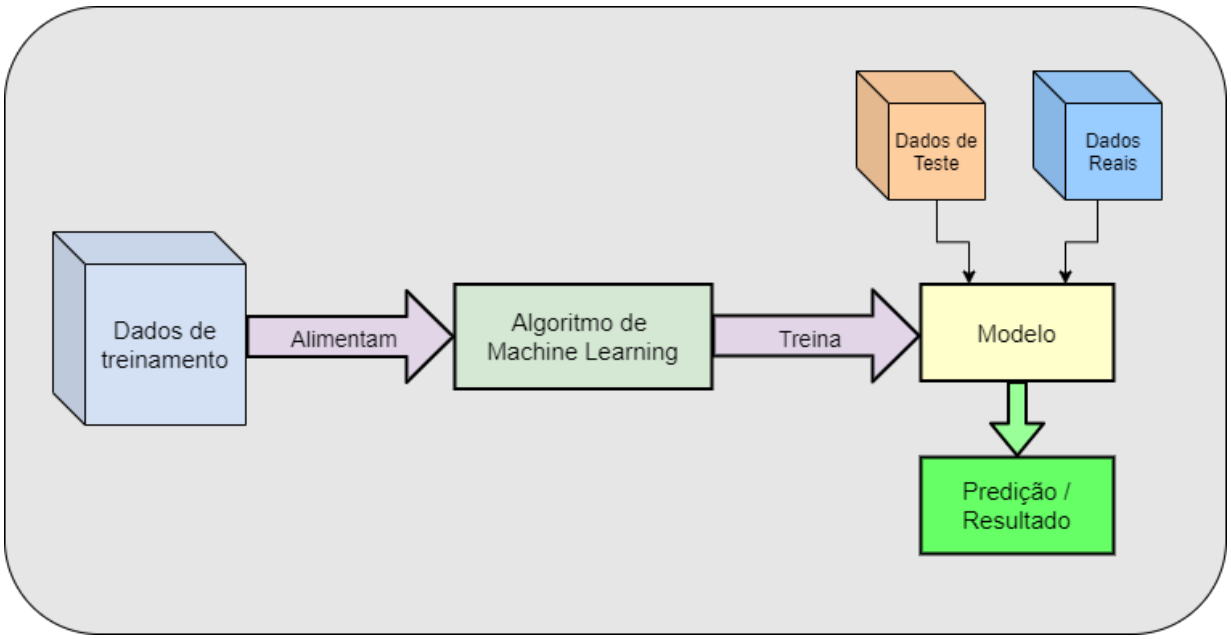

Figura 1.1: Uma esquematização geral de como funciona um algoritmo de *Machine Learning*.

conjunto de treinamento no Capítulo 2, a escolha do algoritmo no Capítulo 3 e finalmente apresentamos algumas possíveis análises que podem ser feitas a partir desse estudo no Capítulo 4.

Capítulo 2

Dados

Em problemas que utilizam aprendizado de máquina, o algoritmo em si é só uma pequena parte do trabalho. Esse modelos aprendem unicamente em cima dos dados que são passados, e por isso o tratamento dos dados tem um papel muito importante. Se fornecermos informações ruins, ou que não condizem com as observações que o modelo será aplicado, obteremos um treinamento ruim dos algoritmos. "*Garbage in, garbage out*" (*entra lixo, sai lixo*) é um lema da área que descreve bem esse problema.

Queremos criar um conjunto de treinamento que represente os dados ao qual o modelo será aplicado. Essa tarefa pode parecer um pouco contraditória, já que para isso precisamos dos dados observacionais. E se temos os dados observacionais, por que não usar eles pra treinar os algoritmos? A explicação é que precisamos fornecer para esses modelos tanto as perguntas (fotometrias) quanto as respostas (parâmetros das populações estelares) para cada uma das galáxias, e essas respostas nós ainda não temos para o conjunto de galáxias observadas.

2.1 A síntese espectral de Ariel Werle

Uma das formas mais tradicionais de estudar as propriedades físicas de objetos astronômicos é através de *Spectral Energy Distribution Fitting* (SED *fitting*). Essas técnicas analisam a distribuição de energia espectral de uma galáxia e a partir de modelos de síntese de populações conseguem estimar valores de massa, metalicidade e outras características das populações estelares das galáxias. Existem diversos modelos e algoritmos diferentes disponíveis nesse meio, cada um com objetivos diferentes. Por exemplo, o STARLIGHT (Cid Fernandes et al., 2005a) é um código de SED *fitting* que descreve o perfil espectral

de uma galáxia através das populações estelares simples que a compõem. Esses métodos funcionam mesmo quando a informação espectral é limitada, como em casos onde se aplicam à fotometria (e.g. Thainá-Batista et al. 2023). Neste projeto, queremos criar um algoritmo que consiga obter boas estimativas das propriedades físicas de galáxias a partir de sua fotometria utilizando técnicas de aprendizado de máquina.

Algoritmos de aprendizado supervisionado se baseiam em olhar conjuntos de dados para criar relações entre as informações descritas. Isso significa que esses algoritmos não levam em conta leis físicas, eles simplesmente encontram relações estatísticas entre os dados. Por um lado isso é bom porque podemos utilizar essa ferramenta para estudar relações mais complexas. Por outro lado é necessário tomar muito cuidado para que o conjunto de treino represente bem tal relação. Se os dados de treinamento forem enviesados, o modelo também mostrará um viés nas suas predições. É necessário criar um conjunto de dados robusto, que tenha dados fotométricos dos objetos além das estimativas das propriedades que se quer estimar, como massa estelar, metalicidade média, idade média, entre outros.

Vamos iniciar a construção desse conjunto de dados utilizando como base o trabalho de Ariel Werle em sua tese de doutorado, Werle (2019). Neste trabalho foram feitas as sínteses espectrais de 137,734 galáxias da STRIPE-82 através de uma nova aplicação do *software* STARLIGHT. Como descrito anteriormente, o STARLIGHT é um programa que extrai propriedades físicas de galáxias ajustando seus espectros com uma combinação linear de espectros de populações estelares simples. Werle usou uma versão modificada do STARLIGHT para gerar sínteses espectrais de galáxias, e sua implementação traz várias inovações em relação às sínteses originais. Em particular, combina os espectros do SDSS com dados fotométricos do Galaxy Evolution Explorer (GALEX, Martin et al. 2005), usando modelos do estado da arte para populações estelares. Estes resultados inovam pois a síntese é feita levando em conta a emissão no ultravioleta (UV) das galáxias, o que melhora a estimativa das propriedades das populações estelares. Essas propriedades foram obtidas para cada um dos objetos da amostra de Werle, então nos resta transformar os espectros em fotometrias (para um dado *redshift*).

Na Figura 2.1 vemos a comparação entre os dados obtidos por Werle et al. (2018) com os espectros do SDSS e do GALEX para três galáxias diferentes. Estes resultados ilustram o enorme ganho na modelagem de espectros com o STARLIGHT quando se inclui na análise, além do espectro óptico, a fotometria UV. A mudança significativa no ajuste no UV leva a

uma melhor determinação das populações estelares, principalmente as associadas a estrelas jovens.

2.2 Adição de linhas espectrais em emissão

Como o objetivo do trabalho de Werle et al. (2018) é a determinação das populações estelares das galáxias, apenas o contínuo dos espectros é ajustado, ou seja, as linhas de emissão são excluídas das análises pois são produzidas pelo gás ionizado, não pelas estrelas. O problema disso é que a inclusão das linhas espectrais é essencial para fotometria (principalmente no caso de bandas estreitas) e devemos considerá-las para criar modelos realísticos para o S-PLUS e outros levantamentos de dados.

Para adicionar as linhas de emissão aos modelos de espectro de Werle et al. (2018), utilizamos o *software* DOBBY (Flórido (2018)). Este é um código que ajusta linhas de emissão assumindo que elas apresentam um perfil gaussiano. Entramos em contato com o grupo responsável pelo DOBBY para nos auxiliar no projeto. Essa comunicação, facilitada pela familiaridade do Ariel Werle com este grupo, forneceu, para cada um dos 137,734 objetos da amostra, um novo espectro sintético contendo as linhas de emissão. Um desses objetos está representado na Figura 2.2.

2.3 Espectro em diferentes *redshifts*

Dado um certo fóton emitido por um objeto com comprimento de onda λ_0 no referencial de repouso, o comprimento de onda λ do fóton observado se relaciona com λ_0 por um fator de $(1 + z)$, onde z é o chamado *redshift*. Em termos matemáticos, definimos a Equação 2.1.

$$\lambda = \lambda_0(1 + z). \tag{2.1}$$

Uma das vantagems de utilizar a amostra de Werle et al. (2018) é que os espectros são ajustados em *redshift* $z = 0$, ou seja, no referencial de repouso. Isso significa que, como conhecemos λ_0, podemos utilizar a Equação 2.1 para aplicar a esse espectro um valor de z e calcular qual seria o espectro desse objeto em diferentes redshifts. Na Figura 2.3 vemos um dos objetos da amostra representado tanto em $z = 0$ quanto em $z = 0.15$.

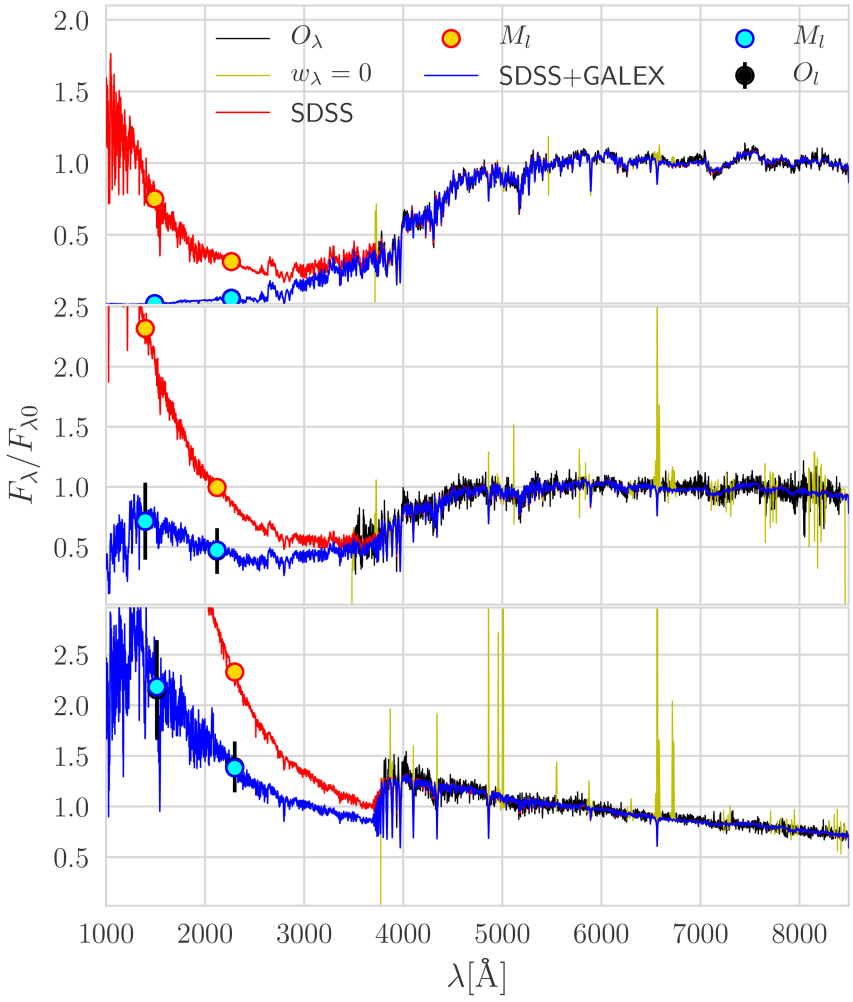

Figura 2.1: Resultados da análise de Werle et al. (2018) para três objetos diferentes, um em cada linha, mostrando o fluxo $F_\lambda/F_{\lambda 0}$ por comprimento de onda λ. $F_{\lambda 0}$ é o valor do fluxo no comprimento de onda de normalização, 5634Å. Em azul está o espectro sintetizado combinando-se a espectroscopia óptica do SDSS com a fotometria do Galex. Em vermelho está a extrapolação para o ultravioleta do ajuste obtido usando-se apenas o espectro óptico. Os fluxos das bandas do Galex são representados por círculos azuis. Os círculos laranja marcam o fluxo esperado no UV quando não se considera a fotometria do GALEX na síntese. Créditos: Werle et al. (2018)

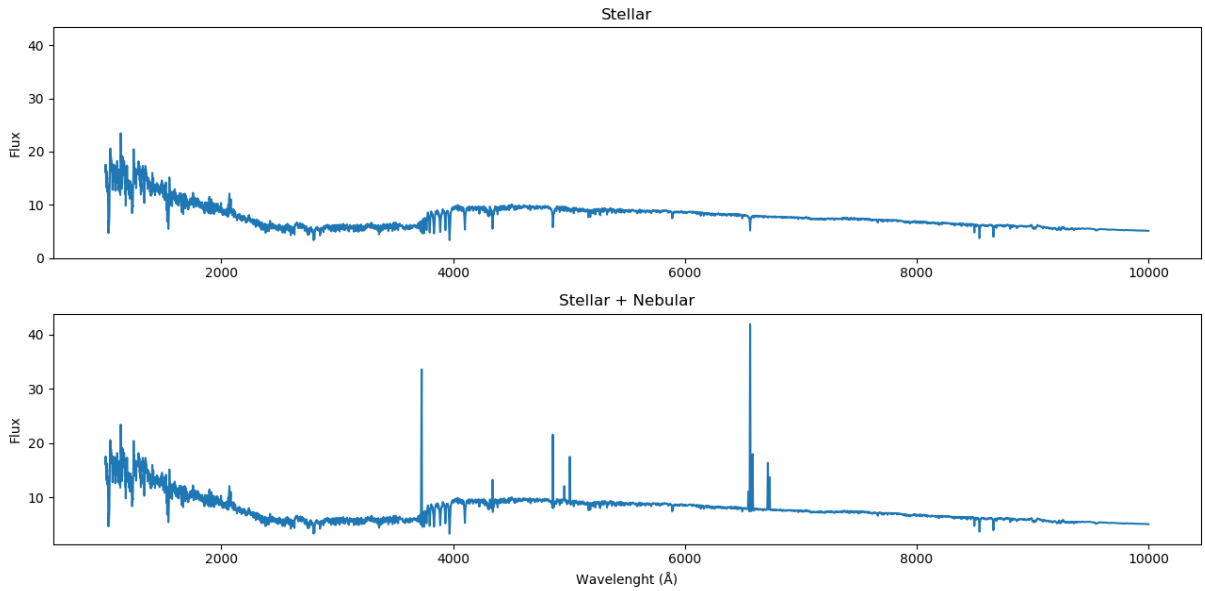

Figura 2.2: Na figura de cima temos um espectro sem linhas de emissão (somente contribuição estelar) da amostra de Werle et al. (2018). Na figura de baixo vemos o espectro com as linhas de emissão adicionadas (contribuição estelar mais nebular) do ajuste do DOBBY.

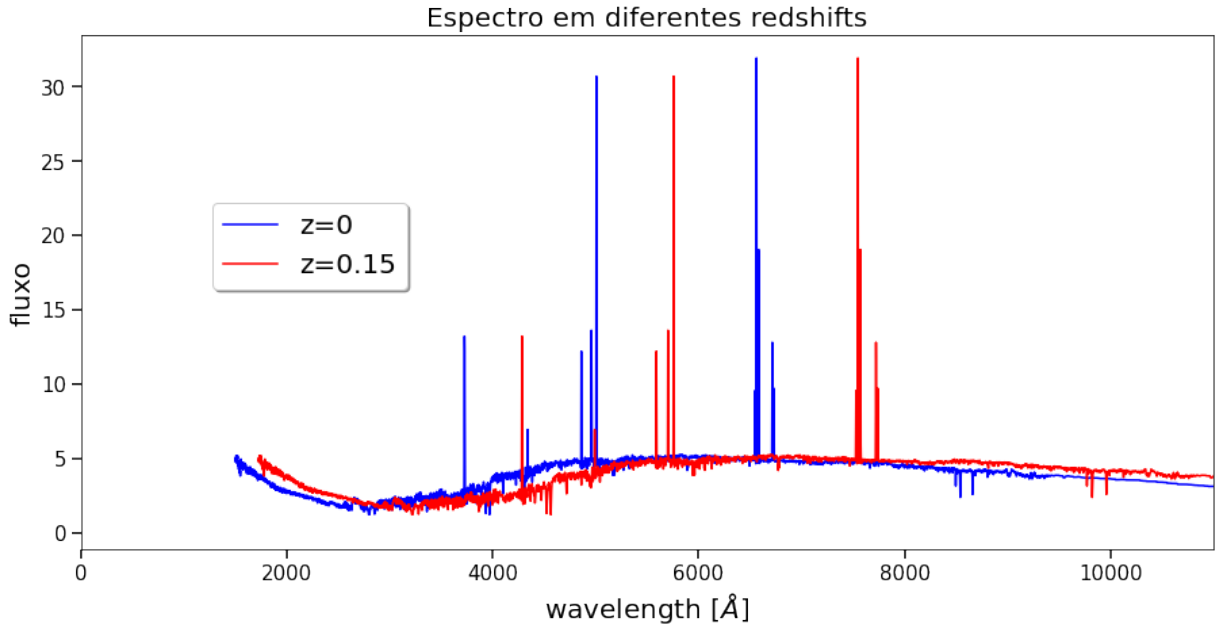

Figura 2.3: Gráfico de fluxo em função do comprimento de onda (*wavelength*) de um espectro. Em azul temos um espectro direto da amostra de Werle et al. (2018) em um redshift no referencial de repouso ($z = 0$) enquanto em vermelho temos esse mesmo espectro em $z = 0.15$ por meio da conta da Equação 2.1.

A ideia é que aplicando diferentes valores de z para um espectro, podemos criar diversos objetos artificiais. Essa técnica é conhecida por *forward modeling*, e é muito utilizada em aplicações de *Machine Learning* para sintetizar novos dados. É importante lembrar que os algoritmos de aprendizado de máquina aprendem em cima unicamente dos dados, e não de relações empíricas. Caso haja interesse em estudar um grupo de objetos em um mesmo *redshift*, é muito valioso poder treinar um modelo a partir de dados de galáxias naquele *redshift*, já que a rede aprenderia mais especificamente as relações entre os parâmetros das populações estelares e as fotometrias independente de uma das variáveis. Isso só vai ser válido caso não haja uma diferença intrínsica entre galáxias de diferentes *redshifts* da nossa amostra de treino. Ou seja, vamos supor que a amostra de galáxias do trabalho de Werle et al. (2018) representa as galáxias das amostras que iremos analisar.

Temos muito interesse em aplicar esses conceitos para aglomerados próximos, com um foco principal em Fornax (Castelli et al., 2021) e Hydra (Lima-Dias et al., 2020) já que podemos assumir que as galáxias pertencentes a esses aglomerados tem todas um mesmo valor de *redsfhit*. Ambos esses aglomerados têm *redshifts* muito baixos (0.0046 e 0.012, respectivamente) e tem grupos de pesquisa que poderiam se beneficiar muito dos parâmetros de populações estelares de cada uma das galáxias dessas estruturas.

Inicialmente, a distribuição de *redshift* da amostra de Werle et al. (2018) está representada na Figura 2.4. Existem poucos objetos com a fotometria de aglomerados próximos (e.g. Fornax), porém temos um limite superior de $z = 0.1$. Isso significa que a nossa amostra contem objetos os quais as suas linhas de emissão vão ficar fora dos filtros estreitos, que é o caso da linha de $H\alpha$ para todos os objetos com $z > 0.02$. Discutimos um pouco mais sobre esse tópico na Seção 4.2.

2.4 Cálculo da fotometria de cada espectro

A fotometria dos espectros depende de quais filtros são escolhidos para serem aplicados. Ela é calculada a partir do fluxo $F(\lambda)$ do espectro e depende do desenho do filtro e fatores físicos do telescópio. Assim, é necessário conhecer também a função de eficiência $E(\lambda)$ de cada filtro. Essa eficiência vem da multiplicação entre a transmissão da óptica do telescópio, a curva de transmissão do filtro, a transmissão atmosférica e a eficiência do CCD. As funções de eficiência de cada filtro são fornecidas diretamente pelos *websites* de

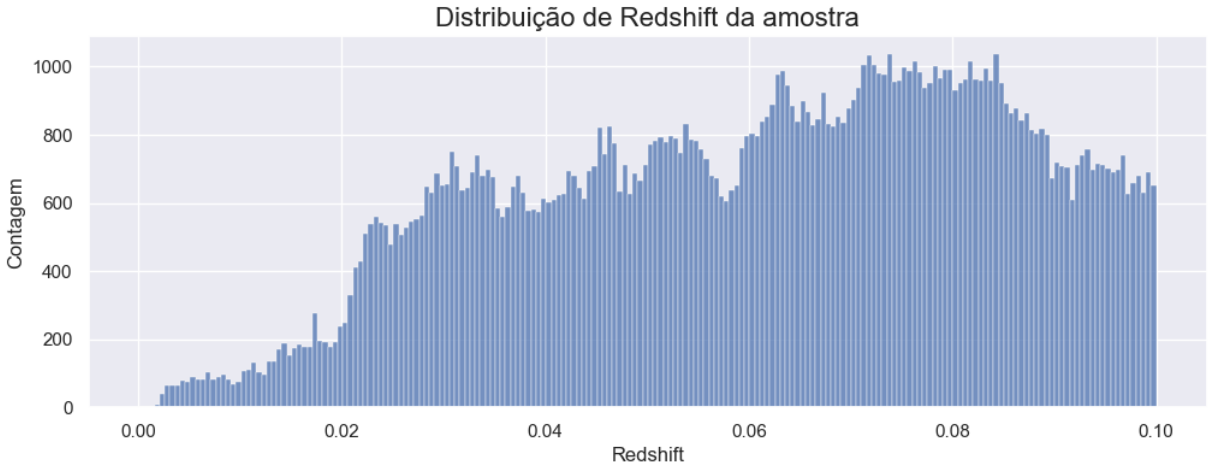

Figura 2.4: Distribuição de *redshift* da amostra de Werle et al. (2018).

cada um dos levantamentos. Na Figura 2.5, mostramos as curvas de eficiência para o S-PLUS.

A magnitude instrumental m_{filtro} em um certo filtro pode ser calculada a partir da Equação 2.2

$$m_{filtro} = \frac{\int \lambda E(\lambda) F(\lambda) d\lambda}{\int \lambda E(\lambda) d\lambda}. \tag{2.2}$$

Podemos também calcular o comprimento de onda efetivo λ_{eff} (ou comprimento de onda central)

$$\lambda_{eff} = \sqrt{\frac{\int \lambda E(\lambda) d\lambda}{\int E(\lambda)/\lambda d\lambda}}. \tag{2.3}$$

Mais detalhes dessas definições encontram-se no manual do Synphot (STScI development Team, 2018).

Usamos as Equações 2.2 e 2.3 para calcular a fotometria de cada espectro sintético nas 12 bandas do S-PLUS. Isso foi realizado a partir do pacote em Python do Synphot, e assim obtivemos como resultado desse código a Figura 2.6 para a galáxia ilustrada na Figura 2.3. Com o código criado, calculamos a fotometria para todas as galáxias da amostra em seus *redshifts* originais.

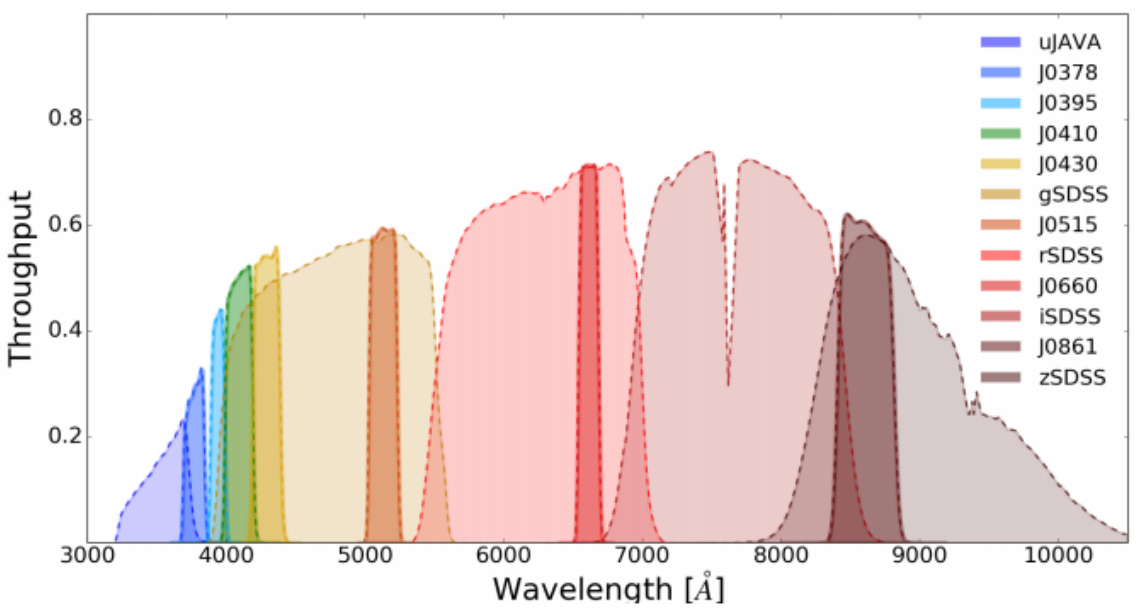

Figura 2.5: Os 12 filtros do S-PLUS em gráfico de comprimento de onda λ por eficiência $E(\lambda)$. Cada filtro é colorido de acordo com a legenda da direita. Créditos: Mendes de Oliveira et al. (2019).

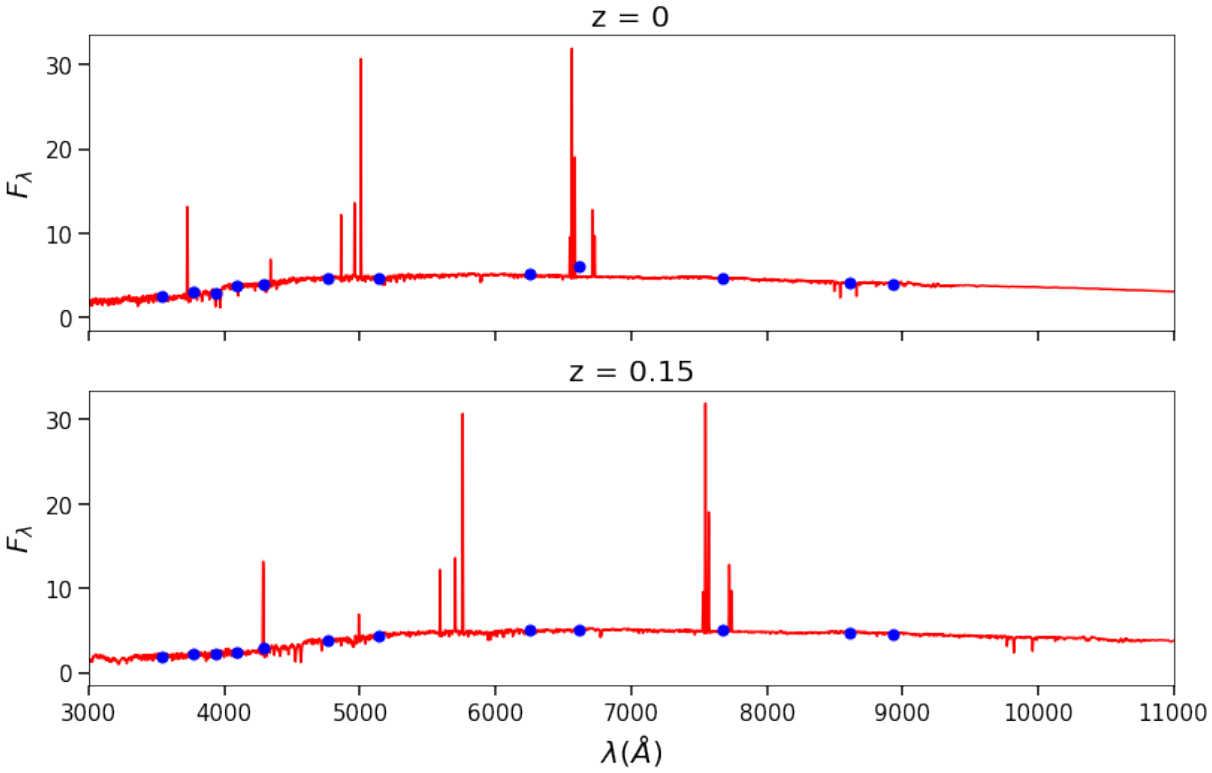

Figura 2.6: Cálculos de fluxo F_λ em função do comprimento de onda λ. As linhas vermelhas representam o espectro enquanto os pontos azuis representam a fotometria calculada para cada filtro. A posição das abscissas dos pontos é dada pelo λ_{eff}. Na figura de cima, o espectro está em redshift de repouso ($z = 0$) enquanto na figura de baixo temos o mesmo espectro em $z = 0.15$.

2.5 Calibração da fotometria com o S-PLUS

As magnitudes instrumentais calculadas com a Equação 2.2 duferen das magnitudes sintéticas por uma constante de calibração que pode ser determinada comparando as magnitudes instrumentais com as magnitudes reais. Além disso, é muito comum que ao calcular magnitudes sintéticas exista um certo desvio sistemático em relação aos dados reais. Isso é visto em vários outros artigos da área (e.g. Lima et al. 2022) e a calibração dessas magnitudes é sempre necessária.

Realizamos um cruzamento entre os dados da síntese de Werle et al. (2018) com as galáxias provenientes diretamente do *Data Release 3* do S-PLUS[1], obtendo assim 6,670 galáxias que têm tanto fotometrias calculadas sintéticamente a partir dos espectros de Werle et al. (2018) quanto fotometrias reais obtidas pelo S-PLUS. A Figura 2.7 mostra, nas curvas de nível, a relação inicial entre a magnitude calculada e a observada. Podemos notar claramente que existe um desvio sistemático entre esses dois dados. É imprescindível que essas magnitudes sintéticas sejam corrigidas, senão a rede predirá valores que não são condizentes. Para arrumá-las, podemos aplicar uma relação linear, transformando todos os dados em cima da linha 1x1. Essas relações lineares são então aplicadas à fotometria sintética de todas as outras 135 mil galáxias da nossa amostra sintética, assegurando que as magnitudes sintéticas estão de acordo com as magnitudes reais.

Um ponto importante a se comentar é que esse ajuste não foi feito usando uma regressão linear ordinária. Acabamos utilizando uma regressão de distância ortogonal (Boggs e Donaldson, 1989) para manter a simetria das variáveis x e y.

2.6 Erros e dados faltantes

É importante criar um conjunto de dados que represente dados reais, e pensando nisso precisamos ter certeza de que os erros e os dados faltantes são considerados. Realizamos um *query* de todos os dados do S-PLUS DR4 dos objetos classificados como galáxias que continham uma medida das bandas *g, r, i* e *z* menor que 21.3. Esse valor foi escolhido por ser o limite superior de magnitude do S-PLUS. Na Figura 2.8 vemos a distribuição de dados faltantes para essa amostra do S-PLUS baseado tanto no filtro quanto no número de dados faltantes por objeto. Também estudamos a correlação de dados faltantes entre os

[1] Query feito pelo site `https://splus.cloud/`

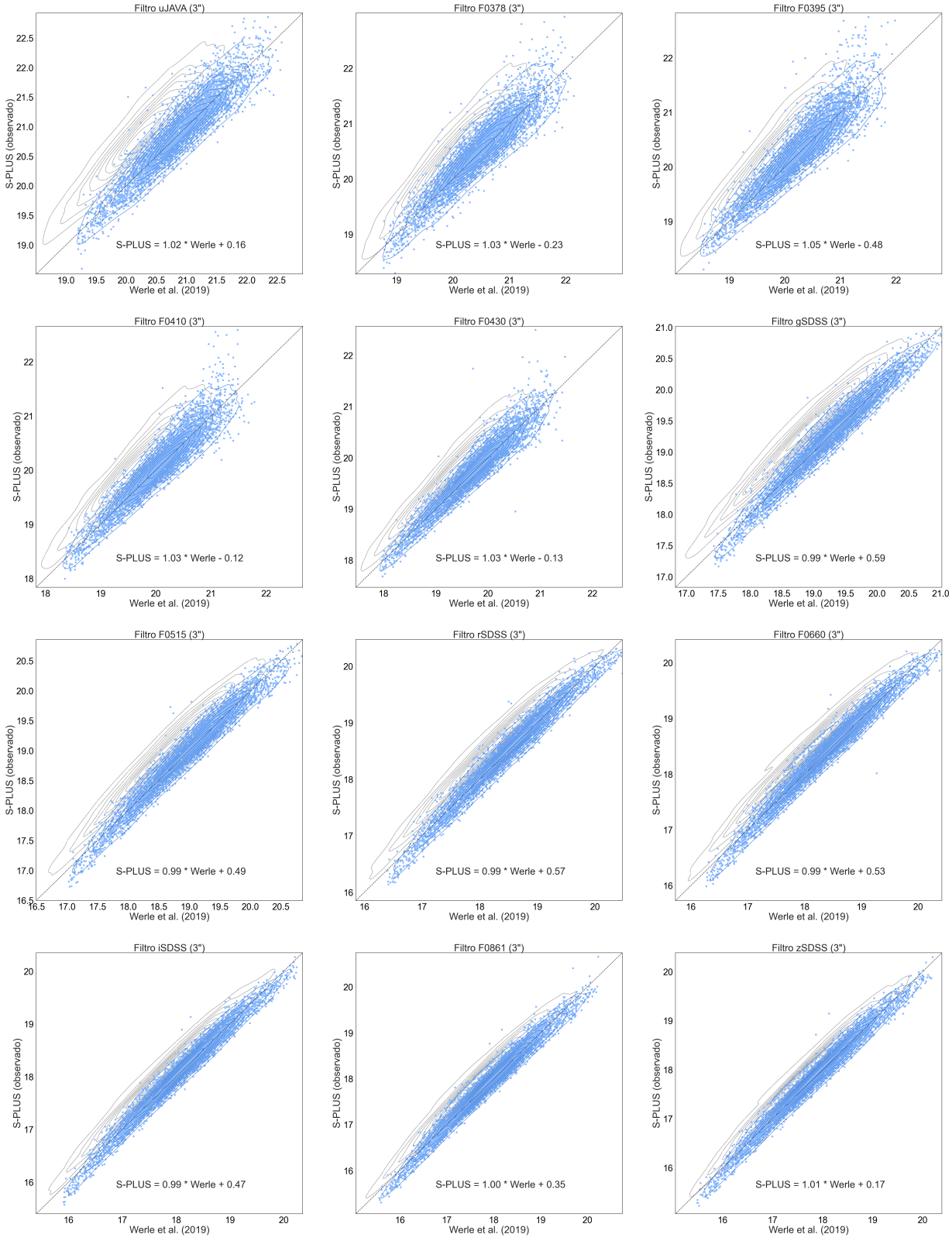

Figura 2.7: Gráficos de calibração para os filtros do S-PLUS. O eixo X são as magnitudes sintéticas calculadas, enquanto o eixo Y são as magnitudes reais vindas diretamente do S-PLUS. As linhas de contorno cinzas mostram a densidade dos dados antes da calibração, enquanto os pontos azuis e as linhas de contorno azuis mostram os dados após aplicar a relação descrita dentro de cada plot. A reta pontilhada representa a linha $X = Y$.

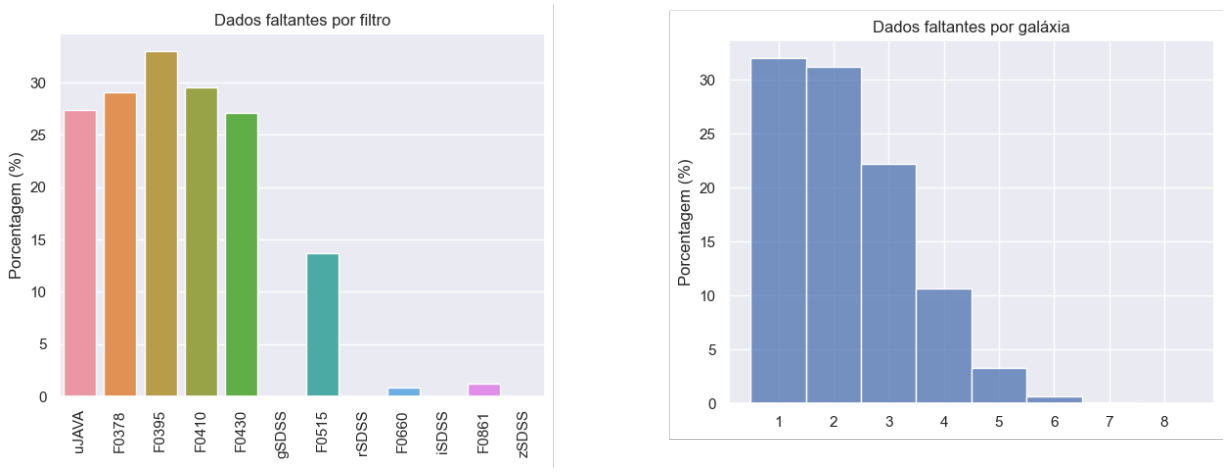

Figura 2.8: Porcentagem de dados faltantes por filtro (esquerda) e porcentagem de dados faltantes por por objeto (direita).

filtros, ou seja, se um objeto tem um dado faltante na região azul, queremos ver se ele tem uma chance maior de ter dado faltante em outra banda na região azul. Essa correlação se mostrou muito baixa em todos os filtros, com exceção dos filtros F0660 e F0861, ambos na região do vermelho.

A partir dessas análises fomos capazes de adicionar dados faltantes na amostra de dados sintéticos de forma que ficasse mais próxima dos dados reais obtidos pelo S-PLUS. As distribuições de dados faltantes foram preservadas, junto com as correlações. Também levamos em conta algumas outras relações que poderiam influenciar os dados faltantes, como o *redshift* e a magnitude na banda r, porém nenhuma delas se mostrou relevante o suficiente.

Apesar de ser um problema simples, o valor escolhido para colocar no lugar de dados faltantes pode ter um impacto no resultado da estimativa. A maior parte dos algoritmos não tem um valor que a máquina entende como faltante (como *Not A Number*), então é necessário substituir por algum número de nossa escolha. Nós testamos diversos valores fixos diferentes (0, 99, -99...) e um método de imputação. Após comparar os resultados não vimos diferenças entre os valores fixos, porém o método de imputação se mostrou um pouco melhor. Acabamos utilizando o algoritmo MICE (*Multiple Imputation by Chained Equations*, van Buuren e Groothuis-Oudshoorn 2011) para a imputação e a aplicamos a todos os dados observados do S-PLUS que tinham um valor acima do valor limite de magnitude (21.3). Esse algoritmo funciona em um processo de 6 passos, como descrito em Azur et al. 2011):

Missing Data	MAE	RMSE
0	0.23	0.36
99	0.23	0.34
-99	0.23	0.33
21.3	0.20	0.30
MICE	0.16	0.26

Tabela 2.1 - MAE e RMSE do conjunto de teste quando substituímos as os dados faltantes por cada valor. MICE corresponde ao algoritmo de imputação. Utilizamos a idade média ponderada em fluxo para a comparação.

1. Passo 1: Uma imputação simples é feita em cada valor faltante, como a imputação da média. Esses valores vão ser substituídos.

2. Escolhemos uma única variável (e.g. Filtro U) e retornamos todos os valores que foram imputados para faltantes.

3. Os valores observados da variável escolhida no Passo 2 são utilizados para treinar um modelo de regressão a partir das outras variáveis do conjunto de dados. Nesse caso, treinamos um algoritmo de Random Forest.

4. Os valores faltantes da variável escolhida no Passo 2 são substituídos pela estimativa da regressão. Quando essa variável for utilizada para a estimativa de outros valores faltantes, tanto os valores observados quanto os imputados serão utlizados.

5. Repetimos os Passos 2-4 pra cada variável que tem dados faltantes. Ao passar por todas as variáveis uma única vez consideramos isso como um "ciclo".

6. Os passos 2-4 são repetidos para um certo número de ciclos.

Repetimos esses processos até um número máximo de 20 ciclos. Apesar de não apresentar resultados tão diferentes, vemos na Tabela 2.1 que ao estimar a idade média ponderada em fluxo para cada valor de *missing data* possível, o MICE acaba se mostrando um pouco melhor do que os outros.

2.7 Os parâmetros das populações estelares

Existem diversos parâmetros provenientes do STARLIGHT que podemos tentar estimar. Separamos alguns principais que podem ser de grande interesse para astrônomos. É possível considerar outros parâmetros (e.g. taxa de formação estelar) dependendo do problema a ser tratado.

- `logmass`: logaritmo da massa estelar da galáxia;
- `Av`: extinção por poeira na banda V;
- `aZflux` & `aZmass`: logaritmo da metalicidade média das estrelas ponderada em fluxo e em massa, respectivamente;
- `atflux` & `atmass`: logaritmo da idade média ponderada em fluxo e em massa, respectivamente;

Para uma definição mais formal desses parâmetros no contexto do STARLIGHT, veja Cid Fernandes et al. (2005b). Além disso, também estimamos as larguras equivalentes de algumas linhas de emissão, incluindo:

- `logHalpha`: logaritmo da largura equivalente de Hα;
- `logHbeta`: logaritmo da largura equivalente de Hβ;
- `logOiii`: logaritmo da largura equivalente de [OIII]5007Å;
- `logNii`: logaritmo da largura equivalente de [NII]6583Å;

Essas linhas foram escolhidas de forma que podemos fazer o diagrama BPT (Baldwin et al., 1981) envolvendo [OIII]/Hβ e [NII]/[Hα].

Capítulo 3

Algoritmo

Vamos enunciar o problema de um modo mais formal: dada a observação fotométrica de uma galáxia, queremos obter os parâmetros das populações estelares daquela galáxia. Isso descreve um problema de regressão, onde queremos partir de um certo conjunto de informações para estimar outros valores contínuos. Essa relação não é óbvia, e uma das melhores formas de lidar com isso é utilizando algoritmos de aprendizado de máquina.

Esses algoritmos não se baseiam em relações físicas previamente estabelecidas, mas tentam encontrar relações matemáticas que descrevem os dados. Isso significa que ao fornecer para um algoritmo desses as perguntas e as respostas, ele deve conseguir aprender a como conectar uma coisa na outra, assim **criando uma relação geral que pode ser aplicada em novas perguntas que ainda não tem respostas**. Esse tipo de algoritmo é conhecido como supervisionado.

Vamos fornecer para os nossos algoritmos o conjunto de dados descrito no Capítulo 2. Alimentamos eles com as 12 magnitudes do S-PLUS mais um valor de *redshift*, e pedimos que ele obtenha os parâmetros de populações estelares que foram definidos na Seção 2.7. Utilizamos um valor de *redshift* como um dos parâmetros de entrada já que esse é obtido das observações e é um fator importante para ajudar o algoritmo a aprender. Aqui utilizamos um valor de *redshift* espectroscópico (z_{spec}), fornecido pelo SDSS, ao invés dos *redshifts* fotométricos (z_{phot}) fornecidos pelo S-PLUS. Posteriormente, discutimos o impacto de utilizar z_{phot} na análise na Seção 4.2.

3.1 A escolha do modelo

Uma forma mais objetiva de olhar para esse problema é comparando diversos algoritmos disponíveis. Em Flores et al. (2021), os autores testam diferentes métodos de *machine learning* clássico e de *deep learning* para estimativa de alguns parâmetros de galáxias. Apesar de utilizarem dados diferentes dos nossos (trabalham com espectros de baixa resolução e não fotometria), decidimos nos basear nos modelos que testaram. Escolhemos 5 algoritmos que pareceram promissores:

- Regressão Linear
- K-Nearest Neighbours Regressor
- Random Forest Regressor
- XGBoost Regressor
- Redes Neurais

Todos os modelos apresentam resultados bons para problemas de regressão. Cada um deles tem alguma vantagem, seja acurácia, tempo de execução ou até facilidade para entender. Pretendemos aqui escolher um melhor modelo levando todos esses fatores em consideração.

Tanto o K-Nearest Neighbours Regressor (KNNr, Cunningham e Delany 2021) quanto o Random Forest Regressor (RF, Louppe 2015) são variações de algoritmos bem conhecidos muito utilizados em problemas de classificação. No caso do KNNr, a regressão é feita via uma interpolação local dos alvos baseado no conjunto de treinamento. No caso do RF, é usado uma combinação de árvores de decisões para estimar uma nova predição. Ambos os algoritmos são relativamente simples, porém mostram uma robustez alta mesmo quando comparado com algoritmos de maior complexidade, como redes neurais.

O XGBoost Regressor (XGB, Chen e Guestrin 2016) tem ganhado muita tração nos últimos anos. É um método baseado em árvores de decisão conhecido por *gradient boosting*, e difere de *Random Forest* pela forma que as árvores são combinadas. Esse método é um dos melhores no ramo de modelos de *machine learning* clássico, ou seja, sem levar em conta algoritmos de *deep learning*. É um algoritmo mais complexo que os anteriores, porém foi

muito adotado nos últimos anos e foi ramificado em outros modelos, como LightGBM (Zhang et al., 2017) e CatBoost (Dorogush et al., 2018).

Métodos de *deep learning* tem dominado o campo de pesquisa por conta de sua robustez, versatilidade e rapidez. Em um mundo onde lidamos com centenas de milhares de dados, redes neurais mostraram a sua habilidade de aprender relações complexas sem conhecimento de priores (na verdade o prior é o próprio conjunto de treinamento). É um método muito popular em problemas de regressão na Astronomia, como a estimativa de *redshift* fotométrico (Lima et al., 2022).

O modelo final será treinado para todos os parâmetros de populações estelares descritos na Seção 2.7, porém escolhemos a massa estelar para servir como base comparativa entre os modelos. Também decidimos utilizar a raiz quadrada do erro médio quadrático (*root mean squared error*, RMSE) como métrica para analisar os modelos testados, junto com o coeficiente de determinação (R^2). O R^2 é uma medida que tenta simular uma "acurácia"para regressão, indo de 0 a 1 e quanto mais próximo de 1, melhor. Toda a análise foi feita em Python. Antes de treinar a rede nós normalizamos todos os dados entre 0 e 1 usando o `MinMaxScaler` da biblioteca `sklearn` (Pedregosa et al., 2011). É importante que os *scalers* sejam salvos e que o conjunto de teste não participe da normalização inicial (evitando assim um vazamento de dados) para se determinar as métricas com ele.

Começamos os estudos em torno de qual modelo utilizar a partir de uma busca em grade simples (*grid search*) variando alguns hiperparâmetros de cada modelo, tal como por quanto tempo deve treinar. Utilizamos 80% (110,187 galáxias) da amostra para treino e 20% (27,547 galáxias) para teste. Os valores das redes neurais são discutidos mais a frente. No caso do XGBoost, o modelo demorava muito tempo para treinar, o que nos impossibilitou de realizar um grid search muito extenso. Os melhores resultados de cada busca estão na Tabela 3.1. Vemos que todos os modelos tiveram resultados bons, até mesmo os mais simples, mas a rede neural acabou sendo o melhor. Resolvemos prosseguir com a rede neural por conta tanto das melhores métricas quanto pela velocidade de treino (ordem de 5 minutos).

Tabela 3.1 - Resultados dos modelos para o conjunto de teste.

Modelo	RMSE	R2	Tempo de Treino
Regressão Linear	0.0196635	0.952068	1 seg.
KNN Regressor	0.0134364	0.977949	20 seg.
RF Regressor	0.0124130	0.981529	5 min.
XGBoost	0.0124973	0.981129	30 min.
Rede Neural	0.0112439	0.984669	5 min.

3.2 A rede neural

A arquitetura da nossa rede neural é relativamente simples, e consiste de uma camada de entrada com 13 neurônios (12 magnitudes do S-PLUS + 1 valor de *redshift*), uma camada de saída com 1 neurônio (um único parâmetro a ser predito) e 4 camadas de ativação ReLU escondidas entre elas com 256 neurônios cada uma (veja Figura 3.1). Utilizar camadas de *dropout* são uma opção boa, já que ajudam a rede a não ter *overfitting* (quando a rede começa a aprender relações específicas ao conjunto de treino). Como não vemos um caso de *overfitting* tão rápido durante o treinamento (algo que é discutido a seguir) decidimos seguir sem incluir *dropout*.

Todo o código da rede foi escrito utilizando o pacote Keras (Chollet et al. (2015)). Este *framework* é intuitivo e fácil de utilizar, e apesar de não ter tantas customizações internas ainda demonstra ótimos resultados.

Aplicamos também um método de validação chamado *K-fold Cross-Validation* (Raschka, 2020). Esse método consiste em separar a base de dados em K conjuntos de mesmo tamanho e treinar a rede K vezes, cada vez utilizando um conjunto diferente de validação. A vantagem é que todos os dados são usados tanto para treinamento quanto para teste. Para isso, divimos os dados em 5 conjuntos diferentes, sempre separando 80% para treino e 20% para validação. Com esse método podemos ter uma noção melhor sobre o comportamento de erro da rede. Treinamos o modelo por 1000 épocas, e os resultados são mostrados na Figura 3.2. Podemos notar que o erro (*loss*) da validação não aumenta, ou seja,o modelo não está aprendendo relações específicas do conjunto de treinamento (*overfitting*). Procuramos definir o número de épocas para o modelo a partir do último momento antes da *loss* começar a aumentar, porém existe um argumento a ser feito sobre o custo-benefício não valer a pena. Apesar de uma *loss* menor ser obtida, o tempo computacional pode ser

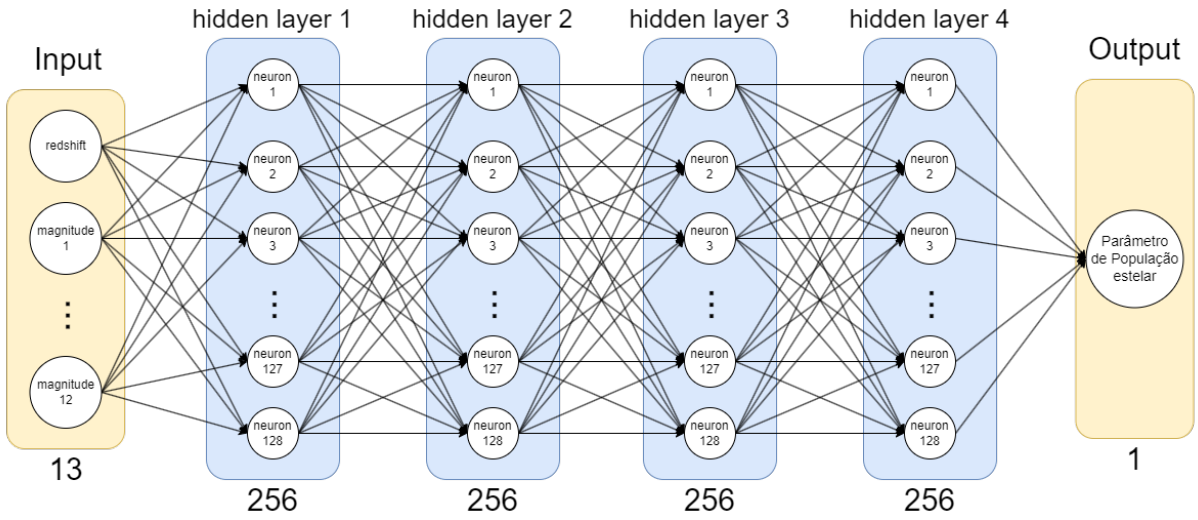

Figura 3.1: Representação visual da melhor rede neural obtida. Ela é composta por uma camada de entrada com 13 neurônios (12 filtros do S-PLUS + 1 valor de redshift) e uma de saída composta por 1 único neurônio (um parâmetro de população estelar). Entre elas existem 4 camadas de ativação ReLU escondidas com 256 neurônios cada.

muito grande. Acabamos decidindo treinar o modelo para 500 épocas.

Uma alternativa é que a estimativa dos parâmetros seja feita de forma conjunta, ou seja, que ao invés de cada rede estimar um único parâmetro, ela consiga estimar todos de uma única vez. Se houver uma certa correlação entre os parâmetros das populações estelares, é possível que isso ajude na estimativa. Durante os nossos testes, vimos que o desempenho piorou, porém talvez uma rede com complexidade maior (mais camadas ou neurônios) seja capaz de estimar melhor todos os parâmetros de uma única vez.

3.3 Estimativa de erros

Redes neurais simples têm um problema intrínsico relacionado a incertezas. Em Gawlikowski et al. (2021) são mostradas diversas tentativas de incorporar estimativas de erros em redes neurais profundas, e algumas delas são bem conhecidas em nossa área. Redes Bayesianas, por exemplo, estão sendo utilizadas na estimativa de redshifts fotométricos (e.g. Lima et al. 2022) ao mesmo tempo que *ensembles* de redes (e.g. de Oliveira et al. 2023). Em nosso caso, vamos utilizar Test-Time Data Augmentation, onde podemos inferir informações sobre a incerteza de uma predição após um processo de *Data Augmentation* (veja mais em Wang et al. 2019). A Figura 3.3 mostra como esse método é aplicado: um valor original é transformado diversas vezes e calculamos sobre cada um sua predição.

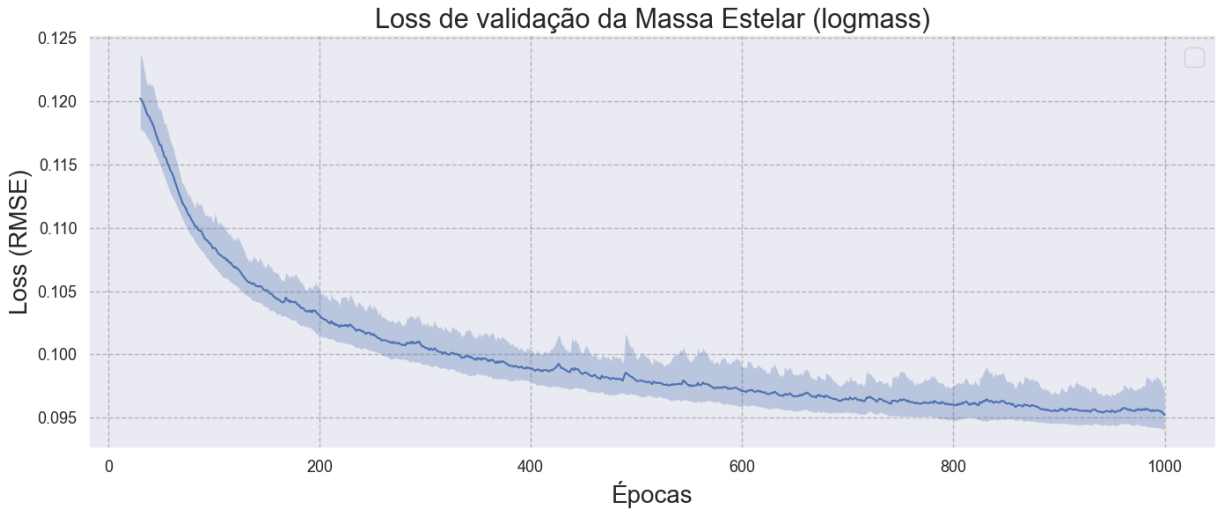

Figura 3.2: Gráfico da raiz do erro médio quadrático (RMSE) do conjunto de validação para cada uma das 1000 épocas. A área sombreada representa os valores máximos e mínimos de um K-Fold *Cross-Validation* com $K = 5$, enquanto a linha representa a média entre os *folds*.

Após combinar todas as predições que foram obtidas a partir de um mesmo objeto original, podemos inferir uma distribuição de probabilidades sobre aquela predição.

Para cada observação do S-PLUS, nós temos acesso ao erro associado a cada uma das magnitudes daquele objeto. Esse erro vem de diversas fontes observacionais, mas o importante é que nós queremos que o erro resultante de cada predição (da massa estelar, por exemplo) leve em conta os erros observacionais. Para isso, treinamos nosso modelo em uma amostra de dados sem erros, porém quando realizamos uma nova predição para uma galáxia assumimos um erro observacional gaussiano e supomos que os erros das diversas bandas são independentes, e fazemos 10,000 amostragens para cada galáxia. Cada uma dessas 10,000 predições é feita com valores de magnitude um pouco diferentes dos anteriores, e ao final teremos uma distribuições de predições como mostrado na Figura 3.4. Utilizamos o erro associado a cada medida de magnitude fornecido pelo S-PLUS. Para medir a incerteza de cada predição, vamos usar o "sigma robusto" (Ivezic et al., 2019) definido na Equação 3.1, onde Q_{75} e Q_{25} são os quartis 75% e 25%. Vale notar que se a distribuição das predições é gaussiana, esta estatística é equivalente ao desvio padrão da gaussiana.

$$\sigma_G = 0.7413 \times (Q_{75} - Q_{25}) \tag{3.1}$$

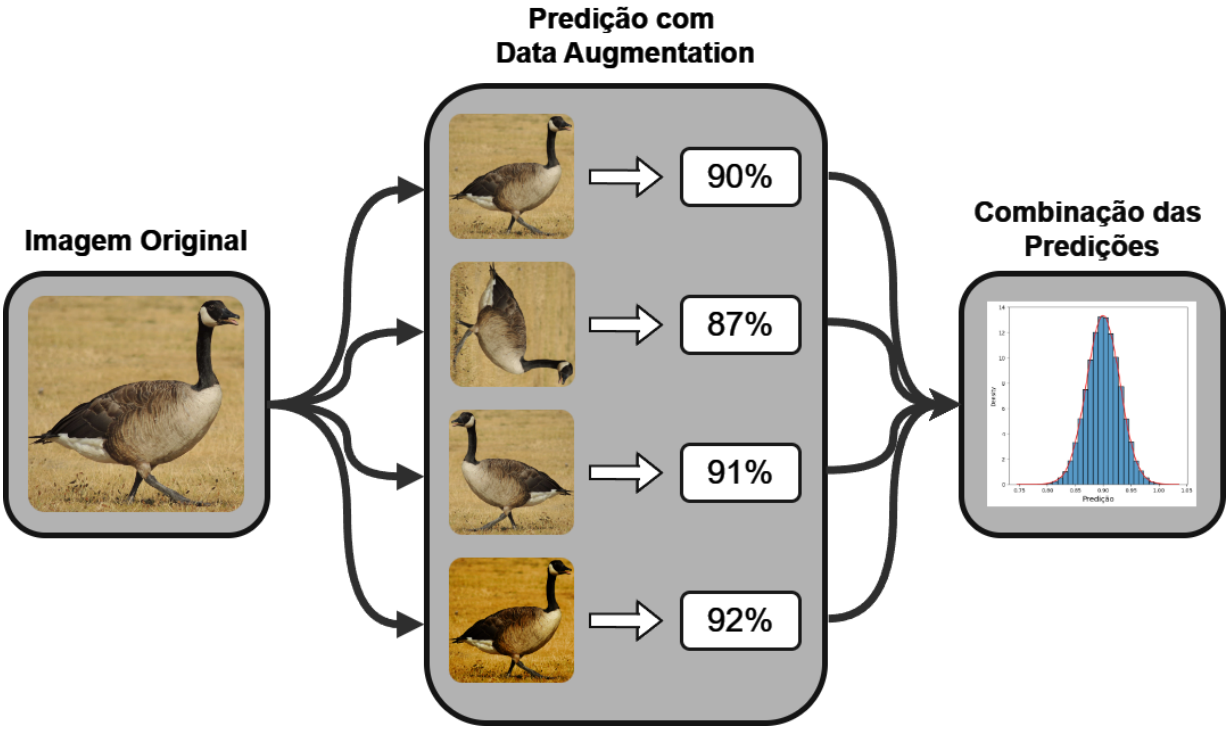

Figura 3.3: Representação visual do método *Test-Time Data Augmentation* para saber a chance de existir um ganso na imagem. Um dado inicial passa por uma transformação diversas vezes, e um algoritmo é utilizado para realizar predições pontuais. Por fim, combinamos as predições pontuais para obter uma distribuição de probabilidades.

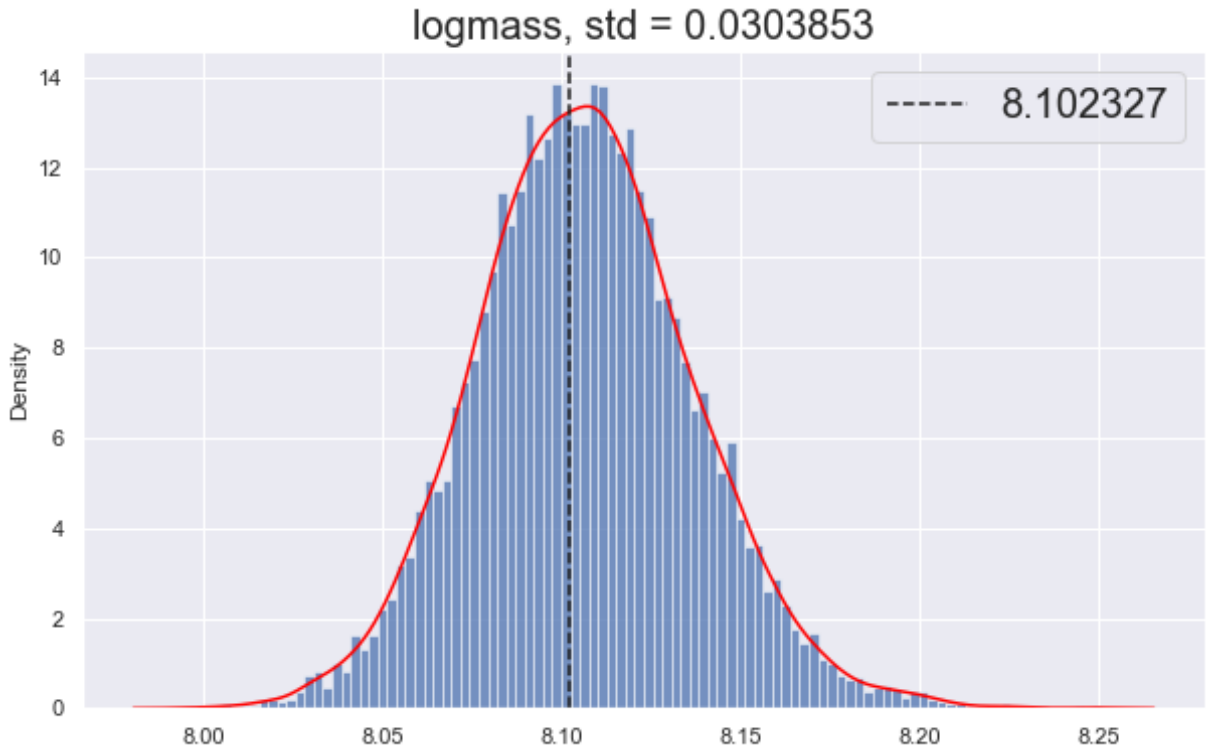

Figura 3.4: Histograma de 10,000 estimativas de massa estelar para a galáxia S-PLUS 0073-0011797 utilizando o método TTA. O eixo-X corresponde ao valor da massa predito. O desvio padrão da distribuição está descrito no título enquanto a linha tracejada corresponde à estimativa sem erros observacionais.

Capítulo 4

Análises

Nosso objetivo é criar uma ferramenta que seja capaz de estimar parâmetros de populações estelares a partir da fotometria de uma galáxia. Isso já foi realizado na Seção 3, então os que nos resta é testar esse modelo. Existem inúmeras aplicações que podem ser feitas, desde análises básicas até as mais completas. Vamos mostrar aqui algumas possíveis aplicações junto com testes que validam a sua robustez.

4.1 Estimativa dos parâmetros

A primeira análise que pode ser feita é examinar a estimativa de cada parâmetro para o conjunto de teste. Podemos aqui ter uma ideia melhor não só sobre a performance do modelo para os dados sintéticos, mas também sobre o que podemos esperar para o erro nos parâmetros. Na Figura 4.1 estão os resultados da predição para um conjunto de teste de 26,000 galáxias. A massa estelar é um parâmetro estimado muito bem com um erro bem baixo (0.06 dex), porém é esperado que seja bem estimada para *redshifts* baixos (Walcher et al., 2010) e não vemos nenhum viés na estimativa. A extinção de poeira mostra uma dispersão maior, porém mesmo com um viés para valores maiores ainda temos um erro baixo. Metalicidades são difíceis de se obter mesmo espectroscopicamente (Thainá-Batista et al., 2023), porém um erro da ordem de 0.2 dex é muito bom utilizando fotometria. Comparando as medidas de metalicidade e idade entre elas mesmas, notamos que as medidas ponderadas em fluxo são ligeiramente mais robustas. É importante notar que apesar de algumas predições mostrarem uma dispersão grande, os erros continuam relativamente baixos. A idade média ponderada em massa (`atmass`), apesar de ser uma das predições com maior dispersão, mostra um erro médio absoluto de 0.16 enquanto sua

ordem de valores é de 6 a 10 dex. Notamos também uma separação entre as populações de galáxias azuis e vermelhas nas predições de idade e metalicidade ponderada em fluxo (a bimodalidade), representadas pelas duas concentrações de galáxias.

Estimamos também as larguras equivalentes das linhas de emissão necessárias para construir o diagrama BPT, com seus resultados na Figura 4.2. As estimativas de $H\alpha$ e $H\beta$ são robustas, com um erro da ordem de 0.14 e 0.1 dex, respectivamente. A estimativa de NII também mostra um erro consideravelmente baixo. Enquanto isso, vemos que a estimativa de [OIII] é um pouco pior do que as outras, e é possível notar um viés onde a rede está subestimando a largura equivalente. Nas estimativas de $H\alpha$ e $H\beta$ vemos uma concentração de pontos com valores de larguras equivalentes muito baixas.

É possível notar, tanto pela Figura 4.1 quanto pela Figura 4.2 que existe um viés em relação aos resíduos da rede. O nosso modelo está superestimando as predições para valores baixos, e depois está subestimando para valores altos. A causa desse problema ainda não é clara, mas parece estar mais relacionada a um desvio sistemático da rede do que algo relacionado aos dados.

Uma comparação próxima que podemos realizar é com Thainá-Batista et al. (2023). Nesse artigo, os autores estimam os parâmetros das populações estelares a partir da mesma base de dados de Werle et al. (2018) utilizando uma versão modificada do STARLIGHT conhecida por AlStar (Algebraic STARLIGHT). Vemos na Tabela 4.1 uma comparação entre o MAE e σ_{NMAD} do conjunto de teste obtidos tanto por nós quanto pelos autores. Essa métrica é calculada pela equação 4.1:

$$\sigma_{NMAD} = 1.4826 \times \text{median}(|x - \text{median}(x)|) \tag{4.1}$$

Decidimos calcular essa métrica para os nossos resultados para ter uma comparação mais fiel já que os autores a utilizam. Também escolhemos a simulação onde a razão sinal por ruído é a maior possível, já que não adicionamos ruído aos nossos espectros sintéticos.

Notamos que o erro da massa estelar é bem próximo em ambos os trabalhos, porém os valores divergem um pouco quando consideramos os outros parâmetros. As idades médias e metalicidades médias tem um erro menor em Thainá-Batista et al. (2023). A extinção por poeira obtém um valor muito próximo de 0 no artigo, o que é algo que não conseguimos replicar. Porém, é importante notar que em Thainá-Batista et al. (2023) os

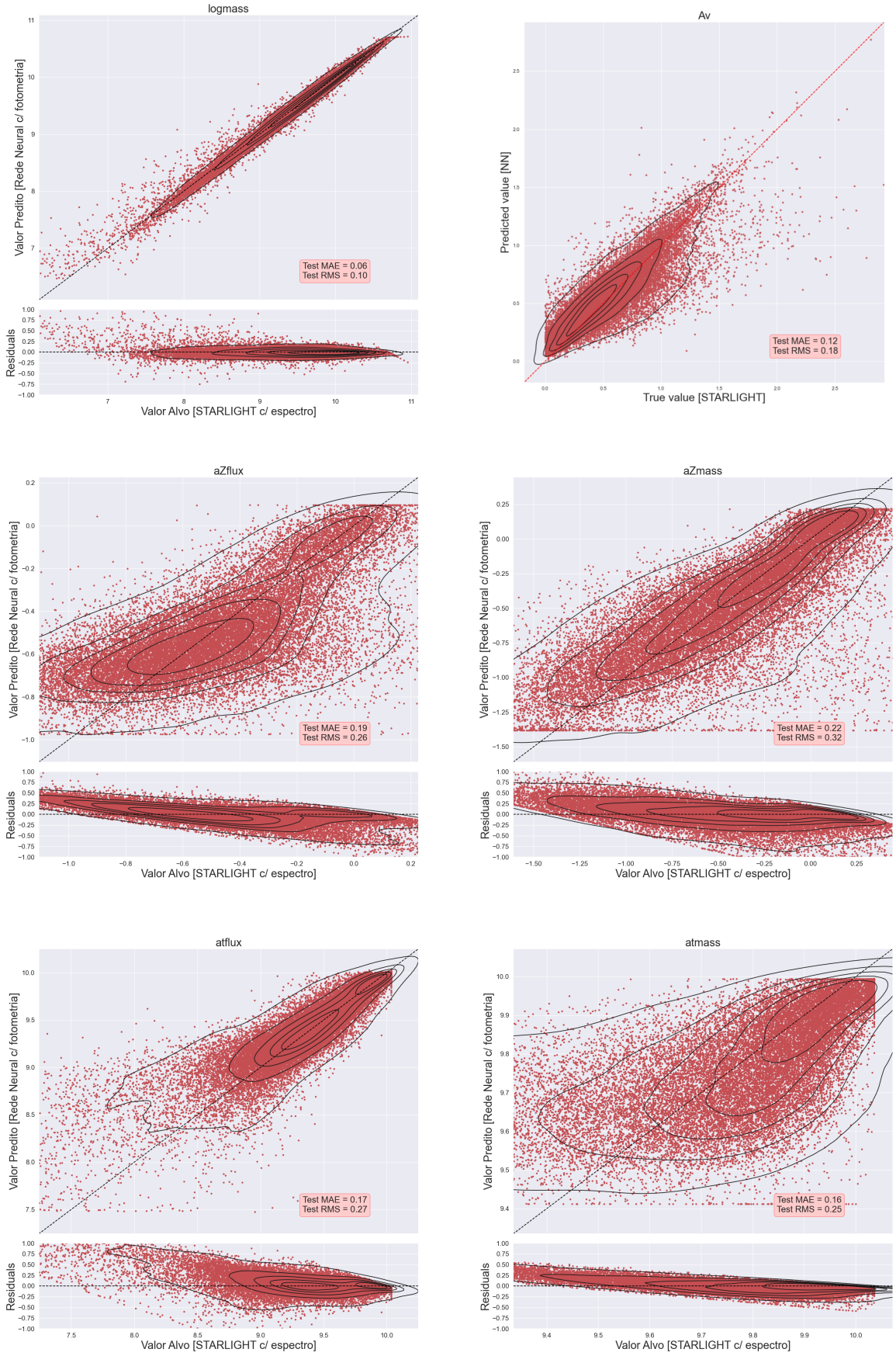

Figura 4.1: Estimativa dos parâmetros de populações estelares. Em cada gráfico, o eixo-X representa o valor obtido espectroscopicamente pelo STARLIGHT, enquanto o eixo-Y representa a estimativa da rede neural. A linha vermelha mostra a identidade $x = y$.

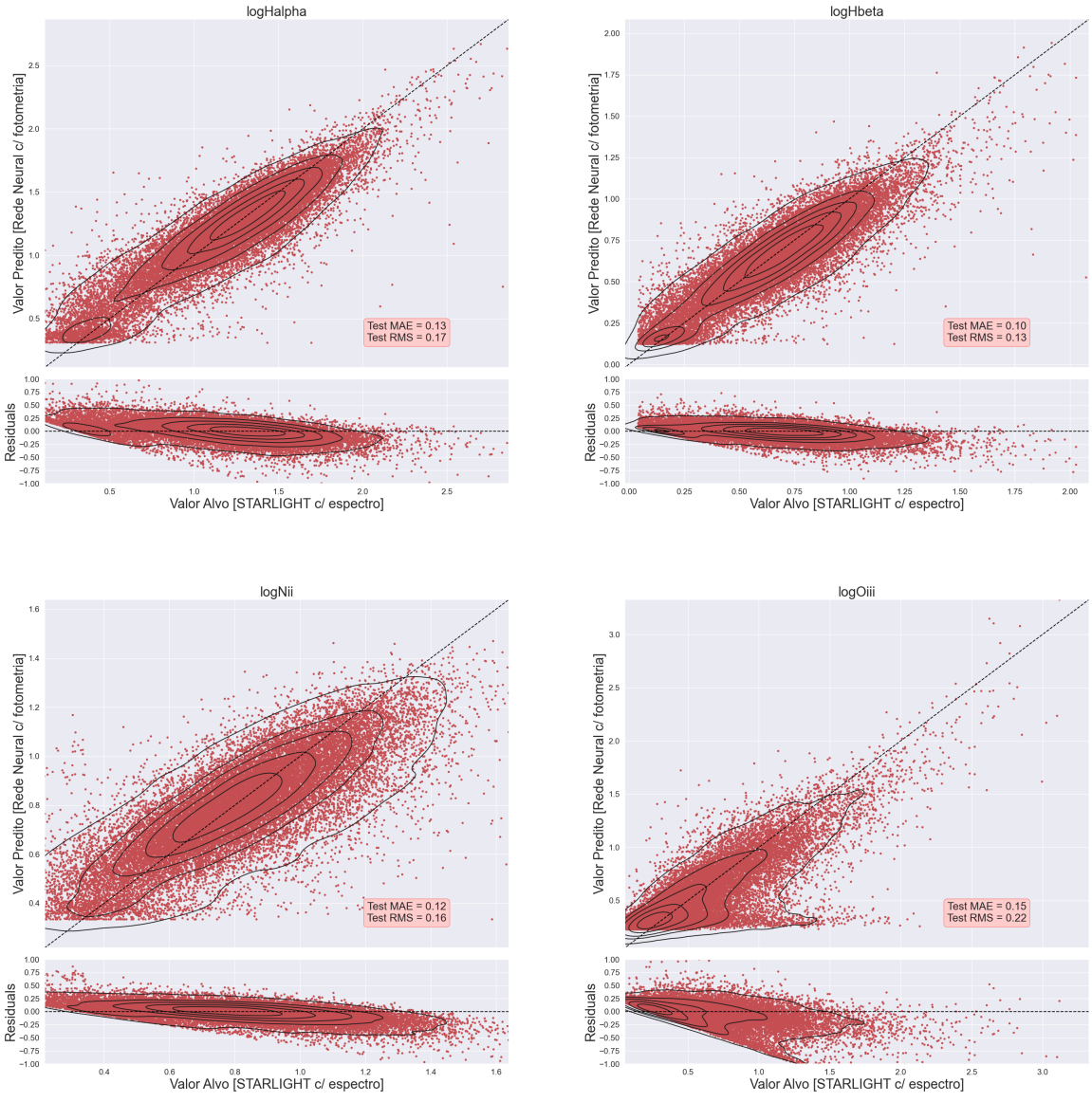

Figura 4.2: Estimativa das larguras equivalentes das linhas de emissão. Em cada gráfico, o eixo-X representa o valor obtido espectroscopicamente pelo STARLIGHT, enquanto o eixo-Y representa a estimativa da rede neural. A linha vermelha mostra a identidade $x = y$.

Propriedade	Esse trabalho	Thainá-Batista et al. (2023)
`logmass` [$M_\odot$]	0.06 ± 0.04	0.04 ± 0.05
`atflux` [anos]	0.16 ± 0.13	-0.01 ± 0.15
`atmass` [anos]	0.16 ± 0.09	0.07 ± 0.09
`aZflux` [$Z_\odot$]	0.18 ± 0.18	0.09 ± 0.17
`aZmass` [$Z_\odot$]	0.21 ± 0.17	0.11 ± 0.17
`Av` [mag]	0.12 ± 0.21	-0.02 ± 0.11
`logHalpha`	0.13	-
`logHbeta`	0.10	-
`logOiii`	0.15	-
`logNii`	0.12	-

Tabela 4.1 - Parâmetro estimado e erros do conjunto de teste para este trabalho e o de Thainá-Batista et al. (2023) para galáxias com S/N = 100. Os valores de erro de cada medida correspondem ao σ_{NMAD} do *bias* das predições.

autores utilizam galáxias com *redshift* até 0.01, enquanto em nosso trabalho utilizamos galáxias até 0.1 em *redshift*. Isso significa que algumas estimativas, principalmente aquelas que dependem das informações das linhas de emissão, podem ser prejudicadas para *redshifts* mais altos. De qualquer forma, essa comparação nos mostra uma boa robustez do nosso modelo mesmo quando comparado a um método mais complexo, com valores de erros bem próximos mesmo para um intervalo de *redshift* bem maior.

4.2 Dependência de redshift

Uma das grandes vantagens em utilizar modelos de aprendizado de máquina para estimativa dos parâmetros de populações estelares ao invés de métodos tradicionais de SED *fitting* é que podemos aplicar o nosso modelo em um intervalo de *redshift* muito maior do que o esperado. Métodos tradicionais, como o STARLIGHT, precisam definir muito bem como cada elemento das populações estelares impacta o espectro final. Enquanto isso, os métodos de *Machine Learning* conseguem aprender diretamente dos dados, e criam-se relações entre as observações que não são tão óbvias. Isso significa que, para objetos mais distantes, é esperado que a estimativa dos parâmetros seja pior, principalmente quando estes são relacionados com linhas de emissão.

A Figura 4.3 nos mostra, na imagem da esquerda, a distribuição de *redshift* no gráfico de predição da massa estelar. Podemos notar que não existe um viés inesperado além da

relação esperada de que quanto maior o *redshift*, maior a massa. A figura da direita nos mostra o erro dessa predição em função do *redshift*, e novamente não vemos nenhum viés, indicando que a precisão de nossas predições é independente do *redshift*.

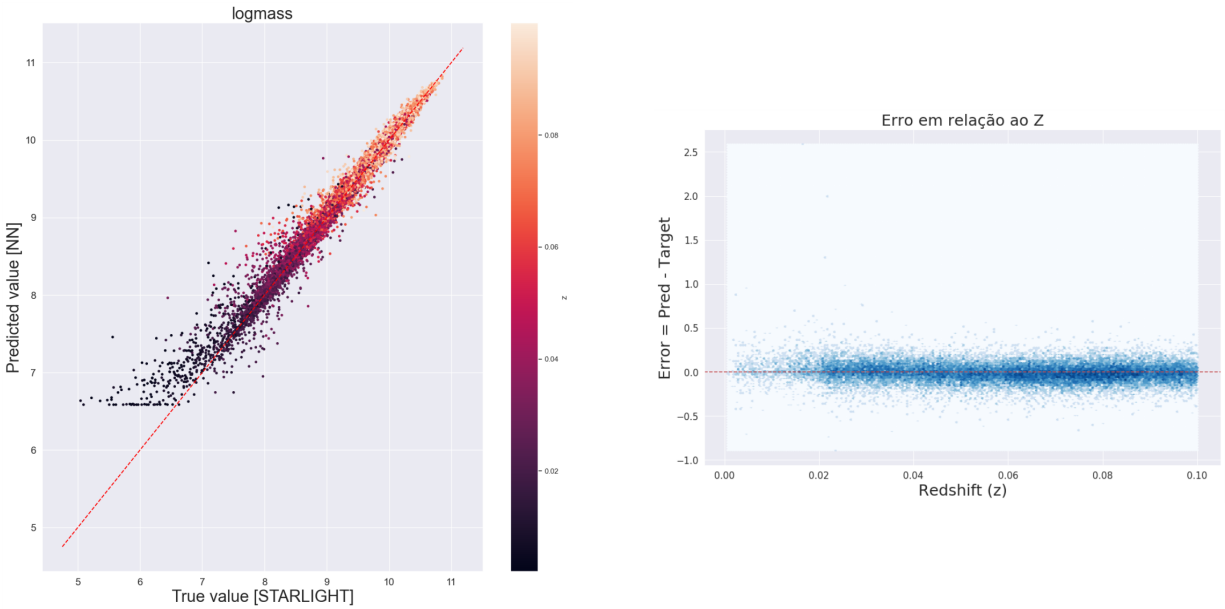

Figura 4.3: Na figura da esquerda, vemos a predição da massa estelar para o conjunto de teste, com os valores pintados dependendo do seu *redshift*. Na figura da direita vemos o erro (*bias*) de cada uma dessas galáxias em relação ao seu *redshift*.

É importante considerar a dependência de um *redshift* bem estimado nas aplicações que pretendemos realizar. *Redshifts* fotométricos (photo-z), apesar do avanço contínuo na área, são consideravelmente menos confiáveis do que estimativas espectroscópicas (spec-z, Lima et al. 2022). Para aglomerados próximos o erro do *redshift* fotométrico pode ultrapassar o seu próprio valor, e isso pode causar um problema na predição dos parâmetros. Na Figura 4.4 mostramos a predição da massa estelar para o mesmo conjunto de galáxias da Seção 2.5. O histograma em azul corresponde a estimativa da massa a partir do *redshift* fotométrico de (Lima et al., 2022), enquanto o histograma em laranja é calculado a partir do *redshift* espectroscópico de Werle et al. (2018). Podemos notar claramente que a dispersão aumenta ao utilizar o photo-z, porém as distribuições continuam centradas nos mesmos valores. Isso significa que predições feitas com *redshift* fotométrico são sim menos confiáveis, porém não por muito.

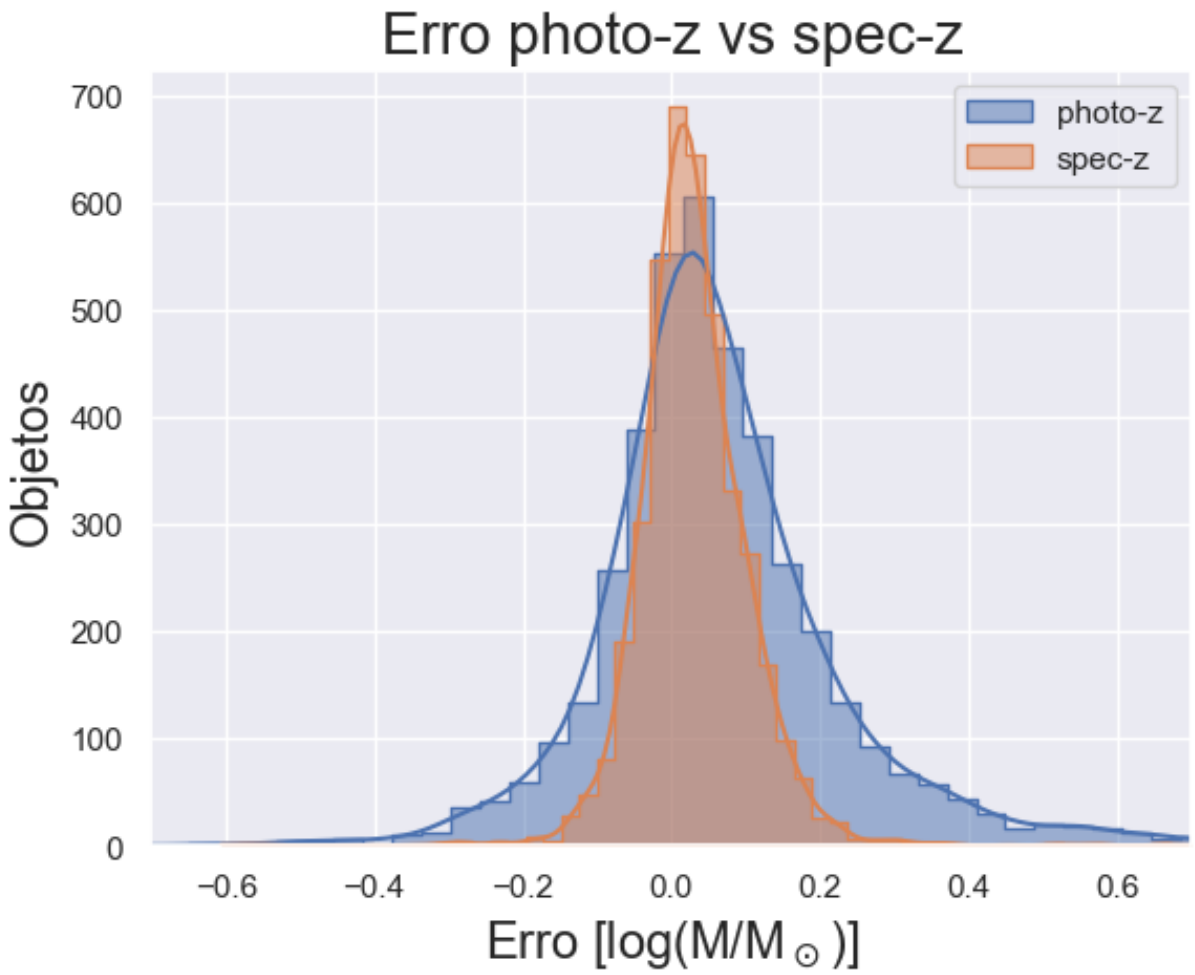

Figura 4.4: Histogramas de erros para predição da massa estelar utilizando o *redshift* fotométrico (azul) e o *redshift* espectroscópico (laranja) como parâmetro de entrada.

4.3 Aplicação em aglomerados próximos: Fornax

O aglomerado de Fornax (Drinkwater et al., 2001) é de grande interesse da comunidade astronômica por ser um dos aglomerados mais próximos da Via Lactea. Estamos participando ativamente de um grupo de pesquisa com interesse nesse aglomerado, e já estamos com dois artigos: um pré-publicado (Castelli et al., 2021) e outro submetido.

Por conta desse grande interesse, decidimos fazer uma aplicação preliminar da rede neural aos objetos de Fornax cujas magnitudes foram medidas pelo S-PLUS. Analisamos aqui uma pequena subamostra de 182 galáxias desse aglomerado fornecida por Haack et al. (2023), de forma que nenhuma delas continha valores faltantes nas bandas *g, r, i, z*. Predizemos, através da rede neural, a idade média ponderada em fluxo (`atflux`) de cada um desses objetos, com os resultados apresentados na Figura 4.5.

Essa aplicação é o que denominamos de "teste de sanidade", onde checamos se a rede apresenta um resultado físico que estávamos esperando. Nesse caso, vemos que quanto mais velha a galáxia, mais vermelha ela é, o que era exatamente o que esperávamos. Porém isso é somente para mostrar o potencial do algoritmo, e que estudos futuros em Fornax (ou em outros aglomerados próximos como Hydra) podem se beneficiar da estimativa dos parâmetros das populações estelares.

4.4 Diagrama BPT

Linhas de emissão são propriedades interessantes para serem estudadas. Apesar do S-PLUS prover filtros estreitos, esses filtros param de ser tão efetivos para objetos mais distantes. Isso significa que a linha de emissão de Hα fica fora do filtro F660 para valores de *redshift* acima de 0.025. Como estimar essas larguras equivalentes de linhas de emissão sem a informação das bandas estreitas é um desafio para os métodos tradicionais, porém com *machine learning* nós conseguimos utilizar a informação dos outros filtros para estimar as larguras equivalentes das linhas espectrais. Na Figura 4.6 podemos ver a estimativa da largura equivalente de Hα para o mesmo conjunto de teste. Assim como na Figura 4.3, não vemos nenhum viés no erro estimado mesmo para $z > 0.03$, o que é inesperado, talvez porque o intervalo de *redshift* seja pequeno.

Utilizando as linhas [NII]6583Å, Hα, [OIII]5007Åe Hβ podemos criar o Diagrama BPT (Baldwin et al., 1981). Esse diagrama é importante para diferenciar linhas de emissão

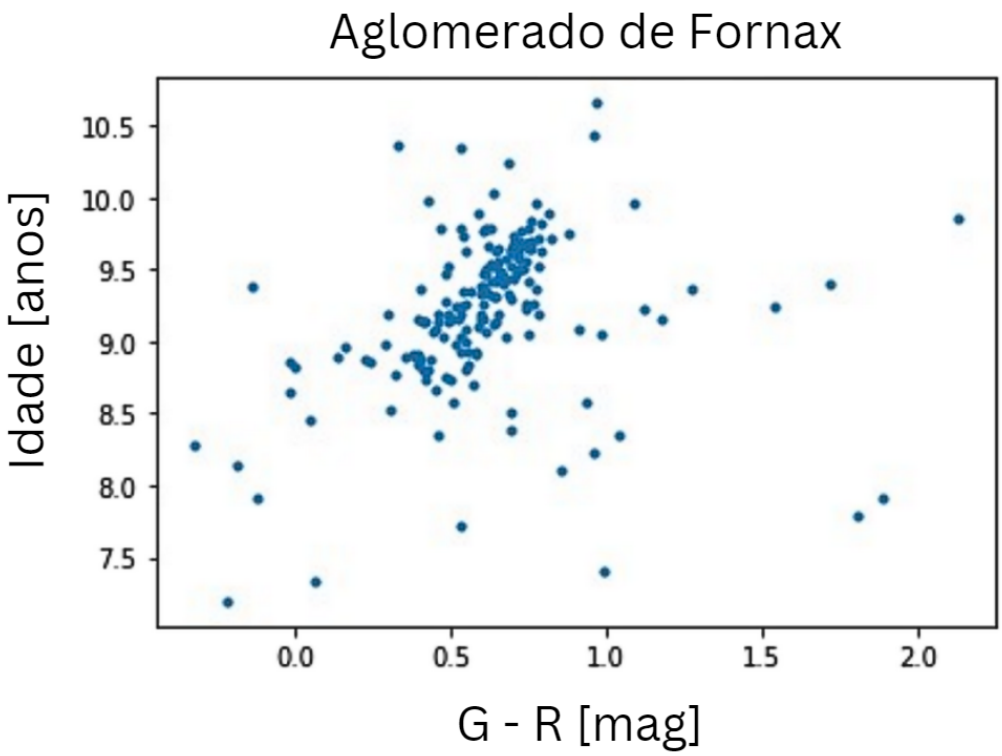

Figura 4.5: Relação idade *versus* cor G-R para um conjunto de 182 galáxias do aglomerado de Fornax.

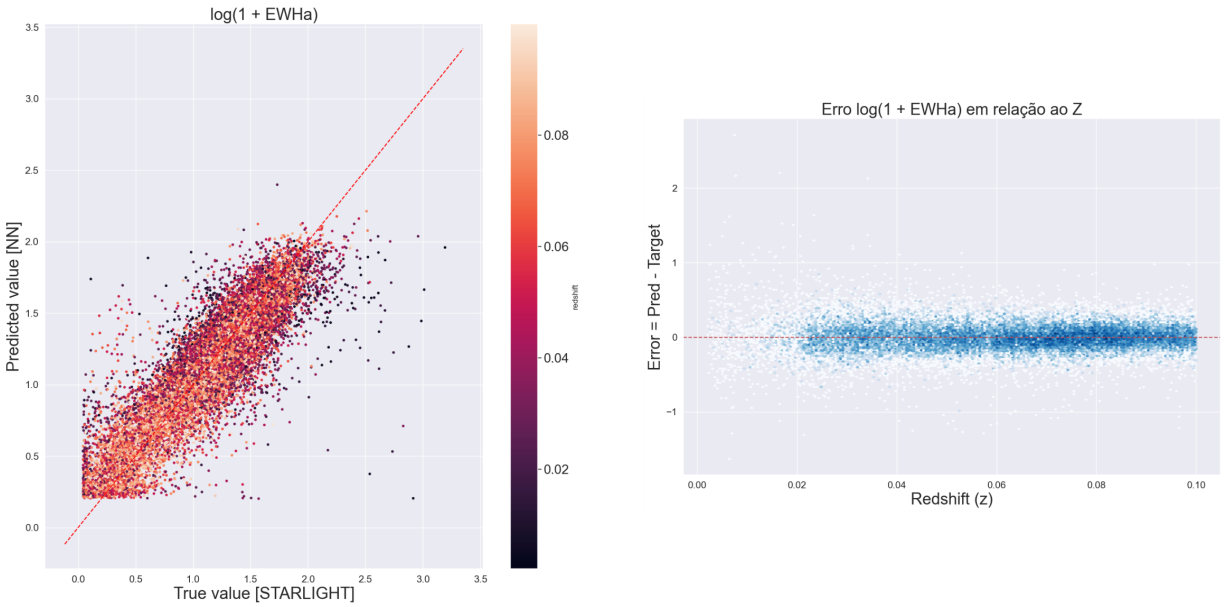

Figura 4.6: Igual à Figura 4.3 porém estimando a largura equivalente de Hα.

produzidas por formação estelar, LINERs e AGNs. Aqui podemos ter uma ideia melhor se a rede neural consegue estimar bem as linhas de emissão para as galáxias do S-PLUS.

A Figura 4.7 mostra o resultado de predição da rede (direita) comparado com o diagrama BPT calculado com dados espectroscópicos (esquerda). Os pontos são coloridos a

partir da posição no eixo-X no BPT espectroscópico (esquerda), de forma que podemos analisar se os pontos que estão em cada uma das asas continuam na mesma posição pela estimativa da rede neural. É possível ver que a rede consegue manter a mesma distribuição de "braços"além de manter todas as galáxias na mesma "posição relativa", ou seja, pontos nos extremos das asas continuam na mesma posição. Isso é extremamente interessante, pois mostra que o nosso modelo é capaz de obter razoavelmente bem a informação das linhas de emissão mesmo para *redshifts* maiores. A partir da nossa estimativa, podemos ter uma noção se uma galáxia é AGN/LINER ou *star forming*, somente a partir da fotometria mesmo em *redshifts* altos.

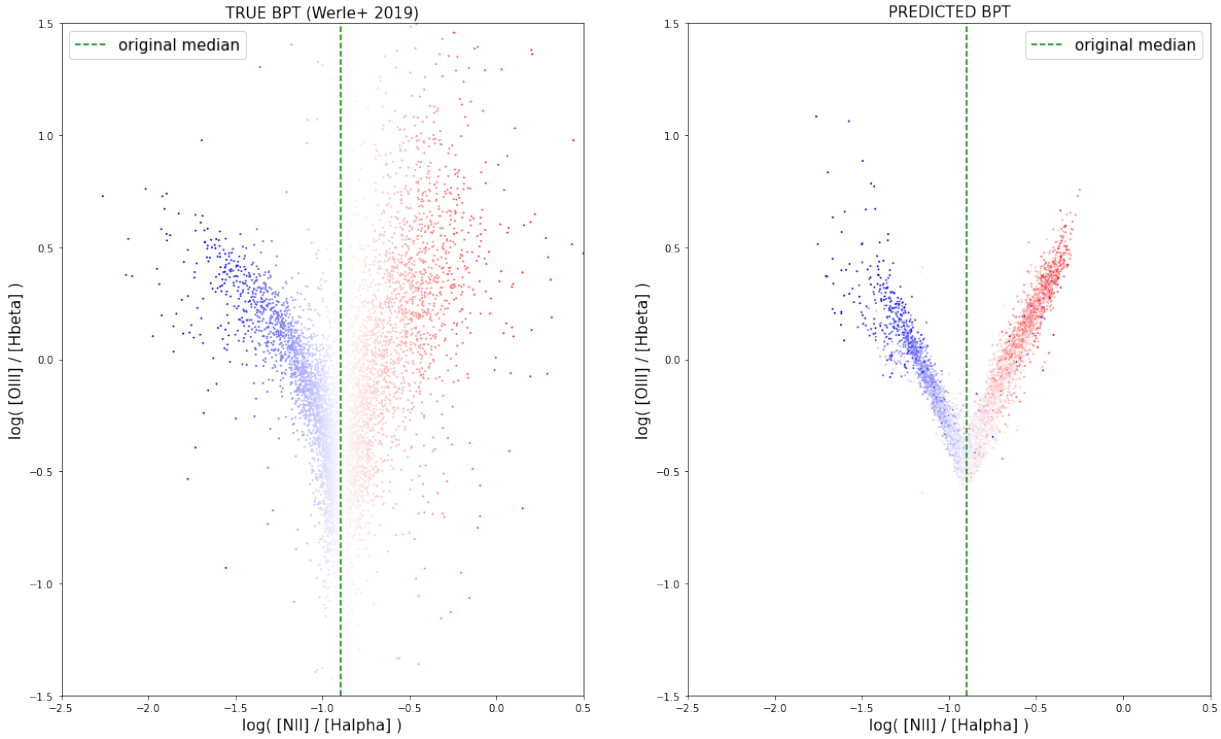

Figura 4.7: Diagramas BPT utilizando dados "reais"provenientes do STARLIGHT (esquerda) e com dados preditos pela rede neural através das magnitudes do S-PLUS (direita). Os pontos são coloridos a partir da posição do eixo-X no diagrama da esquerda.

4.5 *Estudo de importância das features*

Entender como uma rede neural realiza uma predição é algo muito interessante, porém é uma tarefa mais complexa do que parece. Algoritmos de redes neurais são conhecidos por uma baixa interpretabilidade, já que dependem de inúmeras relações não-lineares para chegar em um resultado. Apesar de complexo, essa informação deve estar imbutida dentro

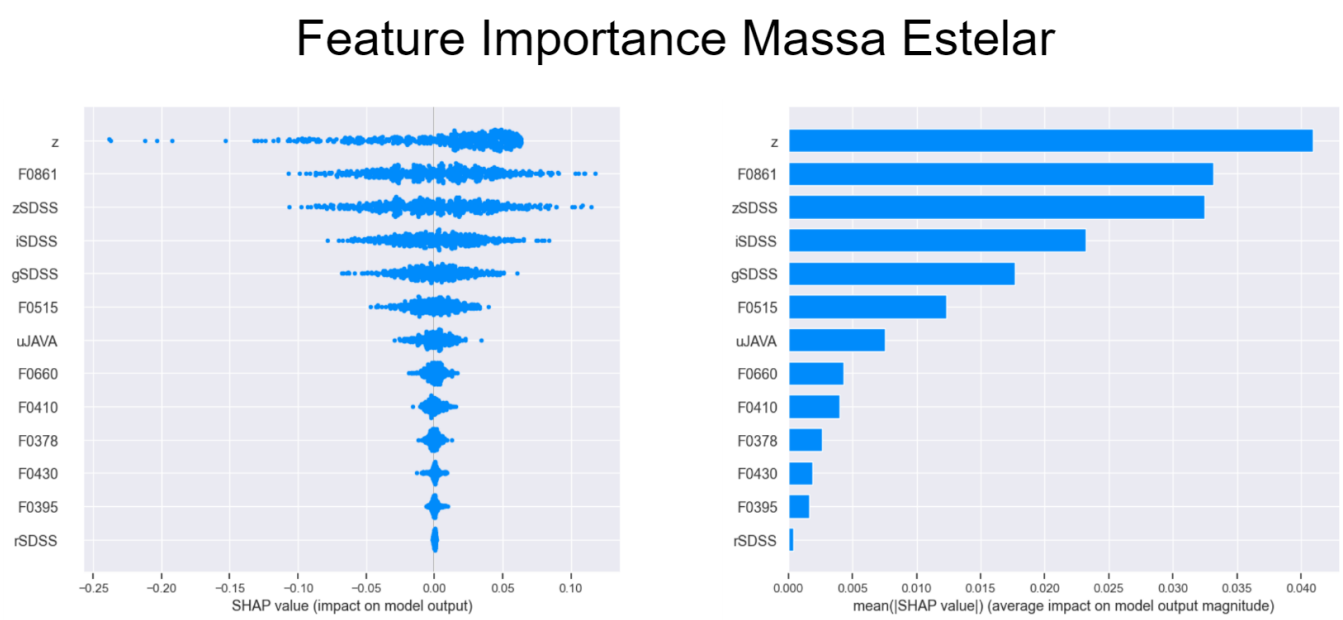

Figura 4.8: Valores SHAP obtidos a partir da rede treinada para obter a massa estelar. Quanto maior o valor SHAP de uma *feature*, maior o impacto dela no resultado final da predição. Na esquerda vemos os valores SHAP, enquanto os histogramas da direita nos mostram a média do módulos dos valores SHAP.

dos pesos da rede, e existem diversos estudos que focam nessa interpretabilidade. Essa análise é raramente feita em trabalhos de astronomia e normalmente cabe somente em artigos específicos de ciência da computação.

Um método possível é que, dada uma rede treinada, nós podemos calcular a importância de cada uma das *features* através dos valores SHAP (*SHapley Additive exPlanations*, Lundberg e Lee 2017). Essa medida, baseada em teoria dos jogos, tenta calcular o impacto de cada *feature* a partir de uma combinação de todos os seus possíveis valores. A biblioteca SHAP em Python nos permite interpretar como os *inputs* da rede afetam o resultado de cada predição. A partir de uma configuração inicial, o código testa diversos valores diferentes para cada uma das *features* da rede e analisa como uma variação na entrada muda o *output*. A grande vantagem de utilizar essa biblioteca é que as análises são feitas em cima de redes já treinadas, enquanto outros códigos fazem essa análise durante o treinamento da rede, o que pode demorar muito.

Podemos ver um dos resultados dessa análise na Figura 4.8 para a interpretabilidade da massa estelar. Nesses gráficos podemos ver as variáveis que tem maior impacto no resultado, e ele nos mostra que o *redshift*, junto com os filtros no vermelho, ditam a massa estelar de um objeto. Esse resultado é esperado mas bem interessante. Análises dessa natureza conseguem mostrar uma boa robustez do modelo.

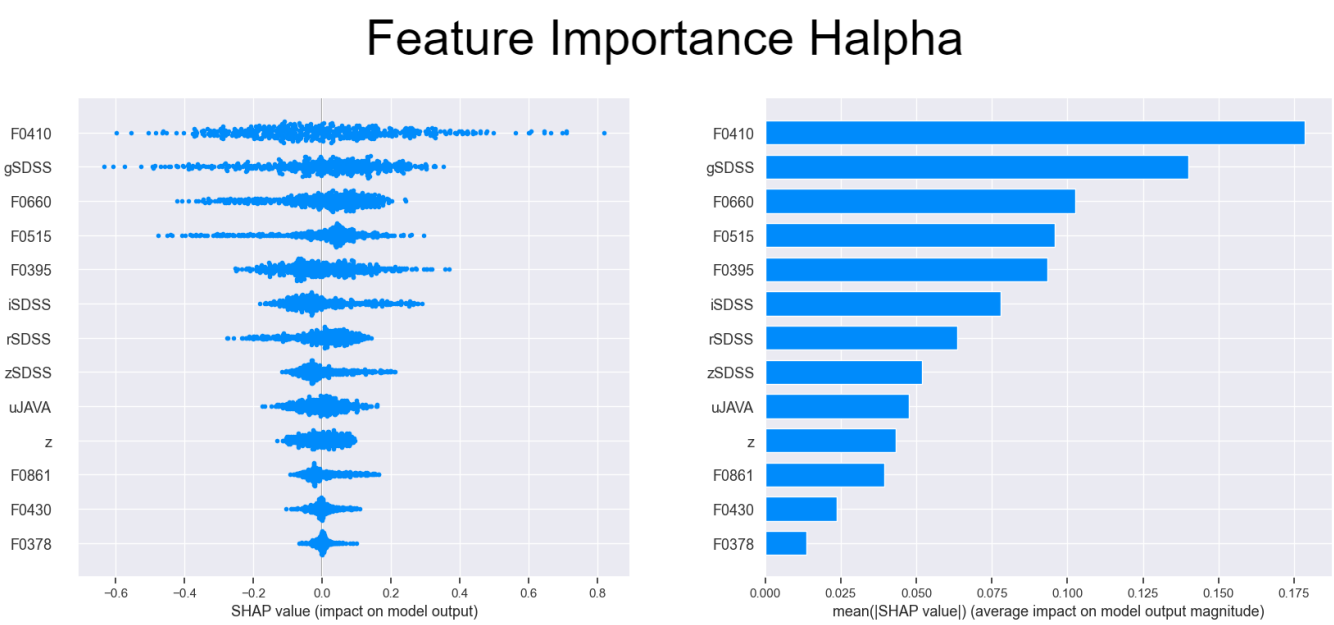

Figura 4.9: Igual a Figura 4.8 mas para a estimativa da largura equivalente de $H\alpha$.

Como discutido na Seção 4.2, obter a estimativa de $H\alpha$ mesmo para *redshifts* altos é muito interessante. Aplicando o mesmo processo que fizemos para a massa estelar, podemos tentar entender como a rede obtém essa predição. Na Figura 4.9 vemos que o filtro F410 é o mais importante, seguido pelo G e F660. Se tivéssemos mais objetos em baixos *redshifts*, é possível que o F660 fosse o filtro mais importante, já que ele ditaria a largura equivalente de Hα para uma porção maior de objetos. Apesar de ter essa informação, entender exatamente o porquê do F410 ser a banda mais importante é uma pergunta sem muitas explicações.

Capítulo 5

Conclusões

A partir das sínteses espectrais de Werle et al. (2018) fomos capazes de criar um conjunto de treinamento completo com fotometria sintética para 137,734 galáxias e que se assemelha às observações do S-PLUS. Após uma calibração extensa e cuidados com *missing data*, geramos uma base de dados relacionando a fotometria desses objetos com os parâmetros de suas populações estelares.

Utilizando técnicas do estado da arte de aprendizado de máquina, treinamos diversos modelos de *Machine Learning* a partir dessa base de dados. Após testar algoritmos como XGBoost e Random Forest, acabamos definindo uma rede neural de *Deep Learning* para as estimativas.

Estimamos os parâmetros de populações estelares com uma boa confiança, mesmo para *redshifts* mais altos. A massa estelar, por exemplo, teve um Erro Médio Absoluto de 0.06 dex. Todos os outros parâmetros também foram obtidos com erros pelo menos uma ordem de grandeza a menos. As larguras equivalentes das linhas espectrais foram encontradas mesmo para objetos onde a informação da linha não está contida na sua banda específica, o que nos impressionou muito (e.g. estimar Hα para $z > 0.02$, que faz a linha cair fora da banda F660).

Mostramos aqui algumas das inúmeras aplicações possíveis que nossa ferramenta pode proporcionar. A estimativa dos parâmetros das populações estelares de uma galáxia é interessante por si só, porém existem diversos estudos que vão usar isso como uma parte de suas pesquisas. Estudos sobre propriedades de membros de aglomerados, por exemplo, podem se beneficiar muito das nossas estimativas.

Pretendemos disponibilizar na internet, junto com a publicação dessa tese, os códigos utilizados para que outros astrônomos possam usufruir. Da mesma forma que transfor-

mamos os espectros de Werle et al. (2018) em um banco de dados para treinar uma rede baseada no S-PLUS, podemos gerar dados similares para outros *surveys* fotométricos, como o J-PLUS e o J-PAS

A astronomia depende cada vez mais de *data science*. O volume de dados que os astrônomos têm acesso hoje é maior do que se poderia imaginar décadas atrás. O *machine learning*, junto com *deep learning*, ganha cada vez mais espaço na área da astronomia, e sua existência já passa de um auxílio aos métodos tradicionais para uma dependência necessária.

Muitos resultados promissores no campo da astronomia fotométrica estão prestes a serem disponibilizados em um curto prazo. Novas fotometrias do S-PLUS, por exemplo, prometem cobrir uma área maior do céu e devem a ir a público em pouco tempo. Como o estudo de *redshifts* fotométricos é de grande interesse, o avanço nessa área é rápido e novos resultados devem aparecer em breve, melhorando as estimativas. Somando ambos esses fatores, assim como evoluções no âmbito geral da astronomia, esse é um projeto que pode ser aprimorado e continuado no futuro.

Referências Bibliográficas

Ait-Ouahmed R., Arnouts S., Pasquet J., Treyer M., Bertin E., , 2023 Multimodality for improved CNN photometric redshifts

Azur M. J., Stuart E. A., Frangakis C., Leaf P. J., Multiple imputation by chained equations: what is it and how does it work?, International Journal of Methods in Psychiatric Research, 2011, vol. 20, p. 40

Balaban S., Deep learning and face recognition: the state of the art, Biometric and surveillance technology for human and activity identification XII, 2015, vol. 9457, p. 68

Baldwin J. A., Phillips M. M., Terlevich R., Classification parameters for the emission-line spectra of extragalactic objects., PASP, 1981, vol. 93, p. 5

Barbosa F. O., Santucci R. M., Rossi S., Limberg G., Pérez-Villegas A., Perottoni H. D., The SDSS-Gaia View of the Color–Magnitude Relation for Blue Horizontal-branch Stars, The Astrophysical Journal, 2022, vol. 940, p. 30

Baron D., , 2019 Machine Learning in Astronomy: a practical overview

Benitez N., Dupke R., Moles M., Sodre L., Cenarro J., Marin-Franch A., Taylor K., Cristobal D., Fernandez-Soto A., Mendes de Oliveira C., Cepa-Nogue J., Abramo L. R., et al. J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey, arXiv e-prints, 2014, p. arXiv:1403.5237

Benítez N., Bayesian Photometric Redshift Estimation, The Astrophysical Journal, 2000, vol. 536, p. 571

Boggs P. T., Donaldson J. R., , 1989 Technical report Orthogonal distance regression https://doi.org/10.6028/nist.ir.89-4197

Böhm V., Kim A. G., Juneau S., , 2023 Fast and efficient identification of anomalous galaxy spectra with neural density estimation

Castelli A. V. S., de Oliveira C. M., Herpich F., Barbosa C. E., Escudero C., Grossi M., Sodre L., de Bom C. R., Zenocratti L., Rossi M. E. D., Cortesi A., Fernandes R. C., Lopes A. R., Telles E., Schwarz G. B. O., Dantas M. L. L., Faifer F. R., Santos A. C., Saponara J., Reynaldi V., Andruchow I., Sesto L., Mestre M. F., de Amorim A. L., de Lima E. V. R., Abboud J. C. R., Cernic V., de Almeida Garcia I. S., , 2021 The Fornax Cluster through S-PLUS

Cenarro A. J., Moles M., Cristó bal-Hornillos D., Marín-Franch A., Ederoclite A., Varela J., López-Sanjuan C., Hernández-Monteagudo C., Angulo R. E., Ramió H. V., Viironen K., Bonoli S., Orsi A. A., Hurier G., Roman I. S., Greisel N., et al. J-PLUS: The Javalambre Photometric Local Universe Survey, Astronomy & Astrophysics, 2019, vol. 622, p. A176

Charlot S., Bruzual G. A., Stellar Population Synthesis Revisited, ApJ, 1991, vol. 367, p. 126

Chen T., Guestrin C., XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , KDD '16, ACM, New York, NY, USA, 2016, p. 785

Chiosi C., Bertelli G., Bressan A., Integrated colours and ages of LMC clusters : the nature of the bimodal distribution of the (B-V) colours., A&A, 1988, vol. 196, p. 84

Chollet F., et al.,, 2015 Keras https://keras.io

Chow J. C. K., Analysis of Financial Credit Risk Using Machine Learning, 2017

Cid Fernandes R., Mateus A., Sodré L., Stasińska G., Gomes J. M., Semi-empirical analysis of Sloan Digital Sky Survey galaxies - I. Spectral synthesis method, MNRAS, 2005a, vol. 358, p. 363

Cid Fernandes R., Mateus A., Sodré L., Stasińska G., Gomes J. M., Semi-empirical analysis of Sloan Digital Sky Survey galaxies - I. Spectral synthesis method, MNRAS, 2005b, vol. 358, p. 363

Cunningham P., Delany S. J., k-Nearest Neighbour Classifiers - A Tutorial, ACM Computing Surveys, 2021, vol. 54, p. 1

de Oliveira F. M. F., dos Santos M. V., Reis R. R. R., , 2023 Data-driven photometric redshift estimation from type Ia supernovae light curves

Dorogush A. V., Ershov V., Gulin A., , 2018 CatBoost: gradient boosting with categorical features support

Drinkwater M. J., Gregg M. D., Colless M., Substructure and Dynamics of the Fornax Cluster, The Astrophysical Journal, 2001, vol. 548, p. L139

Etsebeth V., Lochner M., Walmsley M., Grespan M., , 2023 Astronomaly at Scale: Searching for Anomalies Amongst 4 Million Galaxies

Flórido T. Z., Análise de linhas de emissão em galáxias: o gás difuso ionizado nas galáxias do MaNGA, Universidade Federal de Santa Catarina, 2018, Tese de Doutorado

Gomes J. M., Papaderos P., Fitting Analysis using Differential evolution Optimization (FADO), Astronomy & Astrophysics, 2017, vol. 603, p. A63

Haack R., Smith Castelli V. A., Faifer F., Mendes de Oliveira C., Almeida Fernandes F., Lopes A. R., Detección y medición automática de galaxias ubicadas en la dirección del cúmulo de Fornax en imágenes de S-PLUS, Facultad de Ciencias Astronómicas y Geofísicas, Universidad Nacional de La Plata, 2023, Tese de Doutorado

Heavens A. F., Jimenez R., Lahav O., Massive lossless data compression and multiple parameter estimation from galaxy spectra, MNRAS, 2000, vol. 317, p. 965

Ivezic Z., Connolly A. J., VanderPlas J. T., Gray A., Statistics, data mining, and machine learning in astronomy. Princeton Series in Modern Observational Astronomy, Princeton University Press Princeton, NJ, 2019

Jagtap R., Inamdar U., Dere S., Fatima M., Shardoor N., Habitability of Exoplanets using Deep Learning , 2021, p. 1

Jakka M. S., , 2023 Assessing Exoplanet Habitability through Data-driven Approaches: A Comprehensive Literature Review

Johnson B. D., Leja J., Conroy C., Speagle J. S., Stellar Population Inference with Prospector, The Astrophysical Journal Supplement Series, 2021, vol. 254, p. 22

Jones E., Do T., Boscoe B., Singal J., Wan Y., Nguyen Z., , 2023 Photometric Redshifts for Cosmology: Improving Accuracy and Uncertainty Estimates Using Bayesian Neural Networks

Kaiser N., Burgett W., Chambers K., Denneau L., Heasley J., Jedicke R., Magnier E., Morgan J., Onaka P., Tonry J., The Pan-STARRS wide-field optical/NIR imaging survey. In Ground-based and Airborne Telescopes III , vol. 7733 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2010, p. 77330E

Kollmeier J., Anderson S. F., Blanc G. A., Blanton M. R., Covey K. R., Crane J., Drory N., Frinchaboy P. M., Froning C. S., Johnson J. A., et al. SDSS-V Pioneering Panoptic Spectroscopy. In Bulletin of the American Astronomical Society , vol. 51, 2019, p. 274

Kremer J., Stensbo-Smidt K., Gieseke F., Pedersen K. S., Igel C., Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, IEEE Intelligent Systems, 2017, vol. 32, p. 16

Kuutti S., Bowden R., Jin Y., Barber P., Fallah S., A survey of deep learning applications to autonomous vehicle control, IEEE Transactions on Intelligent Transportation Systems, 2020, vol. 22, p. 712

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

Lima-Dias C., Monachesi A., Torres-Flores S., Cortesi A., Hernández-Lang D., Barbosa C. E., Mendes de Oliveira C., Olave-Rojas D., Pallero D., Sampedro L., Molino A., Herpich F. R., Jaffé Y. L., Amorín R., Chies-Santos A. L., Dimauro P., Telles E., Lopes P. A. A., Alvarez-Candal A., Ferrari F., Kanaan A., Ribeiro T., Schoenell W., An

environmental dependence of the physical and structural properties in the Hydra cluster galaxies, Monthly Notices of the Royal Astronomical Society, 2020, vol. 500, p. 1323

Louppe G., , 2015 Understanding Random Forests: From Theory to Practice

Lundberg S. M., Lee S.-I., A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems , vol. 30, Curran Associates, Inc., 2017

Marshall A., Auger-Williams M. W., Banerji M., Maiolino R., Bowler R., A fresh look at AGN spectral energy distribution fitting with the XMM-SERVS AGN sample, Monthly Notices of the Royal Astronomical Society, 2022, vol. 515, p. 5617

Martin D. C., Fanson J., Schiminovich D., Morrissey P., Friedman P. G., Barlow T. A., Conrow T., Grange R., Jelinsky P. N., Milliard B., Siegmund O. H. W., Bianchi L., Byun Y.-I., Donas J., Forster K., Heckman T. M., Lee Y.-W., Madore B. F., Malina R. F., Neff S. G., Rich R. M., Small T., Surber F., Szalay A. S., Welsh B., Wyder T. K., The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission, ApJ, 2005, vol. 619, p. L1

Meher S. K., Panda G., Deep learning in astronomy: a tutorial perspective, The European Physical Journal Special Topics, 2021, vol. 230, p. 2285

Meijer Q., Lopez M., Tsuna D., Caudill S., , 2023 Gravitational-Wave Searches for Cosmic String Cusps in Einstein Telescope Data using Deep Learning

Mendes de Oliveira C., Ribeiro T., Schoenell W., Kanaan A., Overzier R. A., Molino A., Sampedro L., Coelho P., Barbosa C. E., Cortesi A., Costa-Duarte M. V., Herpich F. R., Hernandez-Jimenez J. A., Placco V. M., Xavier H. S., Abramo L. R., et al. The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters, Monthly Notices of the Royal Astronomical Society, 2019, vol. 489, p. 241

Nousi P., Koloniari A. E., Passalis N., Iosif P., Stergioulas N., Tefas A., Deep residual networks for gravitational wave detection, Physical Review D, 2023, vol. 108

Ocvirk P., Pichon C., Lançon A., Thiébaut E., STECKMAP: STEllar Content and Kinematics from high resolution galactic spectra via Maximum A Posteriori, MNRAS, 2006, vol. 365, p. 74

Parkhi O. M., Vedaldi A., Zisserman A., Jawahar C. V., Cats and dogs. In 2012 IEEE Conference on Computer Vision and Pattern Recognition , 2012, p. 3498

Pat F., Juneau S., Böhm V., Pucha R., Kim A. G., Bolton A. S., Lepart C., Green D., Myers A. D., , 2022 Reconstructing and Classifying SDSS DR16 Galaxy Spectra with Machine-Learning and Dimensionality Reduction Algorithms

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011, vol. 12, p. 2825

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I., Language Models are Unsupervised Multitask Learners, 2019

Raschka S., , 2020 Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning

STScI development Team, 2018 synphot: Synthetic photometry using Astropy

Thainá-Batista J., Fernandes R. C., Herpich F. R., de Oliveira C. M., Werle A., Espinosa L., Lopes A., Castelli A. V. S., Sodré L., Telles E., Kanaan A., Ribeiro T., Schoenell W., , 2023 Estimating stellar population and emission line properties in S-PLUS galaxies

van Buuren S., Groothuis-Oudshoorn K., mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 2011, vol. 45

Walcher J., Groves B., Budavári T., Dale D., Fitting the integrated spectral energy distributions of galaxies, Astrophysics and Space Science, 2010, vol. 331, p. 1

Wang G., Li W., Aertsen M., Deprest J., Ourselin S., Vercauteren T., Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing, 2019, vol. 338, p. 34

Werle A., Analysis of SDSS spectra and GALEX photometry with STARLIGHT: stellar populations and dust attenuation in local galaxies, Universidade Federal de Santa Catarina, 2019, Phd thesis

Werle A., Fernandes R. C., Asari N. V., Bruzual G., Charlot S., Delgado R. G., Herpich F. R., Simultaneous analysis of SDSS spectra and GALEX photometry with starlight: method and early results, Monthly Notices of the Royal Astronomical Society, 2018, vol. 483, p. 2382

Zhang H., Si S., Hsieh C.-J., , 2017 GPU-acceleration for Large-scale Tree Boosting