

**UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE LORENA**

Camila Oliveira Cardoso

Classificação de quasares através do algoritmo t-SNE

Lorena

2021

Camila Oliveira Cardoso

Classificação de quasares através do algoritmo t-SNE

Trabalho de Conclusão de Curso de Engenharia Física apresentado à Escola de Engenharia de Lorena – Universidade de São Paulo.

Orientador: Prof. Dr. Filipe Batoni Abdalla

**Lorena
2021**

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E
PESQUISA, DESDE QUE CITADA A FONTE.

Cardoso, Camila Oliveira
Classificação de quasares através do algoritmo t-SNE /
Camila Oliveira Cardoso ; orientador Filipe Batoni Abdalla.
– Lorena, 2021.
68 p. : il. (algumas color.) ; 30 cm.

Monografia (Graduação em Engenharia Física) – Escola de
Engenharia de Lorena, Universidade de São Paulo, 2021.

t-distributed Stochastic Neighbor Embedding. Quasares.
K Dimensional Tree. Astronomia. I. Abdalla, Filipe Batoni,
orient..

Este trabalho é dedicado a todos os educadores, inclusive meus familiares, que empregaram seu tempo e energia a ajudarem a me tornar uma pessoa melhor.

AGRADECIMENTOS

Aqui começa um registro muito importante, pois em poucos momentos na vida dedicamos tempo a agradecer à todos que fazem da nossa existência algo especial.

Agradeço aos meus pais, Ana Aguiar e Antonio Geraldo, por serem amorosos e terem se dedicado tanto à mim, certamente eu não teria chegado até aqui sem eles. Agradeço pelo meus falecidos avôs pelos anos de vida que estiveram comigo compartilhando um pouco da sua alegria e brilho que tinham em relação à vida. Agradeço a presença maternal das minhas duas avós, que ainda me perguntam a cada vez que conversamos, com todo amor, quando eu volto para casa para vê-las. Agradeço aos meus tios e tias que entre todos sempre cuidaram muito de mim, e agradeço aos meus primos o companheirismo, são como irmãos que eu não cheguei a ter. Agradeço à minha família como um todo, pois com ela nunca me faltou amor.

Agradeço aos meus educadores, do ensino fundamental até a universidade. Essa graduação foi um seguido de desafios constantes, e nessa trajetória eu me alegrei muito por encontrar professores que transbordavam dedicação em ensinar e ajudar os alunos nas suas respectivas caminhadas, e gostaria de mencionar os mais marcantes: Juan Zapata, Morun Neto, Paula Pardal, Fabiano Bargos, Bertha Cuadros-Melgar, Rebeca Bacani, Luiz Eleno e Durval Junior. Um agradecimento especial ao professor Luiz Eleno, pela imensurável ajuda nos últimos anos como coordenador de curso, como orientador de estágio, como professor e como mentor. Agradeço ao professor Carlos Shigue pelos dois anos de projeto de extensão, foi engrandecedor trabalhar com um projeto educacional, e também tê-lo como amigo.

Agradeço à professora Bertha Cuadros-Melgar por ter confiado em mim e nas minhas intenções e ter tornado possível a minha participação no BINGO, que fez com que esse trabalho fosse possível. Com relação ao BINGO, agradeço ao professor Elcio Abdalla por me abrir as portas deste projeto, ao Alessandro Marins por ter me ajudado a achar um caminho dentro dele, à Karin Franzoni por todo apoio e ajuda dada durante a escrita deste trabalho, e a todos os integrantes do grupo de Dados Ópticos pela ajuda mútua e troca de conhecimento que tivemos durante todas as semanas este ano. Agradeço ao meu orientador Filipe Abdalla por me guiar através da execução deste trabalho, sempre sendo muito atento na passagem do conhecimento.

Agradeço aos meus colegas e amigos de trabalho da Shopper que tiveram compreensão e me apoiaram neste encerramento de ciclo.

Aos amigos que eu fiz vivendo em Lorena ao longo de 6 anos, eles foram uma segunda família e passei e ainda passo momentos lindos junto deles, obrigada. Aos amigos

da minha cidade natal, que mesmo após a minha partida, continuam comigo, também só tenho a agradecer. Agradeço à minha companheira pela paciência, amor, ajuda e acolhimento durante esse período desafiante, entre risadas e momentos de cansaço, tê-la ao lado foi muito importante.

“I look up at the night sky, and I know that, yes, we are part of this Universe, we are in this Universe, but perhaps more important than both of those facts is that the Universe is in us. When I reflect on that fact, I look up—many people feel small, because they’re small and the Universe is big, but I feel big, because my atoms came from those stars.”

Neil deGrasse Tyson

RESUMO

CARDOSO, C. O. **Classificação de quasares através do algoritmo t-SNE.** 2021. 68p. Monografia (Trabalho de Conclusão de Curso) - Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2021.

O presente trabalho apresenta uma separação de quasares de outros objetos através do método *t-distributed Stochastic Neighbor Embedding* (t-SNE). O t-SNE trata-se de um algorítimo de redução dimensional, para um conjunto de pontos em altas dimensionalidades onde procura-se preservar a sua distribuição original. Dentre os parâmetros do t-SNE que podem influenciar a distribuição final dos pontos, variou-se a perplexidade e o número de iterações afim de obter um bom resultado. O grupo amostral foi separado em um conjunto de treino e um de teste, e eles compreendem objetos presentes no Dark Energy Survey *Data Release 2* (DES DR2) que tem uma correspondência no levantamento espectroscópico do Sloan Digital Sky Surveys *Data Release 16* (SDSS DR16), onde a classificação contida no SDSS DR16 foi usada como identificação dos objetos em: quasar, estrela ou galáxia. O algorítimo t-SNE foi executado para a amostra de treino, com as seguintes dimensões a serem reduzidas: as magnitudes, as cores e as classificações morfológicas dos objetos. Executou-se 9 reduções distintas variando os parâmetros perplexidade e número de iterações. A melhor redução obtida foi com perplexidade igual a 100 e número de iterações igual a 5000, sendo possível obter um corte que continha 70,18% dos quasares em relação ao total de quasares da amostra. Para testar a eficácia dessa distribuição, submeteu-se a amostra de teste ao algorítimo *K Dimensional Tree* (kD Tree), que identifica qual é o ponto vizinho mais próximo de um determinado ponto em alta dimensionalidade. O kD Tree foi utilizado para identificar em cada objeto da amostra de teste os dois vizinhos mais próximos contidos na amostra de treinamento, e assim realizava-se a média das coordenadas desses dois objetos do treinamento no plano e essa era a respectiva coordenada no plano do objeto da amostra de teste. O mesmo corte foi feito para amostra de teste e resultou em uma composição muito semelhante à obtida para o treinamento, onde continha 76,58% dos quasares em relação ao total de quasares da amostra. O objetivo de fazer uma separação de quasares em um conjunto fotométrico de dados heterogêneos se mostrou alcançável utilizando a ferramenta de redução de dimensionalidade t-SNE. Porém não é possível afirmar um bom resultado quando o modelo for aplicado em um conjunto amostral menos restrito, e se pode melhorar a pureza do corte adotando outros métodos de seleção.

Palavras-chave: *t-distributed Stochastic Neighbor Embedding*. Quasares. *K Dimensional Tree*. Astronomia.

ABSTRACT

CARDOSO, C. O. **Quasar classification using the t-SNE algorithm.** 2021. 68p. Monograph (Conclusion Course Paper) - Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2021.

The present work presents a separation of quasars from other objects using the t-distributed Stochastic Neighbor Embedding (t-SNE) method. The t-SNE is a dimensional reduction algorithm, for a set of points in high dimension where we try to preserve its original distribution. Among the t-SNE parameters that can influence the final distribution of points, the perplexity and the number of iterations were varied in order to obtain a good result. The sample group was separated into a training and a test set, and they comprise objects present in the Dark Energy Survey Data Release 2 (DES DR2) that have a correspondence in the spectroscopic survey of the Sloan Digital Sky Surveys Data Release 16 (SDSS DR16), where the classification contained in the SDSS DR16 was used to identify the objects in: quasar, star or galaxy. The t-SNE algorithm was run for the training sample, with the following dimensions to be reduced: magnitudes, colors and morphological classifications of objects. Nine different reductions were performed, varying the perplexity and number of iterations parameters. The best reduction obtained was with perplexity equal to 100 and number of iterations equal to 5000, being possible to obtain a cut that contained 70.18% of the quasars in relation to the total number of quasars in the sample. To test the effectiveness of this distribution, the test sample was submitted to the algorithm K Dimensional Tree (KD Tree), which identifies which is the closest neighbor point of a given point in high dimensionality. The KD Tree was used to identify in each object of the test sample the two closest neighbors contained in the training sample, and thus the average of the coordinates of these two training objects in the plane was performed and this was the respective coordinate in the plane of the test sample object. The same cut was made for the test sample and resulted in a composition very similar to that obtained for the training, which contained 76.58% of the quasars in relation to the total number of quasars in the sample. The objective of making a separation of quasars in a photometric set of heterogeneous data proved to be achievable using the t-SNE dimensionality reduction tool. However, it is not possible to affirm a good result when the model is applied to a less restricted sample set, and the purity of the cut can be improved by adopting other selection methods.

Keywords: t-distributed Stochastic Neighbor Embedding. Quasars. K Dimensional Tree. Astronomy.

LISTA DE FIGURAS

Figura 1 – Espectro de uma galáxia comum e de uma galáxia do tipo Seyfert.	30
Figura 2 – Espectro de um quasar comparado com as linhas de emissão do hidrogênio.	31
Figura 3 – Imagens comparativas de uma galáxia comum NGC7626 e uma galáxia ativa NGC5548	34
Figura 4 – Bandas de detecção do DES <i>Wide-area</i>	36
Figura 5 – Área do céu coberta pelos sub-levantamentos espectroscópicos disponíveis no DR16.	37
Figura 6 – Densidade de quasares por <i>pixel</i> presentes no SDSS DR16	44
Figura 7 – Os mapas 7a,7b e 7c foram feitos utilizando uma projeção gnomônica, com pixels tendo resolução de $N_{side} = 64$, esquema de ordenamento do tipo <i>RING</i>	46
Figura 8 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 1000.	50
Figura 9 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 2000.	51
Figura 10 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 3000.	52
Figura 11 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 4000.	54
Figura 12 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 5000.	55
Figura 13 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 2 e número de iterações igual a 5000.	57
Figura 14 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 30 e número de iterações igual a 5000.	58
Figura 15 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 50 e número de iterações igual a 5000.	60
Figura 16 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 100 e número de iterações igual a 5000.	61
Figura 17 – Os histogramas com a distribuição dos quasares ao longo do eixo <i>x</i> e do eixo <i>y</i>	63

LISTA DE TABELAS

Tabela 1 – Sistema fotométrico utilizado no DES	32
Tabela 2 – Composição das amostras de treinamento e teste	47
Tabela 3 – Percentual dos diferentes tipos de objetos em relação ao total de objetos contidos no corte α	62
Tabela 4 – Percentual dos diferentes tipos de objetos contidos no corte α em relação ao total de objetos do mesmo tipo contidos na amostra sem o corte. . .	62

LISTA DE QUADROS

Quadro 1 – Descrição dos tributos utilizados dos objetos procedentes do DES. . . 43

LISTA DE ABREVIATURAS E SIGLAS

SNE	<i>Stochastic Neighbor Embedding</i>
t-SNE	<i>t-distributed Stochastic Neighbor Embedding</i>
HEALPix	<i>Hierarchical Equal Area and isoLatitude Pixelization</i>
AGN	<i>Active Galactic Nuclei</i>
DES	Dark Energy Survey
SDSS DR16	Sloan Digital Sky Surveys Data Release 16
DES DR2	Dark Energy Survey Data Release 2
SDSS	Sloan Digital Sky Surveys
KD Tree	<i>K Dimensional Tree</i>
RA	<i>right ascension</i>
DEC	<i>declination</i>
APO	Apache Point Observatory
LCO	Las Campanas Observatory
BSN	Buraco Negro Supermassivo
ΛCDM	Λ <i>Cold Dark Matter</i>
BAO	<i>Baryon Acoustic Oscillations</i>
BINGO	<i>Baryon Acoustic Oscillations (BAO) from Integrated Neutral Gas Observations</i>
SQL	<i>Structured Query Language</i>
PSF	<i>Point Spread Function</i>
HTM	<i>Hierarchical Triangular Mesh</i>

SUMÁRIO

1	INTRODUÇÃO	23
2	DESENVOLVIMENTO	27
2.1	Propriedades intrínsecas e medidas de objetos estelares	27
2.1.1	Sistema equatorial celeste	27
2.1.2	Desvio para o Vermelho	27
2.1.3	Espectroscopia	28
2.1.4	Fotometria	31
2.1.5	Índices de cor	32
2.2	Os objetos encontrados no céu	32
2.2.1	Estrelas	32
2.2.2	Galáxias	33
2.2.2.1	Núcleos Ativos de Galáxias	33
2.2.3	Quasares	34
2.3	Grandes Surveys	35
2.3.1	<i>Dark Energy Survey (DES)</i>	35
2.3.2	<i>Sloan Digital Sky Surveys(SDDS)</i>	36
2.4	Metodologia	37
2.4.1	<i>Hierarchical Equal Area and isoLatitudte Pixelization (HEALPix)</i>	37
2.4.2	<i>Match entre surveys diferentes</i>	38
2.4.3	t-SNE	38
2.4.4	<i>KDimensional Tree</i>	41
2.4.5	Dados pertencentes ao DES	41
2.4.6	Dados pertencentes ao SDSS	43
2.4.7	Pré-processamento e o grupo amostral	44
2.4.8	Análise	47
3	RESULTADOS E DISCUSSÃO	49
3.1	Resultados	49
3.1.1	Análise t-SNE variando o número de iterações	49
3.1.2	Análise t-SNE variando a perplexidade	56
3.1.3	Separação de quasares via análise t-SNE	62
3.2	Discussão	63
4	CONCLUSÃO	65

1 INTRODUÇÃO

Foi com o desenvolvimento da radio astronomia, na década de 50, que foram detectados objetos no céu com grandes distâncias cosmológicas da Terra, que antes eram desconhecidos por não serem facilmente visíveis ao olho humano ou por telescópios óticos (CHAISSON; MCMILLAN, 2013). Em 1963, astrônomos conseguiram uma correspondência à um desses objetos que emitiam em frequências de rádio (um objeto nomeado como 3C 273) e seu respectivo espectro, que causou muito estranhamento por ter linhas de emissões em comprimentos de ondas completamente desconhecidas (SCHMIDT, 1963). Uma descoberta notável (SCHMIDT, 1963) foi quando percebeu-se que essas linhas de emissões eram simplesmente linhas de emissões comuns ao hidrogênio, mas que estavam em comprimentos de ondas distintos dos usuais. Este objeto foi o primeiro quasar a ser identificado, e por esse deslocamento das linhas espectrais do hidrogênio ele ganhou muita importância na área. Quasares, nome precedente do termo em inglês *Quasi-stellar radio sources* (doravante será chamado de quasar ou em plural quasares), são objetos presentes no universo que se caracterizam por serem brilhantes, de maneira bem similar às estrelas, mas com uma grande distância cosmológica. Pode-se mencionar também que diferentemente das estrelas, os quasares estão se afastando da Terra com uma alta velocidade (o que causa o deslocamento das linhas espectrais do hidrogênio) e possuem um espectro de emissão eletromagnético bem distinto delas. Estes objetos tão distintos já foram classificados como diferentes de galáxias ativas, mas atualmente a maioria dos astrônomos definem quasar como o núcleo intensamente brilhante de uma galáxia que o circunda (CHAISSON; MCMILLAN, 2013).

Núcleos Ativos de Galáxias, ou do inglês *Active Galactic Nuclei* (AGN) é o nome dado aos núcleos de galáxias que apresentam um brilho maior do que de galáxias comuns e que possuem uma emissão não-estelar de radiação eletromagnética (CHAISSON; MC-MILLAN, 2013). A fonte de tanta energia de um AGN vem da existência de um Buraco Negro Supermassivo (BNS) no centro da galáxia, que no processo de capturar a matéria ao seu redor acaba convertendo parte da energia potencial gravitacional dela em energia eletromagnética, emitida através do universo (OLIVEIRA, 2017). Os quasares são classificados como AGN, e sabe-se que são objetos que sofrem variações com o tempo, causadas pelas mudanças na disponibilidade de matéria ao seu redor. Quando essa disponibilidade acaba, o BNS continua existindo, porém sem a emissão de luz, e esses quasares mortos podem estar presentes em muitas galáxias próximas. Entender os quasares, sua evolução e distribuição contribui para o entendimento da composição do universo (CHAISSON; MCMILLAN, 2013).

Atualmente, o modelo cosmológico aceito para descrever o universo é o Λ *Cold*

Dark Matter (Λ CDM), que utiliza-se da constante cosmológica Λ (JARVIS *et al.*, 2012) que está associada a entidades cuja essência ainda não compreendemos bem: energia escura e matéria escura. A energia escura seria o elemento faltante para explicar a expansão acelerada do Universo. Já matéria escura interage apenas gravitacionalmente com a matéria visível, e a sua existência é fundamentada pelo comportamento de estruturas como galáxias e aglomerados de galáxias (estruturas de larga escala) (VELTEN *et al.*,). Foram graças às flutuações na matéria escura que as aglomerações de matéria bariônica foram impulsionadas, dando origem à galáxias e aglomerados. Então, pode-se afirmar que por estudar como se distribuem as estruturas de larga escala também é possível entender distribuição de matéria escura no Universo (VELTEN, 2021). Na intenção de gerar resultados para melhorar e comprovar o modelo cosmológico, é possível citar iniciativas como o *Baryon Acoustic Oscillations (BAO) from Integrated Neutral Gas Observations* (BINGO)¹ radio telescópio, que atualmente está sendo construído na região nordeste do Brasil e tem como finalidade mapear estruturas de grande escala, que contribuiria para traçar a distribuição total de matéria no universo (ABDALLA *et al.*, 2021), e o *Dark Energy Survey* (DES)², que tem como principal objetivo caracterizar o setor escuro e testar modelos gravitacionais, examinando também estruturas de grande escala, entre outros (COLLABORATION: *et al.*, 2016).

Pensando na contribuição do estudo e entendimento de quasares, a maneira mais eficiente de identificar um quasar de outro objeto no céu é realizando uma análise espectroscópica. Esta no entanto é uma forma custosa quando se considera o tempo de exposição necessário para registrar fluxos de emissão tipicamente fracos de quasares, lembrando que majoritariamente encontram-se à grandes distâncias (Nakoneczny, S. *et al.*, 2019). Uma das alternativas para essa limitação seria o desenvolvimento de técnicas de separação de quasares utilizando levantamentos fotométricos (técnica abordada na Secção 2.1.4), que em geral conseguem cobrir uma área maior do céu.

Com a importância da determinação das constantes cosmológicas para a formação do universo e a contribuição que o estudo de quasares pode dar para esse tema e outros mais, o presente trabalho tem o objetivo realizar uma seleção de quasares em um conjunto de objetos medido através da fotometria, que contém também galáxias e estrelas. O conjunto de dados fotométricos usados nesta seleção tem origem na base de dados do DES, onde foram escolhidos apenas os que tem uma correspondência no levantamento espectroscópico do *Sloan Digital Sky Surveys Data Release 16* (SDSS DR16) (AHUMADA *et al.*, 2020). Usou-se os seguintes atributos desses objetos: as magnitudes das cinco bandas de detecção, as cores derivadas dessas magnitudes e a classificação morfológica dos objetos também das cinco bandas de detecção.

¹ <https://bingotelescope.org/pt/>

² <https://www.darkenergysurvey.org/>

A metodologia aplicada nesse trabalho consiste em um pré-processamento do grupo amostral, seguido da seleção dos objetos que compõe uma amostra de treino e uma amostra de teste. Em seguida, utilizou-se do algorítimo de *t-distributed stochastic neighbor embedding* (t-SNE) para a classificação dos objetos presentes na amostra de treino. O t-SNE consiste na transformação e redução de uma matriz de alta dimensionalidade, onde é levado em conta a distribuição dos pontos nesse espaço, e se tenta preservar essa distribuição na redução para espaços com 1,2 ou 3 dimensões, isso possibilita a visualização no plano ou no espaço desses pontos, e também a formação de *clusters* que representam a similaridade entre esses pontos (MAATEN; HINTON, 2008). Essa alta dimensionalidade pode ser representada por um conjunto de dados com muitas características diferentes, como por exemplo em um levantamento astronômico que pode conter inúmeras variáveis que caracterizam os objetos (como as magnitudes em diferentes bandas, as cores, o seu *redshift*).

Dentre os parâmetros do t-SNE que podem influenciar a distribuição final dos pontos, variou-se a perplexidade e o número de iterações afim de obter um bom resultado. Após a seleção da melhor distribuição no plano obtida por t-SNE, para testar a eficácia dessa distribuição, submeteu-se a amostra de teste ao algorítimo K *Dimensional Tree* (KD *Tree*) (MANEEWONGVATANA; MOUNT, 1999), que identifica qual é o ponto vizinho mais próximo de um determinado ponto em alta dimensionalidade. O KD *Tree* foi utilizado para identificar em cada objeto da amostra de teste os dois vizinhos mais próximos contidos na amostra de treinamento, e assim realizou-se a média das coordenadas no plano desses dois objetos do treinamento e esta era a respectiva coordenada no plano do objeto da amostra de teste. Espera-se com este trabalho que seja possível obter um panorama da eficiência do uso do t-SNE para separação de quasares dos demais objetos.

2 DESENVOLVIMENTO

Neste capítulo apresenta-se a fundamentação teórica e a metodologia adotada no trabalho. Na fundamentação teórica descreve-se quais são os critérios que podem ser medidos de um corpo celeste e seus métodos, as classes de objetos que podem ser detectados no céu e o então os dois levantamentos astronômicos de onde os dados utilizados neste trabalho se originam. Na metodologia, explica-se os critérios de seleção do grupo amostral e da separação deste em amostra de treino e teste, bem como as ferramentas computacionais empregadas em todos os processos (*t-SNE*, *KD Tree*, *Match*, etc.) e os critérios para análise dos resultados obtidos.

2.1 Propriedades intrínsecas e medidas de objetos estelares

Nessa secção, primeiramente, define-se o sistema de coordenadas utilizado para localizar os objetos no céu, o significado do desvio para o vermelho, propriedade que ajuda a determinar a distância e a velocidade do objeto em relação à Terra, e a técnica de medição espectroscópica e a fotométrica de um corpo celeste.

2.1.1 Sistema equatorial celeste

Para determinar a posição de um astro no céu se faz necessário o uso de um sistema de coordenadas, como o sistema equatorial celeste. Este sistema é fixo na esfera celeste, o equador celeste é tido como plano fundamental e possui duas coordenadas: a ascensão reta α e a declinação δ . A ascensão reta, no inglês *right ascension* (RA), é a medida do arco sob o equador celeste até o meridiano ocupado pelo astro em questão, e tem o ponto de origem onde o equador celeste que cruza com o meridiano passante pelo ponto Áries(o ponto onde o sol toca o equador celeste durante o equinócio de primavera). A Declinação, no inglês *declination* (DEC), é a medida do arco que vai até meridiano que cruza o astro desde o equador celeste. A coordenada RA varia de $0^\circ \leq \alpha \leq 360^\circ$ aumentando para leste e a coordenada DEC varia de $-90^\circ \leq \delta \leq 90^\circ$ (FILHO; SARAIVA, 2004).

2.1.2 Desvio para o Vermelho

Relembrando o conceito do efeito Doppler, sabe-se que quando uma onda é emitida por um objeto em movimento em relação ao seu observador, há uma alteração na frequência da onda medida no referencial do seu observador. Partindo do pressuposto que os objetos estelares estão em movimento aparente com relação à Terra e que a radiação eletromagnética emitida por um objeto nada mais é que uma onda, concluímos que é necessário considerar o efeito doppler quando se mede essa radiação. Por definição o Desvio para o Vermelho (*redshift*) é dado pela Equação 2.1, onde λ é o comprimento de onda medido pelo observador

e o λ_0 é o comprimento de onda emitido pela fonte. Relacionando a velocidade radial do objeto observado com relação ao observador (v_r) e o *redshift*, dado pela Equação 2.2, é possível notar que o z é negativo se a velocidade radial do objeto é negativa, o que acontece quando o objeto se aproxima do observador e quando a velocidade radial é positiva consequentemente z também, então o objeto se afasta do observador (LANDIN, 2002).

$$z = \frac{\lambda - \lambda_0}{\lambda_0} \quad (2.1)$$

Pela Equação 2.2 abaixo, é possível determinar a velocidade relativa desse objeto com relação ao observador, e utilizando-se da Lei de Hubble, e obtermos uma relação entre velocidade relativa e a distância estimada do objeto (LANDIN, 2002). A lei de Hubble é definida pela relação proporcional entre a velocidade relativa de recessão v_r e a distância d de um corpo celeste, representada na equação $v_r = H_0 d$, onde H_0 é a constante de proporcionalidade nomeada constante de Hubble. Com isso podemos concluir que, sendo z o desvio para o vermelho, z representa duas grandezas desse objeto: distância e velocidade (PASACHOFF; FILIPPENKO, 2013).

$$z = \frac{v_r}{c} \quad (2.2)$$

Pela lei de Hubble também podemos concluir que quanto mais distante o objeto, maior a sua velocidade de recessão, e a implicação dessas afirmações é que o universo está expandindo (PASACHOFF; FILIPPENKO, 2013). Como comparação, os objetos no Universo estão se distanciando uns dos outros como em um balão, onde pontos distintos na sua superfície se separam a medida que o balão se enche.

A definição do *redshift* tratado nos parágrafos anteriores, tem origem na consideração dos movimentos peculiares dos astros, aqueles que desconsideram a expansão do universo, com relação à um observador. Entretanto, existe também a definição do *redshift* cosmológico, que é a componente do *redshift* em que se é considerado que o universo está em constante expansão e que um fóton viajando pelo universo tem o seu comprimento de onda expandido, sua frequência diminuída ou desviada ao vermelho, e quanto mais longe o objeto emissor se encontra, maiores são os efeitos da expansão do universo na radiação observada. É importante ressaltar que a componente do *redshift* que se origina dos efeitos da expansão do universo nunca adotam valores negativos e também que superam muito a componente originária dos movimentos peculiares dos astros, sendo em casos de objetos muito distantes essa componente pode-se ser desconsiderada sem grandes perdas (PANNUTI, 2020).

2.1.3 Espectroscopia

Objetos estelares podem ser classificados levando-se em consideração a sua forma quando projetados no céu, porém esse aspecto morfológico traz limitações quanto à

necessidade de se comparar um objeto com o outro de maneira equivalente, pois não se consideram as diferenças nas distâncias relativas dos objetos com a Terra e, com isso, a mudança no seu tamanho aparente e magnitude aparente. Por essa razão, para caracterizar esses objetos, opta-se pelo uso da técnica de espectroscopia ou fotometria, sendo que a primeira técnica traz sobretudo o estudo do espectro eletromagnético proveniente do objeto e a segunda técnica traz, simplificadamente, uma medida do fluxo de luz emitido pelo objeto estudado (LANDIN, 2002).

Observado por Isaac Newton, 1672 e ilustrado por Pink Floyd em um dos seus discos¹, o fenômeno da refração (onde ocorre o espalhamento das ondas) e o fenômeno da difração compõe o princípio da espectroscopia (FILHO; SARAIVA, 2004). Um conceito mais formal é dado: "Espectroscopia é o estudo da luz através de suas cores componentes, que aparecem quando a luz passa através de um prisma ou de uma rede de difração. A sequência de cores formada é chamada espectro" (FILHO; SARAIVA, 2004).

Muito do que se sabe sobre as estrelas é obtido através do estudo do seu espectro. Pelas leis empíricas da espectroscopia formuladas por Gustav Kirchhoff em 1860, sabemos que (FILHO; SARAIVA, 2004):

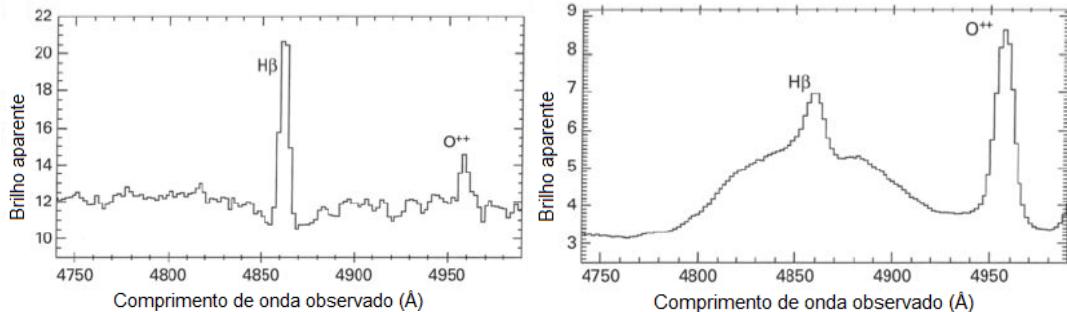
1. corpos no estado líquido, sólido ou gasoso, que esteja quente emite um espectro eletromagnético contínuo;
2. um gás pouco denso emite alguns comprimentos de onda com maior intensidade. Os valores dos comprimentos de onda dependem diretamente dos átomos presentes no gás;
3. um gás frio absorve radiação com alguns valores de comprimento de onda, e essa absorção também depende da sua composição.

A distribuição de intensidade de um espectro contínuo segue aproximadamente a lei de Planck 1901, também conhecida como lei da radiação do corpo negro, que descreve a dependência da distribuição de energia irradiada por um corpo do comprimento de onda e da temperatura (CAVALCANTE; HAAG, 2005). Sabe-se então que a intensidade medida dessa radiação tem uma correlação com a temperatura da fonte. Por isso o estudo espectral da radiação eletromagnética contribui tanto com a caracterização do espaço (FILHO; SARAIVA, 2004).

A espectroscopia foi uma técnica muito importante em grandes descobertas feitas no último século, como por exemplo o reconhecimento das galáxias do tipo Seyfert, 1943, que foram detectadas pelas características presentes em seu espectro óptico, como as Seyfert do tipo I que tem as linhas de emissões permitidas mais alargadas, como se pode ver na Figura 1 (PANNUTI, 2020).

¹ The Dark Side of the Moon, 1 de março de 1973

Figura 1 – Espectro de uma galáxia comum e de uma galáxia do tipo Seyfert.



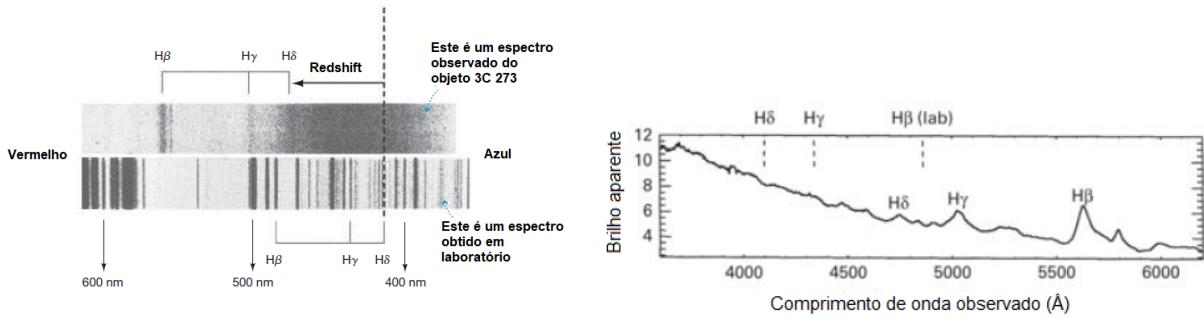
- (a) Parte do espectro do núcleo de uma galáxia comum.
 (b) Parte do espectro do núcleo da galáxia seyfert, NGC 5548.

Fonte: Adaptado de (PASACHOFF; FILIPPENKO, 2013).

A imagem corresponde à parte do espectro do núcleo de duas galáxias, onde o espectro está corrigido para mostrar as linhas espectrais de emissão. Na Figura 1a é mostrada uma galáxia normal. Na Figura 1b é mostrada uma galáxia do tipo seyfert, a NGC 5548, onde as suas linhas alargadas características são causadas pela movimentação de gás em alta velocidade localizado próximo do seu centro.

Os quasares também foram objetos melhores entendidos após análises das fotografias dos seus espectros. Na década de 60, quando observaram um espectro de emissão completamente fora dos padrões do espectro observado para estrelas em um objeto que morfológicamente se parecia à uma, despertou muita curiosidade. Esse espectro continha linha de emissões muito intensas em comprimentos de ondas que não se encaixavam com os comprimentos de ondas vistos no espectro de emissões de gases em repouso, obtidos em laboratório. Até que Maarten Schmidt (SCHMIDT, 1963) percebeu que essas linhas de emissão eram simplesmente as linhas pertencentes ao hidrogênio, ilustrado na Figura 2, que foi aquecido e que sofreu fortemente a influência do efeito Doppler, o que significava que esses objetos apresentavam *redshift* muito grandes, descobrindo assim uma nova classe de objetos (PASACHOFF; FILIPPENKO, 2013).

Figura 2 – Espectro de um quasar comparado com as linhas de emissão do hidrogênio.



(a) Fotografia do espectro óptico do quasar 3C 273 comparado com um espectro de emissão de uma lâmpada quente obtida em laboratório.

(b) Espectro do quasar 3C 273

Fonte: Adaptado de (PASACHOFF; FILIPPENKO, 2013).

Na Figura 2a, na parte superior é mostrada uma fotografia de um espectro óptico de um 3C 273, um quasar, que por ser uma imagem negativa, as linhas negras nesse contexto significam emissões. Para fins de comparação , na Figura 2a na parte inferior é apresentada uma imagem de um espectro emitido por uma lampada quente que continha hidrogênio, hélio, neon e outros elementos. As linhas de emissão $H\beta$, $H\gamma$ e $H\delta$ do hidrogênio de Balmer aparecem difusas e alargadas, também pode-se observar que em comparação com o espectro da lâmpada, as linhas do hidrogênio de Balmer aparecem em valores maiores de comprimento de onda.

Ainda hoje a análise do espectro é a maneira mais segura de identificar um quasar, onde seu espectro é bastante distinto e raramente pode ser erroneamente classificado como outro objeto (PASACHOFF; FILIPPENKO, 2013).

2.1.4 Fotometria

Por definição, a fotometria é a ciência que se mede a intensidade da luz (BURNS, 2021). Primeiramente a fotometria era realizada por meio de comparação visual, impedindo uma acurácia quando as medidas fotométricas eram feitas por mais de uma pessoa ou em ambientes diferentes. Com os anos desenvolveu-se métodos mais modernos e eficazes de quantificação da luz, como por exemplo a medida do fluxo eletromagnético que chega à um detector ou o cálculo da intensidade da luz por contagem de fôtons (BURNS, 2021). O que na prática se mede com esses detectores é o fluxo (F), que é a potência por unidade de área que cruza uma determinada superfície (FILHO; SARAIVA, 2004)

A fotometria na astronomia é utilizada para obter e quantificar a intensidade da radiação eletromagnética vindo do espaço em diversos intervalos de frequência (FILHO; SARAIVA, 2004). Um dos problemas da fotometria é que o fluxo medido em equipamentos fotométricos depende diretamente do conjunto do telescópio/detector e principalmente do filtro aplicado, ou seja, em qual intervalo de comprimento de onda o equipamento está realizando a medição. Com a necessidade de uma padronização para poder comparar dados

medidos em diferentes telescópios foram criados sistemas Fotométricos ou sistemas de magnitudes, que são sistemas que definem quais são as bandas passantes (comprimentos de ondas filtrados) e qual é a sensibilidade do conjunto (FILHO; SARAIVA, 2004).

O levantamento fotométrico utilizado nesse trabalho, DES (ABBOTT *et al.*, 2021), possui seu sistema fotométrico próprio e é caracterizado nas suas especificações. Cada banda fotométrica é um filtro óptico que permite a passagem de um intervalo de comprimento de onda, e essa banda é simbolizada por uma letra. Na Tabela 1 são caracterizados os filtros utilizados no DES.

Tabela 1 – Sistema fotométrico utilizado no DES

<i>Survey</i>	Profundidade	Banda	Comprimento de onda (Å)
DES	10 σ 23,57	g	~4750
DES	10 σ 23,34	r	~6250
DES	10 σ 22,78	i	~7750
DES	10 σ 22,10	z	~9250
DES	10 σ 20,69	Y	~10000

Fonte: A autora

2.1.5 Índices de cor

Os índices de cores podem ser obtidos como a subtração de duas magnitudes do sistema, ou que pode ser dado pela razão entre fluxos de duas bandas distintas (filtros de diferentes cores). A magnitude é uma propriedade aparente que necessita de um ponto de comparação para existir. Lembrando que quanto maior a magnitude, menor o fluxo, quanto menor a magnitude, maior o fluxo (mais intensa a luz que chegará Terra) (FILHO; SARAIVA, 2004)

2.2 Os objetos encontrados no céu

2.2.1 Estrelas

Um objeto cósmico comum, facilmente visto à olho nu ou detectado por telescópios, as estrelas podem ser descritas simplificadamente como outras versões distantes do sol. Elas possuem um ciclo de vida, onde nascem, se desenvolvem e então morrem. As estrelas se constituem de gás, que permanece preso pela ação da sua própria gravidade, e a sua energia vem do processo de fusão nuclear (quando ocorre a união de dois núcleos atômicos, resultado na liberação de energia) que se sucede em seu núcleo (PASACHOFF; FILIPPENKO, 2013).

As propriedades essenciais das estrelas podem ser compreendidas com o conhecimento de algumas quantidades físicas: o seu brilho, sua temperatura (que está relacionada com a cor), a composição química, seu tamanho e massa. Pelo espectro de uma estrela, é possível obter detalhes acerca da sua composição e da sua temperatura. A temperatura da superfície da estrela é o principal fator que influencia na sua aparência espectral, e também pode ser medido através da fotometria, onde é medido a magnitude em diferentes filtros (PASACHOFF; FILIPPENKO, 2013).

2.2.2 Galáxias

Galáxia pode ser definida por ser um grande conjunto de matéria estelar e interestelar, isolado no espaço e com as respectivas partes mantidas juntas pelas suas próprias forças gravitacionais. As características morfológicas de galáxias no céu se distinguem fortemente de estrelas, onde possuem bordas difusas, com algumas galáxias apresentando formas alongadas, enquanto estrelas são nítidas e pontuais. Com relação à essas diferenças visuais das galáxias, foi em 1924 que Edwin Hubble categorizou-as, com base na sua aparência. Elas foram distinguidas em quatro tipos fundamentais: espirais, espirais barradas, elípticas e irregulares (CHAISSON; MCMILLAN, 2013).

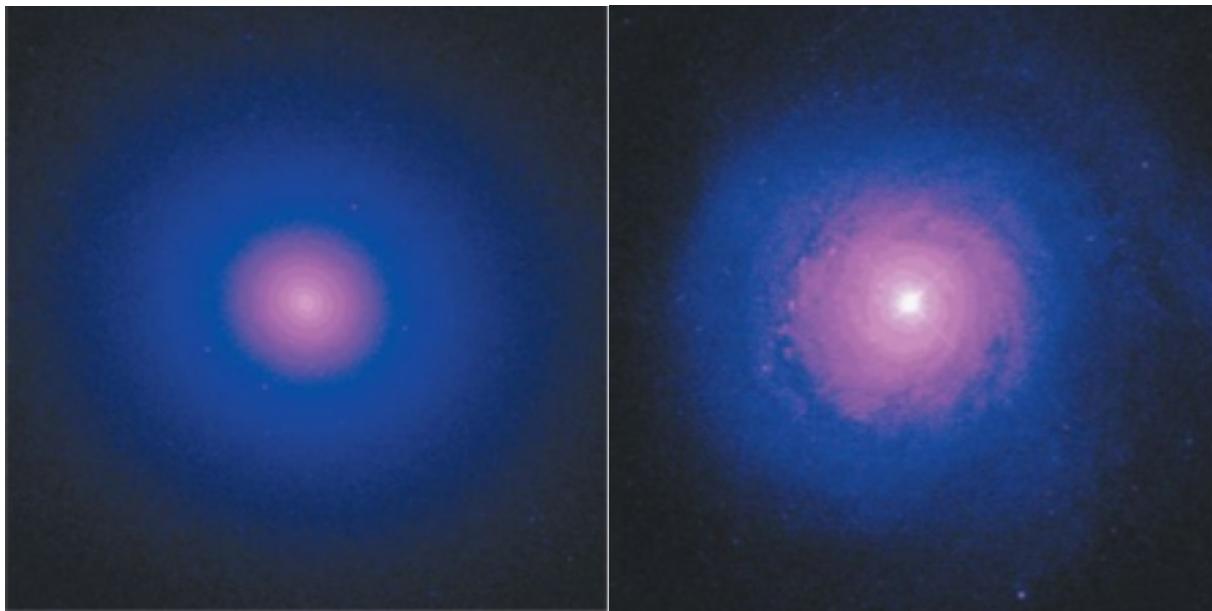
Com relação ao que se pode medir de uma galáxia, o seu brilho e sua cor caracterizam o tipo e a quantidade de estrelas contidas nela, e também podem caracterizar as idades das estrelas contidas nessa mesma galáxia. Obter a informação acerca da luminosidade total de uma galáxia requer integrar o fluxo luminoso de toda galáxia, que é um desafio uma vez que elas apresentam uma borda difusa (FILHO; SARAIVA, 2004).

2.2.2.1 Núcleos Ativos de Galáxias

Conforme visto anteriormente, os AGN tem características distintas em relação à galáxias comuns. Por comparação, nas galáxias usuais geralmente há a presença de um buraco negro no seu centro, consequência do seu processo de formação. Com a presença desse buraco negro, o gás presente na galáxia tende a ir para a região central por conta dos efeitos gravitacionais e com isso resultar na formação de estrelas em seu núcleo, por isso um núcleo brilhante. Já os AGN são um grupo de galáxias que apresentam um brilho anormal em seus núcleos quando comparadas com as do primeiro tipo, pelo fato de o buraco negro ser ativo dado ao influxo de gás neste mesmo buraco negro, e para distingui-las são chamadas de galáxias ativas (PASACHOFF; FILIPPENKO, 2013). Na Figura 3 é possível ver uma galáxia comum e uma galáxia ativa.

Astrônomos observaram várias formas de galáxias que foram consideradas como ativas, cuja a fonte de emissão de energia não são estrelas. É possível citar exemplos de galáxias com núcleo ativo: seyferts, radiogaláxias, quasares, blazares, etc . Os nomes diferentes dados às galáxias ativas levam em consideração a sua morfologia e o tipo de

Figura 3 – Imagens comparativas de uma galáxia comum NGC7626 e uma galáxia ativa NGC5548



(a) Galáxia comum NGC 7626

(b) Galáxia com núcleo ativo NGC 5548

Fonte: Adaptado de (PASACHOFF; FILIPPENKO, 2013).

Imagens do Telescópio Espacial Hubble de galáxias com a mesma morfologia.

radiação emitida (PASACHOFF; FILIPPENKO, 2013).

O que acabou sendo observado quando se tinha em foco quasares, é que ao redor do núcleo super brilhante (o quasar) havia uma estrutura difusa circundante. Este fato levou a conclusão que a soma das partes resultavam na verdade em uma galáxia. Esse mal entendido acontecia por este tipo de galáxia apresentar um núcleo tão brilhante que todo o resto do sistema era ofuscado, e por não ser facilmente perceptível as bordas difusas do restante da galáxia, o quasar era tomado como um objeto pontual (PANNUTI, 2020).

2.2.3 Quasares

O quasar, como vimos anteriormente, é um tipo de específico de AGN muito luminoso, compacto e azulado, e por esse motivo pode ser opticamente confundido com uma estrela. Esses objetos têm um espectro de emissão intenso na frequência do rádio e o seu *redshift* tende a ser grande por apresentar um afastamento da Terra com grandes velocidades (FILHO; SARAIVA, 2004).

É possível observar quasares em quase todo espectro eletromagnético, mas a maior parte destes tem seu espectro de emissão mais intenso no infravermelho. Eles também se encontram à grandes distâncias da Terra (altos *redshifts*) (CHAISSON; MCMILLAN, 2013) e de forma mais espaçada um dos outros, então para detectar-los é necessário que o levantamento tenha as suas bandas de seleção no infravermelho, cubra uma vasta área do

céu e também alcance uma grande profundidade no universo para que se possa chegar à esses elevados *redshifts* (NAKONECZNY *et al.*, 2019).

As características encontradas em quasares diferem dependendo do seu *redshift*. Quando em *redshifts* mais baixos quasares apresentam um aspecto mais azulado do que estrelas típicas, ou seja, tem uma maior intensidade luminosa nos comprimentos de ondas que caracterizam o azul, já em *redshifts* mais altos a sua luz é mais intensa nos comprimentos de ondas que caracterizam o vermelho, resultado do efeito de nuvens intergaláticas que absorvem os comprimentos de ondas mais azuis (PASACHOFF; FILIPPENKO, 2013).

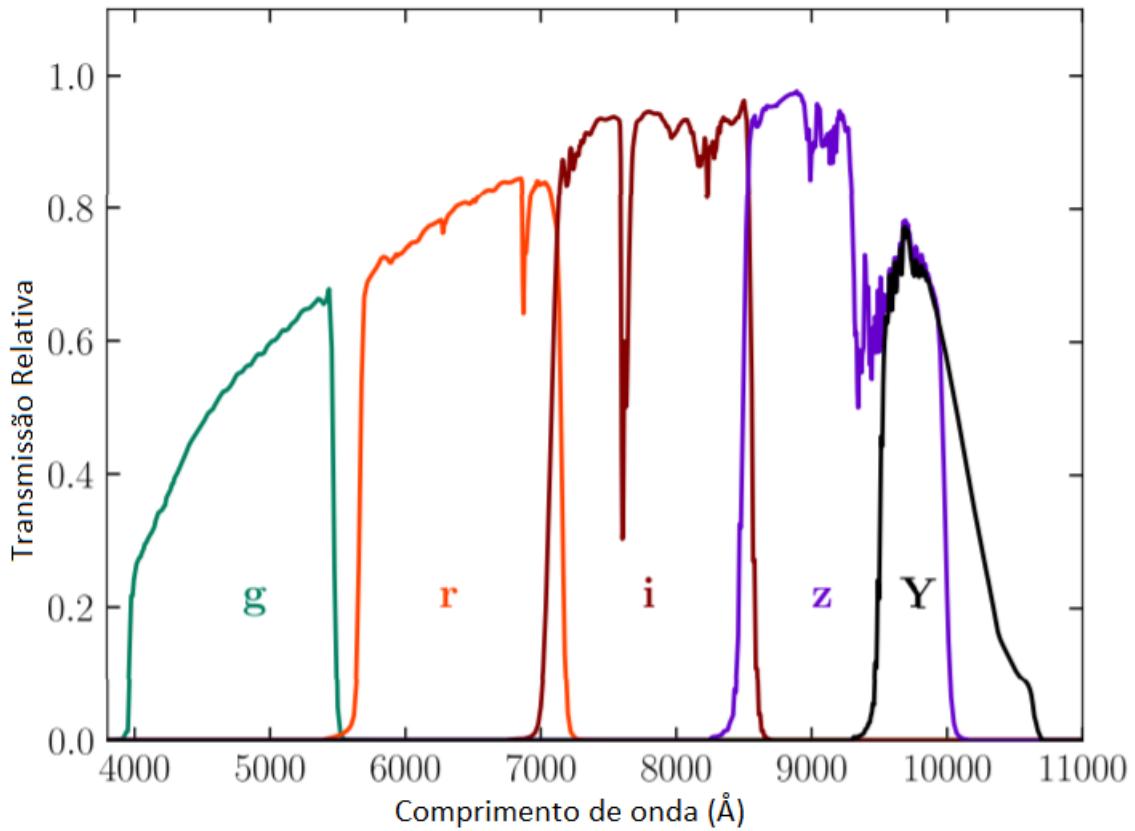
2.3 Grandes Surveys

2.3.1 Dark Energy Survey (DES)

O DES é um levantamento planejado para ter uma área ampla de cobertura, de aproximadamente 5000 graus², mapeando a parte sul da esfera celeste. Ele foi feito para capturar informações do espectro visível ao infravermelho próximo, e utiliza filtros de banda larga escolhidos para melhorar a precisão do redshift fotométrico, precisão essa essencial para atingir os objetivos do projeto de estudar o parâmetro w da equação de estado da energia escura (SÁNCHEZ, 2010).

O DES é composto por dois levantamentos diferentes, sendo o primeiro levantamento chamado de DES *Wide-area* com 5000graus² observados utilizando bandas de detecção g, r, i, z e Y, e o segundo levantamento tendo uma área de observação aproximadamente 27graus² e com bandas de detecção g, r, i e z com o objetivo de caracterizar as curvas de luz de supernovas. No presente trabalho foi utilizado o DES *Wide-area*, onde os limites das bandas de detecção são ilustrados na Figura 4 e os limites nominais de magnitude 10σ são dados para cada banda passante: g = 23.57, r = 23.34, i = 22.78, z = 22.10 e Y = 20.69 (MORGANSON *et al.*, 2018).

Figura 4 – Bandas de detecção do DES *Wide-area*



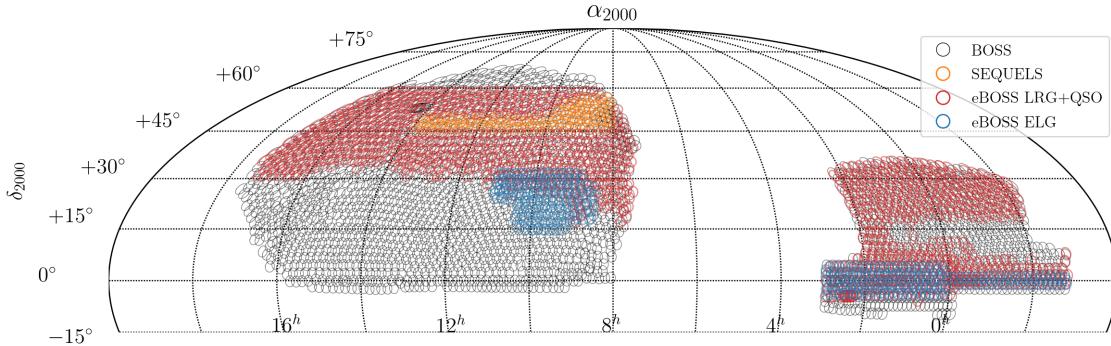
Fonte: Livre tradução de (MORGANSON *et al.*, 2018)

Até o momento, houveram dois lançamentos dos dados processados ao longo da execução do DES. O primeiro lançamento, DES *Data Release 1* (DES DR1), contemplou aquisições feitas de agosto de 2013 à janeiro de 2016. O segundo lançamento, DES DR2 contempla todos os 6 anos de aquisição feitos no DES, incluindo os resultados presentes no DR1, porém reprocessados (MORGANSON *et al.*, 2018).

2.3.2 Sloan Digital Sky Surveys(SSDS)

O SDSS é um levantamento espectroscópico e fotométrico, que inicialmente contava com um telescópio localizado em *Apache Point Observatory* (APO), e desde 2017 conta também com observações desde *Las Campanas Observatory* (LCO). Desde 1998 o SSDS está em operação fazendo registros do céu, onde atualmente se encontra na sua quarta fase e com 16 lançamentos de dados, onde o primeiro aconteceu em 2001, e o mais atual lançamento em 2020. Todos os lançamentos de dados do SSDS são acumulativos, significando que o mais recente lançamento irá conter os dados inéditos mais os dados de todos os últimos lançamentos. Os dados dos *releases* anteriores são reprocessados seguindo *pipelines* mais atualizadas e análises melhoradas, sendo sugerido pelos autores que o acesso seja feito pelo *release* mais atual (AHUMADA *et al.*, 2020).

Figura 5 – Área do céu coberta pelos sub-levantamentos espectroscópicos disponíveis no DR16.



Fonte: (LYKE *et al.*, 2020)

Os sub-levantamentos compreendem: BOSS, SEQUELS, eBOSS LRG-QSO e eBOSS ELG. Mapa centrado em RA = 8h

2.4 Metodologia

Nessa secção apresenta-se as principais ferramentas computacionais utilizadas para processar os dados, seguido das características dos conjuntos de dados escolhidos do DES DR2 e do SDSS DR16 e, por fim, dos critérios de pré processamento, amostragem e análise dos dados.

2.4.1 Hierarchical Equal Area and isoLatitudde Pixelization (HEALPix)

No contexto de objetos encontrados no céu, uma das maneiras de identificar os objetos distintos é utilizando suas coordenadas, que devem ser adequadas para a sua representação na esfera celeste. Outra necessidade que surge também ao pensar nestes objetos é poder dividir ela em pedaços a para entender as características de cada pedaço. Nesse trabalho, utilizou-se da ferramenta HEALPix (GORSKI *et al.*, 1999) que realiza uma repartição da esfera em quadriláteros com áreas iguais, que podem ter formas variadas. Um dos parâmetros da pixelização da esfera é a sua resolução (N_{side}), que através do cálculo $N_{pixels} = 12 * N_{side}^2$ se obtém a quantidade de pixels em esfera está sendo dividida (GORSKI *et al.*, 1999).

A numeração dos pixels seguem uma ordem pré-definida que pode ser um esquema chamado *NESTED* ou um esquema chamado *RING*. No esquema *NESTED* os pixels são ordenados tendo em conta a ordem tomadas pelos pixels formados com a menor resolução. No esquema *RING* os pixels são enumerados do polo norte da esfera até o polo sul e com ordem crescente passando por cada anel formado por pixels que estão em uma mesma latitude (GORSKI *et al.*, 1999).

Ao longo do processamento dos dados neste trabalho, utilizou-se a biblioteca python

Healpy², que apresenta funções adequadas para manipular *pixels* em uma esfera. Em todos os *plots* e no processamento dos dados foi adotado como padrão um valor para a resolução de $N_{side} = 64$, com a ordenação dos *pixels* em um esquema do tipo RING. A escolha dessa resolução é justificada porque nela se tem uma distribuição média de objetos por *pixel* razoável. Ao usar uma resolução menor haveria a perda de detalhes na distribuição e ao usar uma resolução maior haveria o efeito de se ter mais *pixels* contendo nenhum objeto.

2.4.2 *Match* entre *surveys* diferentes

O *Match* é o nome dado, nesse trabalho, para o processo de achar um mesmo objeto em levantamentos astronômicos diferentes. Foi empregado como parâmetro para o *match* as coordenadas dadas por: ascensão reta α e a declinação δ definidos no sistema equatorial de coordenadas. Utilizou-se a Biblioteca python Esutil³, esta tem uma ferramenta que busca correspondências em dois *data sets* diferentes que se baseia no *Hierarchical Triangular Mesh* (HTM), que é um método de divisão da superfície esférica em triângulos iguais pensado para indexar objetos nesses triângulos e facilitar a sua busca. Para definir dois objetos como iguais pela sua posição no céu, foi considerado um erro de 1 segundo de arco, equivalente à 0.0002778 graus, e com a restrição de ter como resultado no máximo 1 objeto (KUNSZT; SZALAY; THAKAR, 2001). A justificativa para o uso de 1 segundo de arco é o valor *point spread function* (PSF) médio das bandas de detecção do DES, que varia de 0,83 a 1,1 segundos de arco (ABBOTT *et al.*, 2021). Esse valor revela o quanto o ponto registrado na imagem sofreu espalhamento, que pode ser admitido como a incerteza na localização do objeto, então escolher para o erro do *match* um valor menor que 1 segundo de arco pode gerar muitos resultados com objetos sem correspondências, e escolher um valor maior pode gerar falsos *matchs*.

2.4.3 t-SNE

Como foi escrito na Introdução, o t-SNE é o algorítimo escolhido para reduzir a dimensionalidade dos objetos fotométricos do DES. Essa alta dimensionalidade desses objetos é dada pelos atributos previamente selecionados (que vão ser detalhados na seção 2.4.5) de cada um deles. Justamente, a ferramenta t-SNE é aplicada em um conjunto de pontos que tem multidimensionalidade com o objetivo de gerar uma redução do espaço dimensional deste conjunto conservando a representatividade deste em um espaço de uma, duas ou três dimensões. É uma técnica que se molda ao conjunto de dados e não converge em um único valor, o que significa que a cada vez em que é executada o conjunto resultante tem um valor distinto (WATTENBERG; VIÉGAS; JOHNSON, 2016).

Para entender a execução do t-SNE, começamos com a parte da sigla SNE que significa *Stochastic Neighbor Embedding*. O SNE primeiro calcula as distâncias euclidianas

² <https://healpy.readthedocs.io/en/latest/>

³ <https://github.com/esheldon/esutil>

entre os pontos em alta dimensão e então converte essas distâncias em similaridade. A similaridade é uma probabilidade condicional de que um ponto x_j seja escolhido como vizinho do ponto x_i . Essa escolha é feita considerando uma distribuição normal da densidade de probabilidade centrada no ponto x_i e com variância igual à σ_i . A similaridade é matematicamente descrita pela Equação 2.3, o termo x_i é um ponto pertencente ao conjunto de dados no espaço original e x_j é um segundo ponto nesse mesmo conjunto de dados (MAATEN; HINTON, 2008).

$$p_{i|j} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (2.3)$$

A perplexidade é um parâmetro escolhido pelo usuário que é fixo, representado na Equação 2.4, onde $H(p_i)$ é a Entropia de Shannon (MAATEN; HINTON, 2008).

$$Perp(P_i) = 2^{H(P_i)} \quad (2.4)$$

A Entropia de Shannon relacionada à um ponto x_i , representada na Equação 2.5, é dada pela soma da similaridade (Equação 2.3) multiplicada pelo logaritmo na base 2 dessa similaridade para cada ponto x_j . O que acontece é que o valor da variância (σ_i) da Gaussiana, é escolhida em função da perplexidade para cada ponto x_i do espaço original. O fato de que cada ponto tenha um valor diferente de variância é apropriado pois regiões diferentes em um conjunto de dados tem densidades de pontos diferentes, e quando a densidade é grande um valor pequeno de variância é apropriado, e quando a densidade é pequena um valor grande de variância é apropriado. É realizada então uma busca binária que procura valores para a variância que resultem na perplexidade definida pelo usuário. A perplexidade pode ser interpretada como um valor diretamente relacionado à dispersão ao redor dos pontos (MAATEN; HINTON, 2008).

$$H(P_i) = - \sum_j p_{i|j} \log_2(p_{i|j}) \quad (2.5)$$

A similaridade citada acima está implícita no termo chamado de Divergência de Kullback-Leibler, representado como $\frac{\delta C}{\delta \gamma}$ na Equação 2.8. Tendo em vista um conjunto de pontos qualquer, pertencente à um espaço M , como representado na Equação 2.6 (MAATEN; HINTON, 2008).

$$X = \{x_1, x_2, \dots, x_n\}, x_i \in R^M \quad (2.6)$$

Esse conjunto de pontos, quando submetido ao algoritmo t-SNE passa por uma transformação no seu espaço, e passa a estar em um espaço R^1, R^2 ou R^3 , como representado na Equação 2.7 (MAATEN; HINTON, 2008).

$$\gamma^{(T)} = \{y_1, y_2, \dots, y_n\}, y_i \in R^{1,2,3} \quad (2.7)$$

Na Equação 2.7, T é um termo chamado de número de iterações, e é um hiper parâmetro que influênciaria muito no resultado da transformação. O termo $\gamma^{(t)}$ pertencente à Equação 2.8, é recalculado para cada valor de t , sendo que t varia de 1 até T (MAATEN; HINTON, 2008).

$$\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta \gamma} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)}) \quad (2.8)$$

Dentro de todos os parâmetros que podem ser utilizados, a perplexidade, o número de iterações e a taxa de aprendizagem são os que influenciam criticamente o resultado final do conjunto. Não existem valores pre-definidos como ótimos para cada um desses parâmetros à serem adotados. Os valores apropriados variam de acordo com o resultado pretendido, que também sofre influências das características do conjunto de dados. Nesse trabalho foram adotados os mesmos valores de perplexidade: 2, 5, 30, 50 e 100, usado nas análises presentes no trabalho (WATTENBERG; VIÉGAS; JOHNSON, 2016). Com relação ao valor do número de iterações, para testar o comportamento dos diferentes valores de perplexidade empregou-se o valor de 5000, esse que no trabalho (WATTENBERG; VIÉGAS; JOHNSON, 2016) alcançou estabilidade na distribuição dos pontos. Para testar essa estabilidade na distribuição do *data set* usado neste trabalho, foram feitas cinco reduções de dimensionalidade utilizando perplexidade 5, valor mínimo sugerido por (MAATEN; HINTON, 2008), e adorando como número de iterações, os valores: 1000, 2000, 3000, 4000 e 5000.

Para evitar que o algorítimo tomasse com uma importância maior atributos que tinham maiores grandezas do que outros, aplicou-se uma normalização no *data set* para que todos os atributos tivessem seus valores em um intervalo de 0 a 1. A Equação 2.9 descreve a normalização, x_o é o valor original do atributo do ponto, x_{min} é o valor mínimo desse atributo em todo o *data set*, x_{max} é o valor máximo desse atributo em todo o *data set* e x_{ef} é o atributo normalizado.

$$x_{ef} = \frac{x_o - x_{min}}{x_{max} - x_{min}} \quad (2.9)$$

Utilizou-se a ferramenta t-SNE disponibilizada pela biblioteca *Scikit Learn*⁴ (PEDREGOSA *et al.*, 2011).

⁴ <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

2.4.4 K Dimensional Tree

Levando em conta o problema envolvido em achar qual é o ponto j mais próximo de um ponto i , em um espaço dimensional N , foi utilizado nesse trabalho o algorítimo de busca chamado K Dimensional Tree (KD Tree), contido na biblioteca Python SciPy⁵. Esse algorítimo baseia-se em árvores de busca binária em dimensões superiores, onde os nós dessas árvores estão associados a um hiper retângulo associado a dos eixos do espaço N . Então cada nó divide os pontos sobre um eixo, e a divisão é feita usando regra do ponto médio deslizante (MANEEWONGVATANA; MOUNT, 1999).

2.4.5 Dados pertencentes ao DES

O acesso aos dados públicos do DES DR2 são disponibilizados na plataforma DESaccess⁶. Os dados foram retirados utilizando uma *query* em linguagem *Structured Query Language* (SQL), onde a tabela de origem dos dados é a **DR2_MAIN_SAMPLE**. Aplicou-se restrições à coleta da amostra para fazer uma seleção por estrelas, essas restrições são sugeridas pela próprio DESaccess, e para essa seleção, são elas:

1. O parâmetro IMAFLAGS_ISO_I = 0, onde esse parâmetro igual a 0 sinaliza que o objeto não apresentou nenhum problema de interferência ou relacionado à sua imagem no momento de detecção;
2. A magnitude na banda I ser menor que 21 ($MAG_AUTO_I < 21$), pois como o levantamento do DES possui profundidade de 23.8 na banda I, esse filtro retira valores próximos do seu limite de detecção que podem conter erros maiores;
3. WAVG_SPREAD_MODEL_I ser maior que -1. O atributo WAVG_SPREAD_MODEL_I é usado para classificação morfológica dos objetos, principalmente entre estrelas e galáxias, na banda de detecção i ;
4. A soma do atributo WAVG_SPREAD_MODEL_I com a multiplicação por 3 do atributo WAVG_SPREADERR_MODEL_I tendo que ser menor que 0,005. O atributo WAVG_SPREADERR_MODEL_I é a incerteza do valor contido no WAVG_SPREAD_MODEL_I.

Visando aplicar o método t-SNE para redução da dimensionalidade dos objetos retirados do DES DR2, selecionou-se 14 atributos, descritos no Quadro 1. O quadro foi elaborado com base na descrição dos atributos contida no artigo de apresentação do DES DR2 (ABBOTT *et al.*, 2021). O raio de Kron citado em um dos atributos é um parâmetro que se usa para definir a abertura circular que captura o fluxo do objeto. A abertura utilizada

⁵ <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html>

⁶ <https://des.ncsa.illinois.edu/desaccess/>

é determinada no primeiro momento de detecção do objeto , que se calcula com base na sua distribuição de luz multiplicado por um fator (HILL *et al.*, 2011). Para a medida MAG_AUTO, o software utilizado (SExtractor (Bertin, E.; Arnouts, S., 1996)), analisa a magnitude de Kron por objeto, essa magnitude é dada pelo fluxo medido do objeto dentro do raio Kron, com um fator de 2,5 (ZHANG *et al.*, 2019). A magnitude definida por MAG_AUTO não passa por um sistema de desavermelhamento, diferentemente da magnitude MAG_AUTO_G,R,I,Z,Y_DERED, e a escolha pela primeira é justificada porque a amostra utilizada não se localiza no equador galático e sim no equador celeste, então qualquer correição de magnitude seria pequena (ABBOTT *et al.*, 2021).

Quadro 1 – Descrição dos tributos utilizados dos objetos procedentes do DES.

Atributo	Descrição
CLASS_STAR_- G,R,I,Z,Y	É um atributo gerado pelo DES que apresenta uma classificação morfológica para o objeto em questão nas respectivas bandas de detecção g,r,i,z e Y, variando entre 0 e 1. O significado desse atributo é que morfologicamente o 0 é uma galáxia e 1 que é uma estrela.
MAG_AUTO_- G,R,I,Z,Y	É um atributo gerado pelo DES que apresenta uma estimativa para a magnitude nas respectivas bandas de detecção g,r,i,z e Y, tendo como base o raio de Kron para um modelo elíptico.
COLOR_g-r	É um atributo gerado localmente fazendo a diferença entre a magnitude na banda G(MAG_AUTO_G) e a magnitude na banda R(MAG_AUTO_R).
COLOR_r-i	É um atributo gerado localmente fazendo a diferença entre a magnitude na banda r(MAG_AUTO_R) e a magnitude na banda i(MAG_AUTO_I).
COLOR_i-z	É um atributo gerado localmente fazendo a diferença entre a magnitude na banda i(MAG_AUTO_I) e a magnitude na banda z(MAG_AUTO_Z).
COLOR_z-Y	É um atributo gerado localmente fazendo a diferença entre a magnitude na banda z(MAG_AUTO_Z) e a magnitude na banda Y(MAG_AUTO_Y).

Fonte: Elaborado pela autora

2.4.6 Dados pertencentes ao SDSS

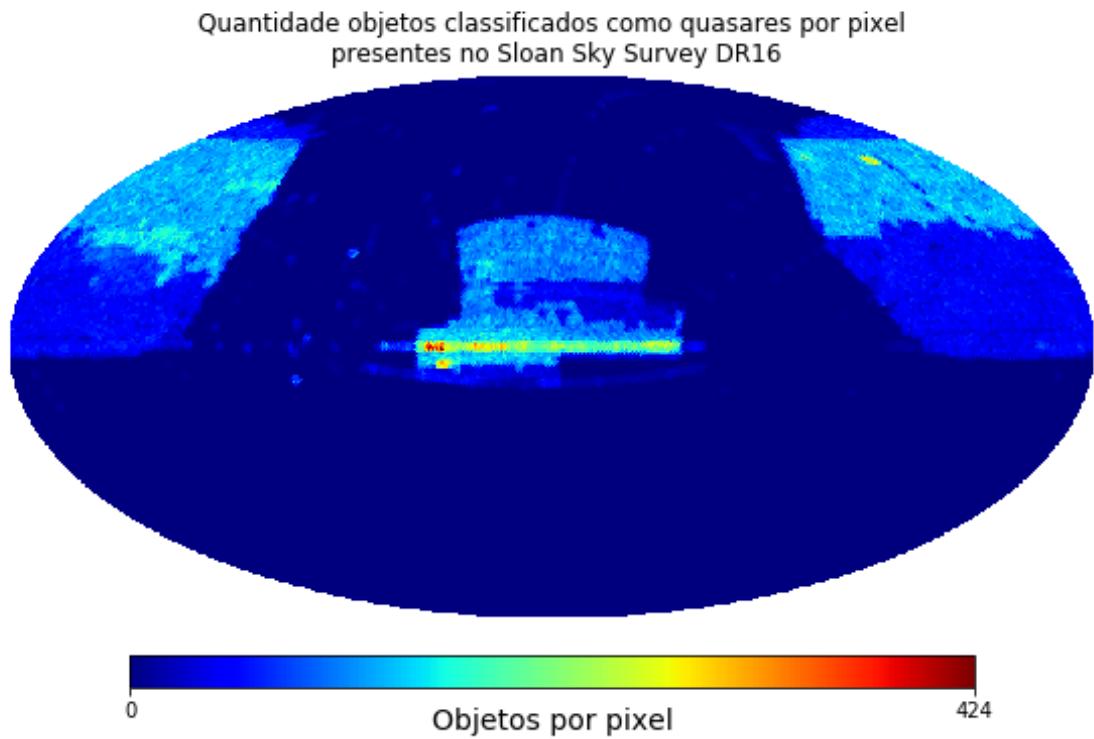
Neste trabalho, utilizou-se dos objetos obtidos por espectroscopia óptica pertencentes ao SDSS *data release 16*(SDSS DR16). Para cada objeto e seu espectro contido nesse levantamento, é disponibilizado um *redshift* e uma classificação espectral (QSO, GALAXY, STAR) alcançada de maneira automatizada, com seleção de amostras para revisão visual do seu espectro (Bolton *et al.*, 2012). Essa classificação de cada objeto foi empregada nesse trabalho na identificação dos objetos obtidos por fotometria provenientes do DES.

Pela plataforma Sky Server ⁷ os objetos detectados por espectroscopia óptica foram retirados da visão **SpecObj** da tabela **specObjAll**, que exclui dados considerados como ruins e duplicados. Na Figura 6 é apresentada a densidade por pixel de todos os quasares obtidos via Sky Server. A projeção utilizada é chamada Mollweide que dentre suas características as principais podem ser descritas como: é uma projeção pseudo-cilíndrica, o tamanho das áreas permanece igual, as distorções aumentam a medida que aumenta

⁷ <http://skyserver.sdss.org/dr16/en/home.aspx>

a distância do ponto central da projeção, e é um tipo ideal para quando se necessita representar o mapa-múndi (USERY, 2017).

Figura 6 – Densidade de quasares por *pixel* presentes no SDSS DR16



Fonte: A autora

O mapa mostra a densidade de quasares presentes nos levantamentos espectroscópicos do SDSS DR16 por *pixel*. Mapa gerado com projeção de Mollweide, com *pixels* tendo resolução de $N_{side} = 64$, esquema de ordenamento do tipo *RING*

2.4.7 Pré-processamento e o grupo amostral

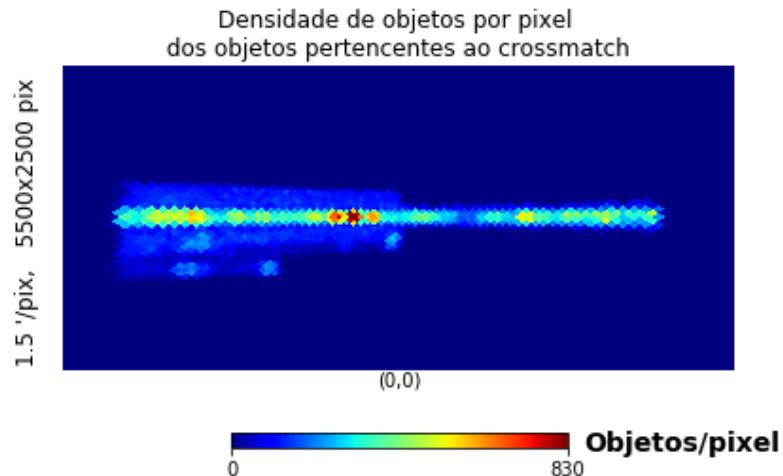
Tendo em foco os dados provenientes do DES DR2 (obtidos de acordo com a descrição na secção 2.4.5), primeiramente eles foram distribuídos por *pixels* e o número de identificação do pixel na esfera, correspondente à cada objeto, foi designado à cada um como atributo. O passo seguinte foi realizar o *match* (técnica descrita na secção 2.4.2) entre os dados do DES e todos os dados espectroscópicos do SDSS DR16. Nesse passo, foram obtidos 177670 objetos que estavam presentes no *data set* do DES e também no *data set* do SDSS, com a sua distribuição por *pixel* representada na Figura 7a. Esta distribuição está ilustrada em uma projeção gnomônica que é azimutal e que tem como ponto de perspectiva o centro da terra (USERY, 2017).

Separou-se o *data set* com base na classificação dos objetos, resultando em três conjuntos distintos de dados: um com quasares, um com galáxias e um com estrelas. Em seguida, os objetos foram agrupados e somados por *pixel*, etapa realizada para cada um dos

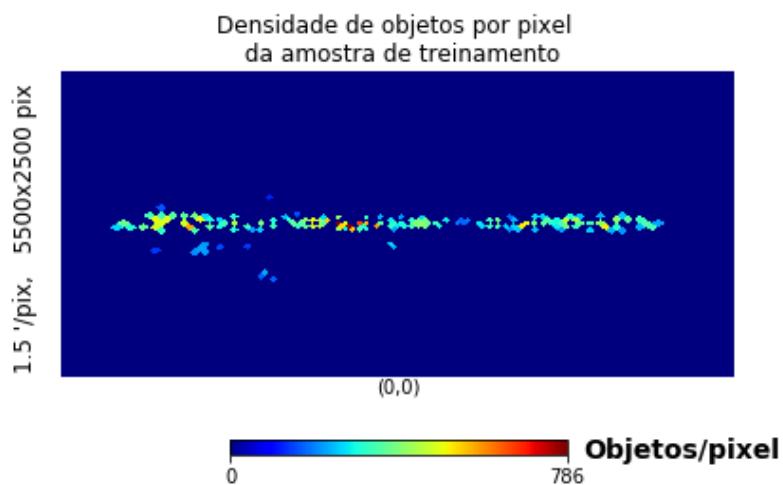
três sub conjuntos. Foram selecionados 20% dos *pixels* que apresentavam maior densidade de objetos, também para cada um dos três sub conjuntos. Após essa seleção, os *pixels* contidos em cada subconjunto foram concatenados, e desse conjunto de *pixels* foi escolhido aleatoriamente 50% dos *pixels* como sendo os *pixels* que continham os objetos para o treinamento e o 50% restante dos pixels que continham os objetos da amostra de teste. Esse seleção foi realizada para garantir homogeneidade na amostragem, ou seja, que fossem selecionados os pixels que mais continham objetos de todos os três tipos.

Na Figura 7b são ilustrados os *pixels* e as respectivas densidades da amostra de treinamento, e na Figura 7c os *pixels* e as respectivas densidades da amostra de teste.

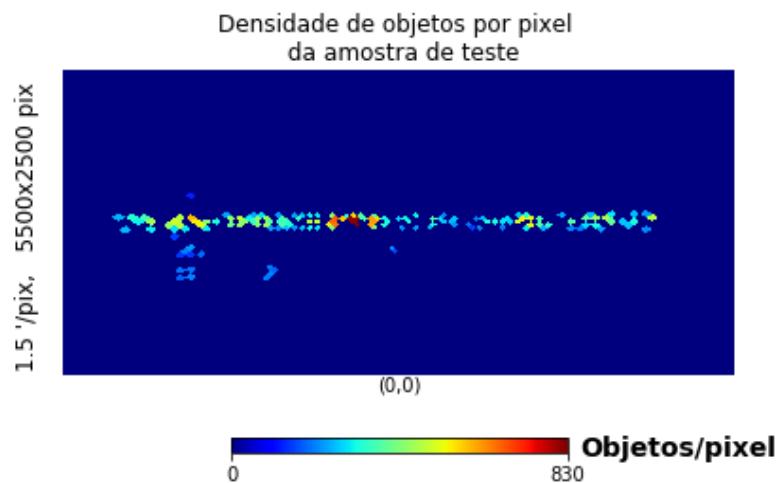
Figura 7 – Os mapas 7a, 7b e 7c foram feitos utilizando uma projeção gnomônica, com pixels tendo resolução de $N_{side} = 64$, esquema de ordenamento do tipo *RING*.



(a) Densidade de objetos por *pixel* do *match* entre SDSS DR16 e DES DR2



(b) Densidade de objetos por *pixel* da amostra de treino



(c) Densidade de objetos por *pixel* da amostra de teste

Essa seleção que resultou nas amostras de treinamento, gerou dois grupos de dados com uma distribuição muito semelhante entre si, como esperado, com números absolutos de quasares, galáxias e estrelas detalhados na Tabela 2.

Tabela 2 – Composição das amostras de treinamento e teste

Classificação dos objetos	número de objetos - Treinamento	número de objetos - Teste
Quasares	18.629	18.251
Galáxias	1.684	1.516
Estrelas	30.610	33.073
Total	50.923	52.840

Fonte: A autora

2.4.8 Análise

Com o objetivo de selecionar a melhor redução espacial feita variando hiper parâmetros do t-SNE, estabeleceu-se dois requisitos. O primeiro requisito consiste em que o corte a ser feito para separar quasares dos demais objetos se trata de um corte único no eixo x do plano cartesiano, que separa a amostra inicial em duas, e a depender da amostra, considera-se o lado esquerdo ou direito do valor de x. O segundo requisito é que o conjunto de dados após esse corte contenha em sua maioria quasares em detrimento dos outros objetos contidos no corte. Desse modo, calculou-se para as 9 distribuições de pontos obtidas com o t-SNE o percentual de quasares em relação ao total de objetos contidos na amostra após o corte, e esse percentual foi calculado para cada ponto x, variando-o de 1 em 1. Foi possível obter uma boa seleção de quasares, com um corte contendo 70,19% do total de quasares da amostra.

Todas as 9 reduções dimensionais realizadas variando os hiper parâmetros de perplexidade e numero de iterações foram feitas a partir da amostra de treinamento.

Após selecionar a distribuição da amostra de treinamento que melhor se podia separar quasares do restante dos objetos, estendeu-se essa distribuição à amostra de teste. Com a impossibilidade de adicionar os objetos pertencentes à amostra de teste na redução dimensional feita com o t-SNE, adotou-se o método KD Tree (técnica descrita na secção 2.4.4) para estender essa distribuição para a amostra de teste. Considerando apenas os atributos utilizados na análise t-SNE da amostra de treino, para cada objeto da amostra de teste, selecionou-se os dois vizinhos mais próximos contidos no conjunto de treinamento, e então tirou-se a média das duas coordenadas x e das duas coordenadas y, e este valor foi adotado como o ponto que representa a redução dimensional deste objeto no plano cartesiano. Foram selecionados dois vizinhos mais próximos pois este é o mínimo de pontos necessários para realizar uma média, e que a cada vizinho mais próximo a mais que se

selecionasse, aumentaria o erro pela possibilidade de pegar objetos com diferenças ainda maiores em seus atributos. Os atributos da amostra de teste foram normalizados assim como os da amostra de treino, antes da aplicação do *KD Tree*.

3 RESULTADOS E DISCUSSÃO

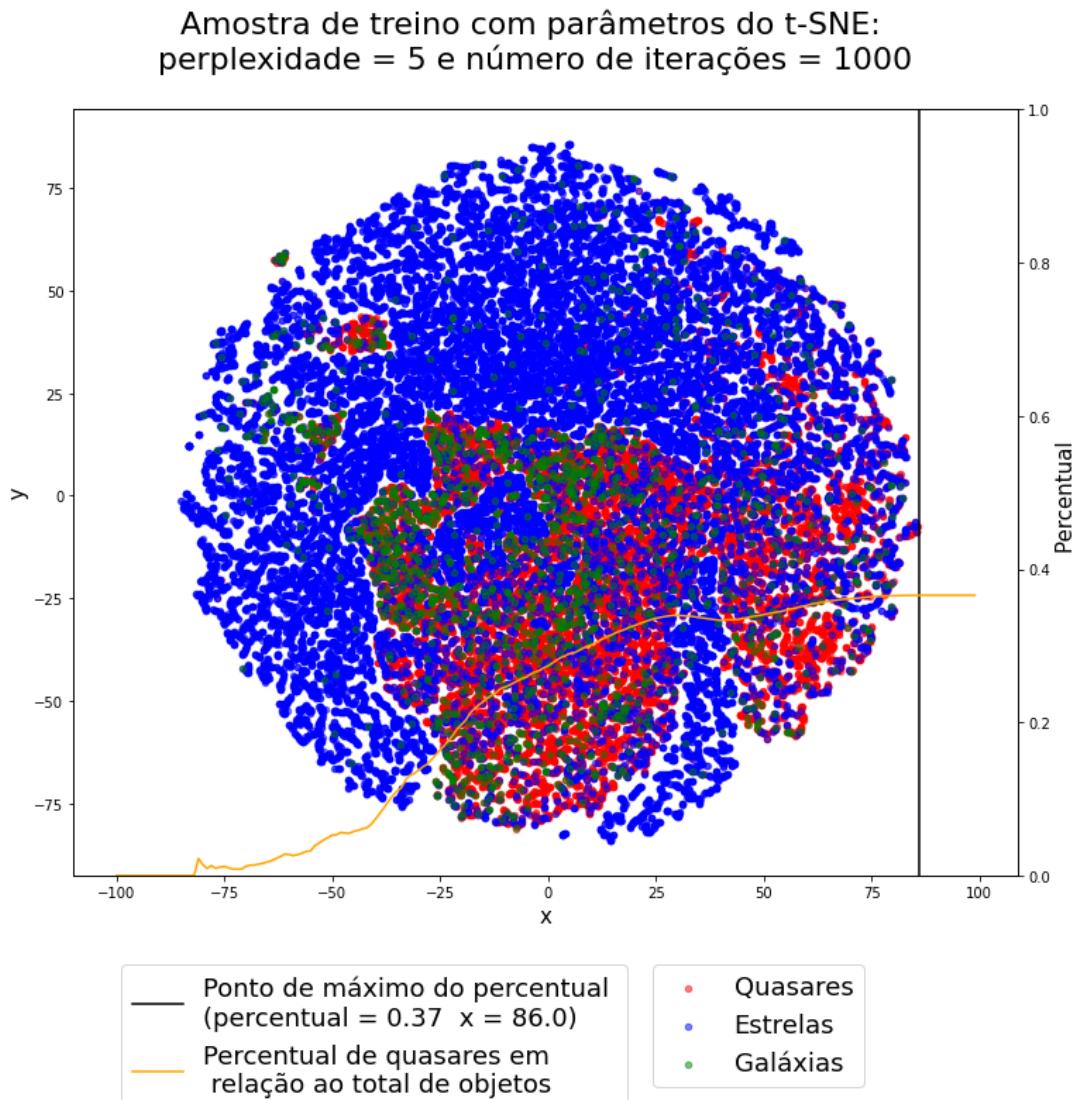
3.1 Resultados

Os resultados estão apresentados em duas partes, onde a primeira se consiste na apresentação das distribuições obtidas através da redução de dimensionalidade por meio do t-SNE. Na segunda parte é apresentado os resultados obtidos aplicando o algorítimo KD Tree na amostra de teste, tendo como base a melhor distribuição obtida via t-SNE utilizando a amostra de treino.

3.1.1 Análise t-SNE variando o número de iterações

A primeira redução dimensional da amostra de treinamento, com perplexidade de 5 e com um número de iterações de 1000, gerou uma distribuição circular com uma aparente formação de dois *clusters* principais, um composto por estrelas e outro composto por quasares e galáxias. Essa distribuição se mostrou ineficaz para separar quasares dos demais objetos, pois o percentual máximo de um corte feito no eixo x foi de 37%. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 8.

Figura 8 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 1000.



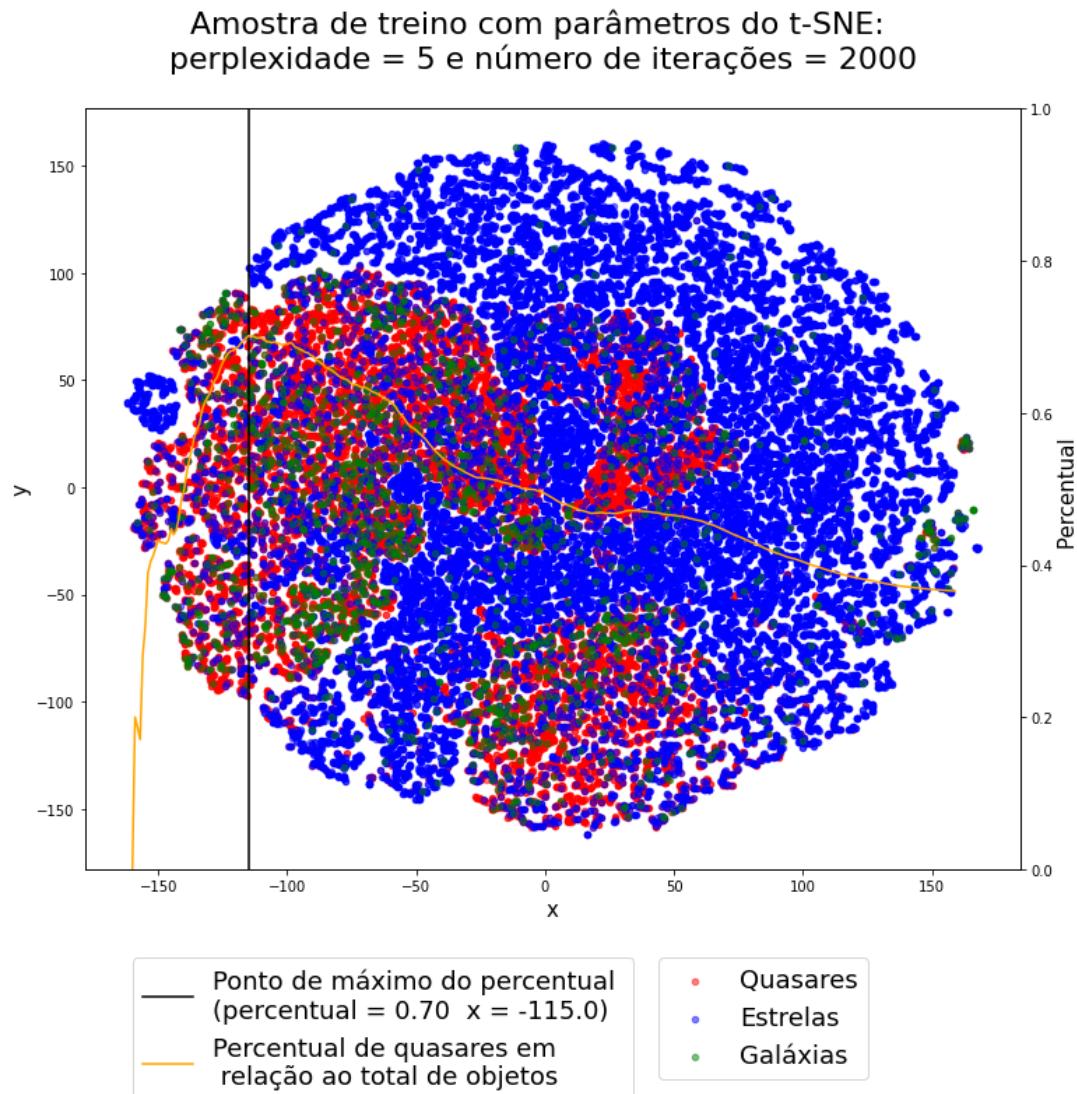
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 5 e 1000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à esquerda do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 5 e com um número de iterações de 2000, apresentou o mesmo padrão circular que a figura com número de iterações igual a 1000, mas com a formação de três *clusters* de quasares em regiões distintas. Essa distribuição também se mostrou ineficaz para separar quasares dos demais objetos, pois o percentual máximo de um corte feito no eixo x foi de 70%, que apesar de ser um percentual alto, o número de objetos contidos na amostra foi pequeno.

Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 9.

Figura 9 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 2000.



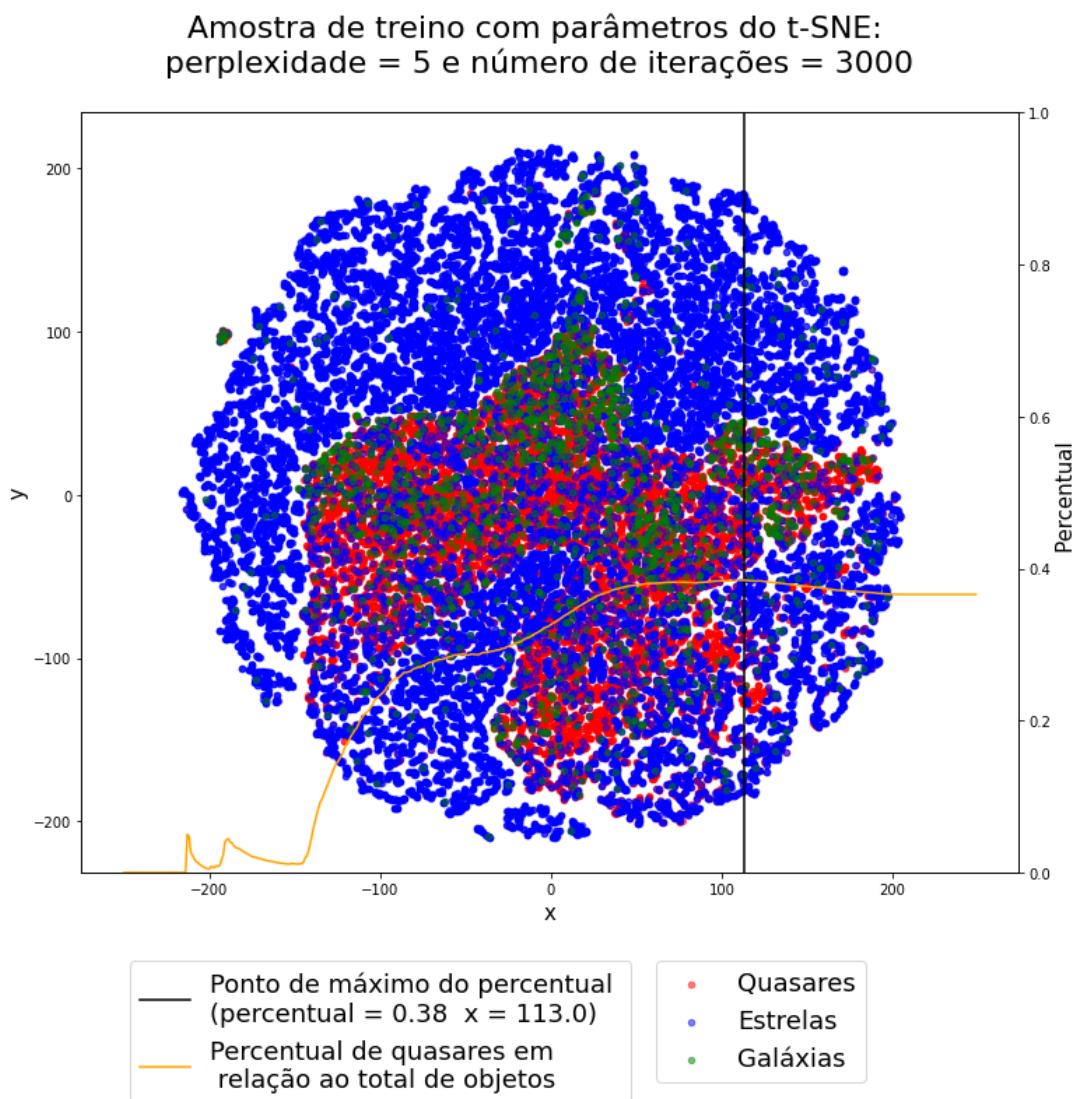
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 5 e 2000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à esquerda do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 5 e com um número de iterações de 3000, um padrão de distribuição circular parecido com que a figura com número de iterações igual a 1000, mas com a formação de um *cluster* de quasares. Essa distribuição também se mostrou ineficaz para separar quasares dos demais

objetos, pois o percentual máximo de um corte feito no eixo x foi de 38%, resultado da localização do *cluster* de quasares, que se localiza na região central da distribuição. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 10.

Figura 10 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 3000.



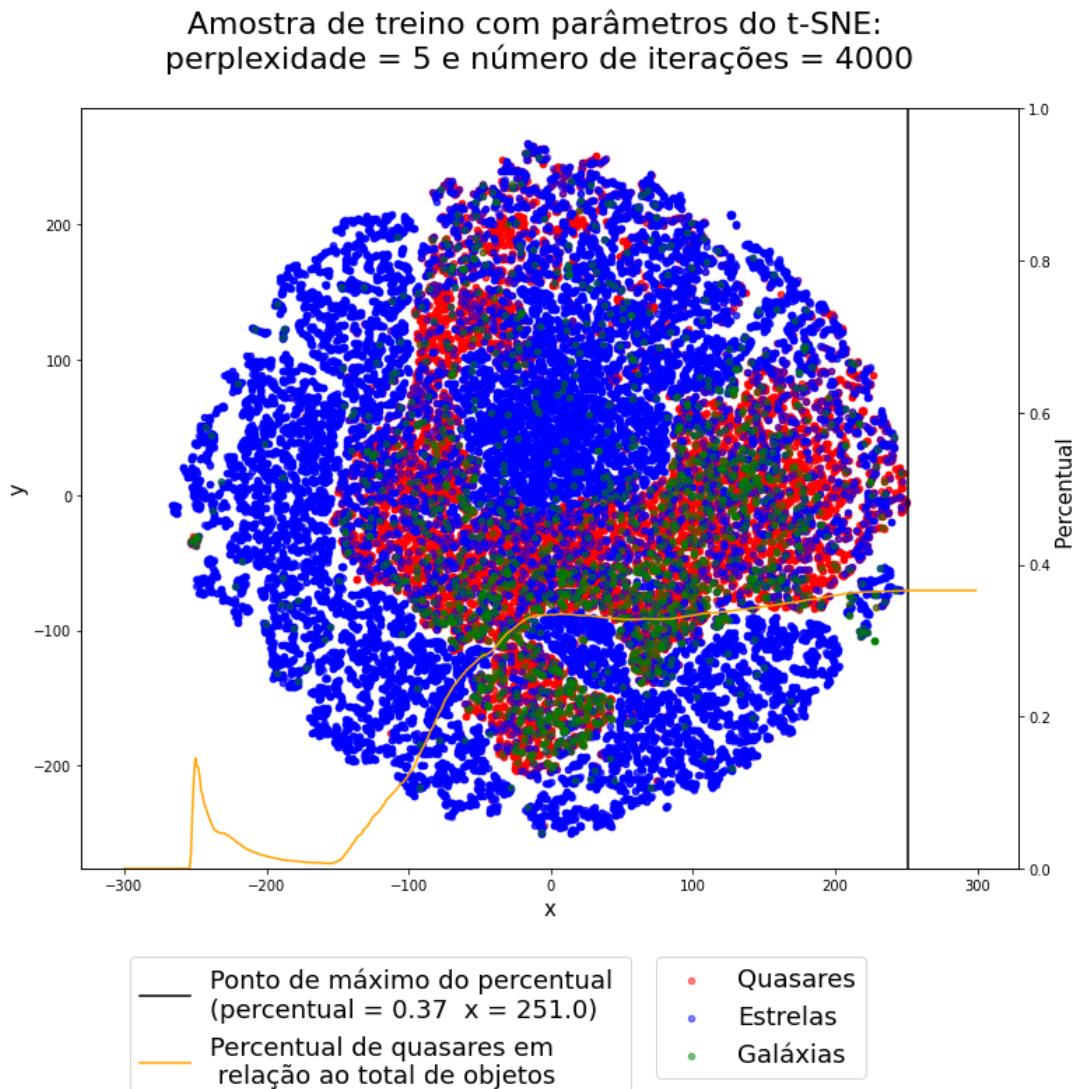
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 5 e 3000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à esquerda do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 5 e com um número de iterações de 4000, apresentou o mesmo padrão circular ilustrado nas últimas

três Figuras 8,9 e 10, mas com a formação de um *cluster* de quasares localizado no centro da distribuição e alongado de forma muito semelhante à Figura 10 . Essa distribuição também se mostrou ineficaz para separar quasares dos demais objetos, pois o percentual máximo de um corte feito no eixo x foi de 37%, resultado da localização do *cluster* de quasares, que se localiza na região central da distribuição. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 11.

Figura 11 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 4000.



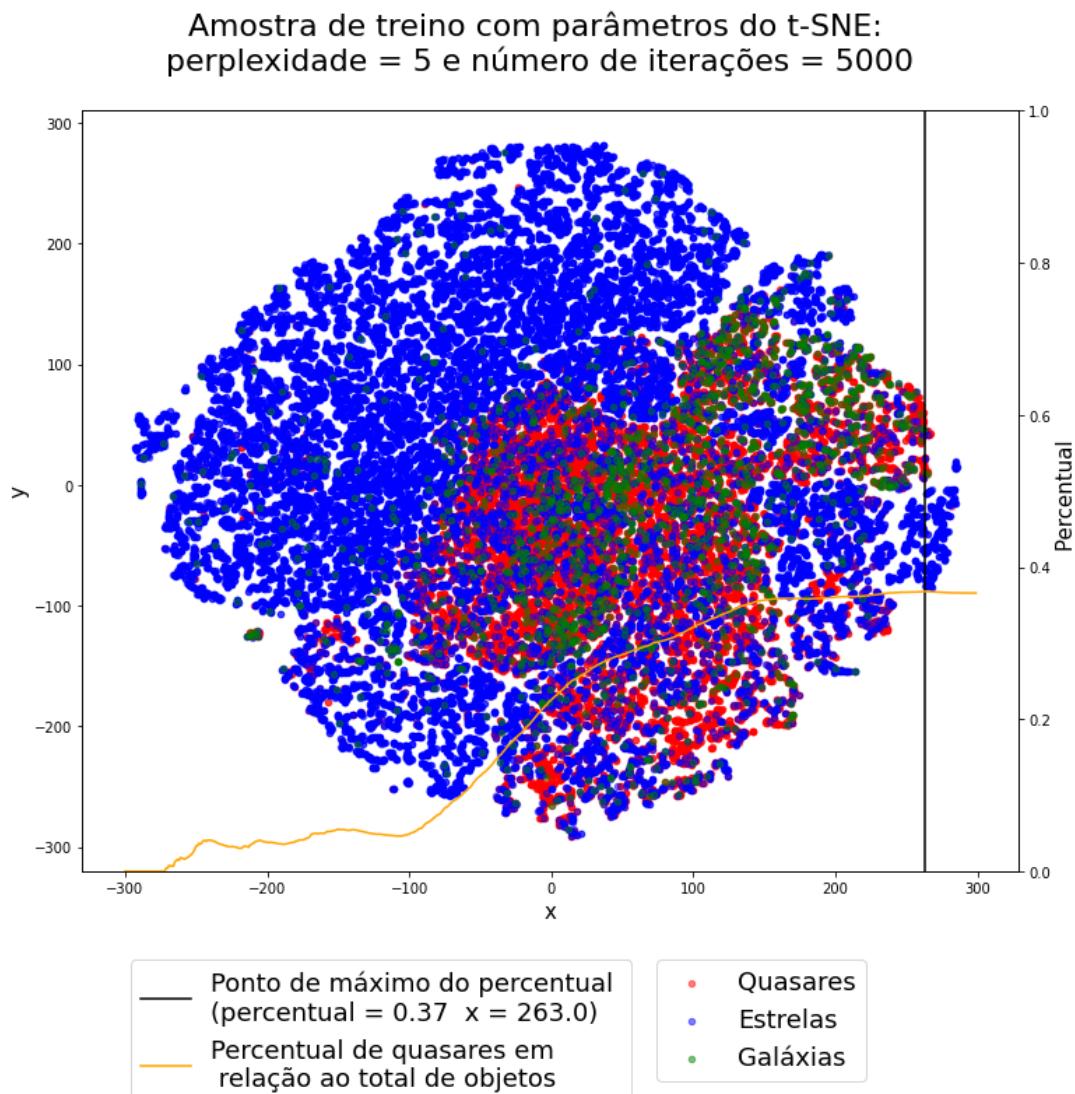
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 5 e 4000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à esquerda do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 5 e com um número de iterações de 5000, apresentou o mesmo padrão circular ilustrado nas últimas quatro Figuras 8, 9, 10 e 11, mas com a formação de um *cluster* de quasares localizado mais à direita da distribuição e com o *cluster* com uma distribuição mais irregular. Essa distribuição também se mostrou ineficaz para separar quasares dos demais objetos, pois o percentual máximo de um corte feito no eixo x foi de 37%. Essa distribuição dos objetos

no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 12.

Figura 12 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 5 e número de iterações igual a 5000.



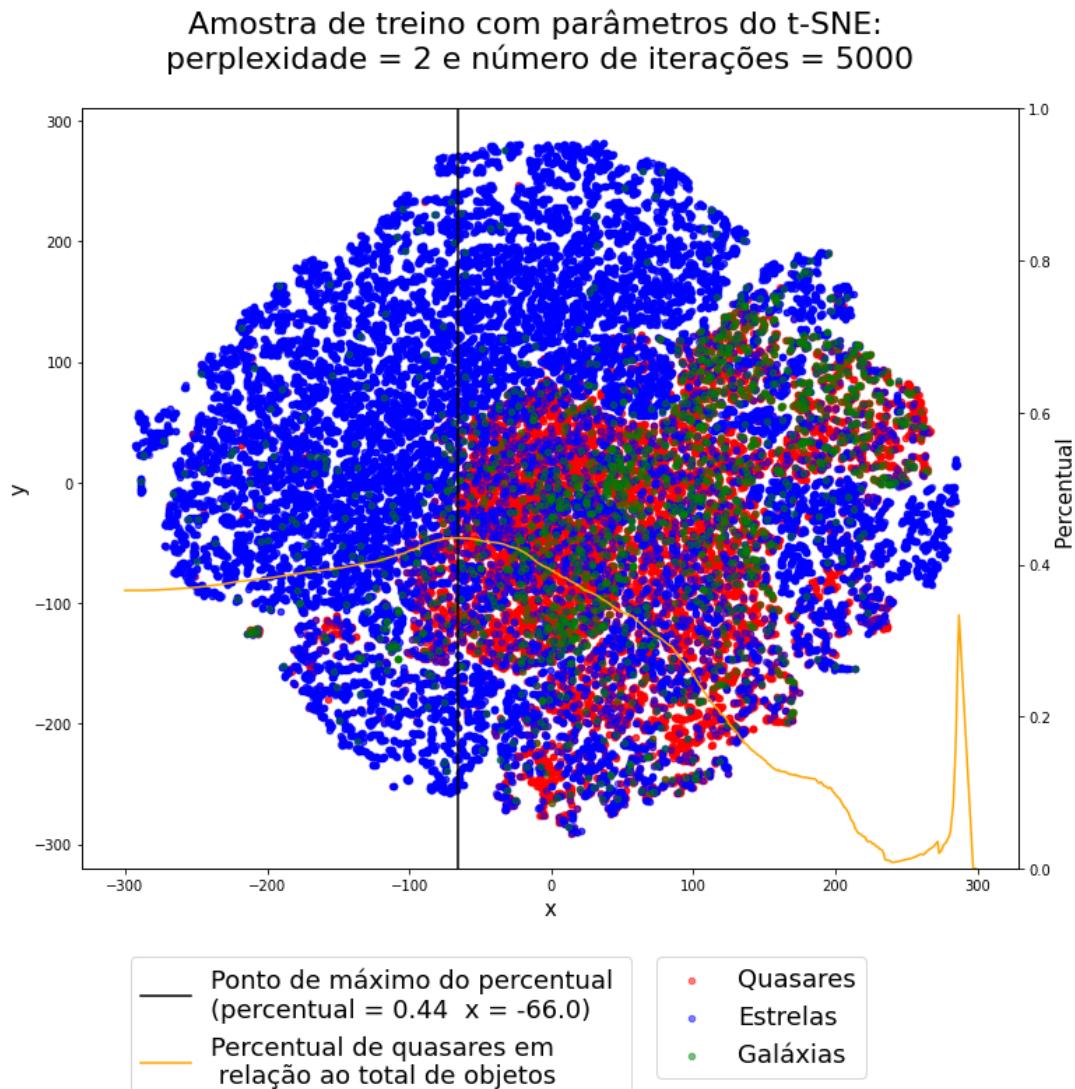
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 5 e 5000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à direita do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

3.1.2 Análise t-SNE variando a perplexidade

A redução dimensional da amostra de treinamento, com perplexidade de 2 e com um número de iterações de 5000, apresentou o mesmo padrão de distribuição da redução feita com perplexidade 5 e número de iterações 5000, ilustrado na Figura 12, também com a formação de um *cluster* de quasares localizado mais à direita da distribuição e com o *cluster* com uma distribuição mais irregular . Essa distribuição também se mostrou ineficaz para separar quasares dos demais objetos, pois o percentual máximo de um corte feito no eixo x foi de 44%. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual de quasares em relação ao total de objetos contidos no corte feito no eixo x está representado Figura 13.

Figura 13 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 2 e número de iterações igual a 5000.



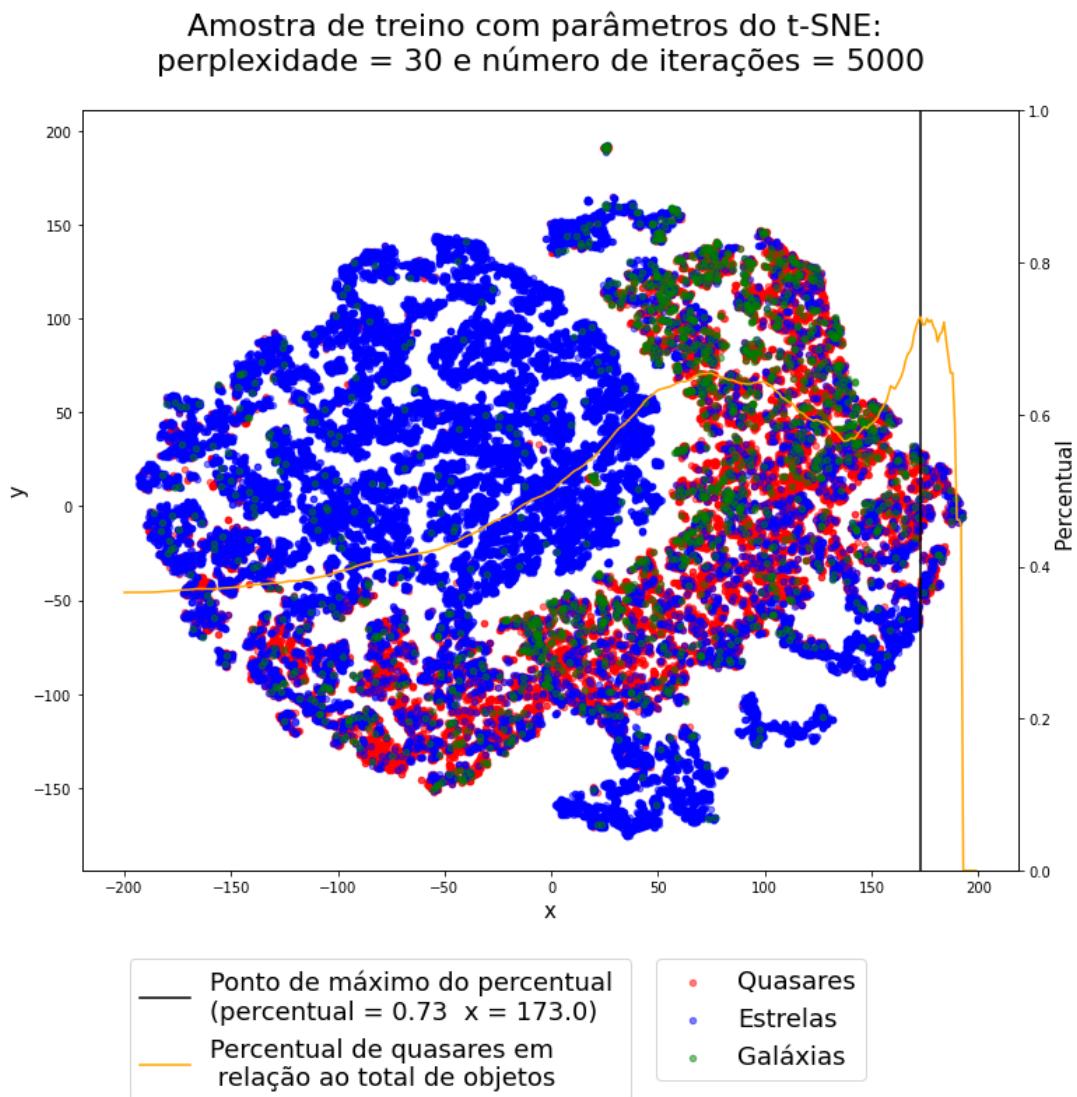
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 2 e 5000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à direita do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 30 e com um número de iterações de 5000, apresentou uma distribuição com maior distinção espacial entre as estrelas, galáxias e quasares, mas com uma sobreposição das galáxias sobre os quasares. Os dados se agruparam mais em *clusters*, e em comparação às distribuições já apresentadas acima, com esse valor de perplexidade os pontos saíram um pouco do padrão circular. Essa distribuição também se mostrou ineficaz para separar quasares dos demais

objetos, apesar do percentual máximo de um corte feito no eixo x ser de 73%, esse ponto de máximo aconteceu em um corte que continha poucos objetos. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 14.

Figura 14 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 30 e número de iterações igual a 5000.



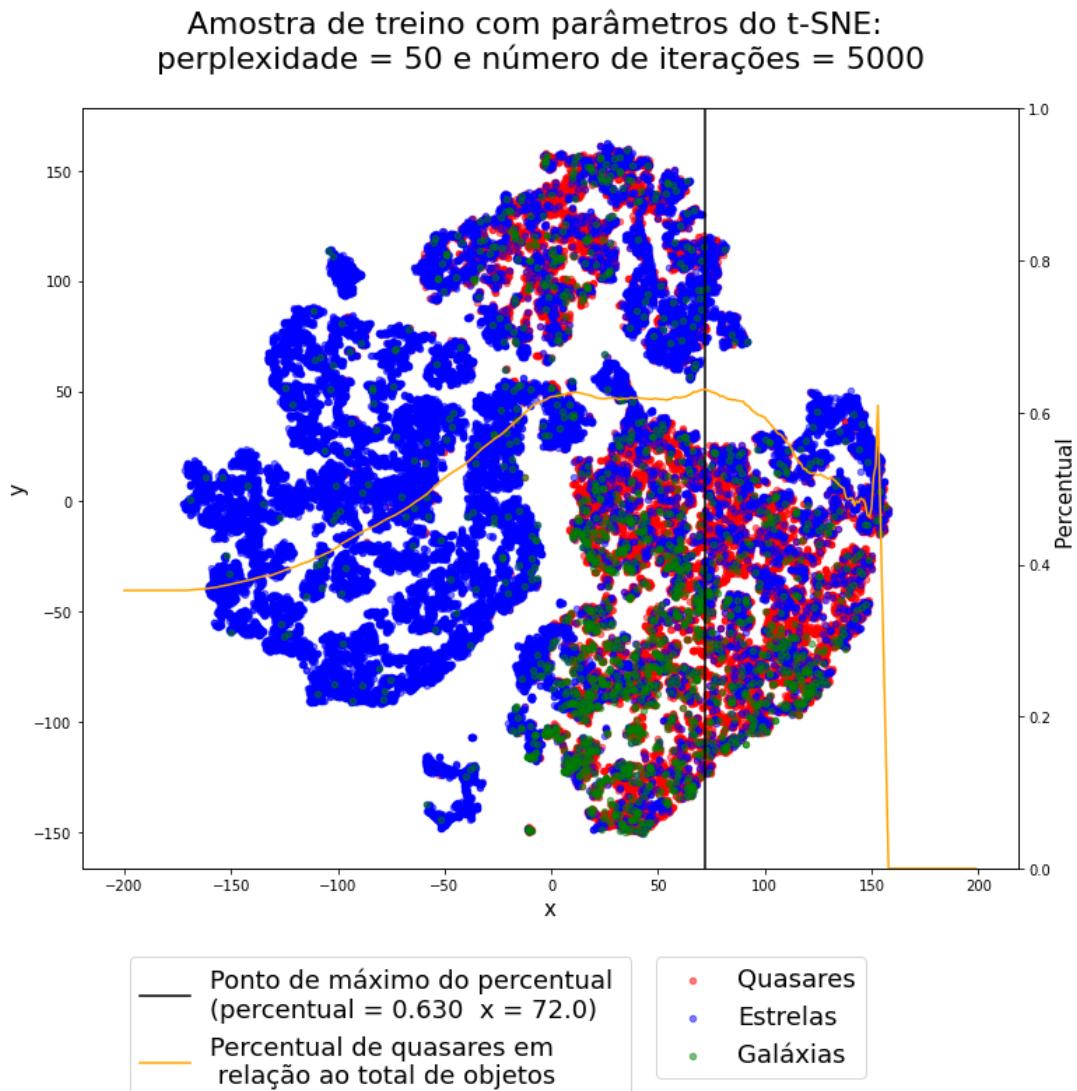
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 30 e 5000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à direita do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 50 e com um número de iterações de 5000, também apresentou uma distribuição com maior

distinção espacial entre as estrelas, galáxias e quasares. Os dados se agruparam mais em *clusters*, três que são facilmente distinguíveis visualmente, onde o mais à esquerda contem majoritariamente estrelas, o que está na parte superior do gráfico apresenta uma mistura entre os três objetos e o que está mais a direita apresenta os três objetos, mas com uma predominância maior de quasares. Essa distribuição também não apresenta distribuição circular, esta vista em perplexidades menores. Nessa distribuição é possível para separar quasares dos demais objetos, com um percentual máximo de um corte feito no eixo x ser de 63%. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 15.

Figura 15 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 50 e número de iterações igual a 5000.



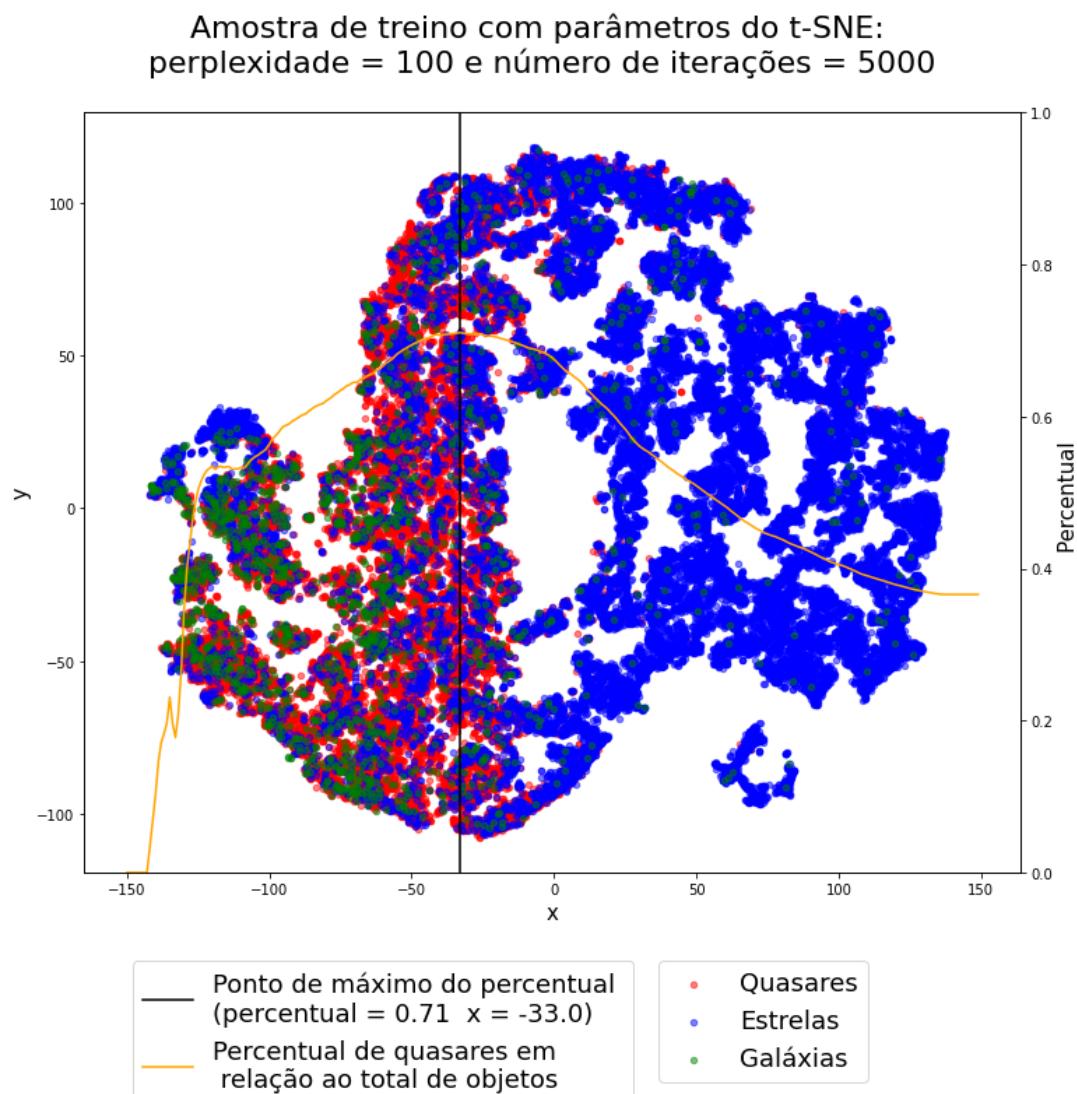
Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 50 e 5000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à direita do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

A redução dimensional da amostra de treinamento, com perplexidade de 100 e com um número de iterações de 5000, apresentou uma distribuição com maior distinção espacial entre as estrelas de galáxias e quasares. Os dados se agruparam mais em *clusters*, onde dois são facilmente distinguíveis visualmente, onde o mais à esquerda contém majoritariamente quasares e galáxias e o que está mais a direita apresenta majoritariamente estrelas. Nessa distribuição é possível para separar quasares dos demais objetos, com um percentual

máximo de um corte feito no eixo x ser de 71%, e foi o corte que conteve mais objetos. Essa distribuição dos objetos no plano, juntamente com a curva que descreve o percentual em relação ao corte feito no eixo x está representado Figura 16.

Figura 16 – Gráfico da amostra de treino após redução dimensional, com parâmetros t-SNE de perplexidade igual a 100 e número de iterações igual a 5000.



Fonte: A autora.

Gráfico de dispersão gerado a partir da redução dimensional realizada com o t-SNE do conjunto de dados de treino, com os valores de perplexidade e o número de iterações respectivamente sendo de 100 e 5000. O percentual apresentado no gráfico foi calculado considerando o total de quasares à esquerda do corte feito no eixo x, com relação ao total de objetos contido nesse corte.

3.1.3 Separação de quasares via análise t-SNE

Tendo em vista o objetivo de realizar uma separação de quasares entre objetos diferentes fez-se um corte simples no plano cartesiano, no eixo x . A redução dimensional t-SNE, que por porcentagem de quasares em relação ao total de objetos contidos no corte e por número de objetos contido no corte, que melhor apresentou essa separação foi o conjunto de treino submetido ao t-SNE com parâmetros: perplexidade igual à 100 e número de iterações igual à 5000, representado na Figura 16. Define-se o corte de seleção α como o processo de eleger os objetos distribuídos no plano cartesiano que estejam localizados em uma coordenada x menor ou igual a -33. Esse corte foi escolhido por ser a coordenada x onde o percentual de quasares em relação aos demais objetos contidos no corte alcançou o valor máximo, de 71%.

Tendo em vista a distribuição da amostra de teste, alcançada empregando KD Tree (como descrito na secção 2.4.8). Calculou-se o percentual do número de objetos em relação ao total de objetos contidos no corte α , para quasares, galáxias e estrelas, e os valores são mostrados na Tabela 3.

Tabela 3 – Percentual dos diferentes tipos de objetos em relação ao total de objetos contidos no corte α .

Classificação dos objetos	Amostra de Treinamento	Amostra de Teste
Quasares	71,02%	70,47%
Galáxias	7,04%	6,31%
Estrelas	21,93%	23,21%

Fonte: A autora

Ainda utilizando o corte α , foi calculado o percentual do número de objetos contidos no corte em relação ao total de objetos de toda a amostra, com os valores esboçados na Tabela 4.

Tabela 4 – Percentual dos diferentes tipos de objetos contidos no corte α em relação ao total de objetos do mesmo tipo contidos na amostra sem o corte.

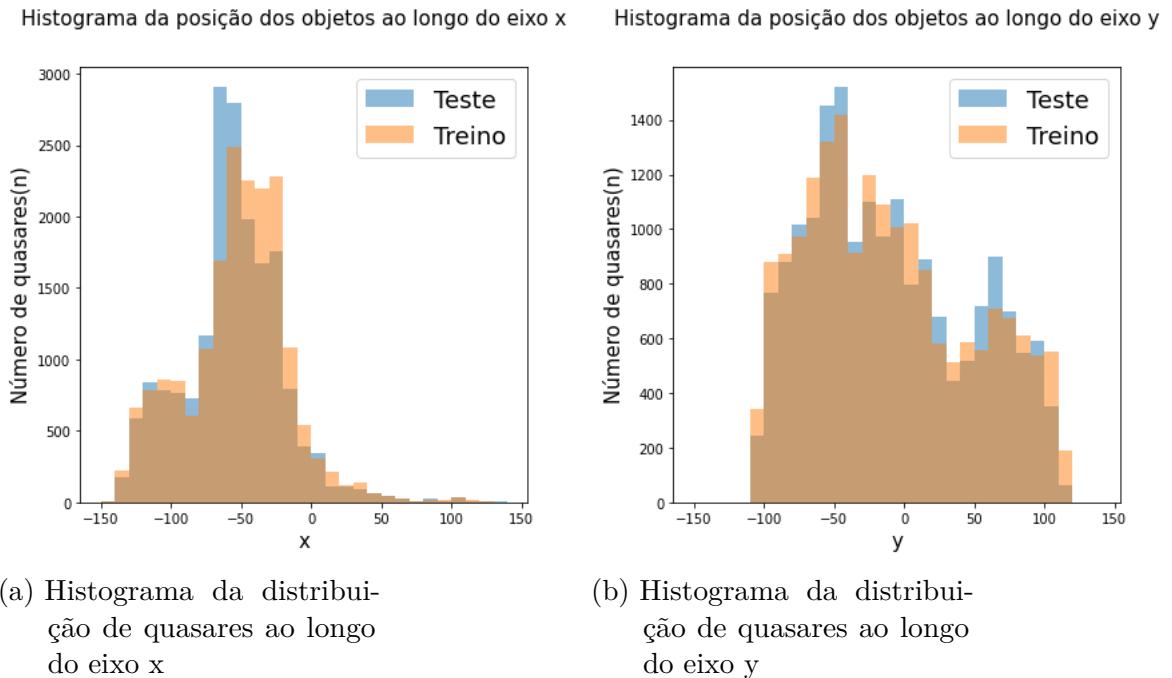
Classificação dos objetos	Amostra de Treinamento	Amostra de Teste
Quasares	70,18%	76,58%
Galáxias	76,95%	82,58%
Estrelas	13,18%	13,92%

Fonte: A autora

Para entender melhor a distribuição final dos quasares na amostra de teste com relação à amostra de treino, foi elaborado dois histogramas, um com a distribuição deles

ao longo do eixo x, ilustrado na Figura 17a, e outro ao longo do eixo y, ilustrado na Figura 17b.

Figura 17 – Os histogramas com a distribuição dos quasares ao longo do eixo x e do eixo y



Fonte: A autora

Os histogramas apresentados na Figura 17a e na Figura 17b foram feitos utilizando um tamanho de intervalo de 10 unidades.

3.2 Discussão

As distribuições dos objetos alcançadas adotando o número de iterações de: 1000, 2000, 3000, 4000 e 5000 e mantendo fixo a perplexidade, se mostraram todas em um padrão circular, com regiões distinguíveis contendo mais estrelas e outras contendo mais quasares. A grande diferença entre essas distribuições foi o raio onde estavam localizados os objetos. Com um número de iteração igual a 1000, 2000, 3000, 4000 e 5000 os objetos podiam ser encontrados, respectivamente em um raio aproximado de 75 unidades, 150 unidades, 200 unidades, 250 unidades e 300 unidades considerando o centro no ponto (0,0).

Já a variação de perplexidade mostrou uma mudança maior na distribuição dos objetos no plano cartesiano. A partir da perplexidade 30 foi possível notar uma melhor divisão espacial entre estrelas e quasares e também os objetos do mesmo tipo ficaram melhor agrupados entre si. A redução de dimensionalidade aplicada à amostra de treino que melhor resultou em uma separação entre quasares e estrelas foi a que tinha como perplexidade 100, e como número de iterações 5000.

Ao analisar todas as distribuições geradas com a ferramenta t-SNE, é possível notar que na região onde se acumula os quasares, também há um acúmulo de galáxias. Esse padrão é confirmado numericamente com a porcentagem de 76% de galáxias contidas no corte α em relação ao total da amostra de treinamento, que é ainda maior na amostra de treino, 82,58%, mostrando que mais de três quartos das galáxias da amostra estavam no corte escolhido para separar quasares de outros objetos. Isso pode acontecer pela redução de dimensionalidade carregar as similaridades nos atributos que destes dois tipos de objetos. Era esperado que os atributos CLASS_STAR_G,R,I,Z,Y fossem suficientes para diferenciar as galáxias de quasares, pois quasares apresentam um perfil de emissão de luz pontual como os das estrelas, diferente do perfil difuso apresentado pelas galáxias.

Com relação à amostra de teste, ela apresentou uma distribuição de quasares no eixo y, Figura 17b, muito semelhante à amostra de treino. A distribuição de quasares no eixo x, Figura 17b, se conservou mais à esquerda no eixo do que a amostra de treino, ponto que fez com que o percentual de quasares contidos no corte α em relação ao total de quasares da amostra tivesse melhor resultado, 76,58%, do que a amostra de treino, 70,18%.

Os resultados do corte α realizado na amostra de treino e na de teste foram muito semelhantes, mostrando a eficácia no uso do método KD Tree para reproduzir a distribuição obtida via t-SNE em um novo conjunto de dados. Utilizar um corte simples apenas considerando o eixo x restringiu consideravelmente a eficiência dos cortes feitos, principalmente no caso onde a distribuição dos objetos apresentava um padrão circular. Um melhor corte poderia ter sido obtido considerando o eixo x e y ou até mesmo a divisão do plano cartesiano em quadrados igualmente espaçados e selecionando os que contivesse um maior número de quasares.

4 CONCLUSÃO

Diante do exposto, o objetivo de fazer uma separação de quasares em um conjunto fotométrico de dados heterogêneos se mostrou alcançável utilizando a ferramenta de redução de dimensionalidade t-SNE. Mesmo havendo uma restrição ferramental do t-SNE para acrescentar novos pontos em uma redução dimensional, a alternativa de utilizar a técnica dos vizinhos mais próximos (*KD Tree*) para inferir as coordenada dos objetos no plano da amostra de teste, se mostrou eficiente. É importante salientar que os resultados obtidos nesse trabalho foram feitos com um grupo amostral muito específico de objetos, selecionando-os em uma região do céu cuja a densidade de objetos por *pixel* é alta. Não é possível afirmar um bom resultado em um grupo de objetos mais generalizado, para isso, mais testes deverão ser realizados.

O corte realizado apenas considerando o eixo *x* pode ser considerado simplista, apesar do bom resultado obtido na separação de quasares de estrelas. Um melhor resultado na separação destes objetos poderia ter sido alcançado considerando também o eixo *y*, ou mesmo dividindo o plano cartesiano em pequenos quadrados e selecionando os quadrados que obtivessem maior presença de quasares. Além disso, a utilização de mais atributos, como por exemplo a razão entre as magnitudes em bandas passantes diferentes, o *redshift* fotométrico (se caso disponível no levantamento) ou também utilização da magnitude com desavermelhamento podem melhorar a distribuição obtida via t-SNE. Para um trabalho futuro, se pretende melhorar essa seleção de quasares com as sugestões anteriores, e estabelecer se o grupo resultante da seleção é puro o suficiente para contribuir com o estudo da distribuição da matéria escura no Universo.

REFERÊNCIAS

- ABBOTT, T. *et al.* The dark energy survey data release 2. **The Astrophysical Journal Supplement Series**, IOP Publishing, v. 255, n. 2, p. 20, 2021.
- ABDALLA, E. *et al.* **The BINGO Project I: Baryon Acoustic Oscillations from Integrated Neutral Gas Observations**. 2021.
- AHUMADA, R. *et al.* The 16th data release of the sloan digital sky surveys: First release from the apogee-2 southern survey and full release of eboss spectra. **The Astrophysical Journal Supplement Series**, IOP Publishing, v. 249, n. 1, p. 3, 2020.
- Bertin, E.; Arnouts, S. Sextractor: Software for source extraction. **Astron. Astrophys. Suppl. Ser.**, v. 117, n. 2, p. 393–404, 1996. Disponível em: <https://doi.org/10.1051/aas:1996164>.
- Bolton, A. S. *et al.* Spectral Classification and Redshift Measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey. , v. 144, n. 5, nov. 2012.
- BURNS, M. S. A practical guide to observational astronomy. CRC Press, 2021.
- CAVALCANTE, M. A.; HAAG, R. Corpo negro e determinação experimental da constante de planck. **Revista Brasileira de Ensino de Física**, SciELO Brasil, v. 27, p. 343–348, 2005.
- CHAISSON, E.; McMILLAN, S. **Astronomy: A Beginner's Guide to the Universe - Eighth Edition**. [S.l.: s.n.]: Pearson Boston, Mass, USA, 2013.
- COLLABORATION:, D. E. S. *et al.* The Dark Energy Survey: more than dark energy – an overview. **Monthly Notices of the Royal Astronomical Society**, v. 460, n. 2, p. 1270–1299, 03 2016. ISSN 0035-8711. Disponível em: <https://doi.org/10.1093/mnras/stw641>.
- FILHO, K. de S. O.; SARAIVA, M. d. F. O. Astronomia e astrofísica. **Rio Grande do Sul: Livraria da Física**, 2004.
- GORSKI, K. M. *et al.* The healpix primer. **arXiv preprint astro-ph/9905275**, 1999.
- HILL, D. T. *et al.* Galaxy and Mass Assembly: FUV, NUV, ugrizYJHK Petrosian, Kron and Sérsic photometry. **Monthly Notices of the Royal Astronomical Society**, v. 412, n. 2, p. 765–799, 03 2011. ISSN 0035-8711. Disponível em: <https://doi.org/10.1111/j.1365-2966.2010.17950.x>.
- JARVIS, M. J. *et al.* The VISTA Deep Extragalactic Observations (VIDEO) survey. **Monthly Notices of the Royal Astronomical Society**, v. 428, n. 2, p. 1281–1295, 10 2012. ISSN 0035-8711. Disponível em: <https://doi.org/10.1093/mnras/sts118>.
- KUNSZT, P. Z.; SZALAY, A. S.; THAKAR, A. R. The hierarchical triangular mesh. In: BANDAY, A. J.; ZAROUBI, S.; BARTELmann, M. (ed.). **Mining the Sky**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 631–637. ISBN 978-3-540-44665-1.
- LANDIN, N. R. Espectroscopia da galáxia ngc 3819 e companheiras. Universidade Federal de Minas Gerais, 2002.

LYKE, B. W. *et al.* The Sloan digital sky survey quasar catalog: Sixteenth data release. **The Astrophysical Journal Supplement Series**, IOP Publishing, v. 250, n. 1, p. 8, 2020.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. 11, 2008.

MANEEWONGVATANA, S.; MOUNT, D. M. **Analysis of Approximate Nearest Neighbor Searching with Clustered Point Sets**. 1999.

MORGANSON, E. *et al.* The dark energy survey image processing pipeline. **Publications of the Astronomical Society of the Pacific**, IOP Publishing, v. 130, n. 989, p. 074501, 2018.

Nakoneczny, S. *et al.* Catalog of quasars from the kilo-degree survey data release 3. **A&A**, v. 624, p. A13, 2019. Disponível em: <https://doi.org/10.1051/0004-6361/201834794>.

NAKONECZNY, S. *et al.* Catalog of quasars from the kilo-degree survey data release 3. **Astronomy & Astrophysics**, EDP Sciences, v. 624, p. A13, 2019.

OLIVEIRA, B. D. d. Influência da radiação de quasares a grandes distâncias. 2017.

PANNUTI, T. G. The physical processes and observing techniques of radio astronomy. **The Physical Processes and Observing Techniques of Radio Astronomy**, Springer, 2020.

PASACHOFF, J. M.; FILIPPENKO, A. **The cosmos: Astronomy in the new millennium**. [S.l.: s.n.]: Cambridge University Press, 2013.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

SÁNCHEZ, E. The dark energy survey. **Journal of Physics: Conference Series**, IOP Publishing, v. 259, p. 012080, nov 2010. Disponível em: <https://doi.org/10.1088/1742-6596/259/1/012080>.

SCHMIDT, M. 3 c 273: a star-like object with large red-shift. **Nature**, Nature Publishing Group, v. 197, n. 4872, p. 1040–1040, 1963.

USERY, E. L. Understanding map projections. In: **The routledge handbook of mapping and cartography**. [S.l.: s.n.]: Routledge, 2017. p. 202–222.

VELTEN, H. Matéria escura e as estruturas cósmicas. **Cadernos de Astronomia**, v. 2, n. 1, p. 70–474, 2021.

VELTEN, H. *et al.* Desvendando a não adiabaticidade da energia escura. Universidade Federal do Espírito Santo.

WATTENBERG, M.; VIÉGAS, F.; JOHNSON, I. How to use t-sne effectively. **Distill**, 2016. Disponível em: <http://distill.pub/2016/misread-tsne>.

ZHANG, Y. *et al.* Dark energy survey year 1 results: Detection of intracluster light at redshift 0.25. **The Astrophysical Journal**, IOP Publishing, v. 874, n. 2, p. 165, 2019.