UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação de Objetos Celestes e Espectros Estelares com Perceptron Multicamadas

Claudio Honorato Junior

Monografia – MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Claudio Honorato Junior

Classificação de Objetos Celestes e Espectros Estelares com Perceptron Multicamadas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dra. Mariana Caravanti de Souza

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Honorato, Claudio Junior

S856m Classificação de Objetos Celestes e Espectros Estelares com Perceptron Multicamadas utilizando o BERT / Claudio Honorato Junior ; orientador: Mariana Caravanti de Souza. – São Carlos, 2025. 56 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. . 2. . 3. . 4. 5. l I. Souza, M. C. d. , orient. II. Título.

Claudio Honorato Junior

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dra. Mariana Caravanti de Souza

Original version

São Carlos

2025

Claudio Honorato Junior

Classificação de Objetos Celestes e Espectros Estelares com Perceptron Multicamadas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Data de defesa: 13 de Setembro de 2025

Comissão Julgadora:

_____

Dra. Mariana Caravanti de Souza
Orientador

_____

Profª. Dra. Solange Oliveira Rezende
VICE-COORDENADORA

São Carlos
2025

AGRADECIMENTOS

Com imensa gratidão, agradeço à minha querida família pelo apoio incondicional, pilar fundamental em toda esta jornada.

Sou imensamente grato ao ICMC pela bolsa de estudos, que tornou possível a realização desta pesquisa.

À minha orientadora, deixo um agradecimento especial pela orientação dedicada, cuja contribuição foi essencial para o desenvolvimento deste trabalho.

À todos que, direta ou indiretamente, colaboraram com este projeto, meu muito obrigado.

RESUMO

Honorato Junior, C. H. 2025. 56p. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este trabalho propõe o uso de Redes Neurais Multicamadas (MLP) para a classificação de objetos astronômicos com base em dados espectroscópicos do Sloan Digital Sky Survey (SDSS). A pesquisa aborda três objetivos principais: (i) classificar objetos celestes em estrelas, galáxias e quasares; (ii) subclassificar estrelas conforme a classificação espectral de Harvard, com base na temperatura superficial. O modelo MLP foi comparado com algoritmos tradicionais como SVM, KNN, Regressão Logística, Random Forest e Naive-Bayes, avaliando seu desempenho por meio de métricas como acurácia, precisão, recall e F1-score. A metodologia incluiu limpeza, normalização e seleção de features espectrais, além de técnicas de validação cruzada para garantir a robustez dos resultados. Os achados mostram que redes neurais profundas podem ser eficazes na análise de espectros astronômicos, oferecendo uma alternativa promissora para o tratamento de grandes volumes de dados no contexto da astrofísica moderna.

Palavras-chave: Inteligência Artificial, Redes Neurais Multicamadas, SDSS, Espectroscopia, Classificação Estelar, Índice D4000, Índices de Lick, QSO.

ABSTRACT

Honorato Junior, C. H. 2025. 56p. Monograph (MBA in Artificial Intelligence and Big Data) – Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2025.

This work proposes the use of Multilayer Neural Networks (MLP) for the classification of astronomical objects based on spectroscopic data from the Sloan Digital Sky Survey (SDSS). The research addresses three main objectives: (i) to classify celestial objects into stars, galaxies, and quasars; (ii) to subclassify stars according to the Harvard spectral classification, based on surface temperature. The MLP model was compared with traditional algorithms such as SVM, KNN, Logistic Regression, Random Forest, and Naive-Bayes, evaluating its performance through metrics such as accuracy, precision, recall, and F1-score. The methodology included spectral cleaning, normalization, and feature selection, as well as cross-validation techniques to ensure the robustness of the results. The findings show that deep neural networks can be effective in the analysis of astronomical spectra, offering a promising alternative for the treatment of large volumes of data in the context of modern astrophysics.

Keywords: Artificial Intelligence, Multilayer Neural Networks, SDSS, Spectroscopy, Stellar Classification, D4000 Index, Lick Indices, QSO.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DR18	Data Release 18 (18ª versão do Sloan Digital Sky Survey)
SVM	Support Vector Machine
D4000	Índice espectral de descontinuidade em 4000 Ångströms
KNN	K-Nearest Neighbors (K-Vizinhos Mais Próximos)
Lick	Conjunto de índices espectrais desenvolvido no Observatório Lick
MLP	Multilayer Perceptron (Perceptron Multicamadas)
QSO	Quasar (Quasi-Stellar Object)
RF	Random Forest
SDSS	Sloan Digital Sky Survey
Z	Redshift (Desvio para o vermelho)
Fe	Ferro
Na	Sódio
Mg	Magnésio
A	Classe espectral de estrelas brancas quentes (7.500–10.000 K)
B	Classe espectral de estrelas azuis muito quentes (10.000–30.000 K)
C	Tipo espectral de estrela rica em carbono (Carbon Star)
F	Classe espectral de estrelas branco-amareladas (6.000–7.500 K)
G	Classe espectral de estrelas amarelas como o Sol (5.200–6.000 K)
K	Classe espectral de estrelas laranja (3.700–5.200 K)
L	Classe espectral de anãs marrons frias (1.300–2.500 K)
M	Classe espectral de estrelas vermelhas frias (≤3.700 K)
O	Classe espectral de estrelas azuis extremamente quentes (>30.000 K)
T	Classe espectral de anãs marrons mais frias que as L (600–1.300 K)
W	Estrelas do tipo Wolf-Rayet

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, os avanços nas tecnologias de observação espacial, impulsionados por telescópios modernos como o Telescópio Espacial James Webb (JWST), têm gerado um volume exponencial de dados astronômicos. A capacidade desses instrumentos de coletar informações detalhadas do cosmos abriu novas fronteiras na pesquisa astronômica, mas também apresentou desafios significativos no que tange ao processamento e interpretação desses dados (BALL; BRUNNER, 2010). A necessidade de categorizar e analisar grandes quantidades de informações para identificar padrões e características específicas de diferentes tipos de estrelas e galáxias representa um dos principais desafios da astronomia moderna (Sharma, K. 2020). A complexidade dos espectros, que carregam assinaturas únicas sobre a composição, temperatura e movimento dos objetos celestes, e a diversidade de objetos presentes no universo tornam essa tarefa ainda mais exigente.

Este estudo busca aplicar técnicas avançadas de Inteligência Artificial (IA) e Big Data para aprimorar a classificação de objetos celestes. Em particular, o uso de redes neurais multicamadas (MLP) apresenta uma abordagem promissora para lidar com o grande volume de dados gerados por telescópios modernos. As MLPs, com sua capacidade de aprender representações complexas de dados, mostram-se particularmente adequadas para extrair informações valiosas de espectros astronômicos (Sharma, K. 2020). A implementação de modelos automatizados para análise de espectros astronômicos não apenas acelera descobertas científicas, permitindo aos astrônomos identificar rapidamente objetos de interesse, mas também democratiza o acesso a informações detalhadas sobre o universo, tornando a pesquisa astronômica mais eficiente e inclusiva.

Neste contexto, este trabalho propõe a utilização de MLPs para a classificação de objetos astronômicos a partir de dados espectroscópicos do Sloan Digital Sky Survey (SDSS). Além disso, pretende-se comparar o modelo de rede neural com classificadores tradicionais, como Máquinas de Vetores de Suporte (SVM), K-Vizinhos Mais Próximos (KNN), Árvore de Decisão, Regressão Logística, Random Forest e Naive-Bayes, a fim de otimizar a identificação de padrões nos espectros de objetos celestes. Essa comparação permitirá avaliar o desempenho relativo das MLPs em relação a métodos consagrados, buscando identificar vantagens e limitações de cada abordagem.

1.1 Objetivo e Questões de Pesquisa

Este trabalho de conclusão de curso propõe desenvolver e avaliar modelos de aprendizado de máquina e aprendizado profundo, com foco em Redes Neurais Multicamadas (MLPs), para a classificação e análise detalhada de objetos astronômicos, utilizando dados espectroscópicos do Sloan Digital Sky Survey (SDSS).

Pretende-se investigar a eficácia de diferentes arquiteturas de MLP e comparar seu desempenho com classificadores tradicionais. O objetivo é otimizar a identificação de padrões complexos nos espectros, que revelam informações cruciais sobre a natureza física e a evolução desses objetos, além disso, a pesquisa busca analisar como diferentes técnicas de pré-processamento e seleção de características impactam a precisão e a eficiência dos modelos na classificação de objetos astronômicos, considerando a complexidade e o volume dos dados espectroscópicos.

Diante deste cenário, as seguintes questões de pesquisa devem ser respondidas:

- Q1: “Qual a capacidade dos modelos de aprendizado de máquina, em especial as Redes Neurais Multicamadas (MLPs), de classificar objetos astronômicos (estrelas, galáxias e quasares) e classificar espectros estelares a partir de dados espectroscópicos do SDSS?”

- Q2: “Qual a capacidade dos modelos de aprendizado de máquina, em especial as Redes Neurais Multicamadas (MLPs), de classificar espectros estelares (O, B, A, F, G, K, M) a partir de dados espectroscópicos do SDSS?”

- Q3: “Como o desempenho dos modelos de aprendizado profundo (MLPs) se compara ao de classificadores tradicionais, como SVM, KNN, Árvore de Decisão, Regressão Logística, Random Forest e Naive-Bayes, na tarefa de classificação de objetos astronômicos?”

Já os objetivos específicos deste trabalho de conclusão de curso incluem:

- Desenvolver e implementar modelos de Redes Neurais Multicamadas (MLPs) para a classificação de objetos astronômicos em três categorias principais: quasares (QSO),

galáxias (GALAXY) e estrelas (STAR). Este objetivo está relacionado às questões de pesquisa Q1 e Q2.

- Realizar a subclassificação de estrelas de acordo com o sistema de classificação espectral de Harvard, utilizando os modelos de aprendizado de máquina desenvolvidos. Este objetivo está relacionado às questões de pesquisa Q1 e Q2

- Comparar o desempenho dos modelos de MLPs com classificadores tradicionais (SVM, KNN, Árvore de Decisão, Regressão Logística, Random Forest e Naive-Bayes) na classificação de objetos astronômicos. Este objetivo está relacionado à questão de pesquisa Q3.

1.2 Organização do Texto

O restante deste trabalho está organizado da seguinte forma:

- Capítulo 2: Dedica-se à exposição dos fundamentos teóricos que sustentam a pesquisa, abordando os conceitos e as tecnologias essenciais para o desenvolvimento e a compreensão dos métodos empregados.

- Capítulo 3: Realiza uma revisão dos trabalhos relacionados, explorando as pesquisas existentes na literatura que abordam temas similares ou complementares, a fim de situar o presente estudo no contexto acadêmico.

- Capítulo 4: Detalha a metodologia adotada, descrevendo as etapas do processo de pesquisa, as técnicas utilizadas e o desenvolvimento dos experimentos conduzidos para alcançar os objetivos propostos.

- Capítulo 5: Apresenta e analisa os resultados obtidos com a aplicação das técnicas, discutindo as métricas de avaliação estabelecidas e sua interpretação à luz das questões de pesquisa.

- Capítulo 6: Conclui o trabalho com uma síntese das principais descobertas, uma discussão acerca dos avanços e das limitações do estudo, e a proposição de sugestões para futuras pesquisas na área.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Classificação em Astronomia

A classificação é crucial na Astronomia para organizar a vasta diversidade cósmica, revelando padrões e processos evolutivos essenciais para nossa compreensão do Universo. Este capítulo explora a aplicação da classificação em astronomia, abordando a Classificação de Objetos Astronômicos, que estabelece as categorias fundamentais dos corpos celestes e seu significado cosmológico, e a Classificação Espectral Estelar, um sistema detalhado que desvenda as propriedades físicas e a evolução das estrelas através da análise de sua luz.

2.1.1 Classificação de Objetos Astronômicos: Fundamentos e Significado Cosmológico

O primeiro objetivo desta pesquisa reside no desenvolvimento de um modelo de aprendizado profundo robusto e preciso, capaz de realizar a classificação de objetos astronômicos em três categorias primárias: quasares (QSO), galáxias (GALAXY) e estrelas (STAR). A distinção e a classificação acurada desses objetos celestes representam um pilar fundamental para a construção de nosso entendimento sobre a natureza, a evolução e a estrutura em larga escala do universo (BALL; BRUNNER, 2010; BORNE, 2013).

Desde os primórdios da astronomia observacional, a catalogação e a classificação dos corpos celestes têm sido atividades cruciais. Inicialmente baseadas em características morfológicas aparentes e no brilho, as classificações evoluíram significativamente com o advento de novas tecnologias e métodos de observação, como a espectroscopia (GRAY; CORBALLY, 2021). A capacidade de distinguir entre diferentes tipos de objetos astronômicos permite aos astrônomos:

- Construir Censos Cósmicos: A classificação sistemática possibilita a criação de inventários detalhados dos constituintes do universo, fornecendo informações cruciais sobre a abundância relativa de diferentes tipos de objetos em diferentes épocas cósmicas.

- Investigar Processos Físicos: Cada categoria de objeto astronômico é regida por processos físicos distintos. Estrelas são fornalhas nucleares onde elementos leves são fundidos em elementos mais pesados. Galáxias são vastos sistemas gravitacionalmente ligados de estrelas, gás, poeira e matéria escura, locais de formação estelar e evolução química. Quasares, por sua vez, são núcleos galácticos

ativos extremamente luminosos, alimentados pela acreção de matéria em buracos negros supermassivos. A classificação correta é o primeiro passo para aplicar os modelos físicos apropriados a cada objeto.

- Mapear a Estrutura em Larga Escala: A distribuição espacial de galáxias e quasares revela a arquitetura filamentar do universo, com vastos vazios e densos aglomerados. A classificação precisa é essencial para estudos cosmológicos que visam entender a formação e a evolução dessas estruturas.

- Rastrear a Evolução Cósmica: Observar a distribuição de diferentes tipos de objetos em diferentes desvios para o vermelho (redshifts) permite aos astrônomos estudar a evolução do universo ao longo do tempo cósmico. Por exemplo, a abundância de quasares atingiu um pico em um determinado período da história cósmica, fornecendo pistas sobre o crescimento dos buracos negros supermassivos e a evolução das galáxias hospedeiras.

A distinção entre quasares (QSO), galáxias (GALAXY) e estrelas (STAR) representa uma classificação fundamental devido às suas naturezas intrínsecas e aos processos físicos dominantes em cada um deles (RICHARDS et al., 2009):

- Estrelas (STAR): São corpos celestes massivos compostos principalmente de plasma, mantidos juntos pela gravidade e cuja principal fonte de energia é a fusão nuclear em seus núcleos. As estrelas variam enormemente em massa, temperatura, luminosidade e ciclo de vida, dando origem à classificação espectral detalhada (Seção 2.1.2). Individualmente, as estrelas são os blocos de construção fundamentais das galáxias e os principais produtores de elementos mais pesados que o hidrogênio e o hélio. Em termos observacionais, estrelas geralmente aparecem como fontes pontuais de luz, com espectros característicos de absorção e emissão dependendo de sua temperatura e composição.

- Galáxias (GALAXY): São sistemas massivos e complexos, compostos por bilhões ou até trilhões de estrelas, juntamente com gás interestelar, poeira cósmica, matéria escura e, frequentemente, um buraco negro supermassivo em seu centro. As galáxias apresentam uma ampla variedade de morfologias, desde espirais bem definidas até elípticas mais amorfas e irregulares. Seus espectros são a combinação da luz de bilhões de estrelas, além de emissões e absorções do gás interestelar. Observacionalmente,

galáxias tipicamente exibem uma extensão espacial definida e estruturas internas complexas.

- Quasares (QSO - Quasi-Stellar Object): São núcleos galácticos ativos (AGN) extremamente luminosos, alimentados pela acreção de matéria em buracos negros supermassivos localizados no centro de galáxias distantes. A enorme quantidade de energia liberada durante esse processo de acreção faz com que os quasares superem em brilho suas galáxias hospedeiras, tornando-os visíveis a distâncias cosmológicas significativas. Seus espectros apresentam linhas de emissão largas e intensas, muitas vezes com grandes desvios para o vermelho devido à expansão do universo, indicando suas grandes distâncias. Inicialmente, foram denominados "objetos quase estelares" devido à sua aparência pontual em observações de baixa resolução, similar a estrelas, mas seus espectros revelaram sua natureza extragaláctica e a presença de processos energéticos não estelares.

2.1.2 Classificação Espectral Estelar

Considerando que este trabalho abordará a classificação espectral estelar por meio de algoritmos de aprendizado de máquina, uma breve discussão sobre o tema se faz necessária neste momento.

A classificação de espectros estelares tem suas raízes no surgimento da espectroscopia, que se consolidou como um campo de estudo essencial na astronomia após a identificação da linha de absorção do sódio no espectro solar por Joseph Fraunhofer em 1814. Em 1860, Robert Wilhelm Bunsen e Gustav Kirchhoff demonstraram que cada elemento químico emite um conjunto único de linhas espectrais. A partir dessa descoberta, a análise espectral foi estendida a outras estrelas com o objetivo de inferir suas composições químicas.

Essa linha de pesquisa evoluiu com o trabalho de Antonia Maury em colaboração com Edward Charles Pickering, publicado em Maury & Pickering (1897). Nessa análise, observou-se uma variação na largura das linhas de absorção, denominada por Maury como "característica c", associada a diferentes magnitudes absolutas.

Entre 1911 e 1914, Annie Jump Cannon catalogou aproximadamente 200.000 espectros estelares manualmente, trabalho que culminou na publicação do catálogo HD (Henry Draper). Em conjunto com Williamina P. Fleming, Cannon desenvolveu a Classificação Espectral de Harvard, fundamentada na similaridade aparente entre os espectros, estabelecendo a sequência "O B A F G K M" (CANNON, 1918). A FIGURA 1 ilustra o perfil espectral em diferentes temperaturas, e a FIGURA 2 demonstra a dependência da intensidade das linhas espectrais com a temperatura.

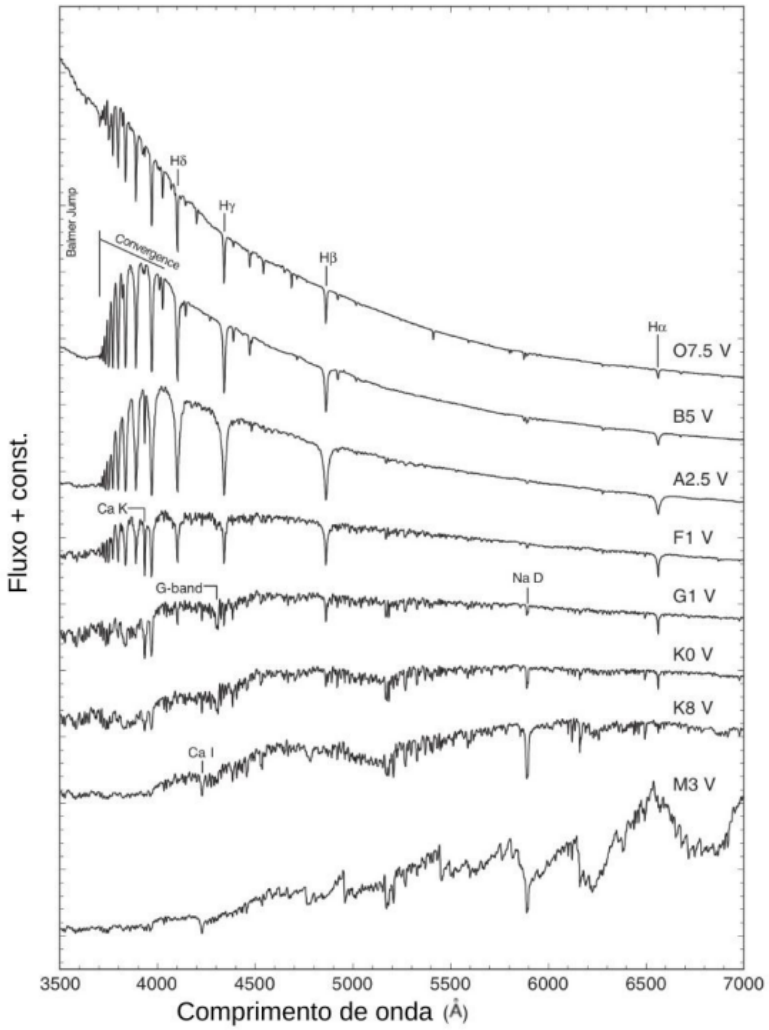


FIGURA 1. Comparação entre diferentes Espectros. FIGURA extraída de Gray & Corbally (2021).

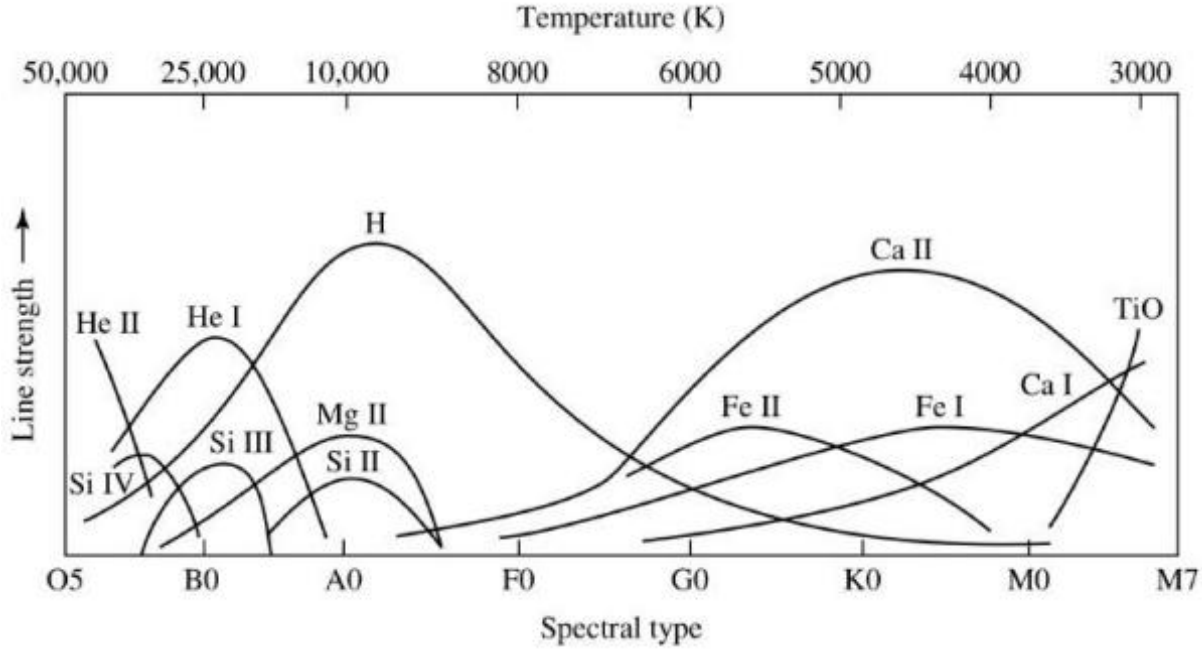


FIGURA 2. Carroll & Ostlie (2017): A dependência da intensidade das linhas espectrais com a temperatura.

Atualmente, compreende-se que as estrelas mais quentes pertencem ao tipo espectral O (estrelas azuis e extremamente quentes, com linhas de hélio ionizado proeminentes), enquanto as mais frias são classificadas como tipo M (estrelas vermelhas e frias, com bandas moleculares de óxido de titânio fortes). Adicionalmente a essa classificação primária, foi implementada uma subclassificação numérica, variando de 0 (estrelas mais quentes dentro de um tipo) a 9 (estrelas mais frias dentro do mesmo tipo). Por exemplo, estrelas B9 (onde as linhas de hidrogênio são mais intensas do que em B0) apresentam temperaturas inferiores às estrelas B0 (onde as linhas de hélio são mais intensas).

Esses estudos iniciais suscitaram uma questão fundamental: qual a origem das significativas diferenças observadas nos espectros estelares? Cecilia Payne, em sua tese de doutorado (PAYNE, Cecilia H. Stellar atmospheres; a contribution to the observational study of high temperature in the reversing layers of stars. 1925. Tese (Doutorado) - Radcliffe College, Cambridge, MA.), considerada um marco na astronomia, elucidou essa questão. Payne demonstrou que as variações nos perfis espectrais não eram primariamente devidas a diferenças na composição química, mas sim às distintas temperaturas dos objetos estelares. A temperatura influencia diretamente a proporção de átomos ionizados e o estado de excitação dos elétrons, resultando na maior abundância de certos elementos nos espectros de estrelas mais quentes.

Posteriormente, Ejnar Hertzsprung e Henry Norris Russel, independentemente, observaram que estrelas pertencentes ao mesmo tipo espectral podiam apresentar magnitudes absolutas distintas. Dentro de uma mesma classe espectral, ambos passaram a denominar as estrelas mais luminosas como gigantes, enquanto Russel introduziu o termo anãs para as menos luminosas (RUSSELL, 1914). Esse conhecimento foi fundamental para a elaboração do Diagrama HR (Hertzsprung-Russel), apresentado na FIGURA 2.1, que revela a relação entre a luminosidade (eixo y), o raio e a temperatura (eixo x) dos objetos estelares.

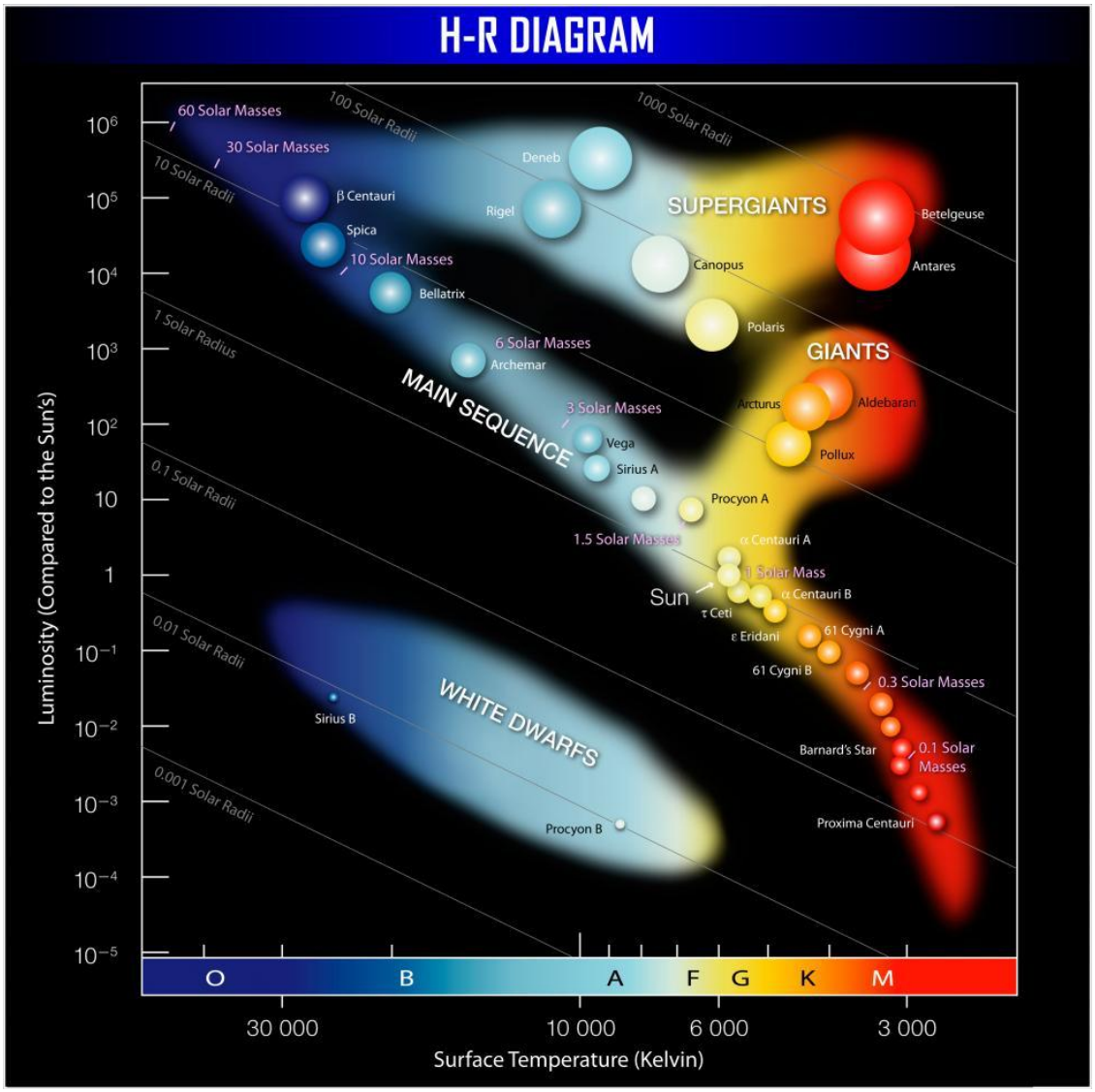


FIGURA 2.1 Carroll & Ostlie (2017): Diagrama de Hertzsprung-Russell destacando a relação entre temperatura, luminosidade e classes espectrais das estrelas.

Hertzsprung, então, investigou se a variação na luminosidade poderia ser detectada nos espectros estelares. Essa questão foi respondida através da análise dos espectros catalogados por Antonia Maury, em colaboração com Pickering. Conforme mencionado, nessa análise observou-se uma variação na largura das linhas de absorção, associada a diferentes magnitudes absolutas. Ademais, em seu trabalho "Radiation from Stars", Hertzsprung, referenciando o estudo de Maury, explorou a existência de estrelas com temperaturas semelhantes, mas com luminosidades e tamanhos distintos: as anãs (pouco luminosas e menores) e as gigantes (mais luminosas e maiores).

Outra forma de evidenciar a interdependência entre temperatura, luminosidade e raio estelar reside na Lei de Stefan-Boltzmann, expressa na Equação 1.1.

$$R = \frac{1}{T_e^2}\sqrt{\frac{L}{4\pi\sigma}} \tag{1.1}$$

Esta equação demonstra que, para uma dada temperatura, a luminosidade de uma estrela deve ser maior quanto maior for o seu raio.

A Classificação Espectral de Harvard permanece em uso até os dias atuais. Contudo, duas novas classes espectrais, L e T, correspondentes a objetos subestelares com temperaturas ainda mais baixas que as do tipo M (como anãs marrons, com atmosferas ricas em metano e vapor d'água), foram incorporadas à sequência original, resultando em: "O B A F G K M L T".

2.2 Conceitos Fundamentais

Para a compreensão dos modelos de aprendizado de máquina aplicados à classificação de objetos astronômicos, torna-se imprescindível a introdução de alguns conceitos fundamentais. Esta seção aborda os pilares teóricos que sustentam a metodologia empregada, definindo *Labels* e *Features*, diferenciando os paradigmas de Aprendizados Supervisionado e Não-Supervisionado, e explicando a crucial divisão dos dados em Conjuntos de Treinamento e Teste. Adicionalmente, serão discutidos os problemas comuns de *Underfitting* e *Overfitting* e as ferramentas de avaliação de desempenho de modelos de classificação, como a Matriz de Confusão e as métricas de Precision, Recall e F1-Score, fornecendo o vocabulário essencial para a análise dos resultados obtidos.

2.2.1 Labels e Features

Em Aprendizado de Máquina, a análise de dados é fundamentada em dois componentes essenciais: os rótulos (*Labels*) e os atributos (*Features*) (GERON, 2019). Estes elementos são definidos da seguinte forma:

- *Labels*: Representam as categorias ou classes associadas a cada espectro, constituindo o alvo da predição do modelo de Aprendizado de Máquina. No contexto deste trabalho, o rótulo é a classificação do objeto astronômico, conforme a coluna class e subclass das Tabela 1 e Tabela 2, que pode assumir valores como "STAR", "GALAXY" ou "QSO".

- *Features*: Correspondem às características ou informações quantitativas extraídas de cada espectro, que o modelo utiliza como base para aprender sobre o problema em questão. A título de ilustração, um dos atributos empregados é a fotometria em bandas “u, g, r, i, z”, que representam magnitudes em diferentes comprimentos de onda

2.2.2 Aprendizados supervisionado e não-supervisionado

Os algoritmos de Machine Learning (ML) podem ser classificados de acordo com a forma como aprendem a partir dos dados. As duas abordagens principais são o aprendizado supervisionado e o aprendizado não supervisionado ((GERON, 2019)). A distinção entre elas está relacionada à necessidade, ou não, de rótulos previamente definidos nos dados.

No aprendizado supervisionado, os algoritmos recebem um conjunto de dados já categorizado, contendo entradas (features) e suas respectivas saídas (rótulos). O modelo aprende a associar padrões específicos dos dados às classes correspondentes. Após o treinamento, um novo conjunto de dados — sem os rótulos visíveis ao modelo — é usado para testar sua capacidade de prever corretamente essas classificações, permitindo o cálculo de métricas como acurácia e precisão.

Já no aprendizado não supervisionado, os dados utilizados não possuem rótulos. O modelo tenta identificar agrupamentos naturais nos dados com base em semelhanças entre os atributos. Esse tipo de abordagem é útil para descobrir estruturas ocultas ou padrões ainda não conhecidos, especialmente em cenários exploratórios.

Neste trabalho, adotaremos exclusivamente técnicas de aprendizado supervisionado com foco em classificação astronômica, em que os objetos já estão identificados como estrelas, galáxias ou quasares. Sendo assim, as seções seguintes tratarão exclusivamente de métodos supervisionados. Para leitores que desejem explorar o aprendizado não supervisionado com mais profundidade, recomendamos a obra de Celebi & Aydin (2016), que oferece uma análise abrangente sobre o tema.

2.2.3 Conjuntos de Treinamento e Teste

No desenvolvimento de modelos de Machine Learning, é essencial dividir os dados disponíveis em dois subconjuntos distintos: treinamento e teste. O primeiro é responsável por ensinar o modelo a reconhecer padrões nos dados, enquanto o segundo é utilizado para avaliar se o aprendizado foi bem-sucedido (GERON, 2019).

O conjunto de treinamento serve para que o algoritmo estabeleça relações entre os atributos dos dados (features) e os seus respectivos rótulos (labels). Por exemplo, ao observar determinados valores em bandas espectrais, o modelo aprende a associá-los com classes como estrela, galáxia ou quasar.

Após esse processo, o conjunto de teste é utilizado para verificar se o modelo é capaz de aplicar corretamente o que aprendeu, classificando dados novos de forma precisa. O desempenho do modelo é medido por métricas como a acurácia, que indica o quão bem o modelo foi capaz de generalizar o conhecimento adquirido.

Uma prática comum é reservar cerca de 70% dos dados para o treinamento e 30% para o teste, embora essa proporção possa variar dependendo do volume de dados disponível e da complexidade do problema. Esse ajuste é facilmente configurável nas bibliotecas de ML utilizadas durante a modelagem.

2.2.4 Underfitting e Overfitting

Em projetos de Machine Learning, dois conceitos fundamentais relacionados ao desempenho do modelo são o *underfitting* e o *overfitting*. Eles representam, respectivamente, os extremos da capacidade de aprendizado: a insuficiência e o excesso de ajuste aos dados ((GERON, 2019)).

O *underfitting* ocorre quando o modelo é incapaz de capturar a estrutura dos dados durante o treinamento. Isso geralmente se deve a uma arquitetura muito simples ou a um número insuficiente de parâmetros. Nessa situação, o algoritmo não consegue identificar padrões relevantes, resultando em baixa acurácia tanto no conjunto de treino quanto no de teste. Em outras palavras, o modelo não aprende o suficiente para realizar classificações eficazes.

Por outro lado, o *overfitting* acontece quando o modelo aprende em demasia os detalhes do conjunto de treinamento, incluindo ruídos e valores atípicos (outliers). Apesar de obter ótimos resultados durante o treinamento, seu desempenho no conjunto de teste é fraco, pois ele não consegue generalizar os padrões para dados novos. O modelo torna-se "especialista" na base de dados de treino, mas falha ao lidar com situações diferentes.

A meta ideal é atingir um equilíbrio entre esses dois extremos. Um bom modelo deve apresentar desempenho consistente em diferentes conjuntos de dados, mantendo acurácia elevada não só no treinamento, mas também na validação.

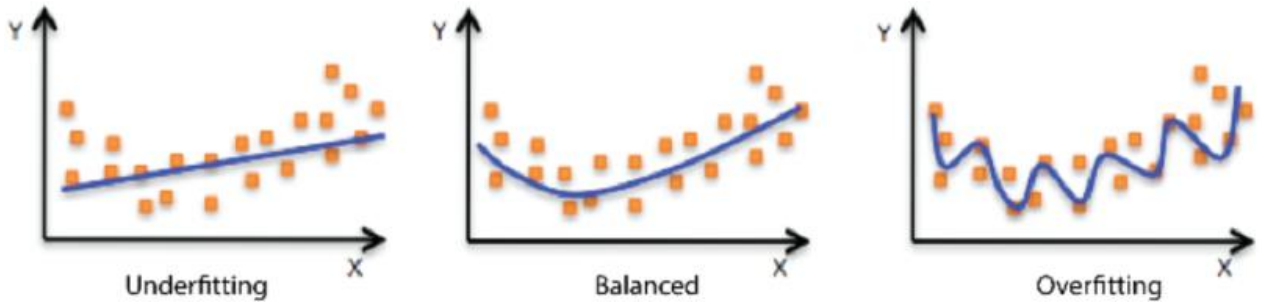


FIGURA 3. Ilustrações que indicam os casos de *Underfitting*, de *overfitting* e um cenário ideal. Fonte: ALVES, J. C (2022)

Esse equilíbrio pode ser visualizado por meio de gráficos que relacionam o erro do modelo com sua complexidade. À medida que a complexidade aumenta, o erro de treinamento tende a diminuir. No entanto, após certo ponto, o erro no conjunto de teste começa a crescer, indicando o *overfitting*. Essa representação ajuda a definir o nível adequado de complexidade que evita tanto o subajuste quanto o sobreajuste.

Portanto, durante o desenvolvimento de modelos, é essencial monitorar as métricas de desempenho em ambos os conjuntos — treino e teste — e buscar um aprendizado que preserve a capacidade de generalização.

2.2.5 Matriz de Confusão

A matriz de confusão é uma ferramenta essencial na avaliação de modelos de classificação em Machine Learning. Ela permite visualizar, de forma detalhada, como o modelo se saiu ao prever as classes dos dados após o treinamento.

Essa matriz é estruturada como um quadro quadrado, onde cada linha representa as previsões feitas pelo modelo, e cada coluna indica os rótulos reais dos dados. Cada célula mostra a quantidade de vezes em que uma classe real foi atribuída a uma classe prevista específica.

Para exemplificar, podemos utilizar um cenário binário com duas classes genéricas: positivo e negativo. Nesse caso, a matriz terá dimensão 2×2, composta por:

- Verdadeiros Positivos (VP): o modelo acertou ao prever a classe positiva;
- Verdadeiros Negativos (VN): o modelo acertou ao prever a classe negativa;
- Falsos Positivos (FP): o modelo previu positivo, mas a classe real era negativa;
- Falsos Negativos (FN): o modelo previu negativo, mas a classe real era positiva.

Essas categorias aparecem organizadas de forma que os acertos (VP e VN) ficam na diagonal principal da matriz — geralmente destacada em verde — enquanto os erros (FP e FN) aparecem fora dessa diagonal, sendo comumente destacados em vermelho.

A matriz ideal seria aquela em que apenas os valores da diagonal principal estivessem preenchidos, ou seja, todas as previsões do modelo coincidem com as classificações reais. Qualquer valor fora da diagonal representa uma falha na predição.

Nas seções seguintes, especialmente no Capítulo 5 – Resultados, serão apresentadas as matrizes de confusão obtidas ao aplicar os modelos aos dados astronômicos. Elas nos ajudarão a compreender melhor o comportamento do modelo em relação às diferentes classes de objetos celestes.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

FIGURA 4. Matriz de Confusão para uma dada classe.

2.2.6 Precision, Recall ou F1-Score?

Ao avaliar o desempenho de um modelo de classificação, é comum começar pela acurácia, que representa a proporção de previsões corretas em relação ao total de amostras analisadas. Essa métrica é definida pela seguinte fórmula:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \quad (2)$$

Em que:

- VP = Verdadeiros Positivos
- VN = Verdadeiros Negativos
- FP = Falsos Positivos
- FN = Falsos Negativos

Entretanto, a acurácia pode ser insuficiente em cenários com classes desbalanceadas, pois não considera separadamente os tipos de erro cometidos pelo modelo. Nestes casos, é fundamental observar outras métricas mais específicas: precisão, recall e F1-score.

- Precisão (*Precision*): Mede a proporção de acertos entre todas as previsões positivas feitas pelo modelo. Em outras palavras, indica quantas das amostras que foram classificadas como positivas de fato pertencem àquela classe.

$$Precisao = \frac{VP}{VP + FP} \quad (3)$$

- Recall (Sensibilidade): Indica a capacidade do modelo de identificar corretamente todos os elementos de uma classe específica. Mede a proporção de verdadeiros positivos em relação ao total de elementos que realmente pertencem àquela classe.

$$Recall = \frac{VP}{VP + FN} \tag{4}$$

- F1-Score: Representa a média harmônica entre a precisão e o recall. Essa métrica é especialmente útil quando se deseja equilibrar os dois aspectos — evitar tanto falsos positivos quanto falsos negativos.

$$F_1 = \frac{2}{\frac{1}{precisao} + \frac{1}{recall}} = 2 \times \frac{precisao \times recall}{precisao + recall} = \frac{VP}{VP + \frac{FN+FP}{2}} \tag{5}$$

Para entender melhor a escolha entre essas métricas, (GERON, 2019) apresenta um exemplo ilustrativo:

“Se você está desenvolvendo um modelo para classificar vídeos seguros para crianças, talvez prefira uma alta precisão (mesmo que o modelo deixe de classificar alguns vídeos bons) a um alto recall (que pode permitir que vídeos impróprios sejam aceitos).” GÉRON (2019)

No contexto deste trabalho, optamos por adotar o F1-score como principal métrica de avaliação, pois buscamos um modelo que minimize ambos os tipos de erro — falsos positivos e falsos negativos —, proporcionando uma classificação mais precisa e equilibrada dos objetos astronômicos.

Com isso, encerramos a apresentação dos fundamentos teóricos de avaliação de desempenho em Aprendizado de Máquina. A seguir, serão detalhados os modelos supervisionados aplicados nesta pesquisa.

2.3 Modelos Utilizados

A presente pesquisa emprega uma variedade de modelos de aprendizado de máquina para a tarefa de classificação de objetos astronômicos. Esta seção detalha os algoritmos utilizados, começando pelo K-Nearest Neighbors (KNN), um método de aprendizado supervisionado não paramétrico. Em seguida, explora-se o Support Vector Machine (SVM), um modelo poderoso para classificação linear e não linear. A Árvore de Decisão e o Random Forest, um ensemble de árvores de decisão, também serão apresentados. Por fim, será introduzido o MultiLayer Perceptron, uma rede neural artificial fundamental para o aprendizado profundo, fornecendo uma visão geral das ferramentas de modelagem que serão aplicadas na classificação dos dados astronômicos.

2.3.1 K-Nearest Neighbors (KNN)

O K-Nearest Neighbors (KNN) emerge como um dos algoritmos de aprendizado de máquina (ML) mais intuitivos e fundamentais, cuja simplicidade reside na sua abordagem direta à classificação. Seu princípio central é a atribuição de classe a uma nova amostra não classificada com base na maioria das classes de seus *K* vizinhos mais próximos no espaço de características, utilizando a distância euclidiana como métrica de proximidade (COVER; HART, 1967).

O processo de classificação utilizando o KNN, ilustrado conceitualmente na FIGURA 4.5, pode ser decomposto nos seguintes passos:

1. Definição do Número de Vizinhos (K): Inicialmente, um valor para o parâmetro *K* é estabelecido. Este hiperparâmetro crucial determina a quantidade de vizinhos com classificações conhecidas que serão considerados na decisão final. A escolha de um valor apropriado para *K* impacta diretamente a capacidade de generalização do modelo, com valores muito pequenos tornando o modelo sensível a ruídos nos dados e valores muito grandes podendo suavizar demais as fronteiras de decisão, levando a um desempenho inferior (COVER; HART, 1967).

2. Cálculo das Distâncias: Uma vez definido o valor de *K*, para cada nova amostra a ser classificada (representada na FIGURA 5 como um ponto de interrogação), são calculadas as distâncias euclidianas entre ela e todas as outras amostras no conjunto de dados de treinamento com classificações previamente conhecidas.

3. Identificação dos K Vizinhos Mais Próximos: Após o cálculo das distâncias, os *K* pontos no conjunto de treinamento que apresentam as menores distâncias em relação à amostra não classificada são identificados como os *K* vizinhos mais próximos.

4. Determinação da Classificação Final: A classificação final da amostra desconhecida é determinada pela classe majoritária entre esses *K* vizinhos mais próximos. Em outras palavras, a classe que aparece com maior frequência entre os *K* vizinhos vota na classificação da nova amostra. Em casos de empate na votação das classes, estratégias como a escolha aleatória ou a ponderação dos votos pela distância podem ser empregadas.

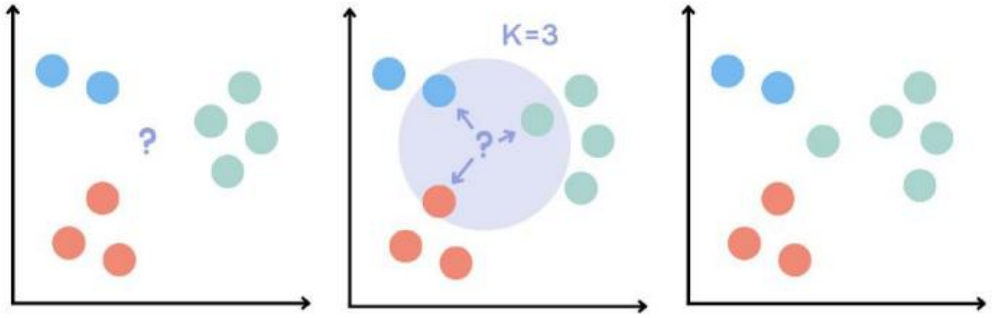


FIGURA 5. Funcionamento do KNN utilizando 3 vizinhos representados no círculo.

2.3.2 Suport Vector Machine (SVM)

O Support Vector Machine (SVM) representa um modelo de aprendizado de máquina de grande versatilidade e eficácia, amplamente empregado em tarefas de classificação devido à sua notável capacidade de adaptação a diferentes distribuições de dados. Uma característica distintiva do SVM é sua aptidão para efetuar separações tanto lineares quanto não lineares entre as classes, flexibilidade essa alcançada através da seleção da função kernel apropriada para delinear o comportamento da fronteira de decisão (QUINLAN, 1986).

Em um cenário simplificado de classificação binária com classes linearmente separáveis, o SVM opera buscando a hiperplano de separação que maximize a margem entre as duas classes. A FIGURA 6 ilustra esse conceito. Essa proximidade as torna suscetíveis a erros de classificação diante da introdução de novas amostras, potencialmente localizadas ligeiramente além dessas fronteiras estreitas.

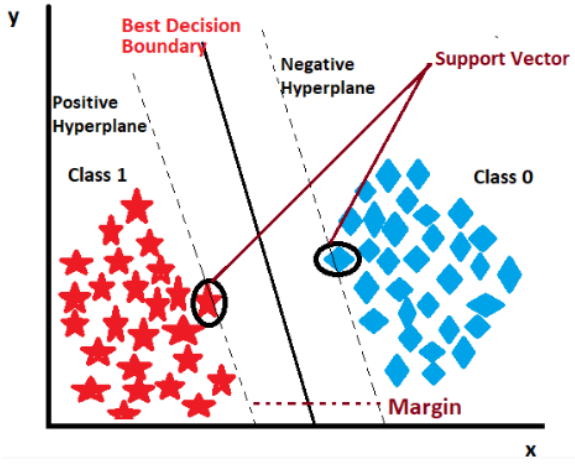


FIGURA 6. Exemplo de separacão de dados de duas classes. Fonte: STATUSNEO, 2023

Em contraste, o SVM fundamenta sua abordagem na identificação e utilização dos chamados *vetores de suporte* (*support vectors*). Esses vetores são representados pelos pontos de dados mais próximos das amostras de classes opostas (evidenciados por círculos na

imagem acima). Os vetores de suporte desempenham um papel crucial na definição da margem máxima de separação entre as classes. O hiperplano ótimo é então construído de forma a equidistar dos vetores de suporte de ambas as classes, estabelecendo a maior distância possível entre elas. Essa maximização da margem confere ao SVM uma robustez inerente contra pequenas perturbações nos dados e contribui para uma melhor capacidade de generalização para amostras não vistas durante o treinamento.

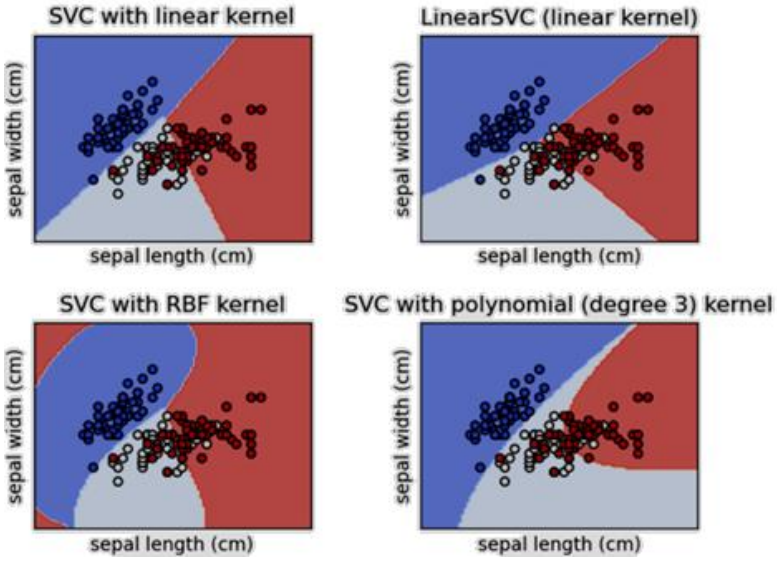


FIGURA 7 – Funcionamento do KNN utilizando 3 vizinhos representados no círculo. Fonte: SCIKIT-LEARN, 2025.

A eficácia do SVM reside, portanto, na sua capacidade de identificar os pontos de dados mais informativos (os vetores de suporte) e construir uma fronteira de decisão otimizada em relação a esses pontos críticos. Ao maximizar a margem, o SVM busca criar um espaço de segurança entre as classes, minimizando o risco de classificações incorretas e promovendo um modelo com maior poder preditivo em dados futuros.

2.3.3 Árvore de Decisão

As Árvores de Decisão constituem uma classe de modelos de aprendizado de máquina intuitivos e poderosos, capazes de realizar tarefas de classificação através da construção de uma estrutura hierárquica de decisões baseadas em regras sequenciais sobre as características dos dados, (GERON, 2019). A representação do modelo assemelha-se a um fluxograma, onde cada nó interno representa um teste sobre um atributo, cada ramo representa o resultado desse teste, e cada nó folha (terminal) representa uma decisão ou uma classificação final.

O processo de classificação com uma Árvore de Decisão envolve percorrer a árvore a partir do nó raiz, seguindo os ramos correspondentes aos valores das características da amostra a ser classificada em cada nó de decisão. Esse processo continua até que se alcance um nó folha, cuja etiqueta define a classe prevista para a amostra.

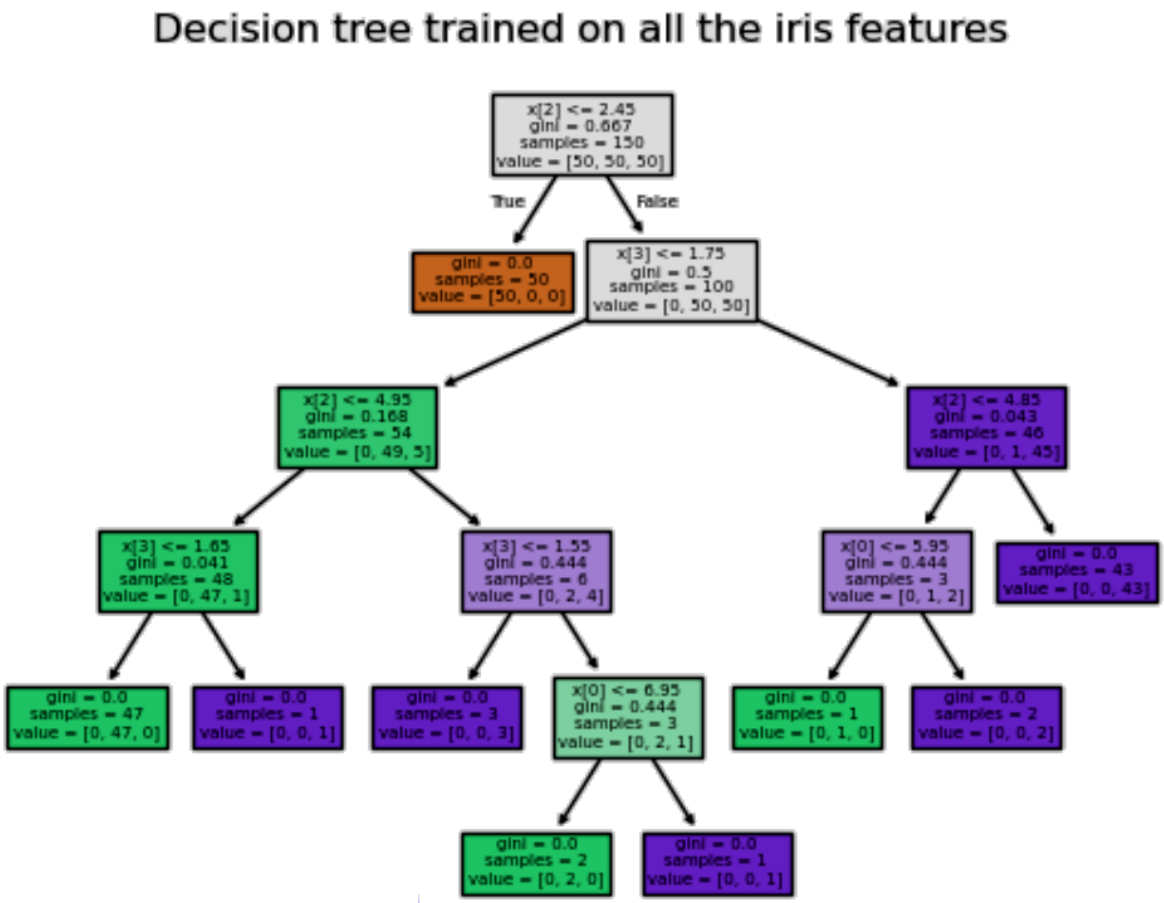


FIGURA 8. Representação de uma arvore com resultados em cada nó folha. Fonte: SCIKIT-LEARN, 2025.

A construção de uma Árvore de Decisão durante a fase de treinamento envolve a seleção recursiva das melhores características para dividir os dados em subconjuntos mais homogêneos em relação à variável alvo (a classe). O objetivo é encontrar as divisões que maximizem o ganho de informação ou minimizem a impureza dos nós resultantes. Métricas comuns para avaliar a qualidade de uma divisão incluem a Entropia e o Índice de Gini.

$$GiniIndex = 1 - \sum_j p_j^2 \qquad Entropy = -\sum_j p_j . \log_2(p_j) \tag{6}$$

Dado que pj é a probabilidade de obtermos a classe j em ambos os casos. O índice de Gini mede o quanto um item escolhido de forma aleatória será classificado incorretamente, e o ´índice de Entropia indica a desordem do conjunto de acordo com a classe pretendida. No

entanto, é importante testar o comportamento de ambos para descobrir qual melhor se adaptará ao problema.

1. Seleção da Melhor Divisão: No nó inicial (raiz) e em cada nó subsequente, o algoritmo avalia todas as características disponíveis e todos os possíveis pontos de divisão para cada característica. A divisão que resulta na maior redução da impureza (ou maior ganho de informação) nos nós filhos é selecionada como a melhor divisão para aquele nó.
2. Criação dos Nós Filhos: Uma vez determinada a melhor divisão, o nó atual é dividido em dois ou mais nós filhos, cada um correspondendo a um possível resultado do teste na característica selecionada.
3. Processo Recursivo: O processo de seleção da melhor divisão e criação dos nós filhos é repetido recursivamente para cada nó filho, até que um critério de parada seja atingido. Critérios de parada comuns incluem atingir uma profundidade máxima predefinida para a árvore, ter um número mínimo de amostras em um nó, ou quando um nó se torna "puro" (contém amostras de apenas uma classe).
4. Atribuição das Classes nas Folhas: Quando o processo de divisão para um nó é interrompido, esse nó se torna uma folha. A classe atribuída a essa folha é tipicamente a classe majoritária das amostras de treinamento que caíram nesse nó.

Uma das principais vantagens das Árvores de Decisão reside na sua interpretabilidade. A estrutura hierárquica de regras é facilmente compreensível e permite identificar quais características são mais importantes na decisão de classificação e como elas interagem para produzir a previsão final. Além disso, as Árvores de Decisão não exigem muitas suposições sobre a distribuição dos dados e podem lidar com dados categóricos e numéricos.

No entanto, as Árvores de Decisão também podem ser propensas ao *overfitting*, especialmente se a árvore for permitida a crescer muito profundamente, aprendendo ruídos e detalhes específicos do conjunto de treinamento que não se generalizam bem para dados não vistos. Técnicas de *pruning* (poda) são frequentemente aplicadas para reduzir a complexidade da árvore, removendo ramos menos significativos e melhorando a capacidade de generalização do modelo.

Em nosso problema de classificação espectral de objetos astronômicos, uma Árvore de Decisão poderia aprender a classificar quasares, galáxias e estrelas com base em limiares em diferentes comprimentos de onda ou em características derivadas do espectro. A estrutura da árvore resultante revelaria quais faixas espectrais ou combinações delas são mais discriminativas para distinguir entre as três classes de objetos, oferecendo *insights* sobre as

diferenças espectrais chave entre eles. A interpretabilidade do modelo permitiria analisar as regras de decisão específicas que levam à classificação de cada tipo de objeto astronômico.

2.3.4 Random Forest

O Random Forest é um algoritmo de aprendizado de máquina poderoso e amplamente utilizado para tarefas de classificação (e regressão), que se baseia no princípio de *ensemble learning*. Em vez de construir uma única Árvore de Decisão, o Random Forest constrói uma "floresta" de múltiplas árvores de decisão independentes e combina suas previsões para obter uma classificação final mais robusta e precisa, Breiman, L. (2001).

A força do Random Forest reside em dois conceitos chave: o *bagging* (bootstrap aggregating) e a seleção aleatória de características. Esses mecanismos introduzem diversidade entre as árvores da floresta, reduzindo a correlação entre elas e, consequentemente, diminuindo a variância e o risco de *overfitting* inerente às árvores de decisão individuais. O processo de construção e classificação com um Random Forest envolve:

1. Criação de Múltiplos Conjuntos de Treinamento (Bootstrap): Para cada árvore, um subconjunto aleatório com reposição do conjunto de treinamento original é criado, introduzindo variabilidade nas amostras de treinamento para cada árvore.
2. Seleção Aleatória de Características: Em cada nó de cada árvore, apenas um subconjunto aleatório de características é considerado para encontrar a melhor divisão, garantindo a não correlação entre as árvores.
3. Construção das Árvores: Múltiplas árvores de decisão são construídas usando os conjuntos de treinamento bootstrap e a seleção aleatória de características, geralmente sem *pruning* extensivo.
4. Agregação das Previsões: Para classificar uma nova amostra, a classe final é determinada pela votação majoritária das previsões de todas as árvores na floresta.

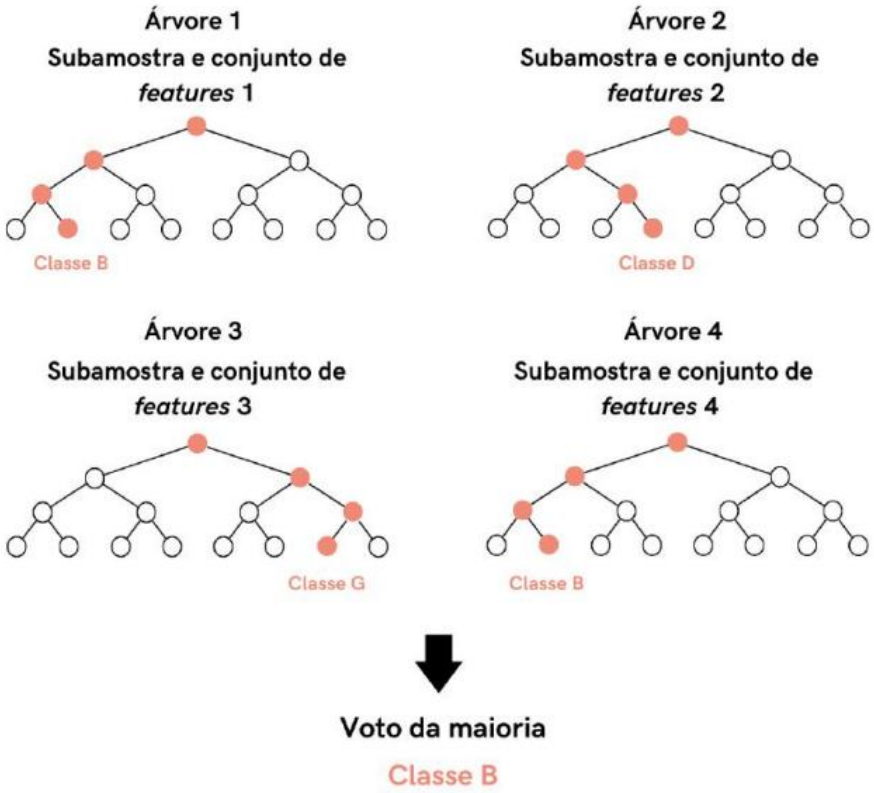


FIGURA 9. Representação de uma classificação com Random forest. Fonte: ALVES, J. C (2022)

A combinação das previsões de múltiplas árvores não correlacionadas leva a um modelo mais robusto e com melhor capacidade de generalização do que uma única árvore de decisão. O Random Forest tende a ser menos sensível a ruídos nos dados e menos propenso ao *overfitting*. Além da alta precisão e robustez, o Random Forest oferece outros benefícios importantes:

- Estimativa da Importância das Características: O algoritmo pode fornecer uma medida da importância relativa de cada característica na tarefa de classificação, baseada em como a precisão do modelo diminui quando os valores daquela característica são aleatoriamente permutados.
- Estimativa do Erro de Generalização (OOB Error): Como mencionado anteriormente, as amostras não incluídas no conjunto de treinamento *bootstrap* para cada árvore (as amostras OOB) podem ser usadas para estimar o desempenho do modelo em dados não vistos, sem a necessidade de um conjunto de validação separado.

Em nosso projeto de classificação espectral de objetos astronômicos, o Random Forest poderia ser aplicado utilizando as características extraídas dos espectros como entrada. A floresta aprenderia a classificar quasares, galáxias e estrelas através da combinação das decisões de múltiplas árvores, cada uma treinada em um subconjunto aleatório dos dados e

considerando um subconjunto aleatório das características espectrais. A capacidade do Random Forest de lidar com dados de alta dimensionalidade e identificar características importantes o torna uma ferramenta valiosa para analisar a complexidade dos espectros astronômicos e distinguir entre as diferentes classes de objetos. Além disso, a estimativa da importância das características poderia revelar quais regiões do espectro são mais informativas para a classificação.

2.3.5 Multi Layer Perceptron

O Multi-Layer Perceptron (MLP) representa uma classe fundamental de redes neurais artificiais *feedforward*, amplamente utilizada para tarefas de classificação complexas devido à sua capacidade de aprender relações não lineares intrincadas entre as características de entrada e as classes de saída, GOODFELLOW, Ian (Deep Learning). Um MLP consiste em múltiplas camadas de nós interconectados (neurônios), incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída.

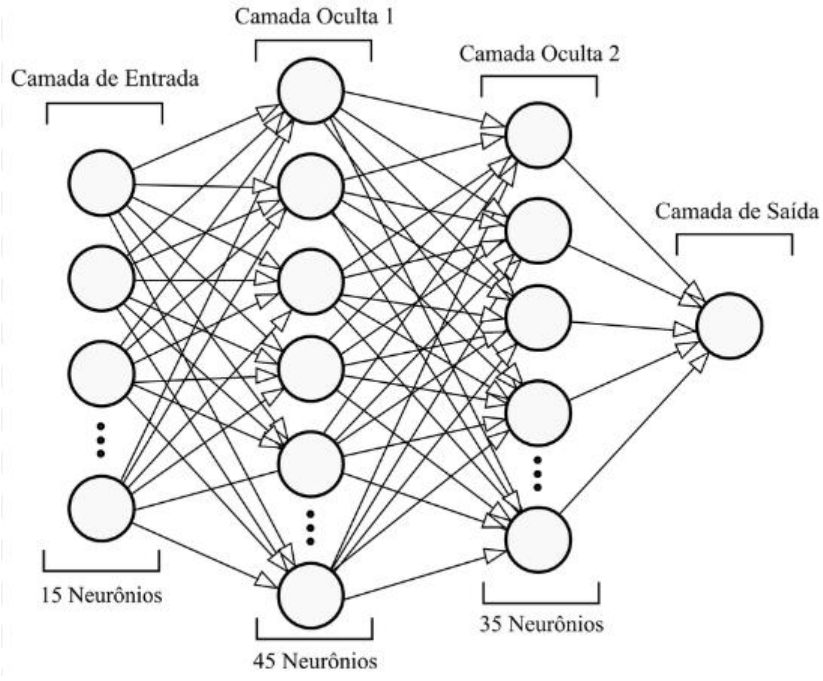


FIGURA 10. Arquitetura de uma MLP com 2 camadas ocultas. Fonte: ResearchGate, 2015

A arquitetura de um MLP permite que o modelo aprenda representações hierárquicas dos dados. Cada neurônio em uma camada recebe sinais dos neurônios da camada anterior, realiza uma transformação linear desses sinais (multiplicando por pesos e adicionando um bias) e, em seguida, aplica uma função de ativação não linear ao resultado. Essa não linearidade é crucial para permitir que o MLP modele relações complexas que não podem ser

capturadas por modelos lineares. O processo de classificação com um MLP envolve as seguintes etapas:

1. Propagação Forward (Forward Pass): Uma amostra de entrada é apresentada à camada de entrada. A ativação se propaga através das camadas ocultas, com cada neurônio calculando sua saída com base nas ativações da camada anterior, seus pesos sinápticos e sua função de ativação. Finalmente, a camada de saída produz uma saída, que no caso de classificação, geralmente representa as probabilidades de a amostra pertencer a cada uma das classes, GOODFELLOW, Ian (Deep Learning). Funções de ativação comuns na camada de saída para classificação incluem a função Softmax (S(xi)) (para classificação multi-classe) e a função Sigmoid (para classificação binária).

$$s\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{7}$$

2. Cálculo da Função de Perda (Loss Function): A saída do MLP é comparada com a classe real da amostra de treinamento utilizando uma função de perda (ou função de custo). A função de perda quantifica o erro entre a previsão do modelo e o valor real. Funções de perda comuns para classificação incluem a Entropia Cruzada (H) e o Erro Quadrático Médio (MSE), GOODFELLOW, Ian (Deep Learning).

$$H = -\sum p(x)\log p(x) \tag{8}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2 \tag{9}$$

3. Propagação Backward (Backward Pass) e Otimização: O erro calculado na etapa anterior é propagado de volta através da rede, camada por camada, utilizando o algoritmo de retropropagação (backpropagation). Durante a retropropagação, os gradientes da função de perda em relação aos pesos da rede são calculados. Esses gradientes indicam a direção na qual os pesos devem ser ajustados para minimizar o erro, GOODFELLOW, Ian (Deep Learning). Um algoritmo de otimização, como o Gradiente Descendente (Gradient Descent) ou suas variantes (e.g., Adam, RMSprop), utiliza esses gradientes para atualizar os pesos da rede de forma iterativa, buscando

encontrar uma configuração de pesos que minimize a função de perda no conjunto de treinamento.

4. Iteração e Convergência: As etapas de propagação forward, cálculo da perda e retropropagação com otimização são repetidas por um número suficiente de épocas (iterações) até que o modelo convirja para um estado onde o erro no conjunto de treinamento (e idealmente em um conjunto de validação independente) seja minimizado.

A arquitetura de um MLP, incluindo o número de camadas ocultas e o número de neurônios em cada camada, bem como a escolha das funções de ativação e do algoritmo de otimização, são hiperparâmetros que precisam ser definidos e ajustados (tunados) para obter o melhor desempenho em um problema específico.

Em nosso problema de classificação espectral de objetos astronômicos, um MLP poderia aprender a identificar padrões complexos nos espectros que discriminam entre quasares, galáxias e estrelas. As camadas ocultas da rede aprenderiam representações abstratas das características espectrais, permitindo ao modelo realizar classificações precisas mesmo em casos em que as relações entre as características e as classes não são lineares ou facilmente discerníveis por modelos mais simples. A capacidade do MLP de modelar não linearidades o torna uma ferramenta poderosa para lidar com a variabilidade e a complexidade dos dados espectrais astronômicos.

.

3 TRABALHOS RELACIONADOS

Nesta seção, são apresentados trabalhos relacionados à proposta desenvolvida neste estudo. A classificação automática de objetos astronômicos tem ganhado destaque na literatura científica com o avanço dos métodos de aprendizado de máquina e o crescimento exponencial de dados provenientes de levantamentos astronômicos como o Sloan Digital Sky Survey (SDSS). Neste cenário, diversas abordagens vêm sendo propostas para lidar com problemas como a distinção entre estrelas, galáxias e quasares, bem como a caracterização física de objetos celestes com base em atributos espectrais e fotométricos. A análise da literatura demonstra a relevância do tema, os desafios envolvidos e as contribuições das abordagens baseadas em inteligência artificial para a astrofísica moderna.

No trabalho de Zhang et al. (2019), os autores propõem uma abordagem de classificação de objetos astronômicos utilizando redes neurais profundas treinadas sobre dados espectroscópicos e fotométricos extraídos do SDSS. A base de dados empregada contém cerca de 300 mil registros, contemplando informações como magnitudes *ugriz*, redshift e medidas espectrais. O modelo desenvolvido utiliza uma arquitetura do tipo MLP (Perceptron Multicamadas), avaliado com validação cruzada de 10-folds. Os resultados demonstraram uma acurácia superior a 97% na distinção entre estrelas, galáxias e quasares.

Em um estudo mais recente, Pérez-Durán et al. (2022) investigaram a utilização de aprendizado profundo com transferência de conhecimento para a classificação de espectros astronômicos. Os pesquisadores utilizaram redes convolucionais (CNNs) aplicadas diretamente aos espectros normalizados, extraídos do SDSS DR16. O modelo foi treinado com espectros de galáxias e estrelas, atingindo uma acurácia média de 95,6%. O diferencial do estudo está no pré-processamento detalhado dos espectros e na normalização por comprimento de onda, além da aplicação de *data augmentation*. Apesar do bom desempenho, o trabalho concentra-se exclusivamente em espectros, sem combinar informações fotométricas e espectrais — estratégia explorada no presente TCC com o objetivo de enriquecer o conjunto de características e aumentar a robustez do modelo preditivo.

Outro trabalho relevante é o de Fustes et al. (2020), que realizaram uma análise estatística e preditiva de objetos do SDSS com foco em identificar padrões em índices espectrais, como D4000, Hδ_A e [O III], com o auxílio de algoritmos de árvore de decisão e Random Forest. A base utilizada foi composta por cerca de 100 mil espectros estelares classificados manualmente, permitindo o treinamento supervisionado dos modelos. O estudo destaca a importância dos índices espectrais como discriminadores de classes estelares, e sugere que seu uso pode aumentar significativamente a precisão de modelos baseados em dados tabulares. Esse trabalho oferece uma base sólida para o uso de variáveis físicas derivadas

de espectros, abordagem que também é adotada neste TCC, em conjunto com atributos fotométricos, para uma classificação mais precisa e explicável.

Em relação aos trabalhos revisados, este TCC se diferencia ao propor uma abordagem que integra múltiplos tipos de dados — espectroscópicos, fotométricos e índices físicos derivados — em uma única arquitetura de rede neural do tipo MLP, aplicada a uma base de dados massiva (com cerca de um milhão de registros combinados). Além disso, é avaliado o impacto de diferentes combinações de *features* sobre o desempenho do modelo, bem como a aplicação de técnicas modernas de normalização e divisão estratificada dos dados. A proposta visa não apenas alcançar alta acurácia na classificação, mas também contribuir para a interpretabilidade dos modelos em contexto astronômico, explorando o potencial dos índices espectrais como marcadores físicos relevantes. Ao realizar essa integração de dados e técnicas, este trabalho contribui para o avanço da aplicação de IA na astrofísica observacional, com foco na reprodutibilidade e escalabilidade dos métodos.

4 MÉTODO DE PESQUISA

Para alcançar os objetivos propostos, esta pesquisa seguirá uma abordagem baseada em aprendizado profundo, integrando técnicas de processamento e análise de grandes volumes de dados astronômicos obtidos do Sloan Digital Sky Survey (SDSS). O fluxo metodológico será composto por oito etapas principais, conforme descrito na FIGURA 11.

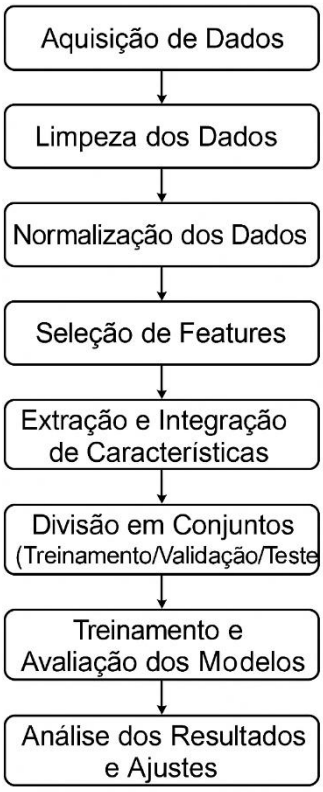


FIGURA 11. Metodologia - Fonte: Autoria Própria, 2025

4.1 Dados Utilizados

Neste trabalho, foram utilizadas duas bases de dados astronômicos provenientes do Sloan Digital Sky Survey (SDSS), cada uma contendo aproximadamente 500.000 registros reais, com informações fundamentais para a classificação e análise de objetos celestes. Essas bases foram obtidas a partir de diferentes tabelas do SDSS, combinando atributos fotométricos, espectroscópicos e físicos. Para fins ilustrativos e para permitir a reprodutibilidade parcial das análises apresentadas, duas amostras representativas desses conjuntos de dados foram

extraídas e encontram-se anexadas neste trabalho sob os títulos: “Classificação de objetos astronômicos – Amostra.xlsx” e “Classificação de espectros estelares – Amostra.xlsx”.

Querys utilizadas:
https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/tree/main/Querys%20Used%20on%20SDSS%20to%20Extract%20Data

Amostras:
https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/tree/main/Data%20Samples

A primeira base, denominada Classificação de Objetos Astronômicos, é composta por dados de aproximadamente 500 mil objetos classificados como estrelas, galáxias e quasares. Cada registro contém coordenadas celestes (RA, Dec), magnitudes aparentes em cinco bandas fotométricas do SDSS (u, g, r, i, z), além de informações do pipeline espectroscópico, como identificador specobjid, redshift espectroscópico, e metadados da observação (run, rerun, camcol, field, plate, mjd, fiberid). A variável-alvo desta base é a classe do objeto astronômico, registrada na coluna class, que pode assumir os valores "STAR", "GALAXY" ou "QSO" (quasar).

A segunda base, intitulada “Classificação de Espectros Estelares”, contém aproximadamente 500 mil espectros estelares, com um foco mais aprofundado nas propriedades físicas das estrelas. Além das magnitudes nas bandas u, g, r, i, z, esta base inclui medidas espectrais de linhas de emissão como [O II] λ3726, [O III] λ5007 e Hα, com seus respectivos fluxos (oii_3726_flux, oiii_5007_flux, h_alpha_flux), métricas de qualidade de ajuste (chisq) e resolução instrumental. Também estão disponíveis valores de redshift, embora, neste contexto, o foco seja na caracterização estelar, como temperatura, metalicidade e idade estimada das populações estelares, a partir de índices espectrais derivados.

As magnitudes fotométricas nas bandas ugriz seguem o sistema AB e são obtidas a partir de medições da intensidade da luz captada por filtros específicos, com sensibilidade a diferentes faixas do espectro eletromagnético. Estas bandas são fundamentais para a construção de diagramas cor-magnitude e para a inferência de propriedades físicas como temperatura efetiva e tipo espectral. No contexto da classificação automática, as diferenças entre essas magnitudes (por exemplo, u - g, g - r) são frequentemente utilizadas como features auxiliares que ajudam a distinguir entre objetos com características espectrais distintas.

As variáveis espectrais relacionadas às linhas de emissão representam fluxos medidos em pontos específicos do espectro óptico, associados a elementos químicos e processos físicos específicos nas atmosferas estelares ou regiões de formação estelar em galáxias. Por exemplo, a linha Hα (hidrogênio alfa) é um marcador clássico de atividade cromosférica e de formação estelar, enquanto as linhas de oxigênio ([O II] e [O III]) estão associadas a nebulosas e regiões ionizadas. As colunas complementares de chi-quadrado e resolução instrumental fornecem contexto estatístico e técnico sobre a qualidade das medições espectrais, permitindo avaliar a confiabilidade dos dados utilizados nos modelos.

Essas duas bases são complementares: enquanto a primeira oferece uma visão global de diferentes tipos de objetos astronômicos e serve como base para tarefas de classificação multiclasse, a segunda permite uma análise mais refinada das estrelas individualmente, com potencial para abordagens de classificação hierárquica ou de regressão voltadas para propriedades físicas específicas. A diversidade e riqueza dos atributos disponíveis tornam essas bases altamente adequadas para aplicações de aprendizado de máquina, especialmente em tarefas supervisionadas que requerem uma combinação de variáveis contínuas, categóricas e derivadas de medidas físicas reais. A utilização conjunta desses dados também permite investigar a performance de diferentes arquiteturas de rede neural e abordagens de feature engineering aplicadas ao domínio da astrofísica.

Por fim, vale destacar que os dados foram extraídos diretamente da plataforma pública do SDSS através de queries SQL personalizadas, elaboradas com base nas tabelas PhotoObjAll, SpecObjAll, galSpecInfo, galSpecLine e outras complementares. Os scripts SQL utilizados para essa extração também estão anexados a este trabalho, com os nomes “Magnetudes_PhotoObj_SpecObj.sql” e “Espectros_PhotoObjAll_SpecObjAll_galSpecInfo_galSpecLine.sql”. Esses arquivos documentam de forma completa os critérios de seleção, filtros aplicados e os joins utilizados para construção das bases, permitindo a replicação da coleta dos dados originais.

4.2 Aquisição e Pré-processamento de Dados

Os dados a serem utilizados serão extraídos do SDSS por meio de comandos SQL aplicados sobre tabelas como galSpecExtra, galSpecInfo, galSpecLine, PhotoObjAll, SpecObjAll, SpecObj e PhotoObj, abrangendo um amplo conjunto de variáveis espectrais, fotométricas e físicas.

4.2.1 Limpeza dos Dados

Nesta etapa, serão removidas entradas que apresentem valores ausentes (NULL), vazios ou inconsistentes, a fim de garantir a integridade do conjunto de dados. Esse processo incluirá:

- Verificação e exclusão de registros com campos obrigatórios ausentes, como redshift (z), magnitude absoluta (absmag_r) ou parâmetros físicos essenciais (mass_tot_p50, sfr_tot_p50);
- Tratamento de variáveis com valores inválidos ou fora de faixas fisicamente plausíveis (como magnitudes negativas ou redshifts negativos);
- Eliminação de duplicatas e registros corrompidos, quando identificados;
- Revisão dos avisos de qualidade fornecidos pelo SDSS, como a flag zwarning, para identificação de entradas que possam comprometer a análise.

Esse processo será conduzido com base em critérios estatísticos e no conhecimento prévio das propriedades físicas e espectrais dos objetos.

4.2.2 Normalização dos Dados

As variáveis contínuas serão normalizadas utilizando o método de padronização estatística (StandardScaler), com média zero e desvio padrão igual a um. Essa transformação é necessária para garantir que todas as variáveis contribuam de forma equitativa durante o treinamento dos algoritmos, evitando distorções causadas por escalas distintas.

4.3 Seleção de Features

A seleção de atributos será orientada por critérios astrofísicos e estatísticos. Com base na análise exploratória das tabelas mencionadas, serão escolhidas características relevantes para a tarefa de classificação, tais como:

- Magnitudes modeladas em bandas ópticas (modelMag_u, g, r, i, z);
- Magnitude absoluta na banda r (absmag_r);
- Redshift espectroscópico (z);

- Parâmetros físicos derivados, como mass_tot_p50, sfr_tot_p50, dn4000, vdisp, fracDev_r, e snMedian_r;

- Intensidades de linhas espectrais importantes, como Ha_6564, Hb_4862, [OIII], [NII].

As variáveis com baixa variância estatística serão descartadas. Também será aplicada uma análise de correlação para eliminar atributos redundantes ou colineares, otimizando o vetor de entrada para os modelos de classificação.

4.3.1 Extração e Integração de Características Físicas e Espectrais

A extração de características será realizada com base nos espectros ópticos de média resolução e nos parâmetros calculados pelo pipeline do SDSS. As informações extraídas serão integradas em um vetor unificado de atributos para cada objeto astronômico, combinando:

- Dados fotométricos multibanda;
- Informações espectrais derivadas de linhas de emissão;
- Indicadores de evolução estelar (ex: índice Dn4000);
- Parâmetros físicos como massa estelar, taxa de formação estelar, concentração de luz, entre outros.

Essa estrutura de dados multidimensional servirá como entrada para os modelos de classificação que serão implementados nas etapas posteriores.

4.4 Treinamento, Validação e Avaliação dos Modelos

Os dados processados serão divididos em subconjuntos para treinamento, validação e teste, respeitando a proporção típica de 70% para treino, 15% para validação e 15% para teste. Será adotada validação cruzada estratificada do tipo *k-fold* (k = 5) para garantir a robustez dos modelos. As métricas que serão utilizadas para avaliar o desempenho incluem:

- Acurácia;
- Precisão (Precision);
- Revocação (Recall);
- F1-Score;
- Matriz de confusão.

Essas métricas permitirão comparar a capacidade de generalização dos modelos aplicados e embasar a escolha da abordagem final para classificação dos objetos astronômicos e todas já foram apresentadas detalhadamente na sessão 2.1.6.

4.5 Interpretação dos Resultados e Ajustes

Após o treinamento e avaliação, os resultados serão analisados criticamente para identificar padrões de erro e oportunidades de melhoria. Poderão ser realizados ajustes nos seguintes aspectos:

- Arquitetura dos modelos de classificação utilizados;
- Seleção e transformação das variáveis preditoras;
- Estratégias de balanceamento de classes, se necessário;
- Aplicação de técnicas de regularização, como *Dropout* ou penalidades L2, para evitar sobreajuste (*overfitting*).

O objetivo será obter um modelo de classificação eficaz, capaz de distinguir com precisão entre estrelas, galáxias e quasares, bem como realizar a subclassificação de estrelas com base em suas propriedades espectrais.

.

5 CONCLUSÕES

5.1 Conclusão Parte 1: Classificação de Objetos Astronômicos

5.1.1 Contextualização

A primeira etapa deste trabalho teve como foco o desenvolvimento de um modelo robusto de aprendizado de máquina capaz de realizar a classificação automática de objetos astronômicos em três categorias fundamentais: estrelas, galáxias e quasares (QSOs). Esta tarefa representa um dos desafios mais críticos e relevantes para a astronomia moderna, especialmente diante da explosão exponencial de dados gerados por levantamentos espectroscópicos de larga escala como o Sloan Digital Sky Survey (SDSS), o Dark Energy Survey (DES) e os futuros projetos como o Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST).

A magnitude deste desafio pode ser dimensionada pela quantidade de dados envolvida: apenas o SDSS, em suas diferentes fases, já catalogou espectros de mais de 4 milhões de objetos astronômicos, com previsões de que futuros surveys possam gerar dados de bilhões de objetos celestes na próxima década. A classificação manual destes objetos, tradicionalmente realizada por astrônomos especializados através de análise visual e interpretação espectroscópica, tornou-se absolutamente inviável em termos de tempo, recursos humanos e custos operacionais.

Além da questão de escala, a automação dessa classificação representa um avanço significativo em múltiplas dimensões: escalabilidade computacional, velocidade de processamento, padronização metodológica, redução de vieses subjetivos e reprodutibilidade científica. A implementação de sistemas automatizados permite não apenas lidar com volumes massivos de dados, mas também garantir consistência analítica, eliminando variações decorrentes de diferentes interpretações humanas e estabelecendo critérios objetivos e quantitativos para a classificação.

Do ponto de vista astrofísico, a distinção precisa entre estrelas, galáxias e quasares é fundamental para diversos campos de pesquisa. Estrelas representam objetos relativamente próximos em nossa galáxia, cujo estudo permite compreender processos de formação e evolução estelar, nucleossíntese e dinâmica galáctica. Galáxias, por sua vez, são sistemas complexos compostos por bilhões de estrelas, gás e matéria escura, cuja análise estatística revela informações cruciais sobre a estrutura em larga escala do universo, formação de estruturas cósmicas e evolução cosmológica. Quasares, os objetos mais energéticos do universo observável, são buracos negros supermassivos em processo de acreção ativa, servindo como laboratórios naturais para física de alta energia e marcos de distância cosmológica.

5.1.2 Metodologia de Pré-processamento e Engenharia de Atributos

O processo metodológico iniciou-se com uma análise exploratória abrangente dos dados, seguida pela seleção criteriosa das variáveis espectrais e fotométricas mais relevantes para a tarefa de classificação. Esta etapa fundamental foi conduzida através de um modelo exploratório baseado em árvores de decisão, que permitiu avaliar quantitativamente a capacidade discriminativa de cada atributo disponível no dataset.

O conjunto inicial de dados continha aproximadamente 20 variáveis diferentes, incluindo magnitudes fotométricas brutas (u, g, r, i, z), índices de cor derivados, parâmetros espectrais, coordenadas celestes e identificadores únicos. A utilização indiscriminada de todas essas variáveis apresentava riscos significativos: aumento desnecessário da dimensionalidade, introdução de ruído, correlações espúrias, aumento do tempo computacional e, potencialmente, degradação do desempenho por sobreajuste.

Para mitigar esses riscos, implementou-se uma estratégia sistemática de seleção de atributos baseada em importância estatística. Utilizou-se um modelo preliminar de Random Forest treinado com o conjunto completo de features, aproveitando a capacidade intrínseca deste algoritmo de calcular a importância relativa de cada variável através da medida de redução de impureza nos nós das árvores de decisão.

5.1.3 Análise de Importância de Atributos

A biblioteca scikit-learn oferece o atributo feature_importances_, que quantifica matematicamente o quanto cada variável contribui para a redução da impureza (geralmente medida pelo índice de Gini ou entropia) durante o processo de construção das árvores. Esta métrica varia entre 0 e 1, sendo que valores próximos a 1 indicam alta relevância discriminativa, enquanto valores próximos a 0 sugerem baixa capacidade de separação entre classes.

O gráfico de importância gerado a partir dessa análise revelou padrões extremamente informativos e consistentes com o conhecimento astrofísico estabelecido. A variável redshift emergiu, de longe, como a mais influente na tarefa de classificação, com uma importância relativa de aproximadamente 0.45, significativamente superior a todos os demais atributos. Este resultado é altamente consistente com a teoria cosmológica, uma vez que o redshift é diretamente proporcional à distância dos objetos e, consequentemente, um discriminador natural entre objetos próximos (estrelas da Via Láctea) e objetos extragalácticos distantes (galáxias e quasares).

Em seguida, destacaram-se as variáveis derivadas de cores espectrais: r_i, g_r e u_g, com importâncias relativas variando entre 0.08 e 0.12. Estes índices de cor são amplamente

reconhecidos na literatura astrofísica por sua forte correlação com propriedades físicas fundamentais dos objetos, incluindo temperatura superficial, metalicidade, idade estelar, tipo espectral e taxa de formação estelar. A relevância destes atributos para a classificação automática valida décadas de conhecimento empírico desenvolvido pela comunidade astronômica.

As variáveis fotométricas brutas (g, z, u, i) também figuraram entre as dez mais importantes, embora com menor peso relativo (importâncias entre 0.03 e 0.06). Embora estas magnitudes sejam dependentes da distância e, portanto, menos discriminativas que os índices de cor, ainda carregam informação valiosa sobre as características espectrais intrínsecas dos objetos.

Conforme esperado, o atributo specobjid, que atua exclusivamente como identificador único dos espectros, apresentou importância praticamente nula (< 0.001), confirmando sua irrelevância para a classificação e validando a robustez da análise.

5.1.4 Seleção Final de Atributos e índices de cor

Com base nesta análise quantitativa, decidiu-se utilizar apenas as seis variáveis mais influentes para compor o dataset final de treinamento. Esta escolha estratégica não apenas simplificou significativamente o modelo, como também aumentou sua interpretabilidade científica, melhorou o desempenho computacional, reduziu o ruído estatístico e minimizou riscos de sobreajuste.

A utilização de índices de cor (diferenças entre magnitudes de bandas espectrais adjacentes) representa uma prática consolidada e teoricamente fundamentada na astronomia observacional. Estes índices oferecem várias vantagens metodológicas cruciais:

Invariância à distância: Ao contrário das magnitudes absolutas, os índices de cor são independentes da distância do objeto (desde que corrigidos para extinção interestelar), tornando-os ideais para classificação de objetos a diferentes distâncias cosmológicas.

Sensibilidade física: Os índices de cor são diretamente relacionados à distribuição espectral de energia (SED - Spectral Energy Distribution) dos objetos, refletindo propriedades físicas fundamentais como temperatura, composição química, idade e taxa de formação estelar.

Robustez observacional: São menos suscetíveis a erros sistemáticos de calibração fotométrica, uma vez que utilizam razões entre medidas tomadas simultaneamente com o mesmo instrumento.

Compatibilidade com modelos teóricos: Podem ser facilmente comparados com predições de modelos de síntese estelar, evolução galáctica e física de quasares, facilitando a interpretação astrofísica dos resultados.

As variáveis selecionadas foram:

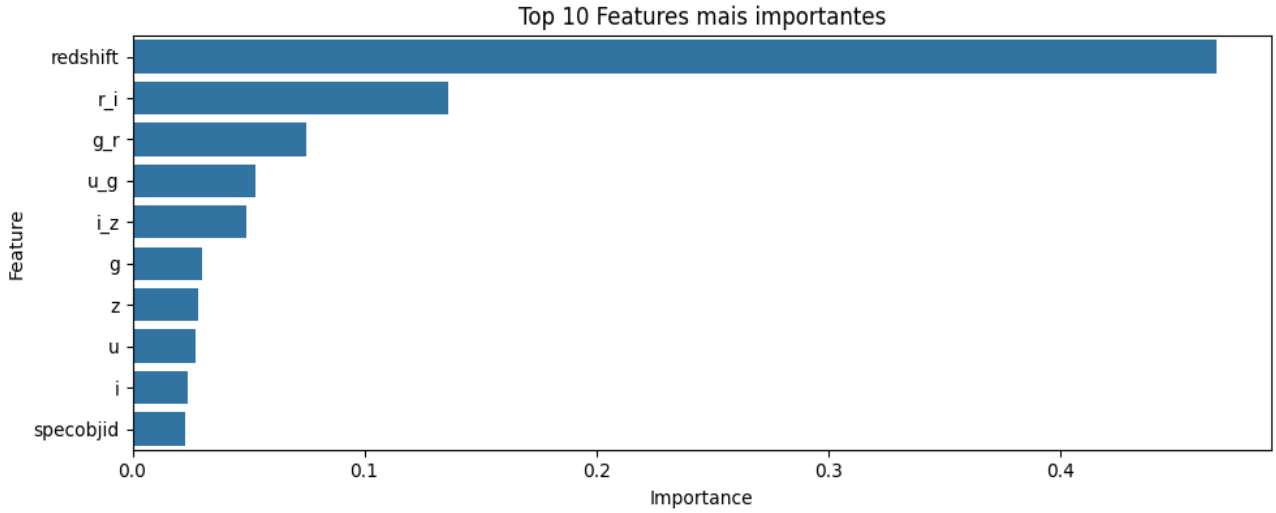


FIGURA 12. Importância das variáveis utilizadas no modelo de classificação espectral estelar. As features foram ordenadas com base na contribuição relativa para a acurácia do modelo. *(Fonte: elaboração própria).*

1. Redshift (z): Representa o desvio espectral para o vermelho causado pela expansão cósmica, sendo diretamente proporcional à velocidade de recessão e, pela Lei de Hubble, à distância luminosa do objeto. Para estrelas locais da Via Láctea, o redshift é essencialmente zero ($z \approx 0$), enquanto galáxias típicas apresentam valores entre $0.01 < z < 1.0$, e quasares podem exibir redshifts extremamente elevados ($z > 1.0$, frequentemente $z > 3.0$). Esta variável é, portanto, o discriminador primário entre objetos locais e extragalácticos.

2. r_i: Diferença de magnitude entre os filtros r (vermelho, $\lambda \approx 620$ nm) e i (infravermelho próximo, $\lambda \approx 750$ nm). Este índice de cor é especialmente sensível à temperatura superficial dos objetos e à presença de características espectrais na região do vermelho, incluindo linhas de hidrogênio, hélio e metais. Para estrelas, varia sistematicamente com o tipo espectral, sendo útil para distinguir entre populações estelares quentes (tipos O, B, A) e frias (tipos K, M). Para galáxias, reflete a composição estelar dominante e a idade da população.

3. g_r: Diferença entre os filtros g (verde, $\lambda \approx 480$ nm) e r (vermelho), tradicionalmente utilizada para estimativas de temperatura, metalicidade e tipo espectral. Este índice é particularmente eficaz para distinguir entre diferentes classes de estrelas e também para separar estrelas de galáxias, uma vez que estas últimas apresentam distribuições de cor características decorrentes de suas populações estelares compostas.

4. u_g: Diferença entre os filtros u (ultravioleta próximo, $\lambda \approx 350$ nm) e g (verde). Este índice é especialmente sensível a objetos com emissão energética no ultravioleta, sendo frequentemente associado a quasares (devido à radiação de buracos negros ativos), estrelas

jovens e quentes (tipos O e B), e galáxias com intensa formação estelar. A banda u, por ser mais suscetível à extinção por poeira, também carrega informação sobre o meio interestelar.

5. i_z: Diferença entre os filtros i (λ ≈ 750 nm) e z (infravermelho, λ ≈ 900 nm). Este índice no infravermelho próximo é valioso para identificar objetos com excesso de emissão em comprimentos de onda longos, incluindo estrelas anãs vermelhas frias, galáxias com alta extinção por poeira, e objetos a altos redshifts cujas características espectrais são deslocadas para o vermelho devido à expansão cósmica.

6. g: Magnitude aparente no filtro g (verde), representando o brilho observado do objeto nesta banda espectral específica. Embora seja dependente da distância, esta magnitude carrega informação importante sobre a luminosidade intrínseca e as características espectrais do objeto na região verde do espectro, incluindo linhas de absorção metálicas e o continuum estelar.

5.1.5 Estratégias de Balanceamento de Dados

O conjunto de dados original apresentava um desbalanceamento severo entre as classes, reflexo da distribuição natural de objetos no universo observável e dos critérios de seleção dos levantamentos espectroscópicos. A classe "estrela" constituía aproximadamente 60% do dataset (≈ 600.000 instâncias), refletindo a facilidade relativa de observar objetos brilhantes em nossa própria galáxia. A classe "galáxia" representava cerca de 30% das amostras (≈ 300.000 instâncias), enquanto os "quasares" constituíam apenas 10% do total (≈ 100.000 instâncias).

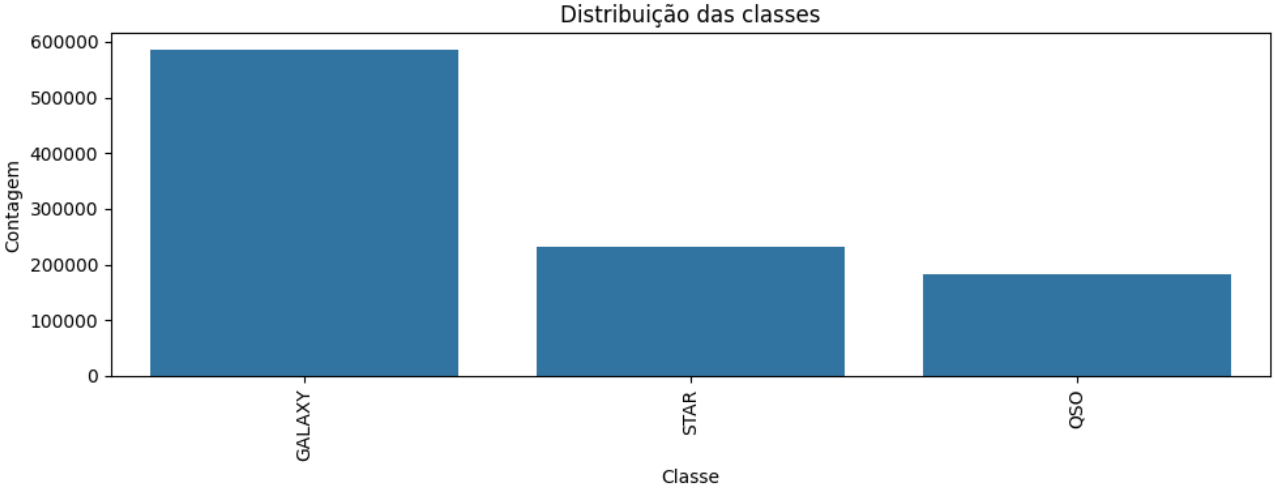


FIGURA 13. Distribuição de classes no dataset original. *(Fonte: elaboração própria).*

Este desbalanceamento representa um desafio metodológico significativo para algoritmos de aprendizado de máquina, que tendem naturalmente a ser enviesados em favor

da classe majoritária. Sem correção, modelos treinados em dados desbalanceados frequentemente desenvolvem estratégias de classificação conservadoras, classificando sistematicamente instâncias ambíguas como pertencentes à classe majoritária, resultando em baixo recall para classes minoritárias.

5.1.6 Implementação da Técnica SMOTE

Para mitigar este problema, implementou-se a técnica SMOTE (Synthetic Minority Over-sampling Technique), uma abordagem sofisticada de balanceamento que gera exemplos sintéticos das classes minoritárias através de interpolação inteligente no espaço de características.

O algoritmo SMOTE opera através dos seguintes passos metodológicos:

Identificação de vizinhos próximos: Para cada instância da classe minoritária, identifica-se os k vizinhos mais próximos (tipicamente k = 5) no espaço multidimensional das características, utilizando a distância euclidiana como métrica.

Interpolação linear: Seleciona-se aleatoriamente um dos vizinhos identificados e gera-se uma nova instância sintética através de interpolação linear entre a instância original e o vizinho selecionado, utilizando um fator aleatório $\lambda \in [0,1]$.

Preservação da distribuição: O processo garante que as instâncias sintéticas mantenham as propriedades estatísticas fundamentais da classe original, incluindo correlações entre variáveis e estrutura topológica do espaço de características.

Balanceamento progressivo: O processo é repetido até que todas as classes atinjam aproximadamente o mesmo número de instâncias da classe majoritária.

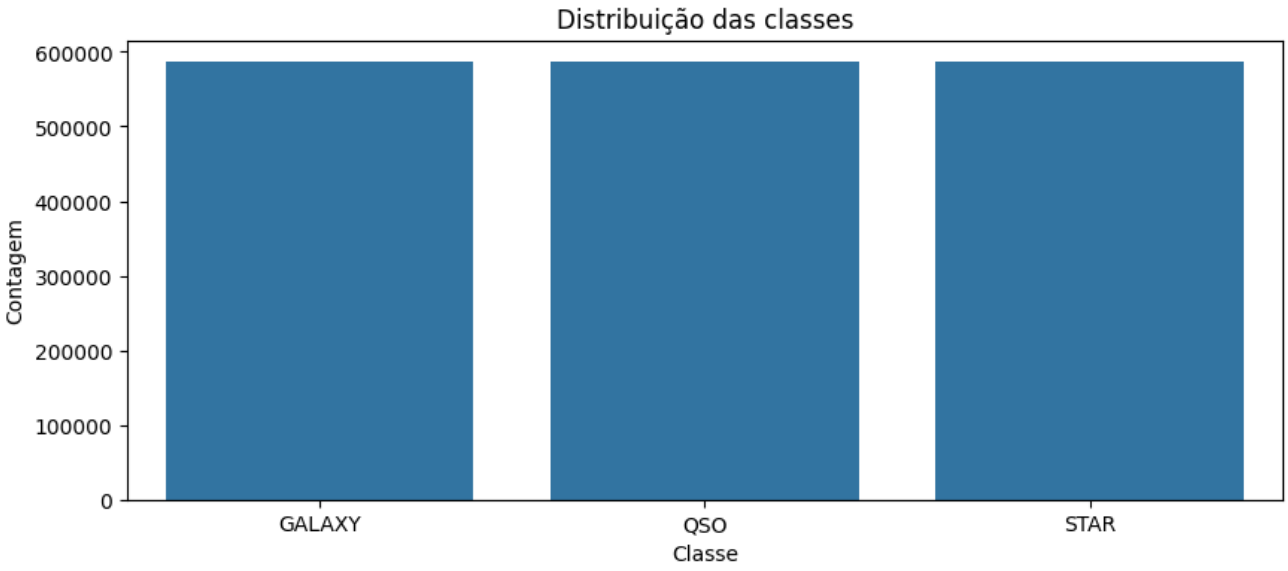


FIGURA 14. Distribuição de classes no dataset após balanceamento. *(Fonte: elaboração própria).*

5.1.7 Resultados do Balanceamento

A aplicação do SMOTE produziu resultados quantitativos impressionantes:

Antes do balanceamento: 999.949 instâncias totais
Estrelas: ≈ 600.000 (60%)
Galáxias: ≈ 300.000 (30%)
Quasares: ≈ 100.000 (10%)

Após o balanceamento: 1.758.261 instâncias totais
Estrelas: ≈ 586.087 (33.3%)
Galáxias: ≈ 586.087 (33.3%)
Quasares: ≈ 586.087 (33.3%)

O aumento de aproximadamente 75% na base de dados foi fundamental para que os modelos pudessem aprender padrões estatísticos significativos de forma equitativa entre as classes. Este balanceamento resultou em ganhos substanciais nas métricas de desempenho, particularmente em recall, precisão e F1-score para a classe dos quasares, que passou de praticamente ignorada pelos modelos iniciais para adequadamente reconhecida.

5.1.8 Estratégia de Treinamento e Validação

A abordagem adotada para o treinamento e a avaliação dos modelos em ambas as partes desta pesquisa seguiu rigorosos princípios de validação estatística, implementando validação cruzada estratificada com 2 folds (CV_FOLDS = 2). Esta escolha, aparentemente conservadora, foi motivada pelo grande volume de dados disponível em nosso dataset, o que permite obter estimativas estatisticamente robustas mesmo com um número reduzido de folds, além de reduzir significativamente o tempo computacional necessário para os experimentos. A estratificação garantiu que cada fold mantivesse aproximadamente a mesma proporção de classes observada no dataset completo, evitando vieses estatísticos decorrentes de partições desbalanceadas. Este procedimento é especialmente crítico em problemas de classificação multiclasse, onde variações na distribuição entre folds podem levar a estimativas enviesadas de desempenho.

5.1.9 Seleção de Métricas de Avaliação

A métrica principal escolhida para guiar o processo de otimização de hiperparâmetros em ambas as fases do projeto foi a F1-weighted (F1 ponderado). Esta é uma versão sofisticada do tradicional F1-score que pondera o desempenho do modelo de acordo com o suporte (número de exemplos) de cada classe. Esta métrica é particularmente adequada para contextos de classificação desbalanceada, pois garante que o modelo seja otimizado considerando tanto o desempenho individual em cada classe quanto a importância relativa de cada classe no dataset, promovendo um equilíbrio robusto entre precisão e recall em contextos multiclasse assimétricos.

5.1.10 Otimização Computacional

Para garantir a eficiência e a robustez dos experimentos computacionais em ambas as partes desta pesquisa, foram implementadas estratégias avançadas de otimização. Todos os modelos foram treinados utilizando execução paralela (n_jobs = -1), aproveitando a totalidade dos núcleos de processamento disponíveis na infraestrutura. Esta abordagem de paralelização demonstrou-se particularmente eficaz para algoritmos baseados em ensemble, como o Random Forest, onde as árvores constituintes podem ser treinadas de forma independente e simultânea, acelerando significativamente o processo. Adicionalmente, otimizações de memória, como o uso de tipos de dados de menor precisão (float32 em vez de float64), estratégias de cache inteligente e técnicas otimizadas de garbage collection, foram aplicadas para gerenciar eficientemente grandes volumes de dados e evitar recomputações desnecessárias.

5.1.11 Metodologia de Busca de Hiperparâmetros

O processo de otimização de hiperparâmetros, crucial para maximizar o desempenho dos modelos, foi conduzido por meio da técnica RandomizedSearchCV. Essa metodologia, escolhida e aplicada consistentemente tanto na Parte 1 quanto na Parte 2 do trabalho, realiza uma busca aleatória e estatisticamente eficiente no espaço de hiperparâmetros predefinido. A escolha do RandomizedSearchCV sobre a busca exaustiva (GridSearchCV) oferece vantagens significativas, incluindo maior eficiência computacional, permitindo a exploração de um espaço de hiperparâmetros muito mais amplo com o mesmo orçamento; flexibilidade estatística, ao incorporar distribuições contínuas e discretas; robustez estatística, minimizando o risco de overfitting nos próprios hiperparâmetros; e escalabilidade, com desempenho que não degrada exponencialmente com o aumento da dimensionalidade do espaço de busca. Esta

uniformidade metodológica entre as fases do estudo assegura a coerência e a comparabilidade dos resultados obtidos.

O modelo Random Forest, que emergiu como o melhor classificador nesta etapa, é um algoritmo de ensemble baseado em árvores de decisão que combina as predições de múltiplas árvores independentes através de votação majoritária (para classificação) ou média aritmética (para regressão). Sua eficácia deriva de dois princípios estatísticos fundamentais:

Bootstrap Aggregating (Bagging): Cada árvore é treinada em uma amostra bootstrap diferente do dataset original, introduzindo diversidade entre os classificadores base e reduzindo a variância do modelo final.

Random Feature Selection: Em cada nó de cada árvore, apenas um subconjunto aleatório das características está disponível para a decisão de split, introduzindo diversidade adicional e reduzindo a correlação entre árvores.

5.1.12 Espaço de Hiperparâmetros Explorado

O processo de otimização explorou um espaço multidimensional de hiperparâmetros cuidadosamente definido com base em conhecimento teórico e experiência prática:

n_estimators: [200, 500, 800] — Controla o número de árvores no ensemble. Valores maiores geralmente melhoram o desempenho até um ponto de saturação, mas aumentam linearmente o custo computacional. O range escolhido equilibra desempenho e eficiência.

max_depth: [None, 20, 40] — Limita a profundidade máxima de cada árvore. None permite crescimento irrestrito até que critérios de parada sejam atingidos. Profundidades limitadas podem prevenir overfitting, especialmente em datasets menores.

max_features: ["sqrt", 0.3, 0.5] — Determina o número de características consideradas em cada split. "sqrt" utiliza $\sqrt{p}$ características (sendo p o número total), uma heurística eficaz para problemas de classificação. Os valores 0.3 e 0.5 representam frações fixas do total de características.

min_samples_leaf: [1, 2] — Especifica o número mínimo de amostras necessárias em cada folha. Valores maiores previnem overfitting ao forçar generalizações em regiões esparsas do espaço de características.

class_weight: ["balanced", None] — "balanced" ajusta automaticamente os pesos das classes inversamente proporcionais às suas frequências, compensando desbalanceamentos residuais mesmo após SMOTE. None utiliza pesos unitários.

5.1.13 Configuração Otimizada Final

Após o processo de busca aleatória, o modelo final convergiu para a seguinte configuração ótima:

python

```
{
 'clf__n_estimators': 500,
 'clf__min_samples_leaf': 2,
 'clf__max_features': 'sqrt',
 'clf__max_depth': None,
 'clf__class_weight': 'balanced'
}
```

Esta configuração revela aspectos interessantes sobre a natureza do problema:

```
[2025-07-14 19:34:18] Iniciando tuning: random_forest
Fitting 2 folds for each of 15 candidates, totalling 30 fits
[2025-07-14 19:35:19] Modelo salvo: /dbfs/FileStore/classification/astronomical_objects/silver/balanceado/modelos/random_forest_best.pkl
Tempo: 58.11s | F1_w: 0.9746 | F1_macro: 0.9682 | Acc: 0.9747 | Prec_w: 0.9745 | Rec_w: 0.9747
F1 por classe: [0.9801 0.9312 0.9933]
Melhores parâmetros: {'clf__n_estimators': 500, 'clf__min_samples_leaf': 2, 'clf__max_features': 'sqrt', 'clf__max_depth': None, 'clf__class
_weight': 'balanced'}
```

FIGURA 15. Métricas e configuração final do modelo escolhido. *(Fonte: elaboração própria).*

n_estimators = 500: Indica que o desempenho continua melhorando com ensembles maiores, sugerindo que o problema se beneficia da diversidade adicional proporcionada por mais árvores.

max_depth = None: A ausência de limite de profundidade sugere que o dataset é suficientemente grande e complexo para suportar árvores profundas sem overfitting significativo.

max_features = 'sqrt': Com 6 características totais, utiliza √6 ≈ 2.4, efetivamente 2 características por split, uma escolha que equilibra diversidade e poder discriminativo.

min_samples_leaf = 2: Previne folhas com apenas uma amostra, reduzindo memorização de outliers e melhorando generalização.

class_weight = 'balanced': Mantém balanceamento ativo mesmo após SMOTE, demonstrando a importância de múltiplas estratégias de balanceamento.

5.1.14 Interpretação da Matriz de Confusão

A matriz de confusão normalizada gerada pelo modelo oferece insights valiosos sobre os padrões de erro e as limitações residuais do classificador. A análise diagonal revela a alta taxa de acerto para cada classe, enquanto os elementos fora da diagonal mostram os tipos de confusão mais frequentes.

As confusões mais comuns observadas foram:

Galáxias ↔ Quasares: Representa a confusão mais frequente entre classes minoritárias, fisicamente justificável pelos núcleos galácticos ativos que podem apresentar características espectrais intermediárias entre galáxias normais e quasares típicos.

Estrelas ↔ Galáxias: Confusões menos frequentes, geralmente envolvendo estrelas peculiares ou galáxias compactas com redshifts muito baixos.

Estrelas ↔ Quasares: Extremamente raras, ocorrendo principalmente para quasares em redshifts muito baixos ou estrelas com características espectrais anômalas.

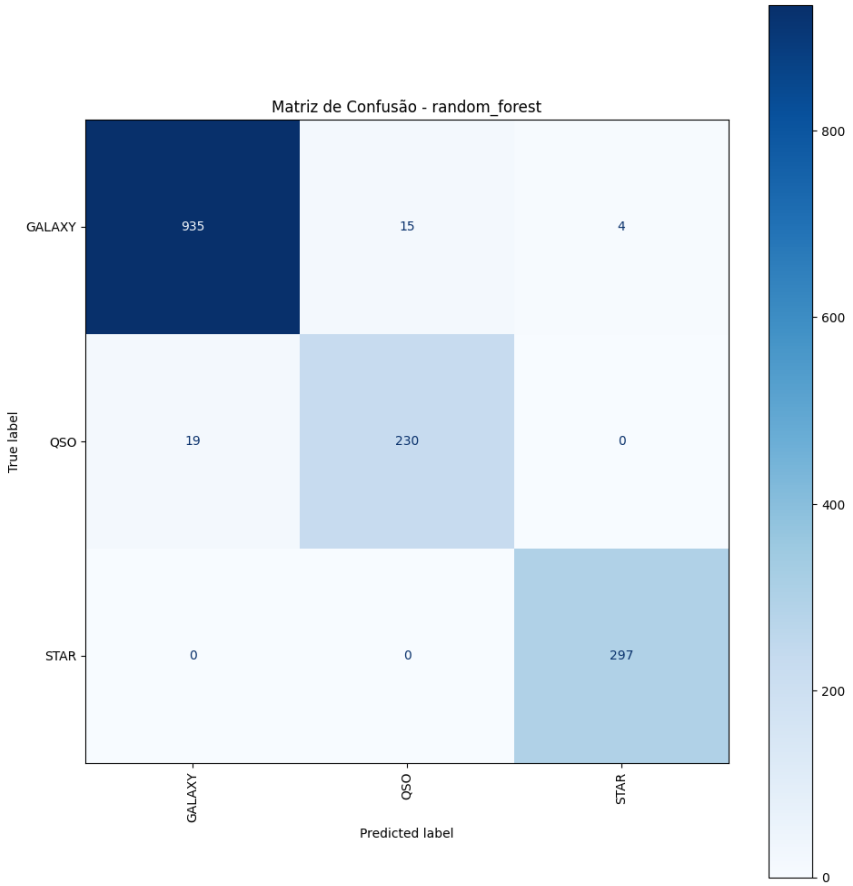


FIGURA 16. Matriz confusão do modelo de classificação de objetos astronômicos. *(Fonte: elaboração própria).*

5.1.15 Parte 1: Conclusão final

A primeira etapa deste trabalho abordou um desafio fundamental na astronomia moderna: a classificação automática de objetos astronômicos em três categorias primárias — estrelas, galáxias e quasares (QSOs). Diante do volume exponencial de dados gerados por levantamentos espectroscópicos de larga escala como o Sloan Digital Sky Survey (SDSS), que já catalogou milhões de objetos, a classificação manual tornou-se inviável. A automação deste processo não apenas garante escalabilidade e velocidade de processamento, mas também padroniza critérios, elimina vieses subjetivos e assegura a reprodutibilidade científica. A distinção precisa entre estes objetos é crucial para diversos campos de pesquisa, desde a evolução estelar em nossa galáxia até a compreensão da estrutura em larga escala do universo e a física de buracos negros supermassivos ativos.

Tendo em vista esse cenário, buscou-se responder à seguinte pergunta de pesquisa (Q1): “Qual a capacidade dos modelos de aprendizado de máquina, em especial as Redes Neurais Multicamadas (MLPs), de classificar objetos astronômicos (estrelas, galáxias e quasares) e classificar espectros estelares a partir de dados espectroscópicos do SDSS?”.

Os resultados obtidos demonstram que modelos supervisionados como Random Forest e MLPs são altamente eficazes na tarefa de classificação astronômica, alcançando métricas de desempenho superiores a 97% em F1-score e acurácia, mesmo em cenários com classes desbalanceadas. A aplicação de técnicas de balanceamento como o SMOTE mostrou-se essencial para garantir o aprendizado equitativo entre as classes, especialmente para os quasares, frequentemente subestimados. Além disso, os modelos conseguiram capturar relações astrofísicas complexas entre variáveis como redshift e índices de cor, evidenciando o potencial dessas ferramentas para acelerar e padronizar processos de classificação espectral em astronomia.

A metodologia empregada iniciou-se com uma análise exploratória e seleção criteriosa de atributos. Utilizando um modelo preliminar de Random Forest, a análise de importância de atributos revelou o redshift como a variável mais influente, com importância de aproximadamente 0.45, um discriminador natural entre objetos locais (estrelas, $z \approx 0$) e extragalácticos (galáxias e quasares, $z > 0$). Em seguida, os índices de cor derivados (r_i, g_r, u_g), com importâncias entre 0.08 e 0.12, destacaram-se por sua forte correlação com propriedades físicas dos objetos, como temperatura superficial e tipo espectral. A seleção final de seis variáveis mais influentes (redshift e os cinco índices de cor/magnitude: r_i, g_r, u_g, i_z e g) simplificou o modelo, aumentou sua interpretabilidade e melhorou o desempenho computacional, validando décadas de conhecimento empírico astrofísico.

Um dos principais desafios superados foi o severo desbalanceamento de classes no conjunto de dados original: estrelas representavam aproximadamente 60% (≈600.000

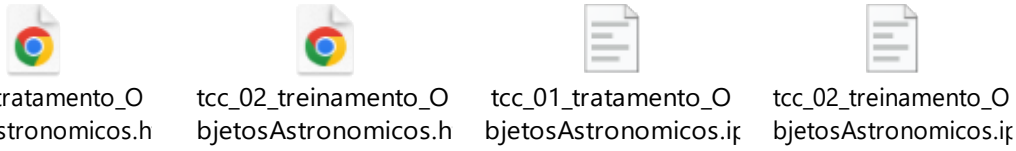
instâncias), galáxias 30% (≈300.000) e quasares apenas 10% (≈100.000) de um total de 999.949 instâncias. Para mitigar esse viés, a técnica SMOTE (Synthetic Minority Over-sampling Technique) foi implementada, gerando exemplos sintéticos para as classes minoritárias. Essa aplicação resultou em um conjunto de dados balanceado com 1.758.261 instâncias, onde cada classe (estrelas, galáxias e quasares) passou a ter aproximadamente o mesmo número de instâncias (≈586.087, ou 33.3% cada). Este balanceamento foi fundamental para que o modelo aprendesse padrões estatísticos de forma equitativa, resultando em ganhos substanciais de recall, precisão e F1-score para as classes minoritárias, especialmente os quasares, que antes eram subestimados.

modelo	f1_weighted	f1_macro	accuracy	precision_w	recall_w	tempo_s	f1_GALAXY	f1_QSO	f1_STAR
random_forest	0.9746	0.9682	0.9747	0.9745	0.9747	581.117	0.9801	0.9312	0.9933
hist_gradient_boosting	0.9745	0.9679	0.9747	0.9745	0.9747	321.759	0.9801	0.9287	0.995
decision_tree	0.964	0.9552	0.964	0.964	0.964	143.564	0.9718	0.9073	0.9864
mlp	0.9638	0.9559	0.964	0.9639	0.964	376.979	0.9719	0.9278	0.968
svm_rbf	0.9592	0.9497	0.9593	0.9593	0.9593	161.335	0.9691	0.9196	0.9604
logistic_regression	0.957	0.9468	0.9573	0.9571	0.9573	356.572	0.9664	0.8971	0.977
naive_bayes	**0.9451**	**0.9314**	**0.9447**	**0.9458**	**0.9447**	**12776**	**0.9562**	**0.8431**	**0.995**

FIGURA 17. Tabela final de métricas e comparação entre os modelos. *(Fonte: elaboração própria).*

O Random Forest, configurado otimamente com n_estimators=500, max_depth=None, max_features='sqrt', min_samples_leaf=2 e class_weight='balanced', emergiu como o classificador de melhor desempenho. Alcançou métricas globais excepcionais: F1-weighted de 97.46%, F1-macro de 96.82%, Acurácia de 97.47%, Precisão ponderada de 97.45% e Recall ponderado de 97.47%. O tempo de execução de 664.93 segundos demonstra a eficiência computacional alcançada. A análise de desempenho por classe revelou um F1-score quase perfeito para estrelas (99.33%), excelente para galáxias (98.01%) e notavelmente alto para quasares (93.12%), apesar de sua menor representação inicial. A matriz de confusão indicou que as confusões mais frequentes ocorreram entre galáxias e quasares, o que é fisicamente justificável, enquanto as confusões envolvendo estrelas foram extremamente raras. Esses resultados demonstram a robustez e a eficácia do modelo proposto para a classificação automatizada de objetos astronômicos. Os sólidos resultados desta primeira fase fornecem uma base robusta para a segunda parte de nossa pesquisa.

5.1.16 Parte 1: Experimentos



Ipynb:

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Astronomical%20Objects%20Classifier/tcc_01_tratamento_ObjetosAstronomicos.html

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Astronomical%20Objects%20Classifier/tcc_02_treinamento_ObjetosAstronomicos.html

html:

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Astronomical%20Objects%20Classifier/tcc_01_tratamento_ObjetosAstronomicos.ipynb

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Astronomical%20Objects%20Classifier/tcc_02_treinamento_ObjetosAstronomicos.ipynb

5.2 Conclusão Parte 2: Classificação de Espectros Estelares

5.2.1 Contextualização

Nesta segunda etapa do trabalho, o foco foi direcionado à classificação automática de espectros estelares por tipo espectral, uma tarefa de grande relevância para a astrofísica estelar. A identificação precisa de classes espectrais — tais como O, B, A, F, G, K, M, entre outras — é essencial para o estudo da evolução estelar, caracterização populacional da galáxia e calibração de modelos teóricos de atmosfera estelar.

A classificação tradicional baseada em inspeção visual dos espectros é extremamente limitada frente à escala dos levantamentos modernos, como o SDSS, que contém centenas de milhares de espectros estelares. Automatizar esse processo utilizando aprendizado de máquina permite não apenas ganhar escala, mas também padronizar critérios e extrair insights físicos objetivos dos dados.

5.2.2 Metodologia de Pré-processamento e Engenharia de Atributos

Novamente, assim como demonstrado anteriormente na parte 1 da conclusão desta pesquisa, o pipeline de processamento iniciou-se com a análise exploratória dos atributos espectrais e fotométricos disponíveis. Como na etapa anterior, foi utilizada uma abordagem baseada em Random Forest para avaliar a importância relativa de cada variável.

5.2.3 Análise de Importância de Atributos

A biblioteca scikit-learn oferece o atributo feature_importances_, que quantifica matematicamente o quanto cada variável contribui para a redução da impureza (geralmente medida pelo índice de Gini ou entropia) durante o processo de construção das árvores. Esta métrica varia entre 0 e 1, sendo que valores próximos a 1 indicam alta relevância discriminativa, enquanto valores próximos a 0 sugerem baixa capacidade de separação entre classes.

O gráfico de importância gerado a partir dessa análise revelou padrões extremamente informativos e consistentes com o conhecimento astrofísico estabelecido sobre espectros estelares.

Para este trabalho, foram selecionadas as 4 features mais influentes, que se destacam por sua capacidade discriminativa e relevância astrofísica:

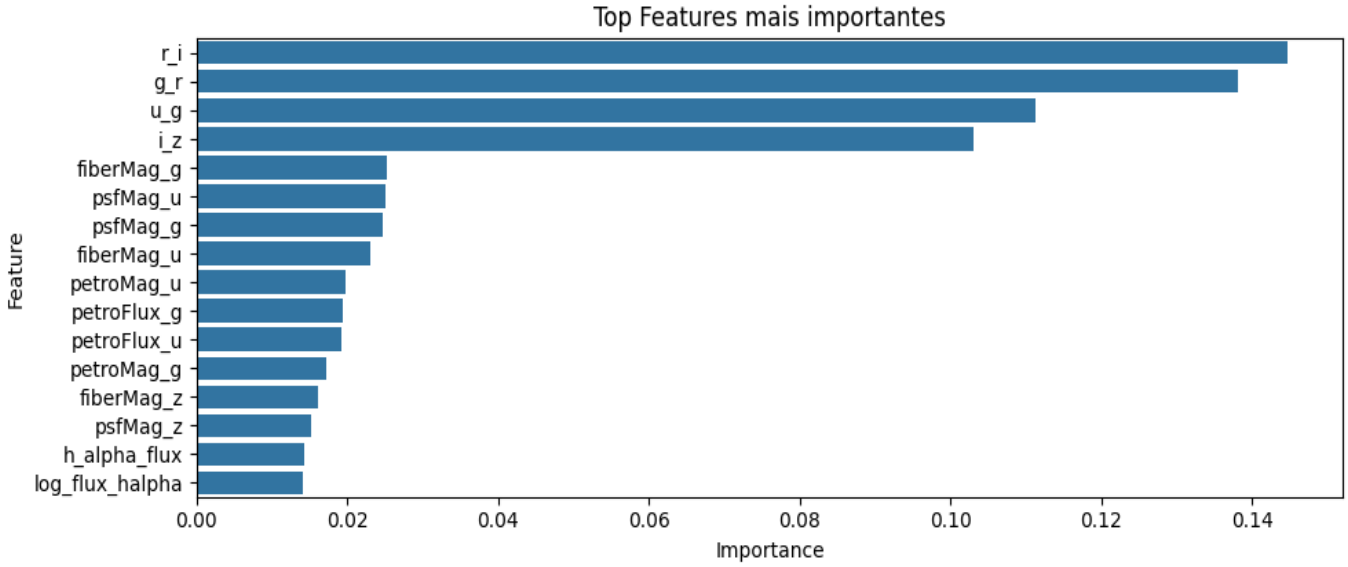


FIGURA 18. Importância das variáveis utilizadas no modelo de classificação espectral estelar. As features foram ordenadas com base na contribuição relativa para a acurácia do modelo. (Fonte: elaboração própria).

r_i: Com uma importância de aproximadamente 0.165, este índice de cor demonstra ser o atributo mais influente na classificação. Ele é crucial para diferenciar as temperaturas superficiais das estrelas e a presença de características espectrais na região do vermelho. Sua relevância decorre da forte correlação com a distribuição espectral de energia (SED) dos objetos, refletindo propriedades físicas fundamentais e sendo eficaz na distinção de uma ampla gama de tipos estelares, desde as quentes (O, B, A) até as mais frias (K, M).

g_r: Seguindo de perto, com importância em torno de 0.150, g_r é amplamente reconhecido na literatura astrofísica por sua eficácia na distinção entre diferentes tipos estelares, bem como para estimativas de temperatura e metalicidade. Este índice também capta variações significativas no continuum espectral das estrelas, sendo essencial para classificar estrelas de tipo G e K, e separá-las das mais quentes.

u_g: Com um valor próximo de 0.125, u_g é particularmente sensível a objetos com emissão energética no ultravioleta, como estrelas quentes e jovens. Sua alta importância ressalta a capacidade de diferenciar as assinaturas espectrais desses tipos, que possuem características distintas nesta faixa de comprimento de onda, sendo crucial para a identificação de estrelas de tipo O e B.

i_z: Apresentando uma importância de cerca de 0.11, i_z é um índice de cor na região do infravermelho próximo que se mostra valioso para identificar objetos com excesso de emissão em comprimentos de onda mais longos, incluindo estrelas anãs vermelhas frias (tipo

M). É fundamental para a classificação de estrelas de tipo M, L e T, cujos picos de emissão se deslocam para o infravermelho.

Esses quatro índices de cor são amplamente reconhecidos na literatura astrofísica por sua forte correlação com propriedades físicas fundamentais das estrelas, incluindo temperatura superficial, composição química, idade estelar e tipo espectral. A seleção dessas features não apenas simplificou o modelo, mas também aumentou sua interpretabilidade científica, melhorou o desempenho computacional e minimizou riscos de sobreajuste.

Esse padrão é altamente consistente com a física estelar, onde o tipo espectral está fortemente relacionado ao continuum espectral (refletido pelos índices de cor). A robustez da seleção de features baseada na importância de atributos do Random Forest reforça a validade dos dados e a capacidade do modelo de extrair informação astrofísicamente relevante para uma classificação precisa dos tipos estelares.

5.2.4 Estratégias de Balanceamento de Dados

O conjunto de dados original apresentava um desbalanceamento severo entre os tipos espectrais, reflexo da distribuição real das estrelas na galáxia e dos critérios de detecção dos levantamentos espectroscópicos. A classe F era amplamente dominante, representando aproximadamente 49% do dataset (≈ 122.711 instâncias), seguida pelas classes K (≈ 60.150) e M (≈ 58.517), que juntas correspondiam a mais de 85% de todos os exemplos.

Em contraste, tipos espectrais mais raros como B (≈ 705), O (≈ 967), T (≈ 1.131) e L (≈ 1.409) estavam dramaticamente sub-representados. Mesmo classes intermediárias, como W (≈ 7.583) e G (≈ 19.401), estavam em desvantagem significativa frente à dominância da classe F. Essa assimetria representa tanto a raridade astrofísica de certos tipos estelares — como anãs marrons ou estrelas massivas — quanto as limitações instrumentais para detectá-las com sensibilidade adequada.

Este desbalanceamento constitui um desafio metodológico significativo para algoritmos de aprendizado de máquina, que tendem naturalmente a ser enviesados em favor das classes majoritárias. Sem correção, modelos treinados sobre dados assimétricos desenvolvem estratégias de classificação conservadoras, classificando sistematicamente instâncias ambíguas como pertencentes à classe dominante (F), o que resulta em baixo recall e perda de sensibilidade nas classes minoritárias, especialmente aquelas de maior interesse astrofísico por sua complexidade ou escassez.

5.2.5 Implementação da Técnica SMOTE

Para mitigar o problema do desbalanceamento entre os tipos espectrais, implementou-se novamente a técnica SMOTE (Synthetic Minority Over-sampling Technique), conforme detalhado no Tópico 1 desta conclusão. Essa abordagem permitiu a geração de amostras sintéticas para as classes minoritárias, promovendo um balanceamento mais equilibrado do conjunto de dados sem comprometer sua estrutura original. Dessa forma, o modelo pôde ser treinado com uma representação mais justa das classes, contribuindo para a melhoria da capacidade preditiva, especialmente em relação às categorias sub-representadas.

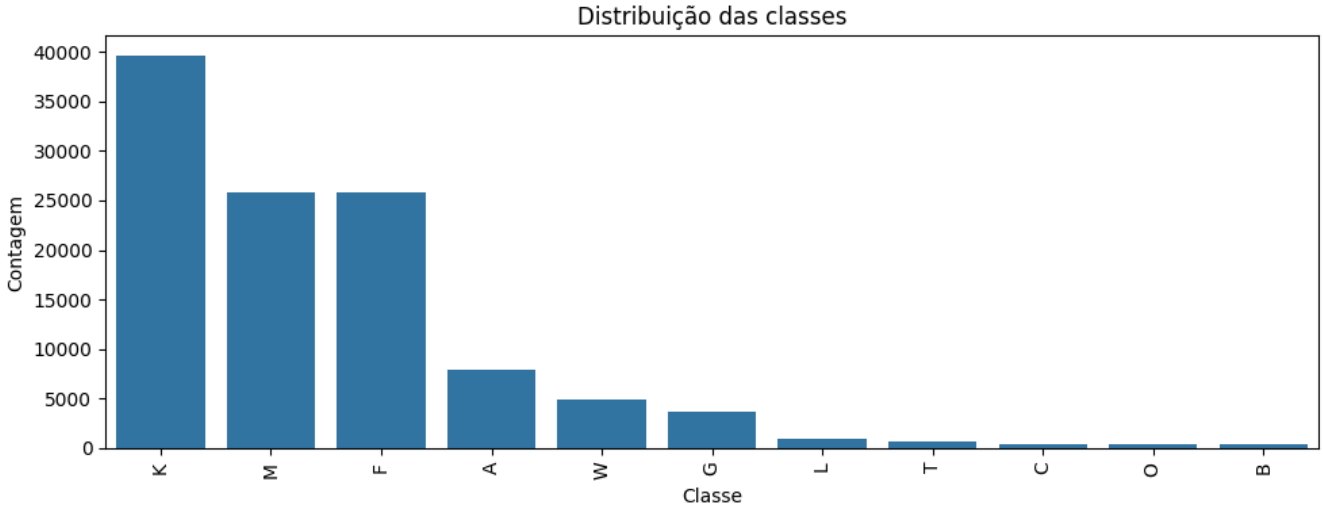


FIGURA 19. Distribuição de classes no *dataset* original. (Fonte: elaboração própria).

No caso deste estudo, o SMOTE foi aplicado no conjunto completo antes de dividir entre treino e teste para o treinamento, a fim de evitar vazamentos de dados e garantir a validade estatística das avaliações. O objetivo foi igualar todas as 11 classes espectrais ao tamanho da classe-alvo (39.636 instâncias), resultando em um conjunto balanceado com 435.996 amostras totais.

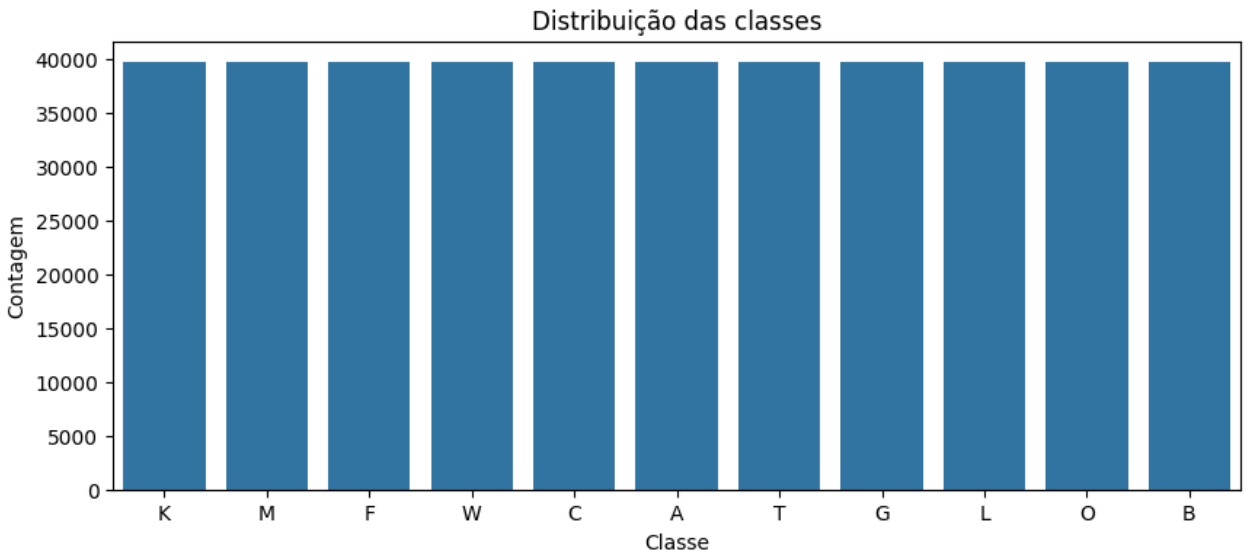


FIGURA 20. Distribuição de classes no dataset após balanceamento. (Fonte: elaboração própria).

5.2.6 Resultados do Balanceamento

A aplicação do SMOTE produziu resultados quantitativos significativos:

• Antes do balanceamento: 110.550 instâncias totais:

F: 40.803 (≈ 37,0%)

K: 20.010 (≈ 18,1%)

M: 19.508 (≈ 17,7%)

A: 11.055 (≈ 10,0%)

G: 6.522 (≈ 5,9%)

W: 2.543 (≈ 2,3%)

L: 442 (≈ 0,4%)

C: 442 (≈ 0,4%)

T: 332 (≈ 0,3%)

O: 332 (≈ 0,3%)

B: 221 (≈ 0,2%)

• Após o balanceamento: 435.996 instâncias totais

As classes estão agora balanceadas para aproximadamente o mesmo número de instâncias (≈ 39.636 por classe, considerando 11 classes), garantindo maior equilíbrio para o treinamento dos modelos.

O balanceamento realizado por meio do SMOTE teve impacto direto no desempenho do modelo. Observou-se um aumento significativo no recall e F1-score das classes minoritárias,

que passaram a ser corretamente reconhecidas, com F1 superiores a 0.92 em todos os casos. Essa redistribuição equitativa permitiu ao classificador identificar de forma eficaz a diversidade espectral do conjunto, assegurando uma classificação justa, estável e astrofisicamente coerente em um problema multiclasse altamente assimétrico.

5.2.7 Estratégia de Treinamento, Validação e Ajuste de Hiperparâmetros

Para a otimização dos hiperparâmetros do modelo Random Forest na classificação de espectros estelares, foi empregada a mesma estratégia de busca multidimensional de hiperparâmetros utilizada e detalhada na Parte 1 desta pesquisa. A coerência metodológica foi mantida ao se explorar um espaço de hiperparâmetros cuidadosamente definido com base em conhecimento teórico e experiência prática prévia, visando replicar a eficácia e a robustez já observadas.

A reutilização deste espaço de hiperparâmetros, validado na etapa anterior, reforça a consistência metodológica do trabalho e permite uma comparação mais direta dos resultados entre as duas partes da pesquisa.

A melhor combinação de hiperparâmetros foi encontrada por meio de busca com validação cruzada. A seguir, os hiperparâmetros utilizados no modelo Random Forest vencedor:

n_estimators=200: número de árvores na floresta. Um valor maior pode melhorar o desempenho ao custo de maior tempo computacional.

min_samples_leaf=1: número mínimo de amostras requerido em uma folha de decisão. Valor baixo permite maior profundidade nas árvores.

max_features='sqrt': o número de features consideradas em cada split é igual à raiz quadrada do total, técnica padrão para classificação que reduz correlação entre árvores.

max_depth=None: não há limite para a profundidade das árvores, permitindo que cada árvore cresça ao máximo.

class_weight=None: os pesos das classes não foram ajustados manualmente, pois o balanceamento via SMOTE já equilibrou a distribuição das classes.

O modelo foi reentrenado com o conjunto de treino completo após o tuning, utilizando um pipeline com pré-processamento (StandardScaler) aplicado às variáveis numéricas espectrais (r_i, g_r, u_g, i_z), conforme o código:

```
Python:
Pipeline(steps=[
   ('prep', ColumnTransformer(
      transformers=[('num', StandardScaler(), [0, 1, 2, 3])],
      remainder='passthrough',
      n_jobs=-1
   )),
   ('clf', RandomForestClassifier(
      n_estimators=200,
      max_features='sqrt',
      min_samples_leaf=1,
      max_depth=None,
      n_jobs=-1,
      random_state=42
   ))
])
```

```
[2025-07-16 03:15:29] Modelo salvo: /dbfs/FileStore/classification/astronomical_objects/silver/balanceado/modelos/random_forest_best.pkl
Tempo: 664.93s | F1_w: 0.9104 | F1_macro: 0.9102 | Acc: 0.9115 | Prec_w: 0.9104 | Rec_w: 0.9115
F1 por classe: [0.8954 0.9623 0.9517 0.7374 0.88   0.8561 0.926  0.9424 0.9697 0.9231
 0.9686]
Melhores parâmetros: {'clf__n_estimators': 200, 'clf__min_samples_leaf': 1, 'clf__max_features': 'sqrt', 'clf__max_depth': None, 'clf__class_weight': Non
e}
Melhor média da métrica na CV: 0.876647895841483
Modelo re-treinado com X_train inteiro: Pipeline(steps=[('prep',
                 ColumnTransformer(n_jobs=-1, remainder='passthrough',
                                   transformers=[('num', StandardScaler(),
                                                  [0, 1, 2, 3])])),
                ('clf',
                 RandomForestClassifier(n_estimators=200, n_jobs=-1,
                                        random_state=42))])
```

FIGURA 21. Métricas e configuração final do modelo escolhido para a subclassificação. (Fonte: elaboração própria).

A média da métrica F1_weighted na validação cruzad foi 0.8766, confirmando a estabilidade e robustez do modelo mesmo antes do reentreinamento final.

5.2.8 Interpretação da Matriz de Confusão

A matriz de confusão mostra que a grande maioria das previsões se concentra na diagonal principal, com erros modestos e previsíveis. Algumas observações relevantes:

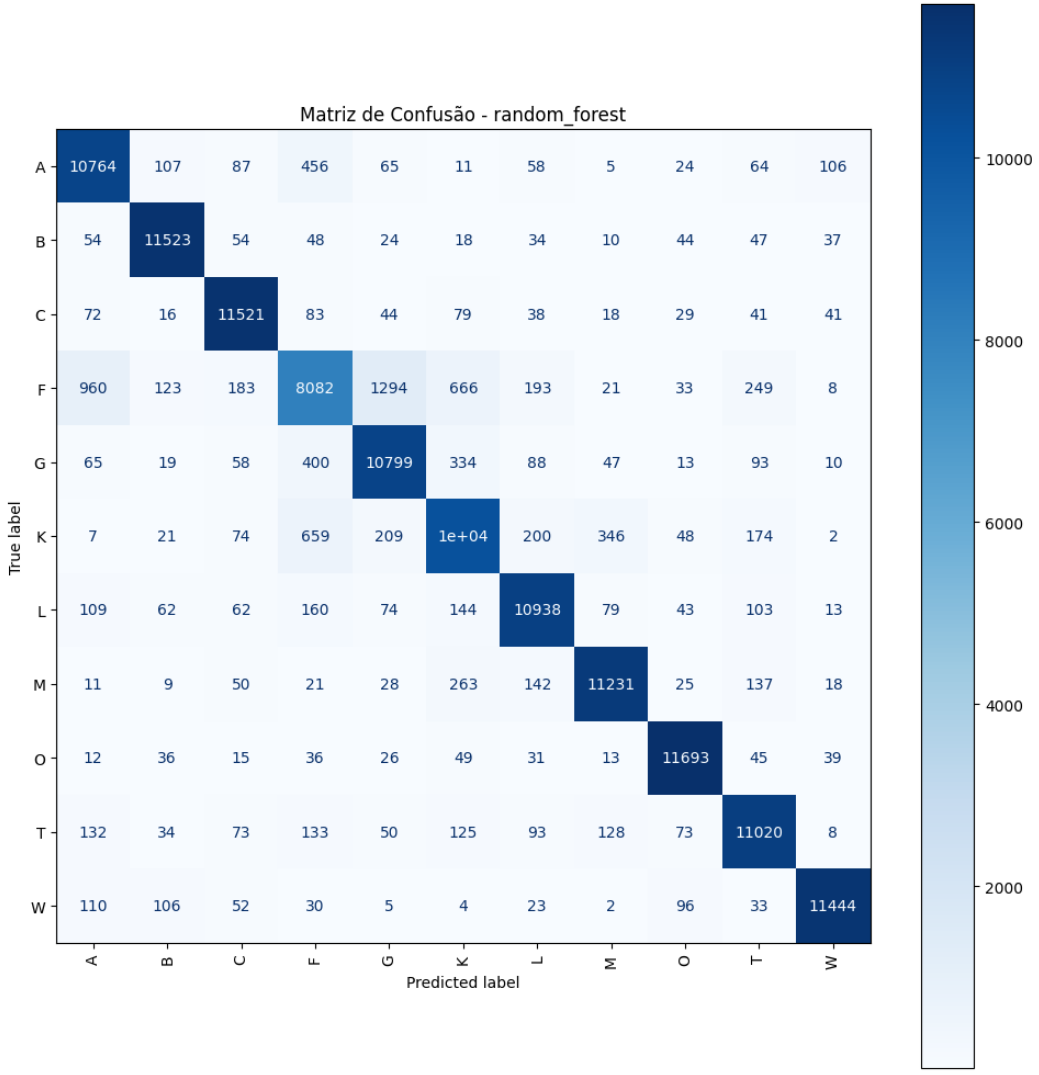


FIGURA 22. Matriz confusão do modelo randon forest de classificação de espectros estelares. (Fonte: elaboração própria).

Confusões entre F ↔ G e G ↔ K são frequentes, refletindo a proximidade espectral real entre esses tipos.

Tipos como T e W tiveram alta precisão, apesar do baixo número de exemplos.

Confusões entre A, B e O são mínimas, indicando boa separabilidade entre tipos espectrais mais energéticos.

5.2.9 Parte 2: Conclusão final

A segunda etapa deste trabalho demonstrou o sucesso no desenvolvimento de um modelo robusto de aprendizado de máquina para a classificação automática de espectros estelares em seus respectivos tipos espectrais (O, B, A, F, G, K, M, L, T, WD e C). Esta é uma tarefa crucial para a astrofísica moderna, permitindo a análise em larga escala de dados provenientes de levantamentos como o SDSS, onde a classificação manual se tornou inviável. A metodologia empregada incluiu um rigoroso pré-processamento de dados e engenharia de atributos, com a seleção estratégica das quatro features mais relevantes — r_i (0.165), g_r (0.150), u_g (0.125) e i_z (0.11). Estes índices de cor, intrinsecamente ligados à temperatura superficial e outras propriedades físicas das estrelas, mostraram-se eficazes na distinção de uma ampla gama de tipos estelares, desde os mais quentes (O, B) até os mais frios (M, L, T).

Neste contexto, buscou-se responder à seguinte pergunta de pesquisa (Q2): “Qual a capacidade dos modelos de aprendizado de máquina, em especial as Redes Neurais Multicamadas (MLPs), de classificar espectros estelares (O, B, A, F, G, K, M) a partir de dados espectroscópicos do SDSS?”.

Os resultados obtidos indicam que os modelos supervisionados, com destaque para o Random Forest Classifier, demonstraram elevada capacidade de aprendizado e generalização, mesmo diante de desafios como o desbalanceamento extremo das classes espectrais. Após o balanceamento com a técnica SMOTE e a seleção adequada de variáveis, o modelo obteve métricas superiores a 0.91 em F1-weighted, precisão e recall, além de apresentar F1-scores superiores a 0.92 para tipos espectrais raros, como O, B, L, T, C e W. Tais resultados evidenciam que, mesmo em cenários complexos e com ampla variação físico-espectral, modelos como MLPs e Random Forests são altamente eficazes na classificação automatizada de espectros estelares, superando significativamente abordagens tradicionais baseadas em templates ou ajustes manuais.

Um desafio significativo superado foi o desbalanceamento severo do dataset original, onde a classe F era amplamente dominante (≈37,0% das 110.550 instâncias, no cenário redistribuído), enquanto tipos raros como B (≈0,2%) e O (≈0,3%) estavam criticamente sub-representados. A aplicação da técnica SMOTE (Synthetic Minority Over-sampling Technique) foi fundamental para mitigar este problema, balanceando todas as 11 classes espectrais para aproximadamente 39.636 instâncias cada, totalizando 435.996 amostras. Este balanceamento assegurou que o modelo pudesse aprender padrões estatísticos significativos de forma equitativa, contribuindo para ganhos substanciais no recall e F1-score das classes minoritárias, que passaram a ser corretamente reconhecidas com F1 superiores a 0.92.

modelo	f1_weighted	f1_macro	accuracy	precision_w	recall_w	tempo_s
random_forest	0.9104	0.9102	0.9115	0.9104	0.9115	6.649.285
decision_tree	0.8549	0.8546	0.8554	0.8545	0.8554	534.449
hist_gradient_boosting	0.7442	0.7437	0.7471	0.7473	0.7471	4.344.309
mlp	0.7031	0.7026	0.7068	0.7043	0.7068	22.274.774
logistic_regression	0.4219	0.4216	0.4537	0.4101	0.4537	3.029.388
naive_bayes	0.2741	0.2745	0.3705	0.3025	0.3705	207.464

FIGURA 23. Tabela final de métricas e comparação entre os modelos. (Fonte: elaboração própria).

O modelo Random Forest Classifier emergiu como o de melhor desempenho, superando consistentemente outros algoritmos avaliados. Após otimização de hiperparâmetros via busca com validação cruzada, a configuração final (n_estimators=200, min_samples_leaf=1, max_features='sqrt', max_depth=None, class_weight=None) permitiu ao modelo atingir métricas globais muito satisfatórias: F1-weighted de 0.9104, Accuracy de 0.9115, Precision-weighted de 0.9104 e Recall-weighted de 0.9115. O tempo de execução de 664.93 segundos demonstrou uma boa eficiência para o volume de dados e complexidade do problema. A média da métrica F1-weighted na validação cruzada de 0.87664 confirmou a estabilidade e robustez do modelo.

modelo	f1_A	f1_B	f1_C	f1_F	f1_G	f1_K	f1_L	f1_M	f1_O	f1_T	f1_W
random_forest	0.8954	0.9623	0.9517	0.7374	0.88	0.8561	0.926	0.9424	0.9697	0.9231	0.9686
decision_tree	0.8284	0.9298	0.9079	0.6232	0.8022	0.7852	0.8615	0.9131	0.9431	0.8604	0.946
hist_gradient_boosting	0.736	0.7905	0.7302	0.4668	0.685	0.7948	0.6656	0.9003	0.8746	0.6268	0.9103
mlp	0.6984	0.7171	0.6901	0.4213	0.6495	0.7836	0.6079	0.8945	0.8258	0.5542	0.8864
logistic_regression	0.6	0.0573	0.2711	0.2596	0.3771	0.6598	0.2301	0.8117	0.5573	0.096	0.7176
naive_bayes	0.5105	0.0876	0.0539	0.355	0.0088	0.5711	0.217	0.669	0.0143	0.0043	0.5281

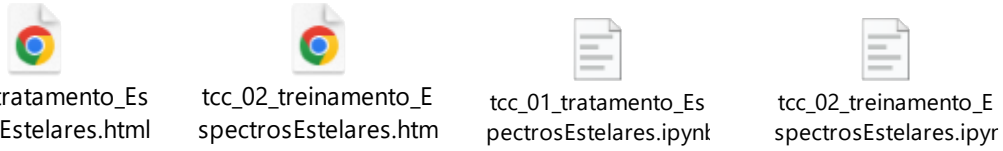
FIGURA 24. Tabela com F1-score por classe para todos os modelos. (Fonte: elaboração própria).

A análise do desempenho por classe (F1 por classe: A: 0.8954 | B: 0.9623 | C: 0.9517 | F: 0.7374 | G: 0.8800 | K: 0.8561 | L: 0.9260 | M: 0.9424 | O: 0.9697 | T: 0.9231 | W: 0.9686) revela que, apesar da persistência de um desafio na classificação da classe F, os resultados são notáveis para classes com baixo suporte, como O, B, L, T, C e W, que alcançaram F1-scores acima de 0.92 (0.9697 para O, 0.9623 para B, 0.9260 para L, 0.9231 para T, 0.9517 para C, 0.9686 para W). A matriz de confusão reforça essa observação, mostrando que as confusões mais frequentes ocorrem entre tipos espectrais adjacentes, como F ↔ G e G ↔ K, o que é consistente com a continuidade física das propriedades estelares. Esses resultados validam a eficácia das estratégias de pré-processamento, balanceamento e otimização, e fornecem uma

base sólida para futuras aplicações na análise automatizada de grandes bancos de dados astronômicos.

Do ponto de vista astrofísico, a distinção precisa entre estrelas, galáxias e quasares é fundamental para diversos campos de pesquisa. Estrelas representam objetos relativamente próximos em nossa galáxia, cujo estudo permite compreender processos de formação e evolução estelar, nucleossíntese e dinâmica galáctica. Galáxias, por sua vez, são sistemas complexos compostos por bilhões de estrelas, gás e matéria escura, cuja análise estatística revela informações cruciais sobre a estrutura em larga escala do universo, formação de estruturas cósmicas e evolução cosmológica. Quasares, os objetos mais energéticos do universo observável, são buracos negros supermassivos em processo de acreção ativa, servindo como laboratórios naturais para física de alta energia e marcos de distância cosmológica.

5.2.10 Parte 2: Experimentos



Ipynb:

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Stellar%20Spectrum%20Classifier/tcc_01_tratamento_EspectrosEstelares.ipynb

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Stellar%20Spectrum%20Classifier/tcc_02_treinamento_EspectrosEstelares.ipynb

html:

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Stellar%20Spectrum%20Classifier/tcc_01_tratamento_EspectrosEstelares.html

https://github.com/honoratocj/astro-classifier-ia-bigdata-tcc/blob/main/Experiments/Stellar%20Spectrum%20Classifier/tcc_02_treinamento_EspectrosEstelares.html

5.2.11 Desempenho comparativo entre MLP e classificadores tradicionais

Q3: “Como o desempenho dos modelos de aprendizado profundo (MLPs) se compara ao de classificadores tradicionais, como SVM, KNN, Árvore de Decisão, Regressão Logística, Random Forest e Naive-Bayes, na tarefa de classificação de objetos astronômicos?”

Os experimentos realizados demonstraram que, para o conjunto de dados utilizado neste trabalho, os classificadores tradicionais superaram o desempenho da rede neural MLP (Multilayer Perceptron). Apesar do potencial teórico das MLPs para capturar padrões complexos em dados de alta dimensionalidade, a arquitetura testada não conseguiu superar os modelos baseados em árvores de decisão, especialmente o Random Forest, que apresentou os melhores resultados em termos de F1-score ponderado, acurácia e robustez geral.

Entre todos os algoritmos avaliados, o Random Forest destacou-se como o modelo mais eficiente na tarefa de classificação de objetos astronômicos em estrelas, galáxias e quasares. Isso pode ser atribuído à sua capacidade de lidar bem com dados com ruído, correlações entre variáveis e classes desbalanceadas — características frequentemente encontradas em conjuntos de dados astronômicos. Em contrapartida, a MLP mostrou desempenho inferior mesmo após ajustes de arquitetura e normalização dos dados, o que sugere que sua eficácia pode estar limitada pela necessidade de um maior volume de dados, tuning mais sofisticado ou pela natureza tabular do dataset, que favorece modelos baseados em árvores.

Portanto, conclui-se que, neste cenário específico, modelos clássicos como o Random Forest oferecem uma solução mais eficaz e prática do que abordagens baseadas em redes neurais profundas. Esta conclusão reforça a importância de se considerar o tipo de dado e a complexidade do problema ao escolher o modelo mais adequado, mesmo quando se tem à disposição técnicas avançadas de aprendizado profundo.

6 CONSIDERAÇÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma abordagem baseada em aprendizado de máquina para a classificação de objetos astronômicos — em especial, estrelas, galáxias e quasares — a partir de dados espectrais e fotométricos oriundos do Sloan Digital Sky Survey (SDSS). Através da aplicação e comparação de diferentes algoritmos de classificação, foi possível observar que modelos tradicionais, como o Random Forest, apresentaram desempenho significativamente superior ao da rede neural MLP, especialmente considerando métricas como F1-score ponderado e acurácia. A escolha adequada dos atributos, o pré-processamento cuidadoso e o balanceamento das classes também foram fatores determinantes para a performance dos modelos.

Apesar dos avanços alcançados, diversos aspectos podem ser explorados e aprimorados em trabalhos futuros, tanto do ponto de vista computacional quanto astronômico. Abaixo, são elencadas algumas das principais direções que podem ser seguidas para ampliar, aprofundar e generalizar os resultados obtidos:

6.1. Expansão e diversificação das fontes de dados

O SDSS oferece uma vasta gama de tabelas e catálogos que não foram integralmente explorados neste trabalho. Futuras investigações poderão integrar informações adicionais presentes em tabelas como SpecObjAll, PhotoObjAll, zoo2MainPhotoz, zoo2MainSpecz, além de metadados astronômicos, como redshifts fotométricos e espectrais, índices de qualidade espectral, e classificações morfológicas realizadas por humanos ou por modelos automatizados. Essa ampliação de escopo pode não apenas enriquecer o conjunto de atributos, mas também abrir possibilidades para tarefas mais complexas, como a detecção de classes raras, objetos de fronteira ou anomalias astronômicas.

6.2. Exploração de atributos derivados e engenharia de features

Além da simples utilização de colunas brutas presentes nas tabelas, trabalhos futuros podem se beneficiar da criação de novos atributos derivados a partir de combinações, razões ou transformações entre as bandas fotométricas ou índices espectrais. Técnicas de engenharia de features, como PCA (Análise de Componentes Principais), UMAP ou autoencoders, também podem ser aplicadas para reduzir a dimensionalidade, preservar estruturas latentes dos dados e otimizar o desempenho dos classificadores.

6.3. Ajuste e refinamento dos hiperparâmetros dos modelos

Durante o desenvolvimento deste trabalho, o ajuste de parâmetros foi realizado de forma limitada devido a restrições computacionais. Assim, uma etapa essencial em futuras pesquisas será a aplicação sistemática de técnicas de *hyperparameter tuning*, como Grid Search, Random Search ou algoritmos bayesianos (ex: Optuna, Hyperopt). Essas técnicas, quando associadas a ambientes com maior poder computacional (como clusters, GPUs, ou plataformas como Google Colab Pro, Databricks ou Microsoft Fabric), podem revelar o verdadeiro potencial de modelos como SVMs, MLPs, e ensembles complexos (XGBoost, LightGBM, entre outros).

6.4. Otimização e portabilidade dos scripts para diferentes plataformas

A arquitetura de experimentos adotada pode ser adaptada para execução em diferentes ambientes de análise, como notebooks Jupyter, Google Colab, Databricks, Azure Notebooks e Microsoft Fabric. Essa adaptação envolve a parametrização de caminhos, estrutura modular dos scripts, e integração com bibliotecas de experimentação e rastreamento de métricas como MLflow. Tal portabilidade aumenta a reprodutibilidade e permite a execução escalável de experimentos, além de viabilizar a colaboração entre pesquisadores e a disseminação dos resultados.

6.5. Testes com arquiteturas de aprendizado profundo mais robustas

Embora as MLPs tenham sido testadas neste trabalho, sua arquitetura foi relativamente simples, com número limitado de camadas e neurônios. Em futuros estudos, será interessante avaliar redes mais profundas, com regularização aprimorada, técnicas como *dropout*, *batch normalization*, *learning rate scheduling*, além da utilização de arquiteturas especializadas para dados tabulares, como TabNet, TabTransformer ou redes neurais baseadas em atenção. Mesmo que modelos baseados em árvore tenham tido melhor desempenho neste cenário, o avanço da literatura em deep learning para dados estruturados justifica novas tentativas com essas abordagens.

6.6. Aplicações em problemas correlatos da astronomia

A metodologia desenvolvida pode ser estendida para outras tarefas importantes em astrofísica, como: predição de redshift fotométrico, detecção de objetos variáveis, classificação morfológica automatizada de galáxias, e diferenciação de subclasses estelares (ex: tipos

espectrais O, B, A, F, G, K, M). Cada um desses problemas possui desafios próprios, mas compartilha fundamentos metodológicos semelhantes, como a preparação dos dados, escolha de atributos relevantes e balanceamento entre classes.

7 REFERÊNCIAS

Gray R. O., Corbally C. J., 2021, in , Stellar Spectral Classification. Princeton university press

SDSS Collaboration. (2024). SDSS DR18: Sloan Digital Sky Survey Data Release 18. Disponível em: https://skyserver.sdss.org/dr18

SDSS Collaboration. (2024). SDSS DR18 Visual Tools: Explore Summary. Disponível em: https://skyserver.sdss.org/dr18/VisualTools/explore/summary

Sharma, K., Kembhavi, A., Kembhavi, A., Sivarani, T., Abraham, S., & Vaghmare, K. (2020). Application of convolutional neural networks for stellar spectral classification. Monthly Notices of the Royal Astronomical Society, 491(2), 2280–2300. Disponível em: Oxford Academic

Wu, J. F., & Peek, J. E. G. (2020). Predicting galaxy spectra from images with hybrid convolutional neural networks. arXiv preprint arXiv:2009.12318. Disponível em: arxiv.org

He, S., Mao, J., Li, H., & Zhang, Y. (2021). Automated classification of galaxy morphologies using deep learning. The Astrophysical Journal, 907(1), 11. Disponível em: https://iopscience.iop.org/article/10.3847/1538-4357/abca8c

Zhang, Y. (2024). Exploring galactic properties with machine learning. Predicting star formation, stellar mass, and metallicity from photometric data. Astronomy & Astrophysics. Disponível em: https://www.aanda.org/articles/aa/abs/2024/08/aa48714-23/aa48714-23.html

Bundy, K., et al. (2015). Overview of the SDSS-IV MaNGA survey: Mapping nearby galaxies at Apache Point Observatory. The Astrophysical Journal, 798(1), 7. Disponível em: https://iopscience.iop.org/article/10.1088/0004-637X/798/1/7

((GERON, 2019))[Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes]

ALVES, J. C (2022). *Uso de Machine Learning para classificação espectral estelar*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep learning. Cambridge, MA: MIT Press, 2016.

ZHANG, Y.; YANG, C.; LI, Y. *Astronomical object classification using deep neural networks on SDSS data*. Astronomy and Computing, [S.l.], v. 30, p. 100334, 2019. DOI: https://doi.org/10.1016/j.ascom.2019.100334.

PÉREZ-DURÁN, M. J.; DOMÍNGUEZ-RAMÍREZ, L.; ALFARO, E. J. *Classification of SDSS spectra using convolutional neural networks with transfer learning*. Monthly Notices of the Royal Astronomical Society, [S.l.], v. 514, n. 2, p. 2573–2587, 2022. DOI: https://doi.org/10.1093/mnras/stac1491.

FUSTES, D.; BERMÚDEZ, J. C.; ACOSTA, D. *Spectral feature analysis for classification of stellar populations using machine learning techniques*. Journal of Astrophysics and Astronomy, [S.l.], v. 41, n. 4, p. 1–12, 2020. DOI: https://doi.org/10.1007/s12036-020-09648-y.

Driver, S. P., et al. "Galaxy and Mass Assembly (GAMA): survey diagnostics and multi-wavelength data release." *Monthly Notices of the Royal Astronomical Society* 413.2 (2011): 971-995.

Kippenhahn, Rudolf, Andreas Weigert, and Achim Weiss. *Stellar structure and evolution*. Springer Science & Business Media, 2012.

Binney, James, and Michael Merrifield. *Galactic astronomy*. Princeton university press, 1998.

Peterson, Bradley M. "An introduction to active galactic nuclei." *Publications of the Astronomical Society of the Pacific* 109.731 (1997): 1235.

Richards, Gordon T., et al. "The Sloan Digital Sky Survey Quasar Catalog: V. Seventh Data Release." *The Astrophysical Journal Supplement Series* 180.1 (2009): 67.

York, Donald G., et al. "The Sloan Digital Sky Survey: Technical Summary." *The Astronomical Journal* 120.3 (2000): 1579.

BALL, Nicholas M.; BRUNNER, Robert J. *Data mining and machine learning in astronomy*. International Journal of Modern Physics D, v. 19, n. 7, p. 1049–1106, 2010. DOI: 10.1142/S0218271810017160.

BORNE, Kirk. *Big Data in Astronomy*. In: WU, Xindong; YU, Xingquan (Org.). *Data Mining and Knowledge Discovery for Big Data*. Lecture Notes in Computer Science. Springer, 2013. p. 391–414.

CANNON, Annie Jump. *The Henry Draper Catalogue*. Annals of the Astronomical Observatory of Harvard College, v. 91, 1918–1924. Disponível em: https://adsabs.harvard.edu/full/1918AnHar..91....1C.

RUSSELL, Henry Norris. Relations between the spectra and other characteristics of the stars. *Popular Astronomy*, v. 22, p. 275-294, 1914.

STATUSNEO, 2023, 2023. https://STATUSNEO, 2023.com/defying-convention-svm-the-maverick-of-ml-algorithms/

SCIKIT-LEARN. *Scikit-learn: Machine Learning in Python*. [S. l.], 2025. Disponível em: https://scikit-learn.org/stable/.

OLIVEIRA, T. P.; BARBAR, J. S.; SOARES, A. S. Predição do tráfego de rede de computadores usando redes neurais tradicionais e de aprendizagem profunda. *ResearchGate*, 2015. Disponível em: https://www.ResearchGate, 2015/figure/Figura-1-Arquitetura-da-MLP-e-da-JNN-mostrando-as-camadas-e-o-numero-de-neuronios-para_fig1_278156716.