Modelo Perceptron Multicamadas para Classificação de Estrelas, Quasares e Galáxias

E Veloso, G Júnior e R Rego
Universidade Federal Rural do Semi-Arido
Pau dos Ferros, Rio Grande do Norte, Brasil
erikyabreu@gmail.com, geovanjunior2017@gmail.com, rosana.rego@ufersa.edu.br

Resumo* – Devido a grande quantidade de dados gerados nos estudos astrofísicos, o processo de classificação dos astros necessita de algoritmos que otimizem essa função. Essa classificação é realizada através da análise de variáveis observadas nos espectros encontrados no espaço, as quais são armazenadas em um banco de dados. Neste trabalho, foi aplicado um modelo de rede neural perceptron multicamadas, buscando classificar os astros em: estrelas, quasares e galáxias. O modelo apresentou 97,32% de acurácia, 97,34% de precisão, 97,33% de sensibilidade e 97,32% de F1-*score.

***Palavras chaves* – estrelas, quasares, galáxias, aprendizagem de máquina.**

I. INTRODUÇÃO

A astronomia é uma área de pesquisa que desde os primórdios detém bastante curiosidade humana e que vem tendo bastante crescimento nos últimos anos devido a evolução dos equipamentos de observação, como por exemplo os telescópios, espectrógrafos e geradores de imagens. Atualmente, já é possível compreender uma quantidade enorme de objetos que são observados no céu e os classificar. Entre eles, pode-se citar estrelas, quasares e galáxias, que vem sendo objeto de estudo de várias pesquisas na área [1].

Uma definição astrofísica aceita para estrelas é qualquer objeto suficientemente massivo que possa inflamar a fusão de elementos em seu próprio núcleo devido as pressões gravitacionais dentro do próprio objeto [2]. A Fig. 1 mostra o aglomerado estelar M13 visto pelo Sloan Digital Sky Survey [1].

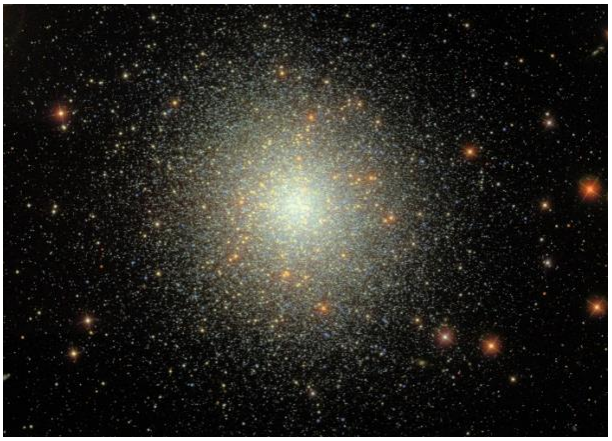

Fig. 1. Aglomerado estelar M13. Fonte: Sloan Digital Sky Survey.

Os quasares, cujo nome vem de Quasi Stellar Radio Sources, foram descobertos em 1960, como fortes fontes de rádio, com aparência ótica aproximadamente estelar, azuladas. Muito provavelmente são galáxias com buracos negros fortemente ativos no centro, como proposto em 1964 por Edwin Ernest Salpeter (1925-2008) e Yakov Borisovich Zel'dovich (1914-1989). São objetos extremamente compactos e luminosos, emitindo mais do que centenas de galáxias juntas, isto é, até um trilhão de vezes mais do que o Sol. São fortes fontes de rádio, variáveis, e seus espectros apresentam linhas largas com efeito Doppler indicando que eles estão se afastando a velocidades muito altas, de até alguns décimos da velocidade da luz. O primeiro a ter seu espectro identificado foi o 3C 273, pelo astrônomo holandês Maarten Schmidt, em 1963. No modelo mais aceito, o buraco negro central acresce gás e estrelas da sua vizinhança, emitindo intensa radiação enquanto a matéria se acelera, espiralando no disco de acresção, e parte da matéria é ejetada por conservação de momento angular [3]. A Fig. 2 mostra o quasar 3C 279, obtida pelo CFHT (Canada-France-Hawaii Telescope) [4].

Fig. 2. Quasar 3C-279. Fonte: Canada-France-Hawaii Telescope.

Galáxias são grupos de estrelas e outros objetos espaciais mantidos juntos pela gravidade e que são classificadas por sua forma. Cada tipo tem características diferentes e uma história de evolução diferente [5]. A Via Láctea, por exemplo, tem centenas de bilhões de estrelas, gás e poeira suficiente para produzir bilhões de estrelas a mais e pelo menos dez vezes mais matéria escura do que todas as estrelas e gás juntos, onde tudo é mantido junto pela gravidade [6]. A Fig. 3 mostra a galáxia espiral M51 e sua companheira mais fraca, disponibilizada pelo Sloan Digital Sky Survey [1].

Fig. 3. Galáxia espira M51. Fonte: Sloan Digital Sky Survey.

II. TRABALHOS RELACIONADOS

A busca por formas de classificação de corpos celestes utilizando modelos de redes neurais também é tópico de produção de vários outros trabalhos na área. Dessa forma, é válido a verificação dos resultados obtidos em outras pesquisas para assim compará-los com os do presente projeto.

O artigo intitulado “Machine Learning Applied to Star–Galaxy–QSO Classification and Stellar Effective Temperature Regression” fez-se o uso de dois bancos de dados disponibilizados pelo Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOSTE) e a quarta fase do Sloan Digital Sky Survey (SDSS-IV), com o intuito de desenvolver uma aplicação de machine learning comumente utilizado, o floresta aleatória. O algoritmo, que combina a saída de várias árvores de decisão para alcançar um único resultado, foi desenvolvido para realizar a classificação dos dados entre estrelas, quasares e galáxias, após a realização de múltiplos testes cegos, foi possível validar a eficiência do algoritmo para categorizar entre estrelas e galáxias, obtendo precisão de classificação superiores a 99%, porém, quando se trata de distinguir quasares o valor é reduzido para parcamente superior a 94% [7].

Possuindo o mesmo intuito do trabalho anterior, classificar as informações repassadas entre estrelas, quasares e galáxias, o artigo denominado “Stellar Objects Classification Using Supervised Machine Learning Techniques” desenvolve algoritmos de aprendizado de máquinas com o auxílio de um banco de dados mais atualizado, disponibilizado pelo Sloan Digital Sky Survey Data Release 17 (SDSS DR17).

O grande diferencial do editorial mencionado anteriormente são os múltiplos modelos de aprendizado de máquina que foram construídos, sendo eles a árvore de decisão, no qual sua composição é similar à de um fluxograma, com etapas muito fáceis de visualizar e entender; o KNN, que busca classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos; o perceptron multicamadas, modelo utilizado no presente artigo; o Naive Bayes, no qual se baseia nas descobertas de Thomas Bayes para realizar as predições; a floresta aleatória, o mesmo aplicado no artigo do Yu Bai; entre outros. O modelo floresta aleatória apresentou o melhor resultado, apresentando aproximadamente 98% de acurácia, e o que obteve pior rendimento foi o Naive Bayes com 91% de acurácia [8].

O artigo "Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects" também faz o uso de distintos modelos de aprendizado de máquina, mais especificamente o Naive Bayes, IB, máquina de vetores de suporte (SVM), perceptron multicamadas e a floresta aleatória, em todos aplicou-se os parâmetros de desempenho para realizar a avaliação posteriormente. Além dos múltiplos modelos, o referido trabalho aborda diferentes bancos de dados para realizar as considerações de qual apresenta um melhor resultado, as medidas de desempenho utilizadas foram a sensibilidade, especificidade e a acurácia balanceada. Levando em consideração o conjunto de dados SDSS DR17, o mesmo utilizado no presente trabalho, os modelos apresentaram acurácia de acordo com a Tabela 1 [9].

Naive Bayes	**IB**	**SVM**	**Perceptron Multicamadas**	**Floresta Aleatória**
0,7219	0,8621	0,9474	0,9738	0,9815

Tab. 1. Acurácia balanceada para o conjunto de dados SDSS DR17 (Adaptação). Fonte: [9].

III. METODOLOGIA

O SDSS (Sloan Digital Sky Survey) é um levantamento astronômico que opera desde 1998, onde os dados obtidos por suas observações são regularmente disponibilizados ao público a fim de contribuir com a comunidade cientifica. Em junho de 2001, o SDSS liberou seu primeiro levantamento de dados (SDSS-I), que contava com mais de 14 milhões de objetos detectados e 54.008 espectros que estavam em acompanhamento. Em 2005, o SDDS completou sua primeira fase de operação, onde ao longo desses 5 anos foram detectados cerca de 200 milhões de objetos celestes e mediu-se o espectro de mais de 675 mil galáxias, 90 mil quasares e 185 mil estrelas. O levantamento segue divulgando dados em fases mais atuais, com o objetivo de responder questões acerca da natureza do universo, a origem da galáxia e dos quasares, e a evolução e formação da via láctea [1].

O banco de dados utilizado nesse projeto consiste em 100.000 observações do espaço feita pelo SDSS no seu lançamento 17 (DR-17) que constitui o final da quarta fase do projeto (SDSS-IV), onde os dados foram obtidos até janeiro de 2020 e abrange mais de um terço de toda esfera terrestre. Nesse ponto, o levantamento do SDSS-IV já contava com 2.863.635 galáxias, 960.678 quasares e 1.021.843 estrelas [1]. Cada observação é descrita por 17 colunas de recursos e 1 coluna de classe, que a identifica em estrela, quasar ou galáxia. Abaixo segue a descrição de cada coluna:

1. Identificador do objeto: o valor único que identifica o objeto no catálogo de imagem usado pelo CAS;
2. α: ângulo de ascensão reta (na época J2000);
3. δ: Ângulo de declinação (na época J2000);
4. u: Filtro ultravioleta no sistema fotométrico;
5. g: Filtro verde no sistema fotométrico;
6. r: Filtro vermelho no sistema fotométrico;
7. i: Filtro de infravermelho próximo no sistema fotométrico;
8. z: Filtro infravermelho no sistema fotométrico;
9. Identificador de execução: número de execução usado para identificar a verificação específica;
10. *Rerun Number*: para especificar como a imagem foi processada;
11. Coluna da câmera: para identificar a linha de varredura dentro da execução;
12. Número do campo: para identificar cada campo;
13. Identificador dos objetos espectroscópicos: ID exclusivo usado para objetos espectroscópicos ópticos (isso significa que 2 observações diferentes com o mesmo identificador dos objetos espectroscópicos devem compartilhar a classe de saída);
14. Classe de objeto: galáxia, estrela ou quasar;
15. Desvio para o vermelho no espectro: valor do desvio para o vermelho baseado no aumento do comprimento de onda da luz;
16. Identificador da placa: identifica cada placa no SDSS;
17. Data de modificação: usada para indicar quando uma determinada parte dos dados do SDSS foi obtida;
18. Identificador da fibra: identifica a fibra que apontou a luz no plano focal em cada observação.

Os dados dividem-se em 59,45% referente a galáxias, 21,59% referente a estrelas, e 18,96% referente a quasares, vendo-se necessário então realizar um balanceamento através de uma implementação de dados por intermédio da Synthetic Minority Oversampling Technique (SMOTE), caso contrário, o modelo poderia ser vítima do paradoxo

da acurácia, na qual seus parâmetros não conseguiriam distinguir adequadamente a classe minoritária das demais categorias, levando-o a acreditar que está alcançando resultados sólidos devido à alta acurácia aparente. Essa falta de diferenciação poderia resultar em sérios problemas, pois as identificações dos casos minoritários poderiam ser cruciais para solucionar a classificação [10].

Além disso, devido ao algoritmo de aprendizado de máquina exigir que os dados sejam representados por valores numéricos, se viu necessário a conversão das classes de valores do tipo "object", no qual cada classe será representada por um vetor de três posições, onde apenas uma delas terá o valor 1 atribuído e todas as outras terão 0.

No modelo computacional do neurônio artificial, apresentado na Fig. 4, os sinais de entrada são representados por um vetor $x = [x_1, x_2, x_3, \ldots, x_N]$ que, ao chegarem ao neurônio, são multiplicados pelos respectivos pesos sinápticos, que são os elementos do vetor $w = [w_1, w_2, w_3, \ldots, w_N]$. O fator "bias" ou viés, representado pela letra "b", é adicionado com o intuito de promover um grau de liberdade maior, no qual este não é afetado pela entrada [11].

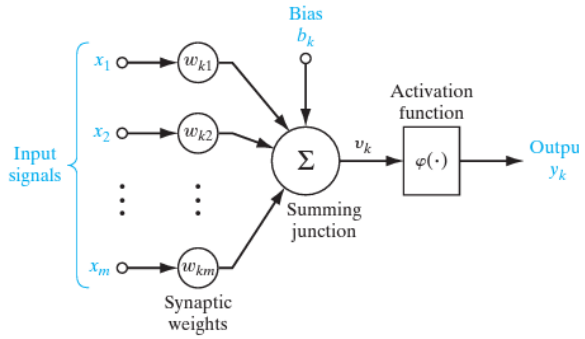

Fig. 4. Modelo computacional de um neurônio. Fonte: [11].

Como observado na fórmula (1), o somatório desses elementos resulta em um valor chamado de "z", que é comumente referido como potencial de ativação [11].

$$z = \sum_{i=1}^{N} x_i w_i + b \qquad (1)$$

O valor z é submetido a uma função matemática de ativação σ, que possui a propriedade de ser não linear. Essa função é responsável por restringir o valor z a um intervalo específico, resultando no valor de saída final do neurônio, denominado y. Diferentes funções de ativação são utilizadas, como a função degrau, sigmoide, tangente hiperbólica, softmax e a ReLU (Rectified Linear Unit). Cada uma dessas funções tem características distintas e é aplicada de acordo com as necessidades do modelo de rede neural em questão [12].

O modelo implementado para a classificação desses objetos caracteriza-se como uma rede neural densa Multilayer Perceptron (MLP), no qual possui uma camada oculta com 17 valores de entrada e 4690 neurônios, como observado na Fig. 5.

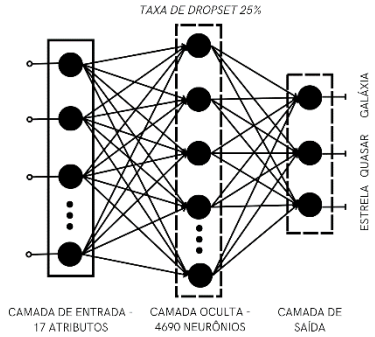

Fig. 5. Modelo de rede neural implementado. Fonte: Autoria própria.

A quantidade de neurônios foi determinada pela fórmula (2), no qual Ni é o número de neurônios de entrada, No é o número de neurônios de saída, Ns é o número de amostras e α é um fator de escala arbitrário, geralmente variando entre 2-10.

$$N_h = \frac{N_s}{\alpha * (N_i + N_o)} \qquad (2)$$

Tendo como propósito encontrar o valor que resultasse no melhor modelo possível, aplicou-se os valores de dois, seis e dez em distintos testes, após a análise dos resultados obtidos observou-se que o mais próximo do ideal para o modelo era um total de 4690 neurônios na camada oculta, valor este obtido aproximadamente com α igual a dois.

Foi implementado no modelo a técnica de regularização Dropout, onde consiste em selecionar e temporariamente desativar de forma aleatória alguns dos neurônios nas camadas ocultas da rede, atualizando os pesos e vieses. Essa técnica foi implementada com o objetivo de permitir que a rede generalize melhor para novos dados, evitando o overfitting. O processo de desativação e ativação dos neurônios, juntamente com as atualizações de pesos e

vieses, permite melhorar o desempenho e a robustez da rede [13].

Tendo como intuito converter qualquer número real oriundo da resultante do somatório realizado nos neurônios em um único valor entre zero e um, aplicou-se a função de ativação sigmoid na camada oculta. Esta função, descrita pela fórmula (3), obtém-se como retorno um valor próximo a 0 quando a entrada são valores pequenos e resulta um valor próximo a 1 quando inseridos números grandes [14].

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

A camada oculta é conectada diretamente com a camada de saída, no qual essa possui três neurônios, sendo responsável por compor o vetor de três posições que representam as classes, caso a saída resulte no vetor y = [1, 0, 0] será classificado como galáxia, o vetor y = [0, 1, 0] representará o quasar e y = [0, 0, 1] corresponderá a estrela.

Por tratar-se da última camada de uma rede neural que possui como propósito realizar uma classificação, softmax foi a função de ativação aplicada nesta camada, sendo descrita pela fórmula (4), para j = 1, …, K.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \quad (4)$$

A função Softmax transforma um vetor de valores em uma distribuição de probabilidade, onde os elementos do vetor resultante estão restritos ao intervalo de zero à um e sua soma total é igual a 1. Essa operação é realizada de maneira independente para cada vetor [14].

Por não exigir um ajuste acurado da taxa de aprendizado e por empregar estimativas adaptativas do primeiro e segundo momentos para ajustar os pesos da rede, demonstrando uma tendência a alcançar o mínimo global de forma mais eficiente, foi implementado o otimizador Adam (Estimativa de Momento Adaptativo) com taxa de aprendizado 0.01 [15].

IV. RESULTADOS

Para a obtenção do modelo de classificação foi necessário realizar algumas etapas. Inicialmente foi realizada a etapa de treinamento, exposto na seção V-A, a etapa de avaliação do modelo está apresentada na seção V-B.

A. Treinamento dos modelos

Tendo como objetivo realizar o treinamento do modelo, o dataset foi embaralhado e dividido em subconjuntos aleatórios com 80% dos dados para treino, onde 20% foram destinados para validação, e 20% para teste.

O treinamento de uma rede neural envolve iterar por várias épocas, ajustando os pesos na busca pela redução de erros. No entanto, não existe um número universalmente ideal de épocas para treinar uma rede neural, se a quantidade for muito baixa, pode ocorrer underfitting, enquanto um número excessivo pode levar ao overfitting [13]

Para mitigar esses problemas a estratégia de Early Stopping foi empregada. Isso envolve calcular a precisão da classificação nos dados de validação, ao final de cada época, no modelo em questão a métrica avaliada foi a entropia cruzada. Quando essa métrica para de melhorar, o treinamento é interrompido, evitando assim o overfitting [13].

A abordagem adotada no modelo foi de parar o treinamento após cinco quedas seguidas da métrica, evitando o risco de encerrar o processo quando ainda há melhorias a serem obtidas, resultando em 31 épocas.

B. Avaliação do modelo

A entropia cruzada é uma métrica clássica amplamente utilizada para abordar problemas de classificação, por esse motivo, foi aplicado no modelo em questão. Nesse tipo de problema, o objetivo é categorizar instâncias em diferentes classes ou grupos, cada um com sua própria distribuição probabilística. Quanto mais claramente definidos esses grupos estiverem, menor será a medida de entropia. A fórmula (5) é utilizada para a realização do cálculo, no qual y_j representa os valores da classe verdadeira, e p_j os valores da predição [16].

$$L = -\sum_{j=1}^{m} y_j * \log(p_j) \quad (5)$$

Tendo como objetivo analisar graficamente a curvatura da função perda do treino e da validação, plotou-se o Gráfico 1, no qual pode-se observar a redução da entropia cruzada em função da quantidade de épocas realizadas.

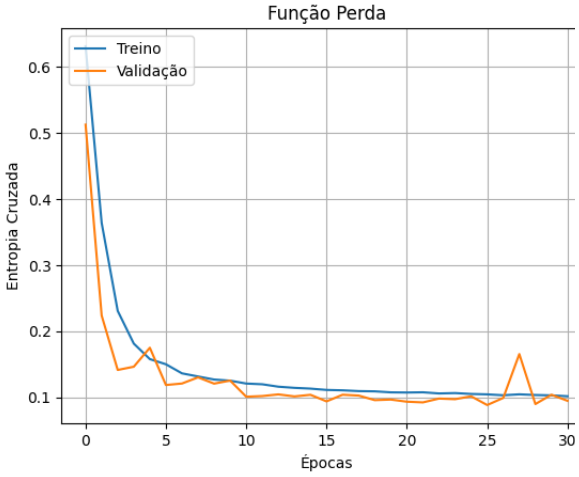

Gráfico 1. Entropia cruzada x Épocas. Fonte: Autoria própria.

Com o propósito de validar a eficiência do modelo de rede neural construído, aplicou-se outras métricas, entre elas a matriz de confusão, a acurácia, a precisão, a sensibilidade e o F1.

A matriz de confusão é uma ferramenta usada, em forma de tabela, que pode ajudar a compreender com que frequência um classificador classifica cada classe. Cada linha da matriz de confusão representa a instância de uma classe real e cada coluna representa a instância de uma classe prevista, sendo que o ideal nessa matriz é que os classificadores considerados ideias contenham contadores elevados na diagonal dos verdadeiros. Na Fig. 6 pode-se avaliar a matriz de confusão da classificação das galáxias, quasares e estrelas, sendo representados por 0, 1 e 2 respectivamente.

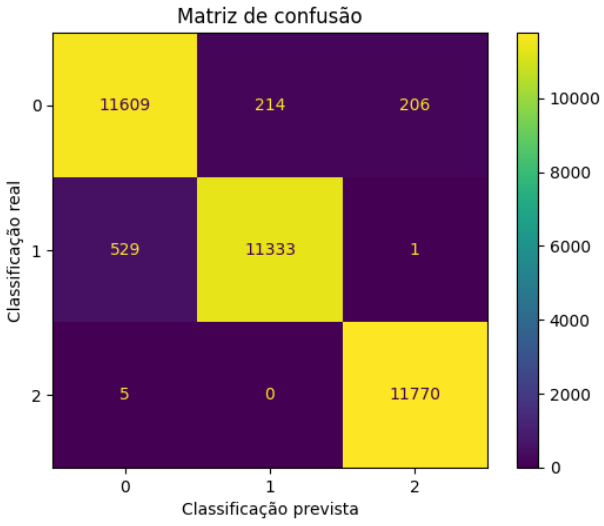

Fig. 6. Matriz de confusão do modelo. Fonte: Autoria própria.

A acurácia é uma medida direta de quão bem o modelo está fazendo previsões corretas em relação ao total de previsões [15].

A precisão é a ferramenta usada para calcular o percentual de predições positivas que estavam corretas, usada normalmente quando se deseja minimizar os falsos positivos [15].

O recall, também conhecido como sensibilidade, é a ferramenta usada para calcular o percentual de valores positivos classificados corretamente [15].

O F1 é a média harmônica entre precisão e o recall. Ele é importante para o balanceamento entre a precisão e a sensibilidade, pois considera tanto os falsos positivos quanto os falsos negativos [15].

Todos os valores das métricas comentadas anteriormente atingidas pelo modelo estão presentes na Tab. 2.

Acurácia	**Precisão**	**Recall**	**F1**
0.97322	0.97340	0.97333	0.97328

Tab. 2. Métricas. Fonte: Autoria própria.

V. CONCLUSÃO

Neste trabalho, foi elaborado o desenvolvimento de uma rede neural de classificação utilizando o modelo perceptron multicamadas. Inicialmente, foi realizada a coleta dos dados, e os ajustes necessários. Posteriormente, o modelo foi construído e analisado através de métricas. O gráfico da função perda permite visualizar que a curva de treinamento e a de validação assemelham-se, finalizando o treinamento com o valor da entropia cruzada menor que 0.10, permitindo afirmar a ausência de erros como overfitting e underfitting. A matriz de confusão possibilita observar valores elevados na diagonal dos verdadeiros, aproximando-se do ideal. Para avaliar a validade de um modelo é fundamental analisar distintas métricas e considerar as implicações de prever falsos positivos e falsos negativos, em todos os cálculos de métricas aplicados, o modelo obteve valores acima de 97,00%.

Portanto, quando comparado com outros algoritmos apresentados em trabalhos relacionados, pode-se observar valores de acurácia deveras semelhantes, porém, o presente trabalho acarreta como diferencial a presença de um balanceamento do banco de dados, aumentando a eficiência do modelo. Dessa forma, tem-se que o alto valor percentual, a semelhança entre os valores obtidos nas métricas, e a equipolência com outros modelos possibilita verificar a eficiência do modelo construído.

VI. REFÊNCIAS

[1] SDSS - Sloan Digital Sky Survey. Disponível em: https://www.sdss.org/.

[2] SUTTER, Paul. What is a star? Space, Nova York, 28, jan. 2022. Disponível em: https://www.space.com/what-is-a-star-main-sequence?fbclid=IwAR2qGwPJ77HFoO1TWSNzAye3_OwXr_0b6AH8ODGy0q36IJ5A5YWgBAZHShE. Acesso em: 13, ago. 2023.

[3] FILHO, K. S.; SARAIVA, M. F. Quasares, 5, abr. 2023. Disponível em: http://astro.if.ufrgs.br/galax/index.htm#quasares. Acesso em: 13, ago. 2023.

[4] CFHT (Canada-France-Hawaii Telescope). Disponível em: https://www.cfht.hawaii.edu/.

[5] MANN, Adam. What is a galaxy? Live Science, 25, ago. 2021. Disponível em: https://www.space.com/15680-galaxies.html. Acesso em: 13, ago. 2023

[6] NASA SCIENCE - National Aeronautics and Space Administration Science. Disponível em: https://science.nasa.gov/astrophysics/focus-areas/what-are-galaxies. Acesso em 13, ago. 2023.

[7] BAI, Yu et al. Machine Learning Applied to Star–Galaxy–QSO Classification and Stellar Effective Temperature Regression. The Astronomical Journal, v. 157, n. 1, 14 de dezembro de 2018.

[8] OMAT, Deen; OTEY, Jood; AL-MOUSA, Amjed. Stellar Objects Classification Using Supervised Machine Learning Techniques. International Arab Conference on Information Technology, Abu Dhabi, United Arab Emirates, p. 1-8, novembro de 2022.

[9] RAMÍREZ, José Luis Solorio et al. Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects. Algorithms, México, junho de 2023.

[10] COOPER, Martin. Turing Talk 2022. ITNOW, v. 64, n. 2, p. 35, maio de 2022.

[11] HAYKIN, Simon. Neural Networks and Learning Machines. (2009). Canadá: Prentice Hall.

[12] CARVALHO, André Ponce de Leon F. de. Redes Neurais Artificiais, 2009. Disponível em: https://sites.icmc.usp.br/andre/research/neural/. Acesso em: 02 de agosto de 2023.

[13] Data Science Academy. Deep Learning Book, 2022. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 18, agosto. 2023.

[14] Layer activation functions. Keras, 2023. Disponível em: https://keras.io/api/layers/activations/. Acesso em: 03 de agosto de 2023.

[15] HARRISON, Matt. Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python. (2019). Brasil: Novatec Editora. Brasil: Novatec Editora, 2019.

[16] CECCON, Denny. Conceitos sobre IA - Fundamentos de ML: funções de custo para problemas de classificação, 2019. Disponível em: < https://iaexpert.academy/>. Acesso em: 20, agosto. 2023.