10.18605/2175-7275/cereus.v17n4p143-152

<< Recebido em: 10.01.2025 Aceito em: 13.06.2025 >>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Classificação de arroz: um estudo de caso utilizando a rede neural MLP

Rice Classification: A Case Study Using MLP Neural Network

José Airton Azevedo dos Santos[1], Aldino Normelio Brun Polo[2], Cidmar Ortiz dos Santos[3]

RESUMO

Este trabalho tem como objetivo analisar o desempenho, na classificação de variedades de arroz, do modelo de rede neural *Multilayer Perceptron* (MLP). A base de dados, obtida da plataforma Kaggle, apresenta 18185 observações das variedades Jasmine e Gonen. Para implementação do modelo de classificação MLP, na linguagem de programação Python, utilizou-se o módulo MLPClassifier da biblioteca Scikit-Learn. Scikit-Learn é uma biblioteca desenvolvida especificamente para aplicação prática de *machine learning*. O treinamento do modelo envolve seleção de hiperparâmetros e validação cruzada para evitar *overfitting*. Para balanceamento das classes utilizou-se a função *resample* da biblioteca Scikit-Learn. O desempenho do modelo MLP foi verificado por meio das métricas *accuracy* e AUC (*Area under the ROC Curve*). Resultados experimentais demonstraram que o modelo MLP apresentou excelentes resultados: *accuracy*=99% e AUC= 0,9869.

Palavras-chave: Classificação; Optuna; resample; Scikit-Learn.

ABSTRACT

This work aims to analyze the performance of the Multilayer Perceptron (MLP) neural network model in the classification of rice varieties. The database, obtained from the Kaggle platform, presents 18185 observations of the Jasmine and Gonen varieties. To implement the MLP classification model, in the Python programming language, the MLPClassifier module of the Scikit-Learn library was used. Scikit-Learn is a library developed specifically for practical application of machine learning. Model training involves hyperparameter selection and cross-validation to avoid overfitting. The resample function of the Scikit-Learn library was used to balance the classes. The performance of the MLP model was verified through the accuracy and AUC (Area under the ROC Curve) metrics. Experimental results demonstrated that the MLP model presented excellent results: accuracy = 99% and AUC = 0.9869.

Keywords: Classification; Optuna; resample; Scikit-Learn.

[1] Doutor em Engenharia Elétrica. Docente do Programa de Pós-Graduação em Tecnologias Computacionais para o Agronegócio (PPGTCA). Universidade Tecnológica Federal do Paraná (UTFPR). E-mail: airton@utfpr.edu.br - https://orcid.org/0000-0002-2568-5734

[2] Discente do Programa de Pós-Graduação em Tecnologias Computacionais para o Agronegócio (PPGTCA).. Universidade Tecnológica Federal do Paraná (UTFPR). E-mail: aldinopolo@alunos.utfpr.edu.br - https://orcid.org/0009-0000-2969-274X

[3] Doutor em Ensino de Ciência e Tecnologia. Docente do Curso de Engenharia de Produção. Universidade Tecnológica Federal do Paraná (UTFPR). E-mail: cidmar@utfpr.edu.br - https://orcid.org/0000-0002-1727-

1. INTRODUÇÃO

O arroz é uma das commodities mais produzidas no mundo. Cerca de 90% da produção mundial concentra-se na Ásia. A China e a Índia são os maiores produtores, juntos produzem 54% do arroz do mundo. O arroz, em termos de produção, fica apenas atrás do trigo e do milho. O Brasil, maior produtor da America latina, tem uma produção anual de aproximadamente 10 milhões de toneladas. O estado do Rio Grande do Sul é o maior produtor nacional de arroz. Produzindo, em média, 8 milhões de toneladas de grão por safra (BASF, 2024).

Segundo Koklu et al. (2021), avaliar ou classificar grãos por métodos manuais pode ser demorado e custoso. Em métodos manuais, o processo de avaliação pode diferir, pois é limitado à experiência dos especialistas em avaliação. Além disso, a tomada de decisão rápida por métodos manuais pode ser difícil quando uma avaliação é feita em larga escala.

Redes Neurais Artificiais (RNAs) podem ser utilizadas para classificar variedades de arroz. A RNA é um método de inteligência artificial muito utilizado, como ferramenta de classificação, em diversas áreas de aplicação. É composta por unidades de processamento simples, chamadas de neurônios, capazes de memorizar, analisar e processar padrões complexos de dados. Podem ser aplicadas para problemas de regressão e classificação (SANTOS, 2021; PINHEIRO et al., 2020; BASTIANI et al, 2018, ANTONIOLLI et al, 2022; PIMENTEL, et al., 2023; SPANCERSKI; SANTOS, 2021a).

Diversos trabalhos, apresentados na literatura, utilizaram métodos de previsão, de séries temporais, em aplicações voltadas ao mercado do arroz. Noorunnahar et al. (2023) compararam os desempenhos, na previsão da produção anual de arroz em Bangladesh, dos modelos ARIMA (*Autoregressive Integrated Moving Average*) e XGBoost (*extreme Gradient Boosting*). Observaram, que o XGBoost, em relação ao modelo ARIMA, apresenta melhor desempenho de previsão da produção anual de arroz. Airlanga et al. (2019) utilizaram os métodos de suavização exponencial e redes neurais para prever a produção de arroz na Indonésia. Concluíram que o método de rede neural é superior, em termos de previsão, ao método estatístico. Já Spancerski e Santos (2021b) proporam, para previsão da produtividade de arroz no Rio Grande do Sul, a rede neural LSTM (*Long Short-Term Memory*). Observaram, neste trabalho, que a rede LSTM forneceu estimativas confiáveis de produtividade do arroz.

Neste contexto, este trabalho tem como objetivo analisar o desempenho, na classificação de variedades de arroz, do modelo de rede neural *Multilayer Perceptron* (MLP).

2. MATERIAIS E MÉTODOS

Nesta seção, apresenta-se uma breve descrição da Rede Neural MLP, da base de dados *Rice Seed Dataset (Gonen&Jasmine)*, bem como as métricas utilizadas.

Rede MLP:

Redes Neurais MLP, baseadas em aprendizado supervisionado, são redes neurais *feed-forward* usadas para resolver problemas de regressão e classificação. A rede neural é composta por uma camada de entrada, uma ou mais camadas ocultas de unidades de processamento e uma camada de saída. Os nós das camadas ocultas recebem sinais, dos nós da camada de entrada, por meio de conexões ponderadas e, após calcular a saída da função de ativação, transmitem a saída para os nós da camada de saída. O treinamento da rede é realizado pelo algoritmo *Backpropagation* (SANTOS; PINTO, 2024; PINHEIRO et al., 2020).

Base de dados e Pré-processamento:

Utilizou-se, para classificação das variedades de arroz, uma base de dados, com 18185 observações, fornecida pelo repositório Kaggle (SEYMA, 2020). Esta base dados, denominada *Rice Seed Dataset (Gonen&Jasmine)*, fornece informações obtidas das variedades de arroz Jasmine e Gonen. A variedade Jasmine tem origem na Tailândia. Esta variedade apresenta níveis médios de rendimento, cerca de 66% das taxas médias mundiais. Já a variedade Gonen, com origem Turca, apresenta-se como a variedade que tem o terceiro rendimento médio do mundo (GURRALA, 2021).

Na sequência, apresentam-se as descrições das variáveis da base de dados (SAAD;ABU-NASSER, 2023).

Area: Área total do grão de arroz.

MajorAxisLength: Comprimento do eixo mais longo do grão de arroz.

MinorAxisLength: Comprimento do eixo mais curto do grão de arroz.

Eccentricity: Caracteriza o quanto o grão de arroz se desvia de um círculo perfeito.

ConvexArea: Área do menor polígono convexo que pode envolver o grão de arroz.

EquivDiameter: Diâmetro de um círculo com a mesma área do grão de arroz.

Extent: Mede a proporção da área do grão de arroz para a área de sua caixa delimitadora.

Perimeter: Captura detalhes finos do formato do grão.

Roundness: Reflete o quão próximo o grão de arroz se assemelha a um círculo perfeito.

AspectRation: Fornece informações sobre o alongamento ou achatamento de grãos de arroz.

Class (Resultado): Um valor binário que indica se a variedade é Gonen (0) ou Jasmine (1).

Os histogramas, das variáveis *Area*, *MajorAxisLength*, *mMinorAxisLength*, *Eccentricity*, *ConvexArea*, *EquivDiameter*, *Extent*, *Perimeter*, *Roundness*, *AspectRation* e *Class*, são apresentados na Figura 1. Uma visão geral dos dados é obtida ao plotar os histogramas das variáveis de entrada e da variável de saída. Por exemplo, pode-se observar que *Class* é uma variável binária, com apenas dois valores possíveis 0 e 1.

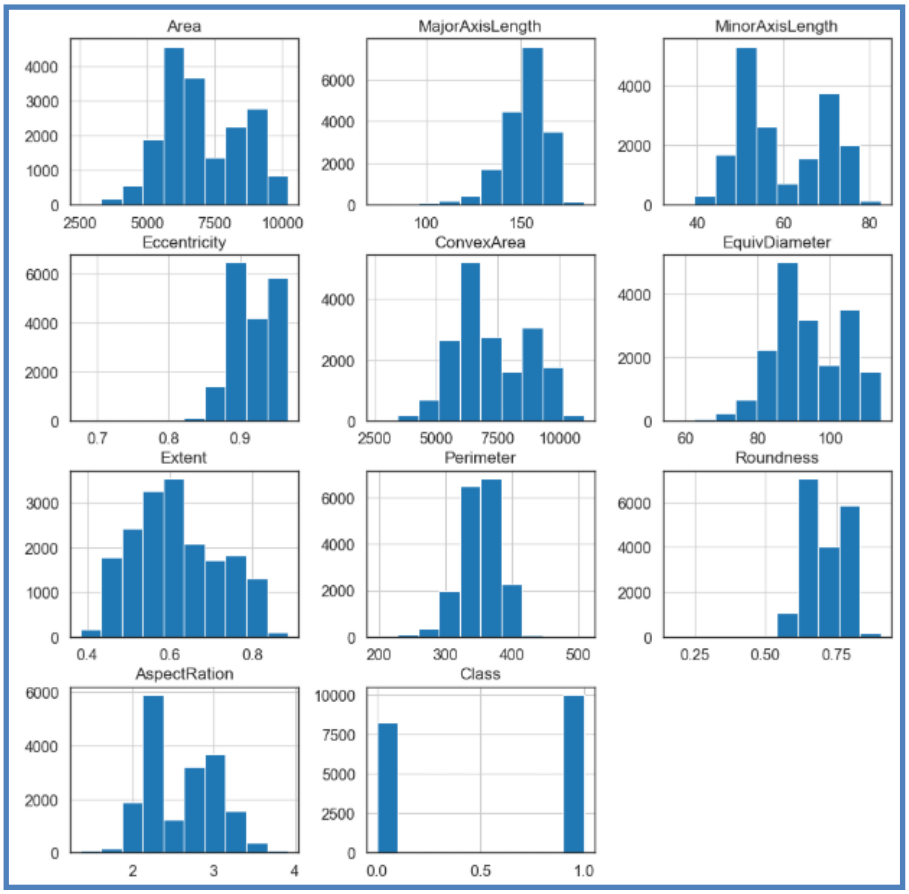

Figura 1. Distribuições das variáveis da base de dados.

Os *boxplots*, das variáveis em estudo, são apresentados na Figura 2. Observa-se, em algumas dessas variáveis, a presença de *outliers*.

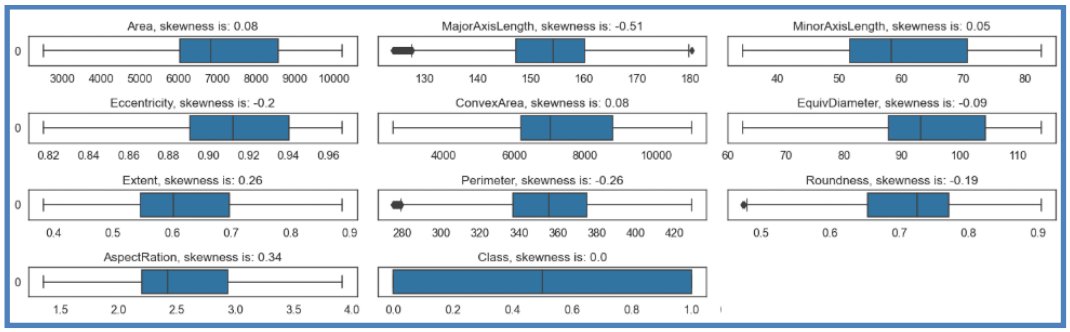

Figura 2. *Boxplots - Outliers*.

Neste trabalho, para suavizar os impactos dos *outliers* nos resultados, substituiu-se, nas respectivas colunas, os valores dos *outliers* pelas medianas de cada uma das variáveis.

Foi observado, na base de dados *Rice Seed Dataset (Gonen&Jasmine)*, um desbalanceamento entre as classes: Gonen (0)=45,1% e Jasmine (1)=54,9% (Figura 3). Dados desbalanceados podem impactar na generalização do modelo de classificação. Utilizou-se, para realizar o balanceamento das classes Jasmine (1) e Gonen (0), a função *resample* da biblioteca Scikit-Learn. A função *resample* cria dados sintéticos para classe minoritária até que as classes estejam em equilíbrio.

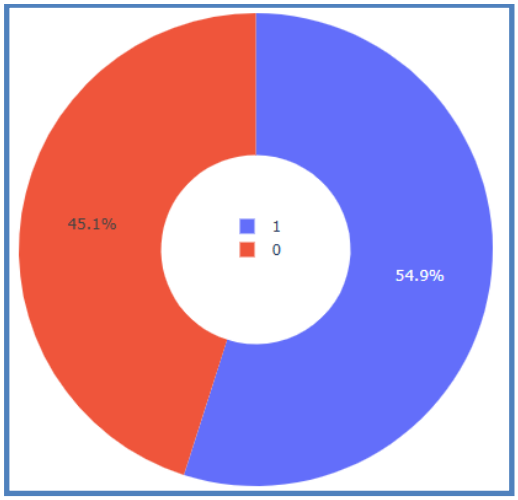

Figura 3. Classes desbalanceadas.

A base de dados, com objetivo de avaliar a precisão da classificação, foi dividida com 70% dos dados para treinamento e 30% para teste. Observa-se também que, neste

trabalho, a validação cruzada *K-fold*, com *k*=10, é adotada para avaliar o desempenho do modelo durante a fase de seleção do modelo.

Métricas:

O desempenho do modelo MLP foi avaliado pelas seguintes métricas (CHANG et al., 2023; MAULANA et al., 2023; BUDHOLIYA et al., 2022):

Accuracy: representa a proporção de amostras classificadas corretamente em relação ao número total de amostras. A equação da métrica *Accuracy* é dada por (Equação 1):

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad 1$$

Onde: TP: Verdadeiro Positivo; TN: Verdadeiro negativo; FP: Falso Positivo e FN: Falso Negativo.

Area under the ROC Curve (AUC): é um valor escalar único que varia de 0 a 1 e fornece um instantâneo do desempenho do modelo. AUC=1 indica o desempenho perfeito.

3. RESULTADOS E DISCUSSÃO

Treinamento:

A otimização dos hiperparâmetros de um modelo melhora significativamente sua capacidade de generalização. Utilizou-se, para otimização dos hiperparâmetros do modelo MLP, a biblioteca Optuna. O Optuna utiliza, para encontrar os melhores hiperparâmetros de um modelo, a técnica de otimização baseyana. O melhor modelo, encontrado pelo Optuna, utilizou o algoritmo de otimização Adam com os seguintes hiperparâmetros (Tabela 1):

Tabela 1. Parâmetros - MLP.

Hiperparâmetro	Intervalo de Busca	MLP
learning_rate	['constant','invscaling','adaptive']	'invscaling'
hidden_layer_size	[(5,10,5),(20,10),(10,20),(50,50),(100,100)]	(10, 20)
alpha	[.3,.1,.01,.001,.0001]	0.001
activation	['logistic','relu','tanh']	'tanh'

Maiores detalhes sobre estes hiperparâmetros podem ser encontrados em Scikit-learn (2024).

Teste:

A Figura 4 fornece um resumo do desempenho, no conjunto de teste, do modelo MLP. Observa-se, por meio dos resultados apresentados, que o modelo MLP apresentou uma acurácia de 99%.

	precision	recall	f1-score
0.0	0.98	0.99	0.99
1.0	0.99	0.98	0.99
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

Figura 4. Desempenho do modelo MLP.

A matriz de confusão fornece um resumo da capacidade do modelo MLP de diferenciar as variedades do arroz (Figura 5). Pode-se notar, por meio da matriz, que 2932 das observações, do Conjunto de Teste, foram classificadas como Gonen e 32 foram classificadas incorretamente como Gonen enquanto eram Jasmine. Como Jasmine foram classificados 2987 observações e 40 observações foram classificadas incorretamente como Jasmine.

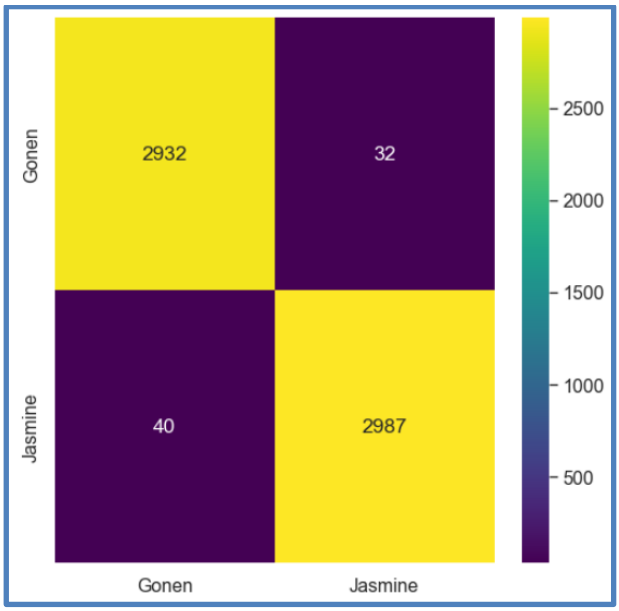

Figura 5. Matriz de confusão – MLP.

A Curva ROC para o modelo MLP é apresentada na Figura 6.

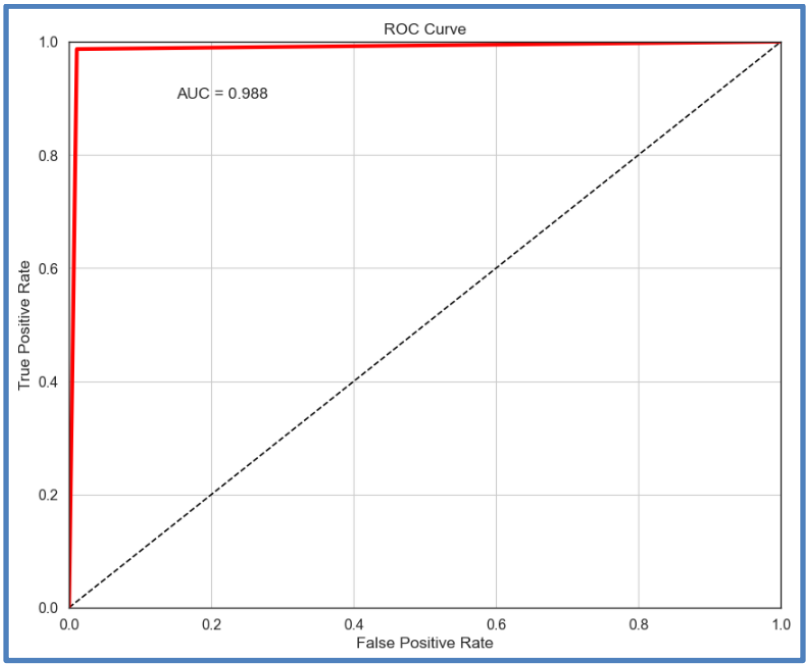

Figura 6. Curva ROC – MLP.

Pode-se observar, por meio do indicador AUC apresentado na Figura 6, que o modelo MLP apresentou muito bom desempenho (AUC=0,988). Lembrando que AUC=1 indica o desempenho perfeito.

4. CONSIDERAÇÕES FINAIS

O arroz, como uma cultura básica primordial em todo o mundo, sustenta bilhões de vidas. A classificação precisa dos tipos de arroz tem imensa importância agrícola, nutricional e econômica. Neste contexto, este trabalho propôs, para classificação de variedades de arroz (Gonen/Jasmine), a utilização de um modelo de rede neural MLP. Para construção desta proposta foram utilizados dados, das variedades de arroz Jasmine e Gonen, fornecidos por meio do repositório Kaggle.

Resultados de classificação do modelo proposto, no conjunto de teste, apresentaram uma acurácia de 99% e um AUC de 0,9869. Estes resultados demonstram o muito bom desempenho, do modelo de rede neural MLP, na classificação das variedades de arroz Gonen e Jasmine. Observou-se também que as métricas *precision*, *recall* e *f1-score* foram consistentemente altas, destacando a robustez do modelo proposto.

Para trabalhos futuros sugere-se: melhorar a interpretabilidade do modelo para fornecer *insights* para utilização em processos de tomada de decisão; pesquisar recursos

adicionais que possam melhorar a precisão da classificação e fazer parceria, com a indústria nacional, para testar este modelo em cenários práticos.

REFERÊNCIAS

AIRLANGGA, G.; RACHMAT, A.; LAPIHU, D. Comparison of exponential smoothing and neural network method to forecast rice production in Indonesia. **TELKOMNIKA**. v. 17, 2019.

ANTONIOLLI, E. A.; SANTOS, J. A. A.; GUEDES, A. L.; PASA, L. A. Previsão do consumo de energia elétrica em um frigorifico: um estudo de caso usando regressão linear, redes neurais e máquina de vetor de suporte. **Tekhne Logos**. 2022.

BASF. **Produção de arroz no Brasil: quais os principais Estados produtores?** Disponível em: https://agriculture.basf.com/br/pt/conteudos/cultivos-e-sementes/arroz/estados-produtores. Acesso em: 10 mai. 2024.

BASTIANI, M., SANTOS, J. A. A., SCHMIDT, C. A P., & SEPULVEDA, G. P. L. Application of data mining algorithms in the management of the broiler production**. Geintec**, v. 8, 2018.

BUDHOLIYA, K.; KUMAR, S.; VIVEK, S. An optimized XGBoost based diagnostic system for effective prediction of heart disease. **Journal of King Saud University - Computer and Information SciencesSharma**. v. 34, 2022.

CHANG, V.; BAILEY, J.; XU, Q. A.; SUN, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. **Neural Computing and Applications**, 2023.

KOKLU, M.; CINAR, Y.; TASPINAR, Y. S. Classification of rice varieties with deep learning methods. **Computers and Electronics in Agriculture**. v. 187, 2021.

MAULANA, A.; FAISAL, F. F. NOVIANDY, T. R.; RIZKIA, T.; IDROES, G. M.; IDROES, G. M.; TALLEI, T. E. Machine learning approach for diabetes detection using finite tuned XGBoost algorithm. **Infolitika Journal of Data** Science, 2023.

NOORUNNAHAR, M.; CHOWDHURY, A. H.; MILA, F. A. A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. **PLoS ONE**. v. 18, 2023.

PIMENTEL, C. C.; SANTOS, J. A. A.; SANDMANN, A. et al. Previsão de geração de energia hidráulica no Brasil: um estudo de caso usando redes neurais artificiais e regressão linear. **Caderno Pedagógico**. 2023.

PINHEIRO, T C.; SANTOS, J. A. A.; PASA, L. A. Gestão da produção de frangos de corte por meio de redes neurais artificiais. **Holos**, v. 2, 2020.

SAAD, A. S.; ABU-NASSER, s. s. Rice Classification using ANN. **International Journal of Academic Engineering Research (IJAER)**. v. 7, 2023.

SANTOS, J. A. A.Aplicação de Redes Neurais na Previsão de Custos de Produção de Frangos de Corte no Estado do Paraná. **Revista FSA**, v. 18. 2021.

SANTOS J. A. A.; PINTO, B. R. Análise comparativa do desempenho de três modelos de previsão de séries temporais do software WEKA. **Revista Cereus**. v. 16, 2024.

SEYMA. S, **Rice Seed Dataset (Gonen & Jasmine)**, Kaggle, 2020. Disponível em: https://www.kaggle.com/datasets/seymasa/rice-dataset-gonenjasmine. Acesso em: 10 dez. 2024.

SCIKIT-LEARN. **Classificador MLP.** 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Acesso em: 10 dez. 2024.

SPANCERSKI, J.; SANTOS, J. A. A. Previsão do preço do açúcar: um estudo de caso usando o software Weka. **Engenharia e Tecnologia**. 2021.

SPANCERSKI, J.; SANTOS, J. A. A. Previsão da produtividade de arroz: uma aplicação de redes neurais recorrentes LSTM. **Revista Cereus**. v. 13, 2021.