http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2023.v15.e414
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 01/08/2023 Revisado: 27/09/2023 Aceito: 14/11/2023

PIXEL-ORIENTED VISUALIZATION FOR EXPLAINING DATA CLASSIFICATION IN A MULTILAYER NEURAL NETWORK

Visualização Orientada a Pixel para Explicação da Classificação de Dados em uma Rede Neural Multicamadas

Marcelo Tenorio[1], Danilo Eler[2]

[1]FATEC, Presidente Prudente, SP. [2]Universidade Estadual Paulista – UNESP, Presidente Prudente, SP.
E-mail: marcelo.tenorio@fatec.sp.gov.br; danilo.eler@unesp.br

ABSTRACT – Accompanying the growth of applications that use Artificial Intelligence, recent research is also growing to explain how these applications work and make them more acceptable to humans. In this context, this paper presents an alternative explanation of the data classification process carried out by an Artificial Intelligence algorithm. The work proposes a pixel-oriented information visualization approach to explain the multilayer perceptron classifier using SHAP. Observing the results obtained in Shapley values, for the Iris dataset composed of four features, the proposed methodology identified one relevant feature, and for the Wine dataset composed of 13 features, the methodology identified six relevant features. The relevant features are those that most influence the classification, this information explains the results as it is possible to understand the reasons for the classifier's successes and errors.
Keywords: Artificial Neural Network; SHAP; XAI; Information Visualization.

RESUMO – Acompanhando o crescimento de aplicativos que utilizam Inteligência Artificial, também crescem pesquisas recentes para explicar o funcionamento desses aplicativos e torná-los mais aceitáveis pelo homem. Neste contexto, este artigo apresenta uma explicação alternativa do processo de classificação de dados realizado por um algoritmo de Inteligência Artificial. O trabalho propõe uma abordagem de visualização de informação orientada a pixel para explicar o classificador perceptron multicamada usando SHAP. Observando os resultados obtidos nos valores Shapley, para o conjunto de dados Iris composto de quatro características, a metodologia proposta identificou uma característica relevante, e para o conjunto de dados Wine composto de 13 características, a metodologia identificou seis características relevantes. As características relevantes são as que mais influenciam na classificação, esta informação explica os resultados pois é possível entender as razões dos acertos e erros do classificador.
Palavras-chave: Rede Neural Artificial; SHAP; XAI; Visualização da Informação.

1. INTRODUCTION

For centuries, machines have become assistants in various human tasks and for decades, such machines have been endowed with intelligence and connectivity. Proof of this is the four industrial revolutions already experienced by humanity (Sakurai; Zuchi, 2018). However, for machines to effectively do certain human activities, many studies have been conducted in the last ten years to make machine learning more acceptable (Arrieta *et al.*, 2020). This acceptance is directly linked to the understanding of how machines make decisions.

There are several ways for machines to learn something, for example, using mathematical or statistical models. However, for many years models inspired by nature, whether human or animal, called intelligent, have been widely applied in machine learning. The initial models that provide intelligence to machines are easier to understand, but recent models compromise the understanding of their behavior. Mastering the behavior of an intelligent system is extremely necessary when discussing ethics, justice, security, trust, and other topics in the social sciences (Miller, 2019). In this context, this paper presents a study related to Artificial Intelligence (AI) and Information Visualization for understanding the results of a data classifier. For that, a new pixel-oriented visualization approach is proposed to aid the explanation of classifiers using SHapley Additive ExPlanations (SHAP). Such an approach shows information about predictions and SHAP information to quantify the contribution of each feature in the classifier predictions. Thus, the main contribution of this paper is a new approach capable of explaining the importance of each feature in the classifier results. In the experiments, it is possible to note how each feature was relevant to the correct prediction and the mistakes.

The other sections of this work are organized as follows: section 2 presents a brief theoretical background; section 3 presents the work methodology; section 4 presents the results and discussions. Finally, section 5 presents conclusions and some suggestions for future work.

2. BACKGROUND

In the last five years, studies in AI have presented important definitions for directing research that seeks to make decisions made by machines more acceptable (Pantelis; Papastefanopoulos; Kotsiantis, 2021).

Currently, AI can be categorized as interpretable or explainable. In interpretable AI, the models (considered transparent) are understood by humans; conversely, in explainable AI, the models (regarded as a black box) need explanation because humans do not understand them. According to Arrieta *et al.* (2020), responsible AI encompasses the characteristics of interpretable and explainable AI. Responsible AI seeks to provide understanding, comprehension, interpretation, explanation, and transparency for models. Responsible AI wants to improve answers to broaden personal and professional applications. Still according to Arrieta *et al.* (2020), transparent models have levels of transparency, can have algorithmic transparency, can be decomposed, or can be simulated. For black box models, there are textual, visual, and local explanations, for example, by simplification or by feature relevance.

There are several methods for explaining black box models, some for specific models, some for any models. SHAP is a method to describe any model (Arrieta *et al.*, 2020). SHAP stands for SHapley Additive ExPlanations, an approach inspired by the Shapley values of cooperative game theory (Malato, 2023). The method evaluates the impact of the features on the result. It's like a football game, where each player contributes to the team. The central concept is that the outcome depends not on a single player but on a group of players. Therefore, SHAP calculates the impact (called Shapley value) of each feature against the output, using combinatorial calculus and retraining the model over the entire combination of features. The average absolute value of a feature's impact on the output is used to measure its importance (Mazzanti, 2020).

According to Marcílio-Jr and Eler (2021), to exemplify how Shapley values help understand a model's forecast, suppose that a model trained to predict house prices predicts that a particular house costs $ 300,000. For this prediction, the model used the following features: pets allowed, one bedroom, size of 100 m^2, and two bathrooms. The average forecast for this scenario was 290,000. The Shapley values show how much each feature contributed to the forecast compared to the mean; they explain the difference between the contributions to the forecast (300,000) and the average forecast (290,000). A possible solution to this problem could be animals allowed contributed 40,000, one bedroom contributed -60,000, size of 100 m^2 contributed 10,000, and two bathrooms contributed 20,000. Note that these values add up to 10,000, the difference between the forecast and the average forecast.

The Shapley value for a given feature is the average contribution of that feature over all possible permutations of the features in the data set. To estimate the Shapley value for the pets allowed feature, you need to calculate the prediction with: pets allowed; pets allowed and size of 100 m^2; pets allowed, size of 100 m^2 and two bathrooms, and so on. A better estimate can be obtained by repeating the sampling process and averaging the contributions.

3. METHODOLOGY

The Python computer language was used for data processing in the development environment Anaconda Navigator 2.0.3, specifically JupyterLab version 3.0.14 (Wang *et al.*, 2023). The libraries used were scikit-learn 0.24.1 (Boisberranger *et al.*, 2023), shap 0.39.0, numpy 1.20.1, and matplotlib 3.3.4 (Lundberg, 2023; Berg *et al*., 2023; Hunter, 2023).

The experiment was applied separately to two data sets, the Iris set with 150 instances and the Wine set with 180 instances; both were randomly divided into 80% of the instances for the training set and 20% for the test set (Lundberg, 2023; Aha *et al.*, 2023).

The Iris dataset has four features (Sepal length, Sepal width, Petal length, and Sepal width) and one output (three types of flowers).

The Wine dataset has 13 features (Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline) and one output (three types of wines). For this set, the data were standardized.

A Multilayer Perceptron classifier with an input layer, a hidden layer, and an output layer was trained, then the SHAP explainer was applied, and the Shapley values were retrieved (Boisberranger *et al.*, 2023; Lundberg, 2023).

The results were organized into tables (one table for each data set) containing the number of rows according to the number of instances in the test set and the columns that are expected class, predicted class, probability of the predicted class, and Shapley value of each feature. These resulting tables were sorted by expected class and predicted class probability columns.

Finally, the resulting values were plotted using the pixel-oriented technique for information visualization (Hunter, 2023). For the Iris set, the images were composed of 30 pixels (instances of the test set) distributed in a matrix of six rows and five columns. For the Wine set, the images were composed of 36 pixels (test set instances) distributed in a matrix of six rows and six columns. To facilitate the visualization of the information, some of the figures received a black marking line, representing the separation of the expected classes.

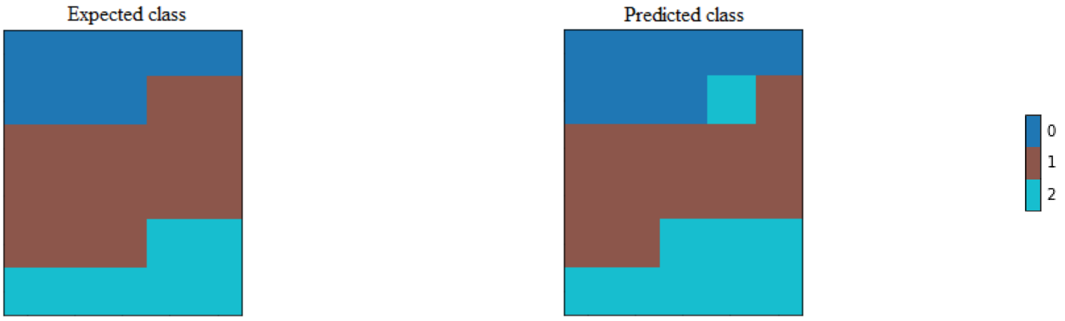
4. RESULTS AND DISCUSSION

4.1. Iris Dataset

The classifier achieved 100%, 96%, or 93% accuracy in many test cases. In some instances, there was 90% accuracy. A 93% accuracy test case was chosen to contain classifier hits and misses, offering diversity in the analysis.

Figure 1 shows the results referring to the classification process, it is possible to observe in the image of the predicted class that there was an error in the classification of two instances of class 1. In other test cases, an error in classifying instances belonging to class 1 was also observed.

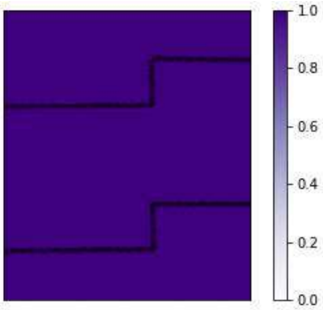
Figure 1. Iris classification



Source: Elaborated by the authors

Figure 2 shows the probability of the instances belonging to the predicted classes. It is possible to observe high probabilities for all instances, representing the classifier's certainty. However, there were two instances of misclassified. Below, the Shapley values will be used to understand what happened.

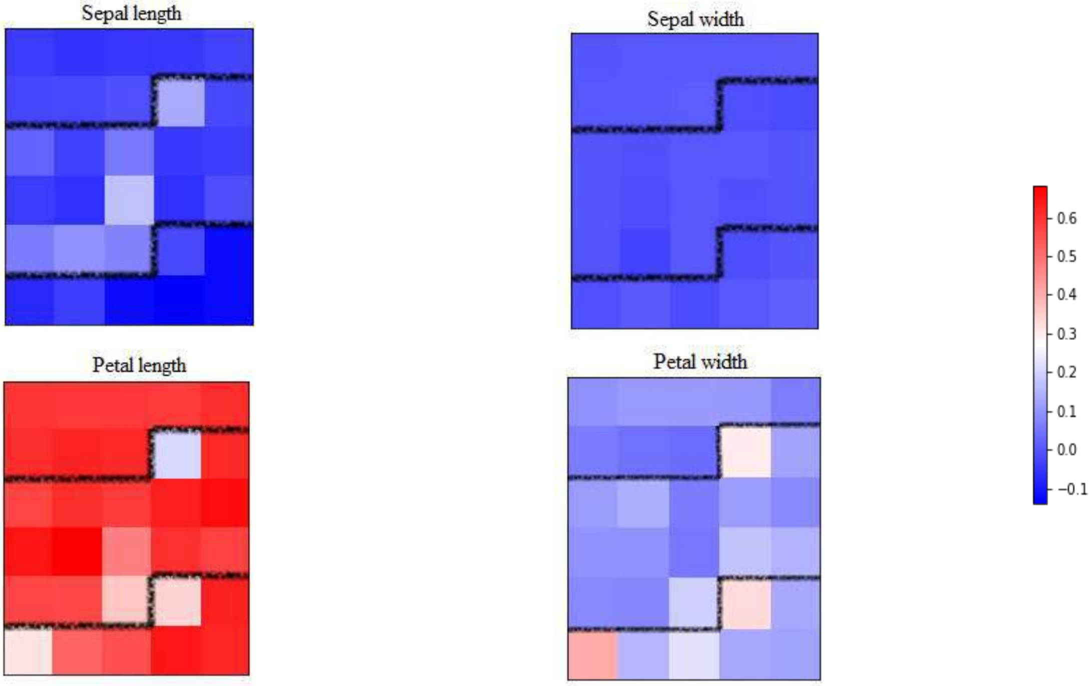
Figure 2. Probabilities of predicted classes



Source: Elaborated by the authors

Figure 3 shows the Shapley value of each feature. It is possible to observe that the Petal Length feature offers a more significant contribution to the classification process, as the Shapley value is higher (red pixels towards the positive numbers) than in the other features. The Petal Width feature also has contribution compared to Sepal Length and Sepal Width features. It is important to note that the misclassified two instances of class 1 have lower Shapley values (blue pixels towards negative numbers) in the Petal Length feature, which is more relevant. The Petal Length feature did not contribute enough for the correct classification for these instances.

Figure 3. Shapley values of the Iris classification

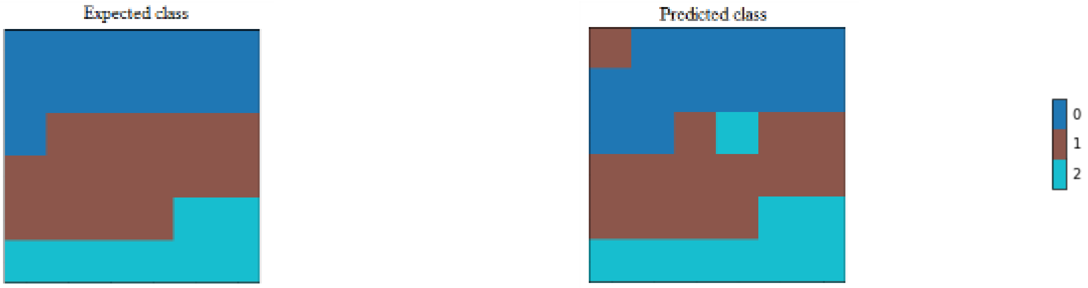


Source: Elaborated by the authors

4.2 Wine Dataset

After many trainings of the classifier, the classification accuracy varied between 90% and 100%. A 91% accuracy test case was chosen for analysis of the results. Figure 4 shows the expected classes of the test set and the classes predicted by the classifier. Note that in three instances, there was a classification error. One instance of class 0 and two instances of class 1.

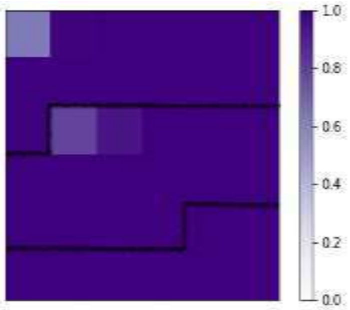
Figure 4. Wine classification



Source: Elaborated by the authors

Figure 5 shows the probabilities of the predicted classes. In this case, it is possible to observe that the three misclassified instances, two of them have a low probability, close to 0.6, and one instance has a high probability, close to 1.0. It is also interesting to note that a correctly classified instance has a probability of around 0.8, unlike the other instances with a probability of 1.0.

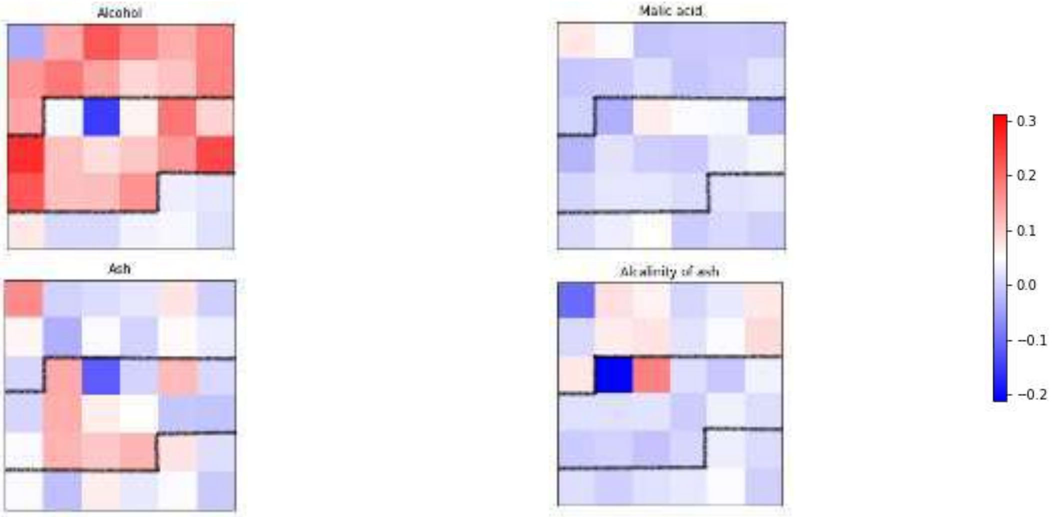
Figure 5. Probabilities of predicted classes

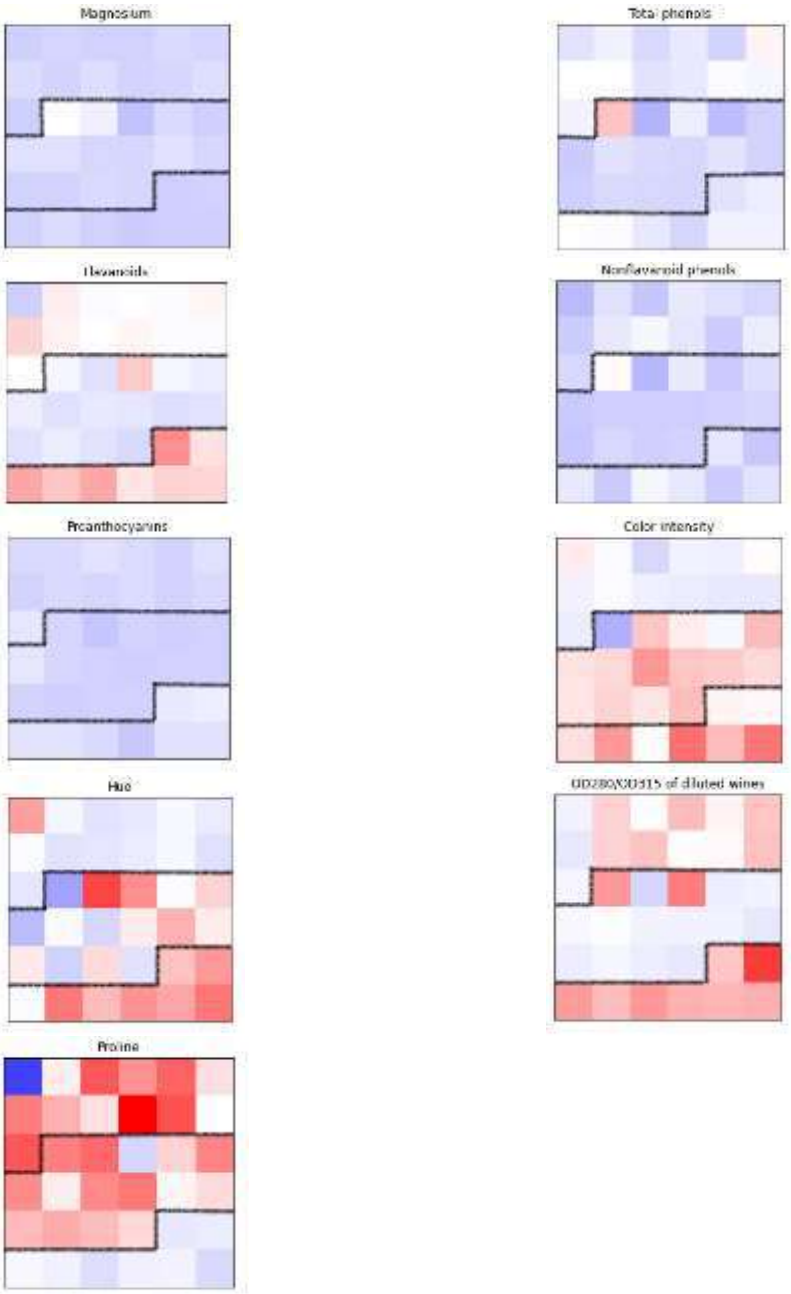


Source: Elaborated by the authors

Figure 6 shows the Shapley value of each feature. For the Wine set, the importance of the feature for the classification varies according to the class. For class 0, the features that most contribute are Alcohol, OD280/OD315 of diluted wines, and Proline. For class 1, the features that most contribute are Alcohol, Color intensity, and Proline. For class 2, the features that most contribute are Flavanoids, Color intensity, Hue, and OD280/OD315 of diluted wines. This phenomenon was not observed in the Iris set, where the same feature (Petal Length) is more important in all classes. Note that in classification errors, the most important features for each class had a low contribution.

Figure 6. Shapley values of the Wine classification





Magnesium

Total phenols

Flavanoids

Nonflavanoid phenols

Proanthocyanins

Color intensity

Hue

OD280/OD315 of diluted wines

Proline

Source: Elaborated by the authors

5. CONCLUSIONS

This paper presents a new approach involving several areas of machine learning, explainability, and information visualization. This paper aimed to apply a new approach based on pixel-oriented information visualization to explain the results obtained in data classification in a multilayer artificial neural network. The SHAP approach was chosen to calculate the relevance of the features and offer subsidies for the explanation. The visualization approach shows the expected and predicted classes, the prediction probabilities, and the SHAP values for each feature.

The results were satisfactory because they met the explainability goals that show how each feature influences the prediction of a classifier. Observing the visualization generated by the pixel-oriented approach, it was possible to explain the behavior of the classifier for both the correct and incorrect prediction.

As further works, it is intended to apply the results of this research from feature selection, considering that the results present the most relevant features in the classification and show how the most pertinent features can impact the classification results.

REFERENCES

AHA, D. *et al*. **UCI Machine Learning Repository**. Available in: https://archive.ics.uci.edu/ml/about.html. Accessed: 2023 May 01.

ARRIETA, A. B. *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82-115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

BERG, S. *et al*. **The fundamental package for scientific computing with Python**. Available in: https://numpy.org/. Accessed: 2023 May 01.

BOISBERRANGER, J. *et al*. **Machine Leaning in Python**. Available in: https://scikit-learn.org/stable. Accessed: 2023 May 01.

HUNTER, J. D. **Visualization with Python**. Available in: https://matplotlib.org/. Accessed: 2023 May 01.

LUNDBERG, S. **Welcome to the SHAP documentation**. Available in: https://shap.readthedocs.io/en/latest/index.html. Accessed: 2023 May 01.

MALATO, G. **How to explain neural networks using SHAP**. Available in: https://www.yourdatateacher.com/2021/05/17/how-to-explain-neural-networks-using-shap/. Published: 2021 May 17. Accessed: 2023 May 01.

MARCÍLIO-JR, W. E.; ELER, D. M. Explaining dimensionality reduction results using Shapley values. **Expert Systems with Applications**, v. 178, p. 115020, 2021. DOI: https://doi.org/10.1016/j.eswa.2021.115020

MAZZANTI, S. SHAP Values Explained Exactly How You Wished Someone Explained to You. **Towards Data Science**, 4 jan. 2020. Available in: https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30. Accessed: 2023 May 01.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences**. Artificial Intelligence**, v. 267,pp. 1-38, 2019. DOI: https://doi.org/10.1016/j.artint.2018.07.007

PANTELIS, L.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods**. Entropy** , v.23, n.1, p., 18, 2021. DOI: https://doi.org/10.3390/e23010018

SAKURAI, R.; ZUCHI, J. D. As Revoluções Industriais até a Industria 4.0. **Revista Interface Tecnológica**, v. 15, n. 2, p. 480-491, 2018. DOI: https://doi.org/10.31510/infa.v15i2.386

WANG, P. *et al*. **The World´s Must Popular Data Science Plataform**. Available in: https://www.anaconda.com. Accessed: 2023 May 01.