


Integrated Data Repository System: Fusion, Learning and Sharing

Jeferson Lopes²^a, Giancarlo Lucca¹^b, Rafael Huszcza²^c, Amanda Mendes²^d,
Eduardo N. Borges²^e, Pablo D. B. Guilherme³^f and Leandro A. Pereira⁴^g

¹*Mestrado em Engenharia Eletrônica e Computação, Universidade Católica de Pelotas, Pelotas, Brazil*

²*Centro de Ciências Computacionais (C3), Universidade Federal do Rio Grande (FURG), Rio Grande, Brazil*

³*Centro de Educação, Humanidades e Ciências Biológicas, Universidade Estadual do Paraná (UNESPAR), Paranaguá, Brazil*

⁴*Eixo Tecnológico de Meio Ambiente, Instituto Federal do Paraná (IFPR), Paranaguá, Brazil*

Keywords: Data Repository, Machine Learning, Deep Learning, Geographic Information System.

Abstract: Currently, an enormous volume of data is being generated from diverse sources, including sensors and social media. Effectively managing this unprecedented scale of data and deriving meaningful insights from these extensive datasets present a significant challenge for computer scientists. In this context, this paper outlines the development and documentation of a project dedicated to actively contributing to these critical data-driven initiatives. The described system integrates the features of a scientific data repository with a suite of data science methods, machine learning tools, and resources for geographic data visualization. By consolidating these functionalities on a single platform, users can streamline their workflow and extract insights from data more efficiently. This integrated approach facilitates seamless transitions from data storage to model training and analysis, fostering collaboration and facilitating knowledge sharing among researchers and practitioners. In this work, we highlight the system's key features, focusing on the datasets repository and the machine learning module as central components of our platform.

1 INTRODUCTION

It is widely recognized that data has been generated at increasingly high rates, especially due to advances in mobile devices and digital sensors, which have greatly facilitated data collection. Hence, large-scale datasets have been curated and are widely employed across various domains (Yaqoob et al., 2016).


Managing this unprecedented volume of data in Big Data era poses a significant challenge for computer scientists (Al Aghbari., 2015). Hence, machine learning algorithms (Tan et al., 2005) have risen as indispensable tools for uncovering intricate patterns in this vast and complex datasets, providing valuable assistance to professionals in diverse data-intensive


fields, including medicine, biology and beyond.


However, to effectively harness the wealth of data through machine learning is a challenge for non-computer science end-users. Hence, platforms must transcend mere data repositories and seamlessly integrate machine learning tools for extracting meaningful insights. Recognizing the expense of organizing databases, a focus on cost-effective solutions involving database management and artificial intelligence is crucial for technological advancement.


The paramount importance of services in this domain is evident in the growth of platforms like Kaggle (Kaggle, 2023) and Hugging Face (Hugging Face, 2023). These platforms succeed in addressing challenges with a versatile approach that spans multiple domains, assisting in the integration of information among stakeholders.


Nevertheless, these feature-rich tools can be intimidating for users who are not familiar with the field of Computer Science, including scientists from other areas, since they tend to require in-depth technical knowledge for operation. Moreover, another chal-


^a <https://orcid.org/0009-0004-9113-3921>


^b <https://orcid.org/0000-0002-3776-0260>

^c <https://orcid.org/0009-0000-1949-9284>

^d <https://orcid.org/0000-0002-5335-166X>

^e <https://orcid.org/0000-0003-1595-7676>

^f <https://orcid.org/0000-0001-7471-6907>

^g <https://orcid.org/0000-0001-6055-8063>

challenge faced by many is the need for local installation of these systems, which may involve complicated configurations and specific hardware requirements.

Therefore, an efficient dataset management platform that integrates data science tools, including artificial intelligence algorithms, yet remains simple and intuitive for all users, is crucial. The developed system, namely Data Symbion Environmental Intelligence¹, aims to encompass all of these aspects. It is a centralized repository where users, both individuals and legal entities, can share their data on the platform. This enables the integration and combination of information from different sources, increasing the chances of extracting useful and valuable insights.

Besides being a dataset repository with integrated machine learning tools that anyone can use, regardless of technical knowledge, our platform is web based, which enables users to access it from anywhere, at any time, with internet connection. The machine learning tools are tailored for creating and utilizing predictive models for both tabular and visual data.

Moreover, our system can be accessed directly through the API, which facilitates seamless integration with data collection systems, streamlining automation processes. The platform also includes a tool for geographic visualization tailored to datasets containing latitude and longitude information, facilitating visual comprehension through interactive maps.

The general objective of this work is to present the development and the structure of the proposed system. The following list outlines the set of specific objectives of the project:

- Offer an efficient and accessible data repository;
- Integrate data from multiple sources;
- Allow users to train and share supervised machine learning models, including, but not limited to, deep learning models for image classification;
- Provide tools for operations on geographic data;
- Allow software and hardware developers to implement graphical interfaces and/or automate processes through the consumption of REST API;
- Provide a user-friendly and welcoming interface accessible to users from diverse domains.

The study is structured as follows: Section 2 provides a review of existing tools for data repositories and machine learning. In Section 3, we begin by comparing these existing tools to our proposed system. We then outline our system's modules and describe its main features. Finally, Section 4 summarizes the contributions of our system and possible limitations.

¹<https://datasymbion.com/>

2 RELATED WORK

Effectively organizing and utilizing the extensive collected data in a systematic manner has a core importance in various fields. Tools focused on database management and artificial intelligence (AI) have shown tremendous potential in recent years (Russell and Norvig, 2009). This section discusses existing software that serve as data repositories, and platforms that provide machine learning tools.

Available solutions for data management such as DSpace², Dataverse³ and ArcGIS⁴, are committed to the FAIR principles (Findable, Accessible, Interoperable, and Reusable). This is crucial to ensure reliability, quality and standardization in the organization and sharing of data (Rocha et al., 2021).

Although repositories play an important role by providing a centralized location for storing and sharing data, effective data management requires a more comprehensive approach (Sayão and Sales, 2022). In the presented context, several platforms offer AI tools, which allow the extraction of valuable insights from collected data, thus supporting decision-making (Hachicha Belghith et al., 2020).

Artificial Intelligence (AI) is a field that focuses on algorithms that bring computers closer to human perception, performing automated tasks. Machine learning (ML) algorithms are AI methods capable of learning and improving from experience (Tan et al., 2005). By training ML models with sufficient amount of information, we can leverage them to recognize patterns in the training data.

Two well recognized and open-source solutions are Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009)⁵ and RapidMiner⁶ (Hofmann and Klinkenberg, 2016), which provide machine learning tools for data mining used to extract valuable insights from datasets. Also, many tools designed for machine learning excel in addressing challenges with a more versatile approach that spans multiple domains. Some emerging tools in this context include Kaggle⁷ (Kaggle, 2023) and Hugging Face⁸ (Hugging Face, 2023). In what follows we discuss some of them:

- **DSpace** – DSpace is a web application, that allows users to publish documents and data, serving as a long-term digital archives system. DSpace

²<https://dspace.lyrasis.org>

³<https://dataverse.org>

⁴<https://www.arcgis.com>

⁵<https://www.cs.waikato.ac.nz/ml/weka/>

⁶<https://altair.com/altair-rapidminer>

⁷<https://www.kaggle.com/>

⁸<https://huggingface.co/>

supports access to all types of digital content including text, images, moving images, mpegs and data sets. Its features make this software a good choice for academic, non-profit, and commercial organizations building open digital repositories. However, despite being free and open source, DSpace requires installation.

- **Dataverse** – Another data repository platform is the Dataverse Project, which is an open source web application to share, preserve, cite, explore, and analyze research data. This tool facilitates data sharing, empowering users to replicate others' work more easily.
- **ArcGIS** – One important application of data repositories are Geographic Information System (GIS), which is a software designed to collect, manage, analyze, and visualize geographic data. A GIS is particularly valuable for integrating and organizing spatial information, covering areas such as marine ecosystems, coastal land use, and economic activities (Randazzo et al., 2021). One important example of GIS is the ArcGIS platform. It is a software that allows secure mapping and spatial analysis, empowering users to unlock geospatial insights. This software is built on scalable and resilient technology, allowing efficient data collection, management, and analysis, thus, facilitating decision-making by easily sharing maps and apps.
- **WEKA** – An open source software that provides a collection of machine learning algorithms for data mining tasks. It encompasses tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Despite its wide range of applications, WEKA not only requires installation of the software itself but also needs a compatible version of Java.
- **RapidMiner** – RapidMiner data science platform is designed for diverse teams to collaboratively generate and share data-driven insights. It caters to various skill sets, from data scientists to business analysts, offering a unified environment. Users can build data and machine learning pipelines with code-free to code-friendly experiences, increasing trust with interactive decision trees and model simulators.
- **Kaggle** – Kaggle empowers users to access hundreds of pre-trained machine learning models for deployment and facilitates exploration, analysis, and sharing of high-quality datasets (Kaggle, 2023). This platform offers a wealth of resources for machine learning enthusiasts, including 296K high-quality public datasets, 2,200 pre-

trained ML models, and a large repository of community-published models, data, and code.

- **Hugging Face** – Hugging Face extends its services beyond providing a platform for model and dataset sharing; it equips users with tools for building, training, and deploying machine learning models based on open-source (OS) code and technologies (Hugging Face, 2023). The platform offers a wide range of libraries and frameworks, including Transformers, Tokenizers, and Datasets, which are designed to simplify the development process and make it more accessible.

3 THE PROPOSED SYSTEM

In real-world applications, it is essential to offer platforms equipped with a comprehensive set of tools featuring convenient and accessible functions. Some key features include:

- **Data repository:** enable users to efficiently, safely, and securely store and manage data.
- **Machine learning:** seamlessly integrate data repositories with AI capabilities to extract knowledge from data.
- **Interactive maps:** offering features like interactive maps for data visualization is essential in geographic systems,.
- **Open Source:** being open source is particularly valuable as it promotes collaboration, innovation, and accessibility in the tech industry.
- **Online access:** online systems that require no installation or complex settings make technology more accessible.

Table 1 presents a comparison of the developed system with the other tools discussed in the previous section considering the aforementioned criteria.

All the presented tools offer valuable services to users, whether for storing and managing data or employing artificial intelligence for data mining. Nevertheless, these systems, whether tailored specifically for the environmental field or designed for more general applications, may be considered overly complex and less intuitive, particularly for users lacking familiarity with the computer science domain.

Our proposed system integrates the functionalities of a scientific data repository with those of a machine learning tool, enabling users to store and share datasets and to employ them for training and using machine learning models. Data Symbiont EI incorporates a GIS module, providing access to interactive maps.

Table 1: Comparison of the discussed tools with the proposal.

Tool	Data repository	Machine Learning	Iterative maps	Open-Source	Online access
DSpace	✓			✓	✓
Dataverse	✓			✓	✓
ArcGIS	✓		✓		✓
WEKA		✓		✓	
Rapid Miner		✓		✓	
Kaggle	✓	✓		✓	✓
Hugging Face	✓	✓		✓	✓
Proposed System	✓	✓	✓	✓	✓

To have a more complete understanding of the project, Figure 1 provides an overview of all screens and features of the proposed platform. It is mapped among screens (green) and system features (yellow).

To register in the system, users are required to provide personal information, including email, name, personal identification number, and password. Once registration is complete, users can access the system and will be redirected to the *Dashboard* page.

This *Dashboard* page presents crucial information pertaining to the system, including details such as the number of institutions with which the user is affiliated and the number of registered *Data Sources*. Also, the system incorporates a guide button to help users easily navigate and access the platform's features.

3.1 Data Repository

The system introduces a comprehensive data repository service, empowering users to securely upload their datasets and customize sharing permissions as desired. This platform offers a structured approach to data storage, ensuring organization and accessibility. These datasets can then be seamlessly utilized across various tools within the system, including data integration, geographic information visualization, and machine learning model training, enhancing the overall functionality and utility of the platform.

Users can securely upload and manage diverse datasets, including structured data such as CSV files, as well as unstructured data like images for deep learning training. Additionally, CRUD (Create, Read, Update, Delete) operations can be adeptly executed on these repositories, thereby ensuring the efficacy of their management and maintenance.

The system offers two types of data storage structures, namely *Data Source* and *Image Source*, allowing users to categorize and structure their information systematically. The *Data Source* is designed for tabular data, such as that found in a spreadsheet. Users can create a *Data Source* instance, define its parameters such as name, description, access type and institution, and upload a CSV file.

The *Image Source* is a structure analogous to the *Data Source*, but instead of being associated with a CSV file, it encapsulates a set of images. For the classification task, the images are organized into folders, each associated with one of the data classes.

To enhance collaboration and data sharing, our platform incorporates customizable sharing permissions. Users can effortlessly adjust access levels, described in Section 3.3, allowing them to share data with others institutions.

3.2 Data Integration

In our system, it is possible to integrate two or more *Data Sources* to produce a new instance, resultant from their combination. This is the so-called *Data Integration Module*. This module allows users to perform two types of integration: by rows or by columns.

Integrating two or more *Data Sources* by row involves concatenating them vertically, stacking one on top of the other. Users can choose which columns to include in the integrated dataset and automatically add a new column to indicate the origin of each sample. Conversely, integrating by column enables users to merge two or more *Data Sources* horizontally. This is especially valuable when the *Data Sources* describe different attributes of the same samples.

After the *Data Sources* are selected, it is necessary to define the type of integration and if it's necessary to remove/maintain duplicated samples. The metadata of the new *Data Source* instance such as name, description and access is also required.

3.3 Data Protection

The system includes several visibility profiles, which enable users to control the accessibility of their *Data Sources*, *Image Sources* and trained models. The three available privacy modes are:

1. **Private:** Only its responsible can access the data/model. So, even when it is linked to an institution, the data/model will not be shared with anyone until it is published.

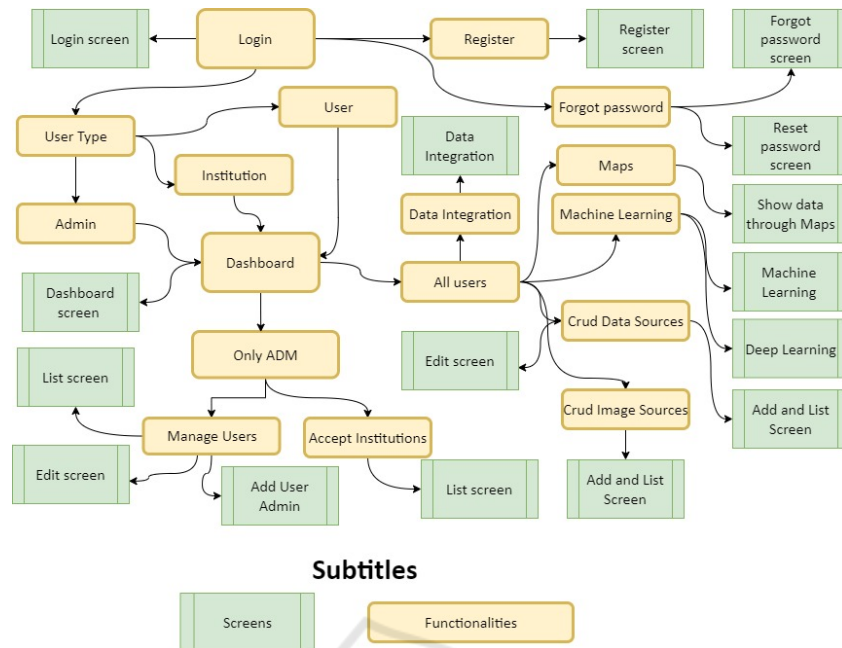


Figure 1: General view of the Project system.

2. **Public:** The data/model will be shared with the users linked to the institution with which it has an active connection. All the other users that are not linked to this institution will not have access.
3. **Open:** All the users, including the ones that are not registered on the system will have access to the data/model.

It is important to note that if a user does not explicitly choose a visibility setting, it is set to the default mode, which is private.

3.4 Geographic Information Visualization

Within our platform, users are empowered to explore and analyze geographical data through an interactive maps section, enhancing their capabilities in biological research and decision-making processes. This module allows users to visualize specific *Data Sources* of interest by selecting them. It can contain a wide range of geographical data.

To ensure accurate representation, users are required to specify the columns that correspond to latitude and longitude coordinates and select a label column that identifies the primary attribute associated with each data point. This functionality allows for the differentiation of individual points, facilitating effective data interpretation and analysis.

In Figure 2, the Geographic Visualization Module is divided into two components. The left screen displays the settings for map creation, showing the selec-

tion of the data source and columns related to latitude, longitude, and the attribute to be visualized. Users can also specify the initial number of samples to be loaded onto the map. The second screen showcases the visualization of data points and their geographical locations, which can be grouped for improved visualization. Furthermore, a button is available for loading additional data onto the map.

3.5 Machine Learning Tools

A model is a computational representation created by a machine learning or deep learning algorithm during the training process. This representation is learned from data and, at the end of training, can be used to make predictions on new input data.

Our system provides a powerful module to allow the persistence and sharing of models trained by users. Thus, shortly after training, in both the Machine Learning and Deep Learning modules, it is possible to save the generated model in the system itself, eliminating the need for the user to save it locally.

These persistence and sharing capabilities offer users greater flexibility, as model training occurs in the background and results—including the model, its parameters, and training status—are automatically saved in this module. This eliminates the need for users to remain on the training page until completion.

In addition to facilitating model sharing among users, this module enhances security, organization, and convenience. It provides a list of all trained models and includes a filter system to simplify searches.

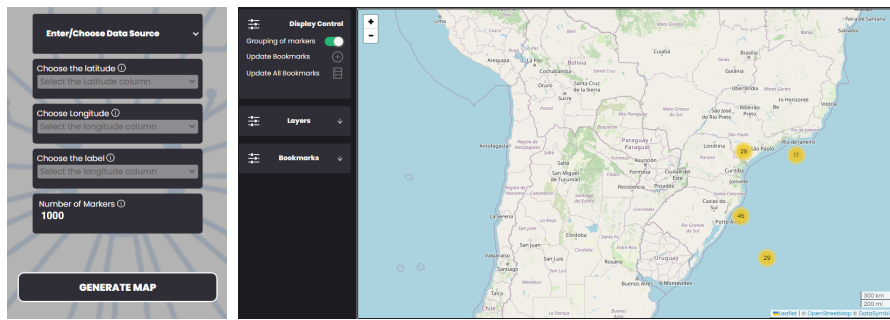


Figure 2: The Geographic Visualization Module.

This tool offers the user the possibility to evaluate the trained models and use them for inference on new data. In order to use the model, the user can manually input the new sample to obtain the model's prediction. This feature allows users to employ their trained models in real-world scenarios.

The *Machine Learning Module* empowers users to extract valuable insights and knowledge from their data. By utilizing advanced analytical techniques, it is possible to uncover patterns, trends, and relationships within datasets. The module allows for experimentation with different approaches, using various algorithms and testing different parameter settings. This functionality enables students and professionals from diverse domains to gain a deeper understanding of their data, empowering them to make informed decisions and drive innovative research.

The system offers a range of machine learning capabilities, including classical algorithms for regression and classification and advanced deep learning algorithms tailored for image classification.

3.6 Classical Machine Learning Algorithms

Classical Machine Learning algorithms are tailored to learn from structured data, such as data in spreadsheet form. Therefore, these methods can be utilized to learn from data stored in the *Data Source* structure provided by our system.

We initially consider explainable methods, to better describe the generated models. In this sense, for the classification problem (Duda et al., 2000), CART (Breiman et al., 1984) and Random Forreast (Ho, 1995) algorithms were implemented. For the regression problem (Hastie et al., 2009), we consider the Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Linear Regression (Weisberg, 2005).

The classification task involves categorizing input data into predefined classes or categories (Sen et al., 2020). The Decision Tree (Breiman et al., 1984) algorithm addresses this task by constructing a series of

rules organized in a tree-like structure. Each leaf node in the tree corresponds to a class label. The algorithm learns these rules from the training data, determining the best splits to classify the data accurately.

Random Forest (Ho, 1995) is a widely used technique for solving classification tasks. It belongs to the ensemble learning methods and works by creating multiple decision trees. The final class prediction is determined through a voting mechanism, where each tree's prediction contributes to the final outcome.

Regression tasks aim to predict a continuous value based on a set of attributes (Sen et al., 2020). For example, predicting the weight of a person based on a set of body measurements is a regression task. One of the algorithms available to address this challenge is the Support Vector Machine (SVM) (Cortes and Vapnik, 1995). SVM aims to find a hyperplane that best fits the data points. The hyperplane is chosen such that it passes through as many points as possible, while still staying within the margin of tolerance.

Another approach to finding the line that best represents the data is by using the linear regression algorithm (Weisberg, 2005), also available in our system. This is a straightforward technique that aims to find the line that, on average, is closest to the data points.

To train a machine learning model, users need to follow these steps:

1. Choose a *Data Source*;
2. Select the type of problem to be solved (classification or regression);
3. Define the target variable to be predicted;
4. Set a random seed for reproducibility;
5. Select the attributes (i.e., columns) from the *Data Source* to be considered;
6. Specify the proportion of the *Data Source* to be employed for model training and evaluation.

3.7 Deep Learning Algorithms

Another important feature in our system is the Deep Learning module. Deep learning, a subset of AI,

The screenshot shows a web-based interface titled "Select Classification Params". It contains several configuration options:

- Image Size:** 224
- Batch size:** 32
- Patience:** 10
- Learning Rate:** 0,001
- Epochs:** 10
- freeze base model:** A dropdown menu currently showing "True".
- Seed:** 42
- New Layer:** A green button to add layers.
- Neurons:** A field to specify the number of neurons in a layer.
- Dropout:** A field to specify the dropout rate, accompanied by a trash icon for deletion.
- Train Model:** A green button to initiate the training process.

Figure 3: Training a Deep Learning model.

focuses on developing models capable of complex learning and pattern recognition tasks. This approach uses deep neural network architectures, composed of multiple layers of interconnected units, to automatically learn hierarchical and abstract representations of data (Goodfellow et al., 2016).

In addition to being widely applied in areas such as natural language processing, recommendation systems, and speech recognition, deep learning models also excel in computer vision. This field of study focuses on developing algorithms capable of extracting knowledge from images and videos to perform tasks such as image classification, object detection, and semantic segmentation (Goodfellow et al., 2016).

The first functionality of this new module is image classification, which consists in categorizing images into predefined groups. Two deep neural network architectures designed for this task have been made available: MobileNetV3 (Howard et al., 2019) and ResNet50 (He et al., 2016). Users can employ these algorithms to train deep learning models on images stored within a *Image Source* structure.

MobileNetV3, initially developed for mobile and edge devices such as smartphones and IoT, strikes a balance between speed and accuracy. This makes it ideal for applications where computational resources are limited (Howard et al., 2019).

In contrast, ResNet50 is a more robust network architecture. Its name, "ResNet," comes from the concept of residual learning, which is central to its design. This approach involves creating shortcuts, known as skip connections, throughout the network. These connections make it easier for information to flow through the network without being lost or distorted, ultimately improving the network's ability to learn and perform complex tasks (He et al., 2016).

In addition to choosing one of the two networks,

the user can also add layers to customize their architecture. After defining the training parameters, as illustrated in Figure 3, the model is trained and made available to the user for inference on new images.

4 CONCLUSION

Effectively managing the vast volume of data generated today is a critical and evolving field in computer science. It is essential to provide end-users from various fields with data repositories integrated with machine learning tools to extract meaningful insights from their data. However, existing tools are often too complex and lack accessibility and intuitiveness for users unfamiliar with computer science.

Our system provides a scientific data repository, enabling users to securely store, share, and utilize datasets for seamlessly training and deploying machine learning models. Two storage structures, Data Source and Image Source, enable organization and sharing of tabular and image data.

The Machine Learning Module offers various algorithms and parameters for users to experiment and better understand their data for decision-making. The Data Integration Module merges datasets, and the GIS module provides an interactive maps section that facilitates geographical data analysis.

While our proposed system offers a comprehensive array of features to tackle various challenges in data management and analysis, it is imperative to recognize its potential limitations. One notable constraint may lie in the scalability of the system, particularly when confronted with exceptionally large datasets. Limited hardware capabilities and software constraints could restrict the size and complexity of

models that can be effectively trained, potentially impeding the exploration of highly dense model architectures or the utilization of extensive datasets.

Future works could focus on enhancing the scalability and efficiency of ML algorithms for processing large-scale datasets stored in scientific repositories. Moreover, leveraging emerging technologies such as federated learning and edge computing holds promise for enabling distributed and real-time analysis could be an interesting feature for the system.

ACKNOWLEDGMENTS

The authors would like to thank Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (3305805/2021-5, 23/2551-0000126-8), Fundação Grupo Boticário (camp.001.2021) and Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná (FA).

REFERENCES

- Al Aghbari., Z. (2015). Mining big data - challenges and opportunities. In *Proceedings of the 17th International Conference on Enterprise Information Systems - Volume 2: ICEIS*, pages 379–384. INSTICC, SciTePress.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall (Wadsworth and Inc.).
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.
- Hachicha Belghith, E., Rioult, F., and Bouzidi, M. (2020). Acoustic diversity classification using machine learning techniques: Towards automated marine big data analysis. *International journal on artificial intelligence tools*, 29(3n04):2060011.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Ho, T. K. (1995). Random decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):832–844.
- Hofmann, M. and Klinkenberg, R. (2016). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.
- Hugging Face (2023). Hugging face - the ai community building the future.
- Kaggle (2023). Kaggle: Your machine learning and data science community.
- Randazzo, G., Italiano, F., Micallef, A., Tomasello, A., Cassetti, F. P., Zammit, A., D'Amico, S., Saliba, O., Cascio, M., Cavallaro, F., Crupi, A., Fontana, M., Gregorio, F., Lanza, S., Colica, E., and Muzirafuti, A. (2021). Webgis implementation for dynamic mapping and visualization of coastal geospatial data: A case study of bess project. *Applied sciences*, 11(17):8233.
- Rocha, R. P. d., Gabriel Junior, R. F., Vanz, S. A. d. S., Borges, E. N., Azambuja, L. A. B., Caregnato, S. E., Pavão, C. M. G., Passos, P. C. S. J., and Felicissimo, C. H. (2021). Análise dos sistemas dspace e dataverse para repositórios de dados de pesquisa com acesso aberto. *Revista Brasileira de Biblioteconomia e Documentação. São Paulo. Vol. 17 (2021)*, p. 1-25.
- Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: a modern approach*. Pearson, 3 edition.
- Sayão, L. F. and Sales, L. F. (2022). Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados. *Palavra chave (La Plata)*, 12(1):171–171.
- Sen, P. C., Hajra, M., and Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 99–111. Springer.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.
- Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., and Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, 36(6):1231–1247.