Development of a Context-Free Data Ingestion Mechanism for AutoML

Gabriel Mac'Hamilton[a] and Alexandre M. A. Maciel[b]
Universidade de Pernambuco, Recife, Brazil

Keywords: Data Ingestion, AutoML, Machine Learning, Human-Computer Interaction.

Abstract: Automated Machine Learning (AutoML) is a technology that simplifies complex data processing and analysis for strategic decision-making by automating machine learning tasks and enhancing the user experience. Data ingestion is a crucial AutoML step that involves collecting external data for machine learning workflows. Typically, AutoML systems include data input modules. However, the lack of a user interface limits the number of users that can utilize it. This work presents the development of a data ingestion mechanism that streamlines and simplifies this machine learning stage into an AutoML framework called FMD. The mechanism underwent three validations: Experimentation in a real-world scenario with two databases from different contexts, evaluation from expert opinions, and usability assessment through a questionnaire using the AttrakDiff method. Following the validations, successful results were achieved in both assessments and in demonstrating the ingestion in various contexts.

1 INTRODUCTION

Amidst an increasingly complex environment, where the shortage of data scientists becomes more evident in the face of high market and academy demands, AutoML systems emerge as a strategic solution for organizations. These systems can reduce the complexity of the model building and optimization processes (Hutter et al., 2019). In this context, AutoML is not just a response to human resource limitations but also a strategy to optimize the potential of available data, accelerate model development cycles, and democratize data science within organizations while also allowing domain experts to work directly with machine learning development (Elshawi et al., 2019).

Data ingestion is of crucial relevance in data science, serving the purpose of consistently and reliably introducing data into the machine learning pipeline (Hapke and Nelson, 2020). According to studies, data wrangling activities, encompassing ingestion, cleaning, and data transformation, consume about 70% of data scientists' time (Saurav and Schwarz, 2016). This underscores the significance of simplifying data ingestion for AutoML systems, as it helps reduce time and effort in this workflow phase. Such simplification allows users to dedicate more time to metadata preparation and model refinement activities, contributing to increased efficiency in machine learning teams (Patel, 2020).

Given the context, this work presents a data ingestion mechanism to alleviate the problems related to the early stages of machine learning practices, domain understanding, and feature engineering. The mechanism proposes a standardized and simplified way to ingest data, using a metadata file called "context file", to help with transforming domain experts' tacit knowledge into explicit knowledge, allowing for better feature selection, allied with a robust data ingestion engine based on the ETL (extract, transform & load) process, and a simple and intuitive user interface (UI).

2 BACKGROUND

2.1 AutoML

AutoML is broadly defined in the literature and commonly describes systems that automate machine learning activities. These systems were motivated by the need to address the limitations and challenges posed by traditional machine learning techniques. These challenges include requiring highly specialized professionals in model development, dependence on domain experts' knowledge tied to a particular model, and introducing human biases, making the models in-

[a] https://orcid.org/0000-0002-3735-190X
[b] https://orcid.org/0000-0003-4348-9291

Mac'Hamilton, G. and Maciel, A. M. A.
Development of a Context-Free Data Ingestion Mechanism for AutoML.
DOI: 10.5220/0013357700003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 2*, pages 581-588
ISBN: 978-989-758-749-8; ISSN: 2184-4992

efficient. AutoML systems can expedite the development of machine learning models, ensure greater optimization, and democratize access to data science for a broader and less specialized audience (Nagarajah and Poravi, 2019). In summary, AutoML aims to find an optimized solution for machine learning applications (Chen et al., 2021).

AutoML systems can automate any stage of machine learning model development. However, according to (Hutter et al., 2019), most systems focus on preprocessing and model tuning, emphasizing hyperparameter optimization, meta-learning, and neural architecture search. The authors also highlight several advantages of using a system that automates these activities, such as reducing human effort in model development, improving algorithm performance, enhancing reproducibility in academic work, and facilitating the reuse of successful models.

2.2 Data Ingestion

The data ingestion is primarily characterized by obtaining and transporting data from an external source to the machine learning workflow. Organizations typically employ this process to optimize data collection, improve data quality and accuracy, and save time and resources. Through the use of data ingestion techniques, it is possible to reduce costly errors in the data collection stage (Hapke and Nelson, 2020).

The main goal of this process is to capture, store, and make data available for future use. Among the methods found in the literature for developing data ingestion, we can highlight batch and streaming. Batch data ingestion is usually performed through ETL (Extract, Transform & Load) routines that collect data from an external source, incorporate it into the workflow, or store it for later use. The batch technique is employed for data that does not need to be consumed in real time. Streaming data ingestion, on the other hand, is used in cases where there is a need for real-time data consumption, requiring specific technologies to support this type of demand (Hlupic and Punis, 2021).

2.3 Software Usability

The graphic interface design of software can determine the success or failure of a product. A tool needs a user-friendly interface to gain user approval and may be replaced by competing options. Therefore, the application of efficient user experience (UX) and user interface (UI) techniques is highly relevant, allowing the development of a valuable system for users (Tidwell et al., 2020).

The efficiency of a UI depends on its intuitiveness and ease of use. Intuitive software is designed to be familiar, with recognizable components and precise interactions, enabling users to apply their prior knowledge and use the interface seamlessly. Due to the need for this familiarity, design patterns are encouraged, allowing users to easily recognize the functionalities displayed in interfaces (Tidwell et al., 2020). In addition to interface familiarity, the overall user experience is a relevant factor in software development. UX is about the visual interface and how users perceive, interpret, and interact with a system. Whalen (2019) emphasizes the need to consider cognitive psychology, thinking patterns, and user expectations when creating effective designs (Whalen, 2019).

2.4 Data Mining Framework - FMD

The Data Mining Framework - FMD is an AutoML system originally developed in 2017 at the University of Pernambuco through various academic works. This achievement is due to its nature as an open-source software, which has been enhanced through multiple projects over time.

The project was initially conceived to enable data mining in virtual learning environments (VLE), with the goal of democratizing data mining activities for users with limited technical knowledge in this field. It was initially named as Visual Educational Data Mining Framework - FMDEV, later being changed to simply Data Mining Framework - FMD. The framework allowed data mining from the Moodle VLE[1] with just a few clicks, providing data analytics and visual graphs to users. The project was developed using technologies such as Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript, and was integrated into Moodle as an HTML block (Gonçalves et al., 2017). The initial architecture of FMD is represented in Figure 1.

The work of (da Silva, 2020) enhanced the project, leading to its current state with a more robust architecture, updated technologies, new functionalities, and a user-friendly interface. The project underwent a refactoring process using Lean Inception techniques and requirements engineering, along with a technology survey through an systematic literature review (SLR). This allowed the development of a more refined architecture based on the original design. At this moment, the project transitioned from being a data mining framework to becoming an AutoML framework, capable of performing automated supervised

[1] https://moodle.org/

machine learning tasks and presenting the results visually, using data related to the educational context.

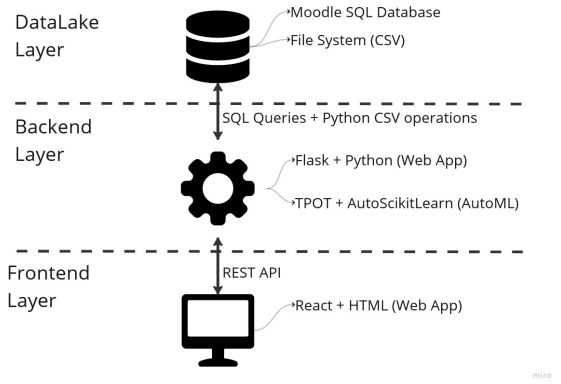

Figure 1: FMD original architecture.

2.5 Related Work

This section presents a comparative analysis of open-source AutoML tools in the market, focusing on usability and data ingestion. Specifically, it examines functionalities identified as essential by (Alves and Maciel, 2023). To select the AutoML tools for comparison, those in the work of (Zöller and Huber, 2021) were considered based on their number of citations and stars on GitHub. The five frameworks selected for benchmarking were: TPOT[2], hpsklearn[3], auto-sklearn[4], H2O AutoML[5], and ATM[6] (Auto Tune Models).

Table 1 compares this project regarding data ingestion with the five most relevant open-source AutoML tools. Observing the table, it becomes evident that, due to their nature being run from the command line, TPOT, auto-sklearn, hpsklearn, and ATM lack most of the functionalities described as necessary for easy data ingestion. Being the only one with a graphical interface, H2O AutoML has the most functionalities. However, the interface primarily focuses on AutoML rather than data ingestion, and data input is limited to data files only. The FMD Data Ingestor stands out as it was developed with a specific focus on the most relevant functionalities outlined in the literature for data ingestion, allowing data input through files or database connections but lacking metadata inference. This absence is mitigated by the "context mapping" functionality, which pre-provides the necessary metadata for specific datasets.

Table 1: AutoML Frameworks Benchmark.

Framework	User Interface	Data Visualization	Multiple data inputs	Metadata Inference
FMD	Yes	Yes	Yes	No
TPOT	No	No	Yes	No
hpsklearn	No	No	Yes	No
auto-sklearn	No	No	Yes	No
H2O *AutoML*	Yes	Yes	No	Yes
ATM	No	No	Yes	No

3 METHODOLOGY

The project's requirements gathering was conducted using Design Science Research (DSR) (Aken, 2004) techniques combined with lean inception (Caroli, 2018) and traditional methods of software requirements documentation. Several meetings were held with project stakeholders for brainstorming and artifact validation, such as user journeys, mockups, and prototypes. Two personas were considered for defining functionalities: the domain expert and the data scientist.

Given that the developed project is premised on integration with an existing AutoML solution, the FMD, the new project architecture must complement the one previously used. Thus, a new layer of data ingestion and processing was added to the original architecture defined by (da Silva, 2020). The project utilizes Javascript for the frontend layer and Python for the backend layer, with an additional data ingestion layer managed by the Pentaho Data Integration Community Edition[7] (PDI-CE) platform, invoked by the backend. A visual representation of the project architecture is presented in Figure 2.

The developed platform presents two distinct and complementary workflows: the "context file" registration and the data source registration as presented on Figure 3. A "context file" is represented by a JSON-format file and contains the necessary metadata for analyzing data from a specific domain. There are two options for data source registration: one for registering data sources from CSV files and another for connecting to a PostgreSQL, MySQL, or Oracle database. After completing the data source registration, the data ingestion begins automatically.

[2]https://epistasislab.github.io/tpot/
[3]https://hyperopt.github.io/hyperopt-sklearn/
[4]https://github.com/automl/auto-sklearn
[5]https://h2o.ai/
[6]https://hdi-project.github.io/ATM/
[7]https://www.hitachivantara.com/en-us/products/pentaho-plus-platform/data-integration-analytics/pentaho-community-edition.html

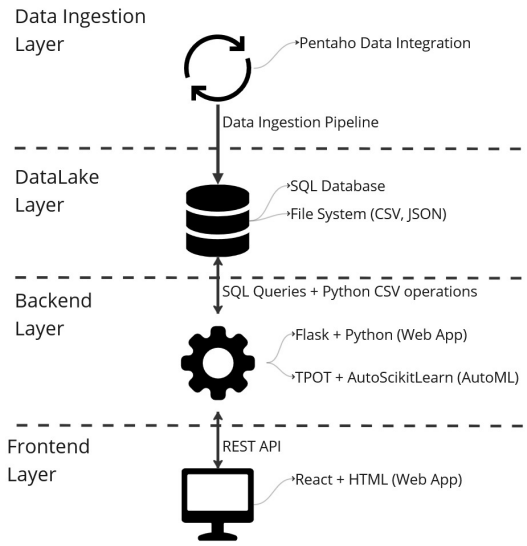

Figure 2: Project architecture.

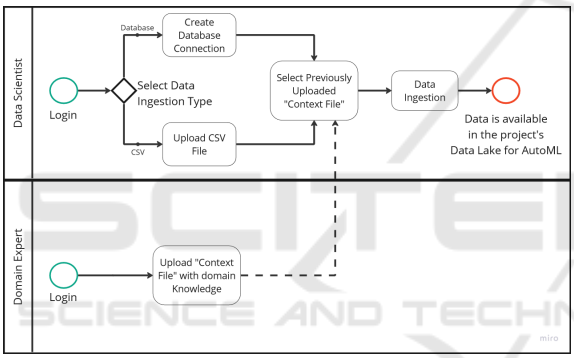

Figure 3: Project workflow.

4 RESULTS

4.1 Data Ingestion Interface for AutoML

Following the developed and validated prototypes, a graphical user interface for data ingestion was created. According to the project workflow, described in Section 3, screens were developed for uploading "Context Files" and for configuring data ingestion.

Figure 4 presents a screen for registering and uploading the context file by a Domain Expert user. After the upload, the transformation json_storage.ktr (Figure 5) is executed, which loads the "Context File" into the project's Data Lake, making it available for selection in the data ingestion configuration area. Besides the upload, the user can also edit the context on the screen.

Figure 4: User Interface for context file upload.

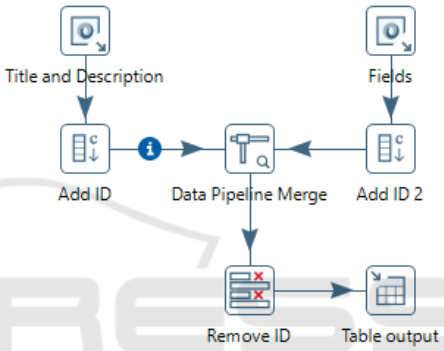

Figure 5: Transformation json_storage.ktr viewed at the visual interface of the PDI-CE.

Figure 6 represents the main screen for registering datasets, where the user will configure data ingestion in two steps, in a wizard interface format that represents a step-by-step guide for easy configuration.

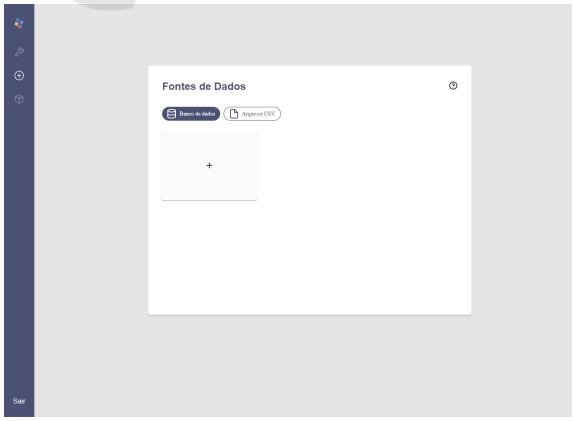

Figure 6: User Interface for the Data Ingestion Configuration.

After the configuration, a series of transforma-

tions are executed in the backend, which collect the registered information, gather the data, and execute the data ingestion into the project's Data Lake. By default, the data is stored in CSV format for better interaction with the AutoML platform; therefore, data transformations are required if it originates from a database connection, performed by the PDI. The first transformation executed is cria_headers_csv.ktr (Figure 7) that generates the CSV metadata for the data file produced at the end of the process. Next, the database data collection transformation, called ingestor_bd.ktr (Figure 8), is executed, and finally, the data is loaded into the Data Lake in CSV format by the transformation ingestor_bd_carga.ktr (Figure 9).

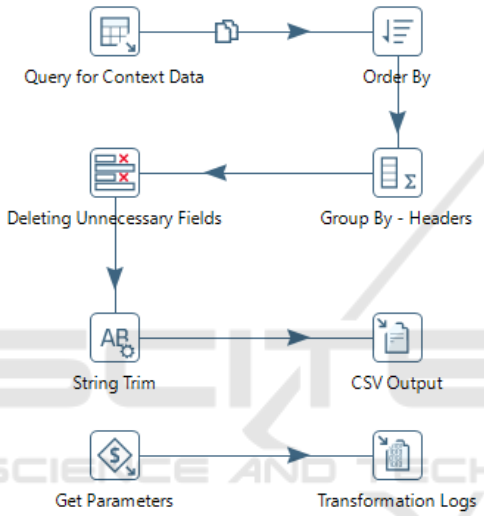

Figure 7: Transformation cria_headers_csv.ktr viewed at the visual interface of the PDI-CE.

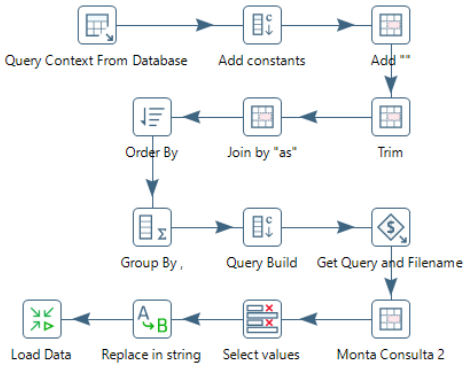

Figure 8: Transformation ingestor_bd.ktr viewed at the visual interface of the PDI-CE.

Figure 9: Transformation ingestor_bd_carga.ktr viewed at the visual interface of the PDI-CE.

4.2 Data Ingestion in a Real-World Scenario

Three context files were created to demonstrate data ingestion capabilities with different data contexts. These contexts are based on existing work in the literature that shows the best features for analyzing specific themes. The themes were: breast tumor classification using the six most important attributes, breast tumor classification using the ten most important attributes, and analysis of student engagement in virtual learning environments.

The two breast cancer contexts were used to demonstrate the ability to use various contexts with the same database; they were also added to the platform through a database connection. The engagement dataset, which was ingested into the platform from a CSV file, demonstrated the platform's capacity to work with widely different contexts (education and health).

The definition of the two contexts related to breast cancer was based on the work of (Ray et al., 2020), which aimed to analyze the best characteristics of the analysis of breast cancer cells to determine whether they are malignant or benign. The same dataset used in that work, the Wisconsin Diagnosis Breast Cancer dataset (WDBC), containing 573 instances, was used for data ingestion. The definition of the engagement context was inspired by (H. R. Macêdo et al., 2021), which sought to determine the most relevant data for analyzing student engagement profiles in virtual learning environments. The engagement dataset was provided by Research Group in Data Science and Analytics (GPCDA), which comprised 30,217 instances. After data ingestion and AutoML execution, the results are shown in Table 2, being (I) Top 6 features for Breast Cancer context, (II) Top 10 features for Breast Cancer context and (III) Features for students engagement context.

Table 2: Performance Metrics Results.

Context	Accuracy	AUC	Recall	Precision	F1 Score
(I)	0.91	0.96	0.93	0.93	0.93
(II)	0.93	0.99	0.91	0.99	0.95
(III)	0.94	0.99	0.92	0.94	0.94

The AutoML system provided the most adequate metrics for classification models. Analyzing the re-

sults, we can conclude that the feature selection based on the context file produced satisfactory values. All models performed very well, with high accuracy, AUC, precision, recall, and F1 score, indicating robust predictive ability.

4.3 Usability Assessment

An opinion survey was conducted through a form to assess the user experience when using the Data Mining Framework, specifically the data ingestion functionality. The questionnaire questions were based on the AttrakDiff method proposed by (Hassenzahl et al., 2000), commonly used in academia to validate usability and system quality aspects from the user's perspective. The survey consists of 28 pairs of opposing words that are used to describe the system in question. Respondents are required to choose the most appropriate description of the system on a scale from -3 to 3.

The study was conducted with 20 technology professionals with various levels of education, ranging from undergraduates (incomplete higher education) to professionals with a master's degree. This range of participants' knowledge levels allowed for the collection of information from both specialists and non-specialists in the field of data science from a sample of 20 participants.

As shown in Figure 10, users classified the data ingestion mechanism into two categories: "desired" and "task-oriented". When classified as "desired," the software likely provides users with a pleasant, aesthetically appealing, and emotionally satisfying experience. This suggests that users positively respond to the software's design, aesthetics, and sensory experience. Additionally, being classified as "task-oriented," as presented in Figure 11, implies that the software is perceived as efficient, functional, and suitable for fulfilling the specific tasks for which it was designed. This demonstrates that users view the software as useful, practical, and aligned with their functional needs. This is a favorable position, representing a positive balance between the pragmatic and hedonic aspects of the system.

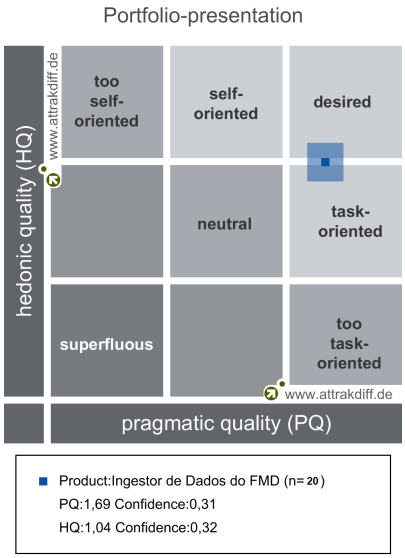

Figure 10: Portfolio of results.

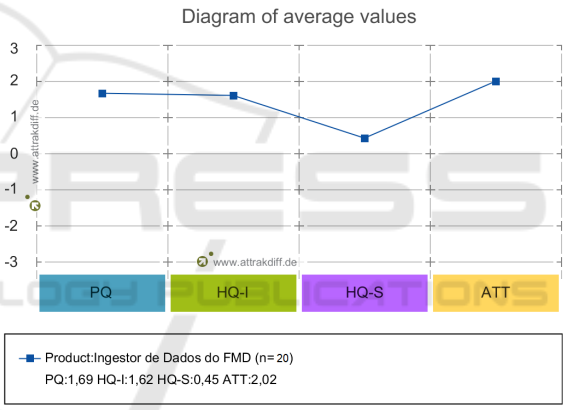

Figure 11: Diagram of average values.

4.4 Expert Opinion Evaluation

Research projects in software engineering often employ the expert opinion method (Garcia, 2010), as scientific development in software engineering has many peculiarities to evaluate the quality of the developed product. However, there are still divergent opinions on using this methodology, as there is no universally accepted framework to deal with expert opinions (Ming Li and Smidts, 2003).

The evaluation process comprised four stages inspired by those outlined by (Garcia, 2010). The first stage involved selecting experts based on credibility, technical knowledge, and availability. Next, opinion elicitation took place by introducing the system to the experts and providing a questionnaire for them to describe their considerations. The third stage involved opinion aggregation, where the overall con-

sensus regarding the obtained responses was assessed. Finally, results analysis was conducted by examining the questionnaire responses.

Analyzing the questionnaire responses provides valuable insights into the relevance of the developed platform and the research area. Based on expert opinions, the functionalities were successfully validated, although the tool has room for improvement in future developments. The highlighted benefits include greater convenience in the ingestion process, a shorter learning curve compared to similar platforms, and the potential for use in different contexts. Additionally, it was emphasized that the framework is suitable for meeting the needs of both technical and non-technical users. As for improvements, suggestions include adding new inputs for structured and unstructured data and options for responsiveness, accessibility, and internationalization of the platform.

5 CONCLUSION

This work presented the development of a data ingestion mechanism customized for AutoML systems, considering their peculiarities and those of their users. The project placed a strong emphasis on the personas identified during the Lean Inception process, guiding the entire development process. For this purpose, the data ingestion mechanism was created and integrated with an AutoML, named the ”Data Mining Framework.” This mechanism enables data input through a simple interface, from generic CSV files and database connections, mapping data contexts.

In terms of usability, various techniques related to the lean inception process were applied to optimize the development of the data ingestion, ensuring that users can perform this machine learning step with a low learning curve. To validate the findings and ensure research reliability, this aspect was assessed through an opinion survey with computer engineering students, along with the expert opinion elicitation process. The survey results confirmed the relevance and effectiveness of the project's identified functionalities, providing a good user experience with a lower learning curve compared to other tools used for the same purpose and the potential for use in different contexts.

Regarding computational intelligence, the project contributes with the development of a data ingestion module for automated machine learning systems, aiming at democratizing data science. The platform allows storing and transforming the tacit knowledge of business area experts into explicit knowledge, assisting in the engineering and selection of ideal attributes for applying machine learning techniques in a specific context. It is essential to highlight the standardization of the data ingestion process and the presentation of an optimized way to input data into the machine learning workflow.

In summary, from a technical point of view, the proposed data ingestion mechanism introduces innovations by automating and simplifying the preprocessing phase, which is often a bottleneck in AutoML workflows. Unlike traditional ingestion tools that require extensive manual intervention, the developed module allows users to access and leverage previously configured domain knowledge within the platform, enabling more informed data preprocessing and feature selection. This approach enhances the reproducibility of data pipelines by embedding provenance tracking and validation mechanisms, ensuring consistency in model training. These advancements contribute to reducing the effort required from users, making AutoML adoption more accessible while maintaining data integrity and reliability.

6 FUTURE WORK

In terms of future work, several areas stand out for enhancing the Data Ingestor. Based on expert opinions, the following future improvements have been identified:

- Expand the data ingestion capabilities to new formats of structured data, such as other database management systems (DBMS) or file formats like XML, JSON, and XLS;
- Expand the data ingestion capabilities to unstructured data, such as images and PDF files;
- Add regionalization functionalities that will enable international contributions to the tool's development, as it is open source;
- Include accessibility functionalities, allowing usage by a broader range of users;

ACKNOWLEDGEMENTS

This paper was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, Fundação de Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Brazilian research agencies.

REFERENCES

Aken, J. E. (2004). Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules. *Journal of Management Studies*, 41(2):219–246.

Alves, G. M. R. and Maciel, A. M. A. (2023). Survey on Data Ingestion for AutoML (S). In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering*, pages 411–414.

Caroli, P. (2018). *Lean Inception: como alinhar pessoas e construir o produto certo*. Editora Caroli.

Chen, Y.-W., Song, Q., and Hu, X. (2021). Techniques for Automated Machine Learning. *ACM SIGKDD Explorations Newsletter*, 22(2):35–50.

da Silva, R. (2020). Desenvolvimento de uma Solução de Aprendizado de Máquina Automatizado Integrável a Múltiplos Ambientes Virtuais de Aprendizagem. Master's thesis, Universidade de Pernambuco, Recife.

Elshawi, R., Maher, M., and Sakr, S. (2019). Automated Machine Learning: State-of-The-Art and Open Challenges. arXiv:1906.02287 [cs, stat].

Garcia, V. C. (2010). RiSE reference model for software reuse adoption in brazilian companies.

Gonçalves, A. F., Maciel, A. M., and Rodrigues, R. L. (2017). Development of a data mining education framework for visualization of data in distance learning environments. *International Conferences on Software Engineering and Knowledge Engineering*.

H. R. Macêdo, P., B. Santos, W., and M. A. Maciel, A. (2021). Análise de Perfis de Engajamento de Estudantes de Ensino a Distância. *RENOTE*, 18(2):326–335.

Hapke, H. and Nelson, C. (2020). *Building Machine Learning Pipelines: Automating model life cycles with tensorflow*. O'Reilly Media, Sebastopol, CA.

Hassenzahl, M., Platz, A., Burmester, M., and Lehner, K. (2000). Hedonic and *ergonomic quality aspects determine a software's appeal*. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 201–208, The Hague The Netherlands. ACM.

Hlupic, T. and Punis, J. (2021). An Overview of Current Trends in Data Ingestion and Integration. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1265–1270, Opatija, Croatia. IEEE.

Hutter, F., Kotthoff, L., and Vanschoren, J., editors (2019). *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham.

Ming Li and Smidts, C. (2003). A ranking of software engineering measures based on expert opinion. *IEEE Transactions on Software Engineering*, 29(9):811–824.

Nagarajah, T. and Poravi, G. (2019). A Review on Automated Machine Learning (AutoML) Systems. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–6, Bombay, India. IEEE.

Patel, J. (2020). The Democratization of Machine Learning Features. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 136–141, Las Vegas, NV, USA. IEEE.

Ray, S., AlGhamdi, A., AlGhamdi, A., Alshouiliy, K., and Agrawal, D. P. (2020). Selecting Features for Breast Cancer Analysis and Prediction. In *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 1–6, Las Vegas, NV, USA. IEEE.

Saurav, S. and Schwarz, P. (2016). A Machine-Learning Approach to Automatic Detection of Delimiters in Tabular Data Files. In *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1501–1503, Sydney, Australia. IEEE.

Tidwell, J., Brewer, C., and Valencia, A. (2020). *Designing interfaces: Patterns for effective interaction design*. O'Reilly Media, Sebastopol, CA.

Whalen, J. (2019). *Design for How People Think: Using Brain Science to Build Better Products*. O'Reilly Media, Sebastopol, CA.

Zöller, M.-A. and Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, 70:409–472.