

An Educational Data Mining Model based on Auto Machine Learning and Interpretable Machine Learning

Gabriel Novillo Rangone
School of Engineering
and Environmental Sciences
National University of Villa Mercedes
Villa Mercedes, Argentina
gnovillo@unvime.edu.ar

German Antonio Montejano
Faculty of Physical, Mathematical
and Natural Sciences
National University of San Luis
San Luis, Argentina
gmonte@unsl.edu.ar

Ana Gabriela Garis
Faculty of Physical, Mathematical
and Natural Sciences
National University of San Luis
San Luis, Argentina
agaris@unsl.edu.ar

Carlos Andres Pizarro
School of Engineering
and Environmental Sciences
National University of Villa Mercedes
Villa Mercedes, Argentina
cpizarro@unvime.edu.ar

Walter Ruben Molina
School of Engineering
and Environmental Sciences
National University of Villa Mercedes
Villa Mercedes, Argentina
wrmolina@unvime.edu.ar

Abstract— This paper proposes a new Data Mining Educational Model for knowledge extraction with a minimum presence of data scientists, a scarce and determinant human resource for the application of machine learning in Data Mining. The model allows generating data sets semi-automatically, obtaining an optimized algorithm through automatic machine learning (AutoML) and explaining the results with interpretable machine learning (IML). It is applied in the field of University Education and implements a process with three general stages (Data Analysis and Integration, Data Modeling and Results Evaluation) and is validated by means of a friendly prototype to non-expert users with data obtained from an Argentine Public University. With this proposal we aim to allow universities to draw conclusions on complex problems, requiring a minimum number of data science experts and providing a framework for both end users and legal entities to be informed of the results generated.

Keywords— Automated Machine Learning, Educational Data Mining, Interpretable Machine Learning, Model proposed for Educational Data Mining.

I. INTRODUCTION

Educational Data Mining (EDM) is a discipline concerned with developing methods to explore the unique and large volumes of data from educational environments and using those methods to better understand students and the institutions in which they learn. In recent years, there has been a great deal of research showing how EDM techniques have been used to improve academic performance, adaptive systems, educational quality, dropout rates, dropout, procrastination, and institutional management at all educational levels (initial to university) and for all modalities (face-to-face, online, hybrid). Some of the most important references are the "International Conference on Educational Data Mining" that has been held since 2008, and the "Journal of Educational Data Mining" that has been published since 2009. The achievements were made possible due to the advancement of three determinant components such as computational power, Big Data and Artificial Intelligence (AI) [1].

Educational Big Data draws on data from classrooms, educational institutions, ministries, curricula, pedagogical models, adaptive systems, virtual campuses, massive online courses and learning analytics, which together with AI techniques such as machine learning (ML) and Deep learning, allow scientists to obtain positive results for a large number of educational problems [2]. Different international organizations and prestigious authors, dedicated to the continuous improvement of education, have proposed advice and recommendations on how to best leverage AI technologies to achieve the 2030 Agenda and promote process designs that facilitate the implementation of knowledge extraction techniques for educators or non-expert users in the area of data mining [3],[4].

ML is a discipline belonging to AI that seeks that machine can "learn" or generalize knowledge from a set of experiences, without the need to be programmed in an explicit or traditional way. Different machine learning systems can be observed, such as supervised, unsupervised and reinforcement learning systems to perform classification, regression, clustering, association rules, and dimensionality reduction tasks, among others [5],[6].

A large majority of educational problems can be solved by applying supervised machine learning systems for classification and regression tasks. This can be verified by analyzing the reviews on educational data mining [17].

In order to apply ML to data from educational big data, it is necessary to have data scientists to perform very specific tasks such as data munging, feature preprocessing, machine learning algorithm selection, selection of the most optimal hyperparameters, visualization of the results and their interpretability and explainability. Given that data scientists are scarce due to the variety of cognitive skills they must possess and in addition to the high demand for them in the financial, commercial and industrial fields a shortage of this human resource has been generated in the educational field, making it difficult to achieve the democratization of educational data mining by applying machine learning to extract knowledge about complex educational problems..

AutoML provides methods and processes to make ML available to ML non-experts, to improve the efficiency of ML, and to accelerate ML research [7]. While using AutoML enables good predictions, the models generated are often complex and difficult to interpret. Interpretability means explaining a particular decision to users, including understanding the main tasks that affect the results and knowing the patterns, rules and features that an algorithm learns. This has great relevance to different stakeholders, including data scientists, end users and legal bodies to ensure that they are not break ethical or legal rules.

IML is an area of data science that develops methods for interpreting ML models [8]. There are some methods for intrinsically interpretable models and others for post hoc interpretation (model agnostic). Another classification groups methods for interpreting local or global models; that is, it explains an individual or general prediction of model behavior [9].

Following the research lines promoted by the International Organizations in Education and the European Commission on AI [10], and considering the shortage of data scientists in ML, a new model for EDM is proposed for supervised ML system and classification and regression tasks, which can be executed without data scientists or with a minimum presence of them. This undertaking is achieved by articulating state-of-the-art techniques and tools such as AutoML, IML, and the automation of part of the Extraction, Transformation and Load (ETL's) scripts.

This work is structured as follows. In the next section related work is discussed. Then the state of the art in educational data mining is shown. Then the proposed model is presented, followed by the validation and its results. Finally, conclusions and future work are presented.

II. RELATED JOBS

Numerous ML studies applied to the university environment can be found in the scientific community.

Castrillón et al. [11] design a methodology using AI techniques for training a system that classifies a new student into a series of predetermined categories of academic performance. The classification allows for the early identification of students with potential academic performance problems.

Carrizo et al. [12] show a line of research in which they apply DM to create a model that predicts the number of engineering graduates from a university, and identifies which patterns determine student graduation.

Istvan et al., in the same path, propose a line of research to obtain an Attrition Indicator Model, using DM techniques [13]. The model is incorporated into a computer system that allows the early detection of university student dropout.

Liu et al. [14] describe a predictive model, analyzing the behavior of students at a university during 2017. They built an ML system that automatically finds an optimized model for prediction. While these approaches make use of ML to find solutions in the university setting, none of them fully utilize AutoML techniques, as proposed in this paper.

Bianco et al. [15] apply AutoML in EDM projects to automate the process of algorithm selection and hyperparameterization, also with the goal that non-expert

users in data science can train a model and draw conclusions for decision making. The framework uses AutoML and includes model interpretability, generating knowledge patterns for decision making in educational environments.

Tsiakmaki et al. [16] study the effectiveness of AutoML for predicting student learning based on their participation in educational platforms. They limit the search space to the models used to achieve interpretability of the results.

Unlike the previous ones, the present work proposes to jointly apply the techniques to semi-automate ETLs (extract, transform & load), to automate the machine learning stage and to automate the interpretability of results and models obtained for both educational classification and regression problems.

III. STATE OF THE ART IN EDUCATIONAL DATA MINING

One of the first works in EDM is presented by Minaei-Bidgoli et al. who presented an approach to predict students' final grade, based on features extracted from data recorded in a web-based educational system. Among the most recent publications is the proposal by Akpinar et al. [18] who analyze student strategies in blended courses using click-through data.

EDM at the university level also has a significant body of related work. Panizzi [19] shows a systematic mapping of literature. He notes that most come from Asia, followed by North America and Europe. The most studied topics are academic performance, student profiles, educational quality, university dropout, followed by student procrastination, cooperative education, and student employability.

AutoML was introduced with the Auto-WEKA framework [20]. The intention was to help non-expert users effectively identify ML algorithms and hyperparameter settings suitable for their applications. Since 2014, both commercial and open source AutoML frameworks and tools were developed for AutoML, which fall into four major groups (centralized, distributed, cloud, and Neural Architecture Search - NAS), most of which are open source, except for cloud. The programming languages in which they are mostly built are Python, Scala, Java and R.

IML can be applied at different stages of the AutoML model development process. The first is in the model inputs, this can include summarizing the main characteristics of a dataset, finding the representative or critical points and determining the relevant characteristics of a dataset. The second is in modeling, which can be either white-box or black-box, a classification that will depend on its simplicity, transparency and explainability. The third is in a posteriori modeling, which helps to understand the most important features of a model, how those features affect predictions, how each feature contributes to the prediction, and how sensitive a model is [8].

The methods used for IML can be classified according to several criteria: intrinsic or post hoc, model-specific or model-independent, local or global, etc. However, the present work focuses on the intrinsic or post hoc criterion. This criterion distinguishes whether interpretability is achieved by restricting the complexity of the ML model (intrinsic) or by applying methods that analyze the model after training (post hoc).

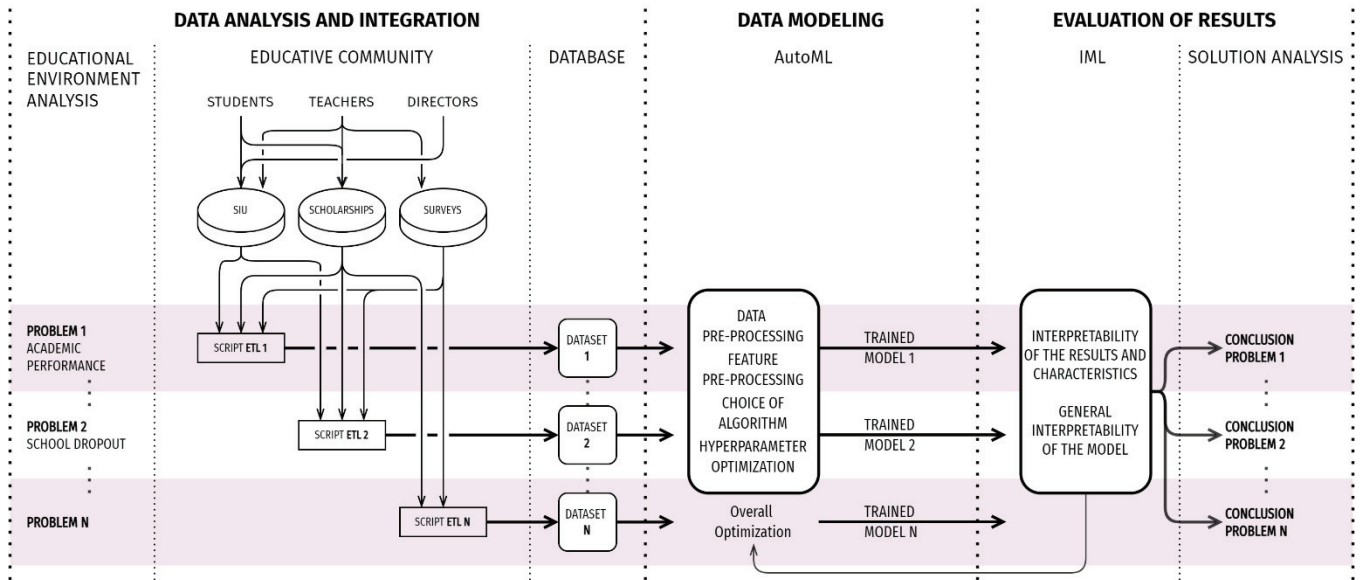


Fig. 1. Proposed Model.

Some of the most commonly used methods of a model-independent explanation system are: the partial dependence graph method, the individual conditional expectation method, the cumulative local effect graph, and the Shapley additive explanation

IV. PROPOSED MODEL

As shown in *Figure 1*, the model, which is instantiated in the domain of University Education, implements a process with three general stages or sub-processes: 1) Data Analysis and Integration, 2) Data Modeling and 3) Evaluation. of Results.

The Data Analysis and Integration stage begins with the evaluation of the educational environment by an interdisciplinary team and is divided into 4 particular stages:

- 1) Definition of the problem to investigate (dropout, procrastination, breakdown, or student performance, among others),
- 2) Configuration of the objectives to be met, eg. In the case of studying desertion, the objective may be to know the cause or causes that most influenced (distance from the University, age, schedules, grades, etc.),
- 3) Definition of the necessary and accessible data for the institution,
- 4) Acquisition of data from the different sources that the educational community feeds through applications, software from the institutions themselves, and software for general use, such as surveys, scholarships, University Information System (SIU), etc.

From here, some tasks are automated using ETL scripts, and others are handled by the analyst. The analyst will be able to select the data from the different sources, carry out the sampling, filter columns or rows, and group data, while the tasks of eliminating missing, duplicates, inconsistencies, and outliers (3 sigma, interquartile ranges, etc.).

Once these tasks are completed, a dataset (separate, complete, consistent and updated data set that serves as the basis for training an ML algorithm) is created for each educational problem and stored in a database.

The Data Modeling Stage begins with the acquisition of the data set or dataset to be trained for the problem posed and the definition of the type of problem to be solved (classification or regression), continues with the optional inclusion of data preprocessors and preprocessors of characteristics, and the inclusion or exclusion of specific algorithms, in addition to the resampling strategy; and concludes with the determination of resource constraints such as time, memory, cores, etc. This stage is mainly developed by a data analyst with general knowledge of ML or a Data Scientist with a minimal presence or with little experience, since practically all the particular tasks are automated.

The Results Evaluation stage begins with the acquisition of the metrics of the trained model. You can choose to analyze the metrics and conclude in a traditional way or activate the Interpretability methods. If the activation is carried out, the type of methods must be selected, whether they are interpretable (linear regression, logistic, decision tree, RuleFit, etc.) or independent of the model (PDP, function interaction, SHAP, etc.). The Interpretability options are activated according to the settings chosen in previous stages. The model also includes the option to perform Interpretability of the end-to-end pipelines that AutoML previously performed, allowing information to be obtained to understand and debug the trained models. The conclusion of the problems is performed by a group of multidisciplinary professionals of the educational institution that evaluate the magnitude of the characteristics (causes) that influence the result.

V. VALIDATION

To validate the proposal, a prototype was developed that implements the stages mentioned above and the model was evaluated in the field of an Argentine public university.

Trained Models

Search Q ≡ ⌵ ⌶












ID	Date	Model	User	Actions
3	16/07/2021	Student Profile by Performance in Health Sciences Careers	admin	   
2	15/07/2021	Dropout Prediction for Licenciatura en Obstetricia 3 rd Course	admin	   
1	12/07/2021	Dropout Prediction for Analista de Sistemas 1 st Course	admin	   

Fig. 2. EDM Web based tool Main Screen.

The architecture of the prototype is divided into two large layers: the front-end and the back-end, as shown in Figure 2. The front-end contains the view and the prototype controller. The view provides the graphical interface for the management of requirements as well as for the interaction between the data analyst and the instance automation process.

We coded the front-end in Java Script, using Bootstrap and Angular frameworks in addition.

In the back-end of the application there is the database manager, the ETL to generate the specific intermediate database, the datasets and the AutoML and IML processes. In particular, we use Python libraries to run the processes autosklearn, numpy, pandas, pdpbox, matplotlib, seaborn, scipy, shap and sklearn.

The virtual machine, where the back-end is hosted, works with GNU-Linux operating system; specifically, it uses the Ubuntu 12.04 LTS (long-term support) distribution. Additionally, it utilizes Python 3.7, a C++ compiler compatible with C++11, and the SWIG interface compiler.

VI. RESULTS

Tests have been carried out on the initial prototype. A desktop computer was brought with a 5.3 Ghz Intel Core i9-10900K processor, 16 GB of RAM, 20 MB of cache and a 2-terabyte hard drive.

The execution of the prototype was carried out with data provided by the Secretariat of Innovation and Technological Articulation of the National University of Villa Mercedes (UNViMe), within the framework of the Research Project "Search for Knowledge applying Data Science in the Domain of Education". Being a university with a few years of life, it dedicated itself to working with the Faculty of Health Sciences, and more specifically with data from students who entered in 2019 and 2020 in the Bachelor of Nursing career because it is the career with higher attendance of students.

The educational problem raised was the loss of regularity (less than 2 subjects passed on May 2 of the year following

their admission) by the students, for which a group of data was analyzed, among which the academic situation stands out of each of the subjects of the first year of the career (9 in total), if the student has a job, if he/she is financially supported by the family, if he/she has children, the academic level of the parents, and the difference in years between received high school and started university (delay).

For more details, the code for the execution of these results can be found in an open repository [21].

The constructed dataset has the following characteristics:

1) 258 instances or rows (students who entered the Bachelor of Nursing in 2019 and 2020, excluding readmissions).

2) 15 characteristics or columns (10 characteristics referring to the student's academic situation and 5 characteristics referring to their social situation).

3) The number of regular students is 160 (61.7%), while the number of students who did not regularize is 98 (38.3%).

The tests have allowed us to extract these results:

1) The model used was inherited from the AutoSklearnClassifier class of the Autosklearn library, allowing the model to perform meta-learning and construction of sets (25 sets) without restricting the preprocessing of data and functions. Cross validation was set up with 5 folds. The model was trained on all available classifiers (14 algorithms), no one in particular was excluded or included. The training time was 72,000 seconds with steps of 1,000 seconds for each algorithm.

2) The classification algorithm that obtained the best accuracy score (0.9794) was the gradient boosting. In addition, complementary metrics such as recall, f1 (average precision and recovery), and error rate are calculated, yielding results of 0.9828, 0.9811, and 0.0259, respectively. For the calculation of precision, recall and f1, the libraries were imported from autosklearn.metrics, and for the calculation of error rate, a function was created from autosklearn.metrics.make_scorer.

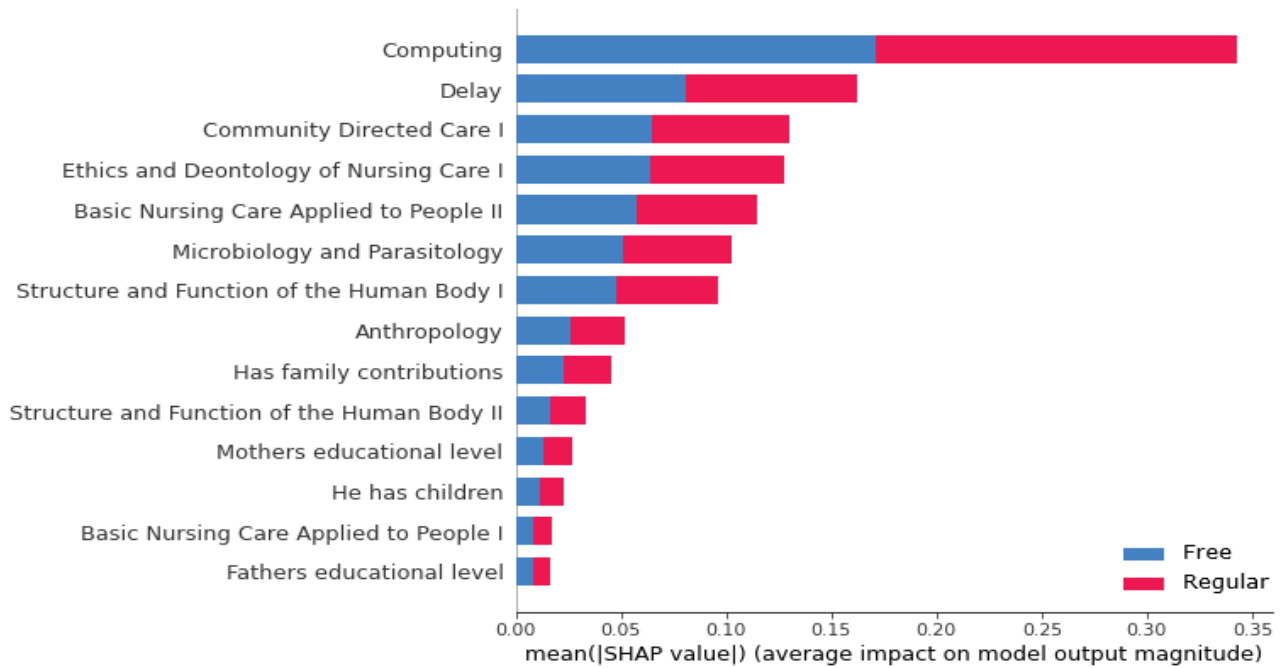


Fig. 3. Shap chart of Important Characteristics of UNViMe 2019/2020 students.

3) The Interpretability stage allows us to know automatically which are the most influential characteristics in

Important characteristics of UNVIME 2019/2020 students

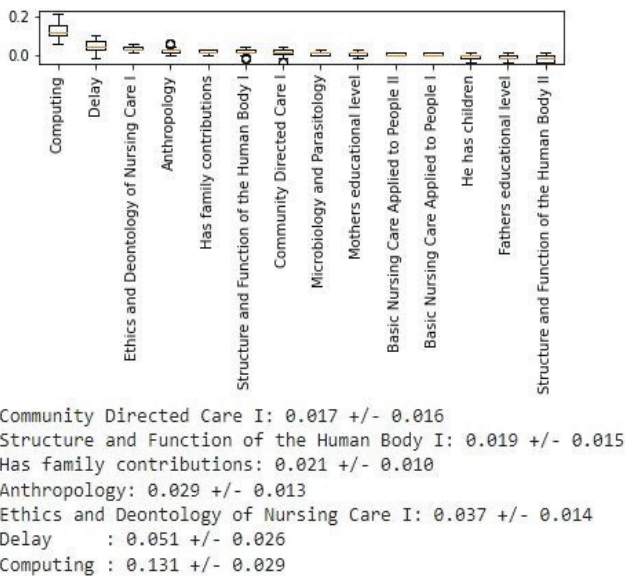


Fig. 4. Distributions of Feature Importance Values.

the result. In this case, it was decided to apply model-agnostic methods since the algorithm that achieved the best performance (gradient boosting) is not within the interpretable models and to achieve complementarity in the analysis, it was decided to apply a tool for global model-agnostic methods (permutation importance) and a tool for local model-agnostic methods (SHAP). While global methods describe the average behavior of a machine learning model, local methods achieve explanation through each instance and its predictions.

4) In the case of Shap (Figure 3), the summary diagrams method was obtained to make a bar graph of the

characteristics and their assigned weights of the students of the Bachelor of Nursing career with whom the model was trained. It is observed that the characteristics Computing (first year subject) and Delay (difference in years between graduating from high school and starting university) are the most influential for being a regular student with 35% and 17% respectively. As additional data, the subjects with the highest number of postponements were ordered as follows (1st: Microbiology and Parasitology, 2nd: Basic Nursing Care Applied to People I, 3rd Computing).

5) Starting from the sklearn.inspection library (Figure 4), the Permutation Importance method was used to discover the features that generated the greatest error when their original values were perturbed. This method provides reliable information when the model to which it is applied has a low error (0.0259 in this case). The information used, as in the case of Shap, was the output of the model trained for the dataset of the Nursing Degree Course. The most influential characteristics for this type of global method were also the Informatics subject and the delay characteristic.

6) The model also proposes a particular stage of General Interpretability that informs about the models and in this way to know the channelings that had a better performance. In this case, the Pipeline Profiler library was pulled, and the most successful pipeline was found to use a gradient boosting algorithm with data and feature preprocessing. This information is valuable when the model needs to be retrained because the results are not optimal.

VII. CONCLUSION AND FUTURE WORK

This work has presented an educational data mining model based on AutoML and IML, which allows university educational institutions to make conclusions on complex problems. The model makes it possible to minimize the need for data science experts both for the choice of models and their hyperparameters and for their interpretation, while

providing a framework for both end users and legal entities to be informed about the generated results.

The proposal has been validated through a prototype, which has allowed testing with a data set from an Argentine university.

Future work will investigate the possibility of automation for unsupervised machine learning that allows the model to include the possibility of solving educational problems of clustering and association.

In addition, the automatic balancing of the dataset is expected to be optionally included in the data analysis and integration stage since it is available for the modelling stage.

The latest version of Auto-Sklearn released in 2020 is being evaluated, as well as AutoKeras for structured data for inclusion in the tool.

REFERENCES

- [1] A. Merceron, R. S. Baker, M. Chi, A. M. Olney, A. Rafferty, and K. Yacef, "Editorial Acknowledgment", *JEDM*, vol. 14, no. 1, p. i-ii, Jun. 2022.
- [2] B. Williamson, *Big data en educación*. San Sebastián de los Reyes (Madrid): Ediciones Morata, S. L. 2018. [En Línea] Disponible en: <https://elibro.net/es/lc/unvime/titulos/119511>
- [3] United Nations Educational, Scientific and Cultural Organization, "Beijing Consensus on Artificial Intelligence and Education", Beijing, People's Republic of China, 2019.
- [4] Instituto Internacional para la Educación Superior en América Latina y el Caribe, "Towards universal access to higher education: international trends", Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 7, place de Fontenoy, 75352 París 07 SP, Francia, 2020.
- [5] P. Meseguer González y R. López de Mántaras Badia, *Inteligencia artificial*. Madrid: Editorial CSIC Consejo Superior de Investigaciones Científicas, 2017.[En Línea] Disponible en: <https://elibro.net/es/lc/unvime/titulos/42319>
- [6] M. Awad and R. Khanna, "Machine Learning," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineer and System Designers*, Berkeley, 2015, pp. 1-18.
- [7] F. Hutter, "AutoML", 2022, [online] Available: <https://www.automl.org/>
- [8] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Jun. 2018, [online] Available: <https://christophm.github.io/interpretable-ml-book/>.
- [9] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning", 2018, [online] Available: <https://arxiv.org/abs/1702.08608>.
- [10] European Commission, "New rules for Artificial Intelligence – Questions and Answers," April 2021, [online] Available: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683. [Accessed 30 01 2022].
- [11] O. D. Castrillón, W. Sarache and S. Ruiz-Herrera, "Predicción del rendimiento académico por medio de técnicas de inteligencia artificial," *Formación universitaria*, vol. 13, no. 1, pp. 93-102, Febrero 2020.
- [12] C. Carrizo, J. Saldarini, G. Ribotta, F. Cardona and J. I. Marotti, "Modelo para predecir la cantidad de graduados de Ingeniería de UTN aplicando técnicas de minería de datos," en *XIX Workshop de Investigadores en Ciencias de la Computación*, Buenos Aires, 2017, pp. 325-330, Abril 2018.
- [13] R. Istvan and V. Lassagna, "Sistema Informático para la Detección Temprana de Deserción Estudiantil Universitaria: Estudio sobre ingresantes de la UTN Regional La Plata", *Innovación y Desarrollo Tecnológico y Social (IDTS)*, vol. 1, no. 2, pp. 1-15, Noviembre 2019.
- [14] R. Liu and A. Tan, "Towards Interpretable Automated Machine Learning for STEM Career Prediction", *JEDM*, vol. 12, no. 2, pp. 19–32, Aug. 2020.
- [15] S. Bianco, S. Martins, H. Amatriain and H. Merlino, "Propuesta de automatización para proyectos de minería," en *XXV Congreso Argentino de Ciencias de la Computación (CACIC)*, Río Cuarto, 2019, pp.366-375, Octubre 2019.
- [16] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis and O. Rago, "Implementing AutoML in Educational Data Mining for Prediction Tasks," *Applied Sciences*, vol. 10, no. 1, p. 90, 2019.
- [17] J. Morales Carrillo, V. Trujillo Utreras, S.Cevallos Molina and M. Santana Cedeño, "Data mining in education: a literary review", *ESPAMCIENCIA*, vol. 9, no E, pp.14-20, Mayo 2019..
- [18] N.-J. Akpinar, A. Ramdas and U. Acar, "Analyzing Student Strategies In Blended Courses Using Clickstream Data" in *Proceedings of the 13th International Conference on Educational Data Mining*, Fully virtual conference, July 2020.
- [19] M. Panizzi, "Establecimiento del estado del arte sobre la Minería de Datos Educacional en el Nivel Superior: Un Estudio de Mapeo Sistemático," *Revista de Investigaciones Científicas de la Universidad de Morón*, vol. 2, no. 4, pp. 51-60, Julio 2019.
- [20] L. Kotthoff, C. Thornton, H. Hoos, F. Hutter and K. Leyton-Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA" In *JMLR*, pp.1–5, April 2017
- [21] G. Novillo Rangone, *EDM Tool*, Github Repository, 2022, [Online] Available: <https://github.com/UNViMe/edm-tool>