Propuesta de automatización para proyectos de minería de datos educativa

Santiago Bianco, Sebastian Martins, Hernán Amatriain, Hernán Merlino

Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información
Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús
Remedios de Escalada, Buenos Aires, Argentina.
santiago.bianco.sb@gmail.com, smartins089@gmail.com, hamatriain@gmail.com,
hmerlino@gmail.com

Resumen. En los últimos años es cada vez más frecuente que establecimientos educativos cuenten con diversos sistemas de información. La aplicación de técnicas de minería de datos permite hacer uso de la información generada por estos sistemas para mejorar la calidad en la enseñanza de las instituciones. Tanto es así que existe una disciplina dedicada exclusivamente a esto denominada Minería de Datos Educativa. No obstante, esta práctica requiere de una experticia en el campo de la explotación de información que no muchos miembros de la comunidad educativa poseen, lo que dificulta su aplicación. Para subsanar este problema, en este trabajo se propone un marco de trabajo automatizado, demostrando que se pueden conseguir resultados iguales o mejores a los que se obtendría aplicando procesos de minería de datos tradicionales y sin necesidad de conocer en detalle el funcionamiento de los algoritmos que se aplican.

Palabras clave: Tecnología Informática Aplicada en Educación, Automated Machine Learning, Educational Data Mining.

1. Introducción

La Minería de Datos Educativa (EDM, por sus siglas en inglés) se define como "una disciplina emergente, relacionada con el desarrollo de métodos para explorar los tipos únicos de datos que provienen del entorno educativo y el uso de esos métodos para entender mejor a los estudiantes y al entorno en el que aprenden" [1]. Las principales categorías en las cuales las líneas de investigación se han centrado son [2]: análisis y visualización de los datos, detección de comportamientos no deseados, modelado del estudiante, proveer recomendaciones a los docentes, administrativos y/o responsables académicos, predicción del rendimiento de los estudiantes, entre otros. Se señala, además, que las investigaciones en los últimos años se han centrado principalmente en el estudio del comportamiento de los estudiantes en sistemas educativos. Existe, además, una incipiente tendencia hacia el análisis de la información para ayudar a dichos sistemas, y la potencial mejora de algunos aspectos de la calidad de la educación y de los procesos de aprendizaje y el análisis del comportamiento de los estudiantes en cursos y carreras universitarios [3].

Cada vez existen más algoritmos y herramientas para implementar las distintas técnicas de minería de datos, aunque generalmente estas se centran en la eficiencia y precisión más que en la facilidad de uso. En este contexto, Romero y Ventura señalan como líneas de investigación de interés, el diseño de procesos o métodos que faciliten a educadores y/o usuarios no expertos en el área de minería de datos la implementación de las técnicas de extracción de conocimiento [2].
El objetivo de los algoritmos de automatización de aprendizaje automático (del inglés Automated Machine Learning, AML o AutoML) es generar un modelo predictivo evitando realizar manualmente algunas tareas iterativas que requieren un conocimiento específico del área de la ciencia de datos, como la selección del modelo apropiado de acuerdo al problema, la optimización de sus hiperparámetros y la selección de atributos relevantes [4].
En base a estos dos últimos aspectos, en el presente trabajo se explora la aplicación de AutoML en proyectos de EDM, de manera que se pueda automatizar el proceso de selección de algoritmos y su hiperparametrización, así como también gran parte del pre-procesamiento de datos. De esta manera se busca que usuarios no expertos en el área de la minería de datos y ciencia de datos sean capaces de entrenar un modelo en base a sus datos y puedan extraer conclusiones que sirvan de soporte para la toma de decisiones.

2. Estado de la cuestión

En las últimas décadas se han desarrollado numerosas investigaciones aplicando minería de datos para la resolución de problemas en el dominio de la educación: reglas de asociación [5], agrupamiento [6], clasificación [7], entre otras. Estas pueden utilizarse, por ejemplo, para describir el comportamiento de poblaciones estudiantiles, como apoyo a la toma de decisiones, identificación de causales de deserción y abandono y gestión de aulas virtuales [8]. Incluso también han empezado a utilizarse técnicas de minería de procesos y minería de datos no convencionales, como minería de grafos, para resolver problemas de las instituciones educativas [9] [10].
El inconveniente es que en todas las investigaciones antes mencionadas se cuenta con la inclusión en el equipo de un experto en el área de minería de datos o alguna disciplina afín, algo que encarece el proceso y dificulta su aplicación a las instituciones educativas que se interesen en este tipo de soluciones. En un proceso de EDM tradicional como el que se muestra en la figura 1, es necesario realizar tareas que involucran un conocimiento del dominio de la educación, así como también otras que son propias del área de minería de datos y requieren un cierto grado de experticia para ser ejecutadas correctamente. Este subconjunto de tareas incluye el pre-procesamiento de datos, selección de algoritmos y refinación iterativa del modelo obtenido, como puede observarse en detalle en la figura 2.
Si bien existen herramientas que facilitan la manipulación de datos, análisis algorítmico, visualización de datos y ejecución de procesos de explotación de información [11], estas siguen requiriendo tener un alto grado de experticia en el área de la ciencia de datos.

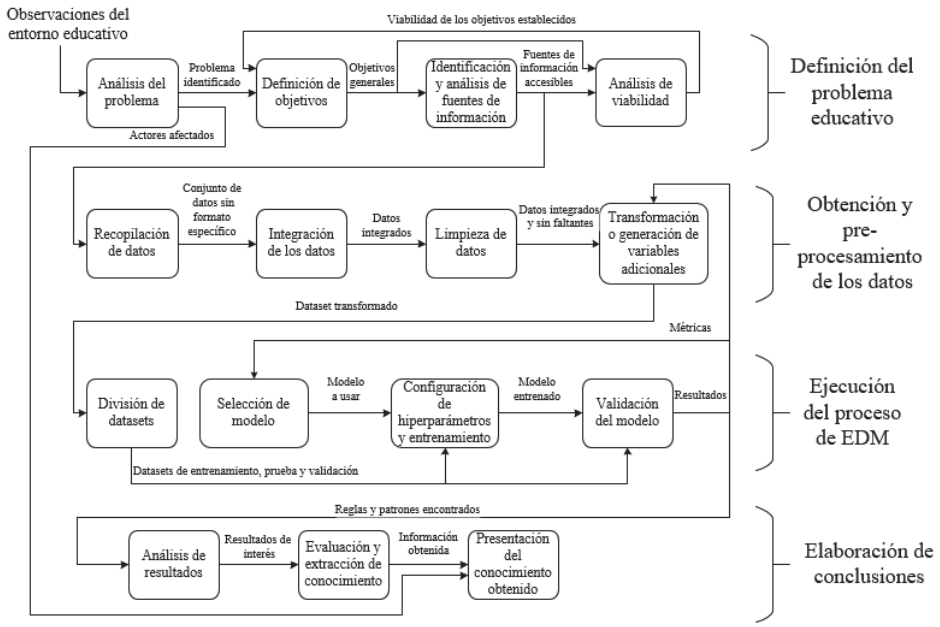

Fig. 1. Actividades de un proceso de EDM tradicional.

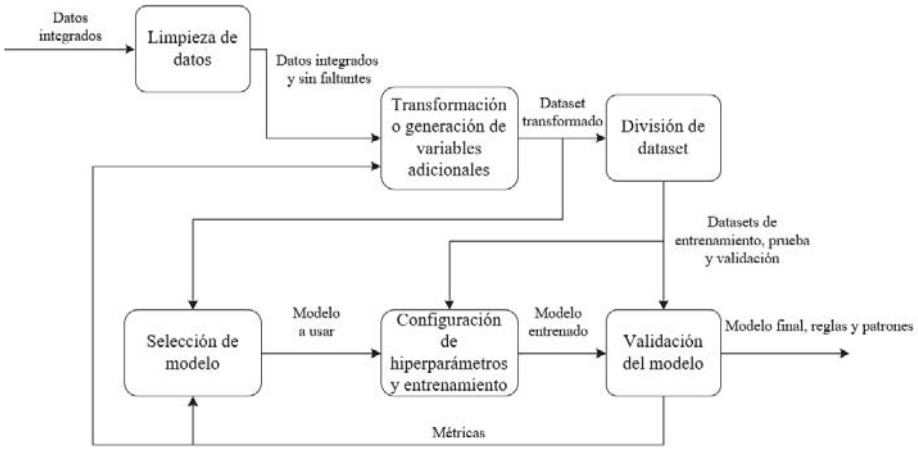

Fig. 2. Detalle de actividades propias de un proceso de minería de datos.

En otros dominios, como respuesta a esta situación se comenzó a trabajar con herramientas de Automated Machine Learning para la búsqueda de patrones en grandes masas de información [12] [13]. AutoML es un campo emergente que estudia los métodos por los cuales se pueden optimizar y automatizar los procesos de entrenamiento de los algoritmos de aprendizaje automatizado. Esta idea puede ser representada como se muestra en la figura 3, un proceso que recibe un conjunto de datos integrados y como salida provee la lista de transformaciones que se deben aplicar a los

datos, el modelo más óptimo que puede aplicarse y sus hiperparámetros, todo agrupado en un pipeline que puede ser ejecutado con una línea de código o mediante consola. Lo único que hay que hacer es proveerle un nuevo conjunto de datos y realizará la transformación y predicción. Se pueden utilizar tanto para clasificación como para regresión y existen varias herramientas desarrolladas que utilizan diversos enfoques para lograrlo, como redes bayesianas [14] [15] o algoritmos genéticos [16], lo que evita probar a fuerza bruta las distintas combinaciones de transformaciones y configuraciones de los modelos. Esto último también evita el refinamiento iterativo que debe realizarse en un proceso de EDM tradicional, en el cuál se valida el modelo y, en caso de no lograr una buena métrica, se lo debe modificar, elegir otro distinto o cambiar alguna transformación de los datos.

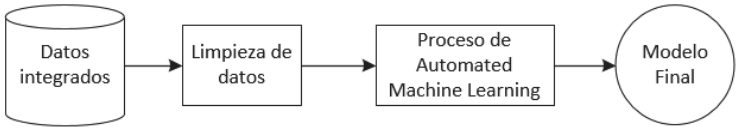

Fig. 3. Ejemplo AutoML pensado como un proceso.

Si bien el uso de AutoML simplifica la obtención de modelos predictivos de calidad, también genera que estos modelos sean más complejos y difíciles de interpretar. Para problemas como la clasificación de imágenes o texto, generalmente no es necesario saber las características del modelo implementado sino que solamente interesa la salida del mismo. Es decir, importa más que el algoritmo haga una clasificación correcta antes que saber el por qué decidió realizar esa clasificación, por lo que la generación de modelos complejos no es un inconveniente. No ocurre lo mismo en dominios en los cuales lo que se busca es, más allá de una correcta clasificación o predicción, encontrar una explicación del porqué de esas situaciones. El área de la educación se incluiría dentro de este segundo grupo. En los procesos de EDM es necesario explicar los modelos obtenidos para conseguir un análisis más profundo del problema educativo. Esto puede verse, por ejemplo, a través del estudio de la deserción universitaria. No solo se busca predecir tempranamente si un estudiante es propenso a abandonar una carrera o no, sino que también se pretende determinar las causas por las que los estudiantes abandonan sus estudios. De esta manera se pueden determinar en dónde asignar recursos para prevenir estas situaciones en un futuro y mejorar la calidad de la institución educativa. Por lo tanto, para poder implementar técnicas Automated Machine Learning en la educación es necesario, además, agregar una etapa de interpretación del modelo obtenido, para extraer reglas, identificar cómo influyen los distintos atributos en el resultado y definir la interrelación entre los mismos.

3. Descripción del problema

Según lo expuesto previamente se identifica la necesidad de aplicar un proceso de minería de datos en entornos educativos que no requiera conocimientos específicos en el área de explotación de información y afines. Esto podría subsanarse con la aplicación de Automated Machine Learning para la etapa de extracción de conocimiento, pero

estos procesos generalmente proveen, no sólo modelos generalmente difíciles de interpretar, sino también distintas transformaciones a los datos que pueden derivar en la creación de nuevos atributos, los cuáles deben poder rastrearse y relacionarse con los atributos originales para generar conclusiones valiosas.
La interpretación de modelos complejos de caja negra es un área de creciente interés [17], por lo que se han generado algunas técnicas que, combinadas adecuadamente con AutoML, serían capaces de generar un proceso de EDM aplicable por cualquier usuario no experto. En base a esto surgen las siguientes preguntas de investigación:

- ¿Es posible aplicar Automated Machine Learning en entornos educativos y obtener modelos predictivos con una efectividad similar a los de un proceso tradicional de EDM?
- De poder obtener esos modelos, ¿es posible agregar una etapa posterior de interpretabilidad que posibilite la comprensión de los patrones generados para que sirvan como apoyo a la toma de decisiones?
- De resultar ambas respuestas positivas, ¿Se puede generalizar el proceso descrito en un marco de trabajo que permita automatizar los proyectos de EDM?

4. Solución propuesta

Para resolver el problema antes descrito se propone utilizar AutoML junto con herramientas de interpretación de la forma en la que se muestra en la figura 4.

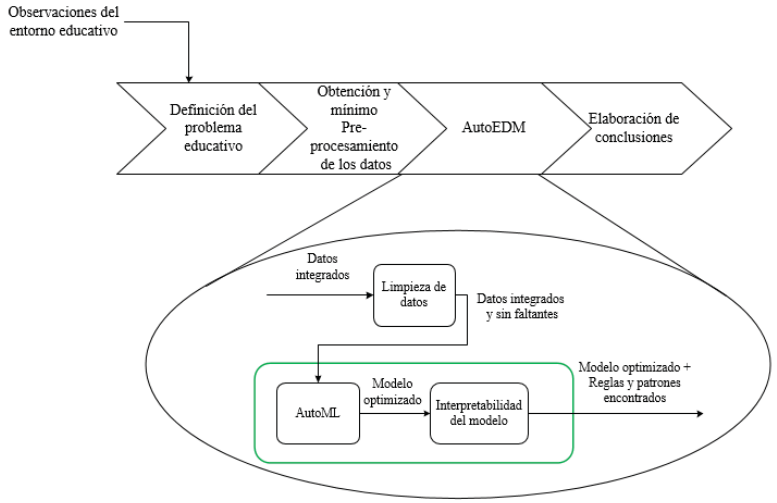

Fig. 4. Proceso de AutoEDM propuesto.

Las reglas y patrones encontrados pueden ser utilizados para la elaboración de informes y conclusiones con respecto a los datos que ayuden al proceso de toma de decisiones en la institución o aula. El modelo optimizado que se obtiene puede ser usado posteriormente para predecir el atributo clase en nuevos casos.
Este framework de trabajo posibilitaría la simplificación de cualquier proceso de EDM, evitando los problemas inherentes a los procesos de minería de datos tradicionales,

como selección del algoritmo apropiado, optimización de hiperparámetros y transformación de los datos. De esta manera, se puede concentrar el trabajo en las áreas de análisis del dominio y elaboración de conclusiones, aspectos que pueden ser realizados por un usuario con conocimientos en el área educativa.

5. Prueba de Concepto

Para probar qué tan factible es la utilización de AutoML en la educación, se contrasta un proyecto aplicando esta tecnología sobre los mismos datos utilizados en un proyecto de EDM tradicional, desarrollado por este grupo de investigación [18]. La idea es comparar los resultados obtenidos para demostrar que no sólo se simplifica el proceso sino que el modelo final presenta mejores métricas para el conjunto de datos de prueba. Se analizará puntualmente el caso de la deserción universitaria, describiendo las causas por las que los alumnos pierden la regularidad en la carrera. La base de datos cuenta con los atributos descritos en la tabla 1. El atributo clase a analizar será "Es Regular" y se utilizarán todo el resto de los atributos para su descripción, tal como se hizo en el estudio original.

Tabla 1. Descripción de los atributos de la base de datos utilizada.

Variable	Tipo	Valores	Distribución
Edad al primer año de cursada	Discreto	18 a 64	μ=25,31 σ=4,64
Discapacidad	Booleano	Sí	97,78%(1409)
		No	2,22% (32)
Diferencia entre el egreso del secundario e ingreso a la carrera	Discreto	0 a 37	μ=3,54 σ=4,16
Tipo de colegio secundario	Nominal	Técnico	12,77% (184)
		Bachiller	55,59% (801)
		Comercial	31,37% (452)
		Sin datos	0,28% (4)
Trabaja	Booleano	Sí	35,39% (510)
		No	64,61% (931)
Categoría últimos estudios del padre	Ordinal	0	27,34% (394)
		1	38,03% (548)
		2	27,62% (398)
		3	6,18% (89)
		4	0,83% (12)
Categoría últimos estudios de la madre	Ordinal	0	24,64% (355)
		1	29,49% (425)
		2	30,19% (435)
		3	14,84% (185)
		4	2,85% (41)
Es Regular	Booleano	Sí	16,38%(236)
		No	83,62% (1205)

Para la implementación de la etapa de AutoML se propone el uso de una librería denominada TPOT [16], la cual está implementada en el lenguaje de programación Python. Dicha librería fue elegida por su simplicidad de uso y porque utiliza algoritmos

genéticos para optimizar los distintos modelos que prueba, los cuales presentan mejores resultados con respecto a las otras estrategias [19].
Como se mencionó anteriormente, existe una creciente demanda para poder interpretar modelos de Machine Learning de caja negra, por los cuales se han desarrollado varias alternativas [20]. Por la naturaleza de los casos de EDM, y en base a trabajos ya realizados por este grupo de investigación [21] [22], las alternativas más viables que se pueden aplicar son SHAP [23], por su aplicabilidad en cualquier tipo de modelo, y el método de interpretación propuesto por Satoshi Hara [24] ya que con él se logran reglas interpretables simples tanto para clasificación como para regresión y no requiere configuración extra. Para la prueba se utilizaron solamente los estimadores de TPOT cuyos modelos finales incluyeran árboles de decisión, simples o ensamblados. Esto puede realzarse a través del diccionario de configuración. Para más detalles, el código para la ejecución de este experimento puede encontrarse en un repositorio abierto [25].

6. Resultados obtenidos

Como primera medida para comparar los resultados se utilizó la métrica de exactitud. Sin embargo, como se trata de una muestra desbalanceada (16% regulares vs. 84% no regulares), se aplicaron varias métricas para medir la calidad del modelo generado. El resumen comparativo puede observarse en la figura 5.

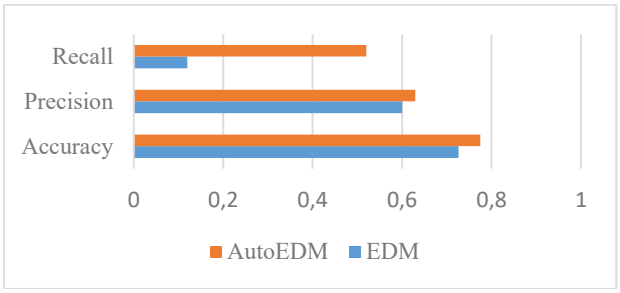

Fig. 5. Métricas para la clase positiva

Puede verse que para el mismo conjunto de datos y para el mismo problema de minería de datos, el modelo de AutoEDM presenta mejores resultados que utilizando el enfoque tradicional. Sin embargo, a esto hay que agregarle algo de interpretabilidad, como se mencionó anteriormente, para poder utilizar la información que brinda el modelo, además de predicciones. Esto se consiguió de distintas formas. Primero, utilizando una implementación del algoritmo defragTrees [24], obteniéndose reglas simples que explican el patrón identificado por el modelo obtenido. En la figura 6 se ejemplifican las reglas generadas para la clase con menor cantidad de ejemplos.
En dicha figura puede observarse que las variables influyentes para que un alumno mantenga la regularidad son la edad al primer año de cursada, Xt6 y Xt2. Estas dos últimas tienen esos nombres ya que son variables derivadas de los atributos originales.

```
[Regla  5]
y = 1 when
        Edad primer anio < 0.333236
        Xt6 < 0.000355
[Regla  6]
y = 1 when
        0.333330 <= Edad primer anio < 0.334549
        Xt2 < 0.000301
```

Fig. 6. Extracto de reglas obtenidas con defragTrees.

Para poder identificar cuáles son, se utiliza la matriz de correlación identificando la variable original con mayor relación (figura 7). Observando los colores más altos y más bajos de la escala de la derecha se identifica que se corresponden con las variables "Categoría último estudio madre" y "Trabaja".

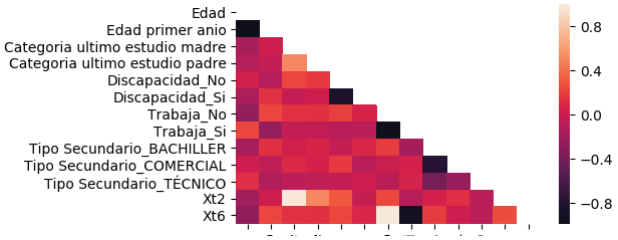

Fig. 7. Matriz de correlación

Es importante señalar que luego de la ejecución del módulo de AML, además de generarse los dos atributos antes mencionados (Xt2 y Xt6), se estandarizaron los valores de los atributos para optimizar el comportamiento de los modelos, perdiendo interpretablidad. La solución propuesta permite convertir los valores transformados a su valor original obteniéndose el resultado que se muestra en la figura 8. Aquí se ve como la edad, los estudios de la madre y el trabajo influyen en el atributo clase.

```
[Regla  5]
Regular = SI when
        Edad primer anio < 27
        Categoría último estudio madre < 1
[Regla  6]
REGULAR = SI when
        25 <= Edad primer anio < 27
        Trabaja = NO
```

Fig. 8. Reglas reinterpretadas del modelo

Finalmente, el algoritmo SHAP permite identificar los atributos con más incidencia en el modelo. En el ejemplo propuesto, se obtuvo como resultado que las características más influyentes para determinar la regularidad de un estudiante son, en orden: a) la edad al primer año de la cursada, b) si es estudiante trabaja y c) los estudios de la madre, coincidiendo con lo mostrado en las reglas antes mencionadas.

De los resultados obtenidos en la prueba de concepto no solo se verifica la funcionalidad de la propuesta, permitiendo ejecutar de manera más simple un proceso

de EDM, sino que además se consiguen mejores resultados. También se consigue mostrar que, a pesar de generar modelos complejos, puede agregársele técnicas de interpretación de modelos de caja negra para permitir una posterior elaboración de conclusiones que brinden mayor conocimiento sobre las variables relevantes para el proceso de toma de decisiones.

7. Conclusiones y Trabajo Futuro

En el presente trabajo se identificó la necesidad de contar con mecanismos de abstracción para los procesos de EDM que permitan a los usuarios no expertos en el área de minería de datos llevar a cabo un proyecto de este estilo. Como solución se presentó autoEDM, un marco de trabajo que involucra Automated Machine Learning en combinación con técnicas de interpretabilidad para modelos de caja negra.
Para validar este framework, se ejecutó con un conjunto de datos utilizado en un proyecto de EDM tradicional como prueba de concepto, comparando los resultados obtenidos en ambos proyectos. A raíz de esta prueba se verifico la correcta integración del proceso propuesto combinando herramientas de AML y de interpretación permitiendo generar patrones de conocimiento que sirvan como apoyo a la toma de decisiones en entornos educativos.
En este sentido, la presente investigación presenta evidencias iniciales sobre la posibilidad de diseñar pipelines genéricas para cualquier proyecto de minería de datos educativa que puedan ser utilizadas por usuarios no expertos.
Como futuras líneas de investigación se propone:

- Desarrollar nuevos casos de validación con otros proyectos de EDM.
- Ampliar la aplicación en distintos tipos de problemas (por ejemplo aquellos que requieren modelos de regresión).
- Evaluar otros algoritmos de interpretabilidad no analizados en esta investigación, como LIME [26].
- Desarrollar un ambiente integrado mediante una interfaz que permita la ejecución íntegra del proceso de autoEDM.

8. Referencias

[1] IEDMS (2019). International Educational Data Mining Society. www.educationaldatamining .org. Página vigente al 29/07/2019.
[2] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
[3] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM| Journal of Educational Data Mining, 1(1), 3-17.
[4] Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Tin Kam Ho, Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. 2015 International Joint Conference on Neural Networks (IJCNN), 1-8.
[5] Bose, R., & Sugumaran, V. (1999). Application of intelligent agent technology for managerial data analysis and mining. ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 30(1), 77-94.

[6] Chrysostomou, K., Chen, S. Y., & Liu, X. (2009). Investigation of users' preferences in interactive multimedia learning systems: a data mining approach. Interactive Learning Environments, 17(2), 151-163.
[7] Su, J. M., Tseng, S. S., Lin, H. Y., & Chen, C. H. (2011). A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments. User modeling and user-adapted interaction, 21(1-2), 5-49.
[8] Prince Sattam Bin Abdulaziz University, Osman Hegazi, M., & Abugroon, M. A. (2016). The State of the Art on Educational Data Mining in Higher Education. International Journal of Computer Trends and Technology, 31(1), 46-56.
[9] Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining: Survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(1).
[10] Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. Acm Sigkdd Explorations Newsletter, 13(2), 3-6.
[11] Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for Educational Data Mining: A Review. Journal of Educational and Behavioral Statistics, 42(1), 85-106.
[12] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.
[13] Avasarala, B. R., Day, J. C., & Steiner, D. (2016). System and method for automated machine-learning, zero-day malware detection U.S. Patent No. 9,292,688. Washington, DC: U.S. Patent and Trademark Office.
[14] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 847-855). ACM.
[15] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2019). Auto-sklearn: Efficient and Robust Automated Machine Learning. In Automated Machine Learning (pp. 113-134). Springer, Cham.
[16] Olson, R. S., & Moore, J. H. (2019). TPOT: A tree-based pipeline optimization tool for automating machine learning. In Automated Machine Learning (pp. 151-160). Springer, Cham.
[17] B. Kim, R. Khanna, and O. Koyejo. Examples are not enough, learn to criticize! Criticism for interpretability. Advances In Neural Information Processing Systems, pages 2280–2288, 2016b.
[18] Bianco, S., Martins, S., Rodríguez, D., & García Martínez, R. (2017). Ingeniería de explotación de información aplicada a la gestión universitaria: caso licenciatura en sistema Universidad Nacional de Lanús. In XII Congreso de Tecnología en Educación y Educación en Tecnología (TE&ET).
[19] Zutty, J., Long, D., Adams, H., Bennett, G., & Baxter, C. (2015, July). Multiple objective vector-based genetic programming using human-derived primitives. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation (pp. 1127-1134). ACM.
[20] Hall, P., & Gill, N. (2018). An Introduction to Machine Learning Interpretability-Dataiku Version. O'Reilly Media, Incorporated.
[21] Kuna, H., García Martínez, R., Villatoro, F. (2009). Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología 5: 39-44.
[22] Díaz, L., Martins, S., Garcia-Martinez, R. 2015. Descubrimiento de Patrones Socio-económicos de Población Estudiantil de Carreras de Ingeniería Basado en Tecnologías de Explotación de Información. Proceedings X Congreso de Tecnología en Educación y Educación en Tecnología. Pág. 306-315. ISBN 978-950-656-154-3.
[23] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 4765–4774).
[24] Hara, S., & Hayashi, K. (2016). Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach. arXiv:1606.09066
[25] Repositorio del framework autoEDM (2019). https://github.com/santibianco/autoEDM/ Página vigente al 29/07/2019.
[26] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.