ISSN: 1692-7257 - Volumen 1 – Número 39 - 2022

Revista Colombiana de
Tecnologías de Avanzada

Recibido: agostoe 23 de 2021
Aceptado: octubre 2 de 2022

COMPARACIÓN EMPÍRICA DE DOS MODELOS DE APRENDIZAJE AUTOMÁTICO GENERADOS MEDIANTE PROCESOS DIFERENTES

EMPIRICAL COMPARISON OF TWO MODELS OF MACHINE LEARNING GENERATED THROUGH DIFFERENT PROCESSES

MSc. Alveiro Rosado Gomez[1], PhD. Liliana Calderón Benavides[2], PhD. Jorge Andrick Parra[2]

Universidad Francisco de Paula Santander Ocaña[1]
Universidad Autónoma de Bucaramanga[2]

E-mail: aarosadog@ufpso.edu.co,{japarra, mcalderon}@unab.edu.co

Resumen: El aprendizaje automático, viene demostrando un potencial en la construcción de modelos que representan el comportamiento que existe en los datos, estos modelos son utilizados en diferentes áreas del conocimiento para optimizar las decisiones que se toman, el aprendizaje automático automatizado, es un campo de trabajo creado para satisfacer la demanda de herramientas que permitan la construcción de modelos de forma precisa, ágil y rápida y que estén disponibles para personas que no dominan la estadística y la tecnología. Esta investigación hace la comparación empírica del comportamiento de dos modelos generados utilizando herramientas de aprendizaje automático normales y automatizas para clasificar las personas desmovilizadas que pueden abandonar el proceso de reintegración. Existe coincidencia en el algoritmo que se utilizó, también en varios de los atributos que se emplearon y la evaluación sugiere que, para el conjunto de datos utilizado, el aprendizaje automático automatizado, tiene mejor rendimiento que el tradicional.

Palabras clave: Aprendizaje automático, aprendizaje automático automatizado, estudio empírico, inteligencia artificial.

Abstract: Machine learning has been showing potential in the construction of models that represent the behavior that exists in the data, these models are used in different areas of knowledge to optimize the decisions that are made, automated machine learning is a field of work created to satisfy the demand for tools that allow the construction of models in a precise, agile and fast way and that are available for people who are not fluent in statistics and technology. This research makes the empirical comparison of the behavior of two models generated using normal and automated machine learning tools to classify demobilized people who may leave the reintegration process. There is agreement in the algorithm that was used, also in several of the attributes that were used and the evaluation suggests that, for the data set used, automated machine learning has better performance than traditional.

Keywords: Machine learning, automated machine learning, empirical study, artificial intelligence.

1. INTRODUCION

La Inteligencia Artificial (Artificial Intelligence, AI), viene tomando participación en las actividades diarias de la humanidad, mejorando su calidad de vida y ayudándolo con tareas repetitivas que requieren hacer una toma decisiones, esta automatización simula el comportamiento humano mediante el aprendizaje que las maquinas hacen por medio de la experiencia y basados en reglas (Borana, 2016), (Han, Kamber, & Pei, 2012). Una de las áreas de la AI que viene tomando importancia es el aprendizaje automático (Machine Learning, ML), el cual proporciona técnicas y algoritmos que le permiten a las maquinas aprender el comportamiento que tienen los datos y generar modelos que lo representen (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019), (Nguyen, y otros, 2019), (Jiménez-Carvelo, González-Casado, Bagur-González, & Cuadros-Rodríguez, 2019).

Dentro de la generación de los modelos de ML, se busca que sean robustos y exactos, que permitan hacer la descripción o clasificación, con los niveles de error mínimos; para lógralo es necesario adelantar un proceso de ML, para seleccionar y adaptar los datos, para que funcionen con el algoritmo que se quiere aplicar y que permita encontrar los patrones ocultos en los datos (Han, Kamber, & Pei, 2012). Este proceso es cíclico y se debe adaptar continuamente los atributos y comparar los resultados de los algoritmos para seleccionar el más exacto (Telikani, Tahmassebi, Banzhaf, & Gandomi, 2021), (Hernández Royett, J et al, 2018).

Una de las evoluciones recientes del ML es el aprendizaje automático automatizado (AutoML), el cual, reduce el grado de conocimiento y embebe la forma en que se desarrolla el proceso de ML, siendo trasparente para el usuario las tareas internas que este tipo de solución realiza. Estas herramientas tienen como objetivo reducir el grado de conocimiento técnico y estadístico que debe tener un experto del dominio para construir modelos de ML (He, Zhao, & Chu, 2020). El AutoML, proporciona un conjunto de tareas automáticas que permiten realizar el procesamiento de datos, la ingeniería de características, la optimización de hiper-parámetros y la optimización del modelo sin necesidad de la intervención humana (Lakshmanan, Robinson, & Munn, 2020).

Esta investigación hace una comparación empírica del comportamiento de dos modelos; el primero realiza un procesamiento tradicional de ML para tener el modelo con la mayor exactitud para los datos utilizados, el segundo paso fue generar el modelo con las herramientas de AutoML llamada *Auto ViML*, luego de tener el modelo con esta herramienta se realizó la comparación entre los modelos generados de las dos formas, en cuanto a métricas en la matriz de confusión y el número de características seleccionadas.

2. METODOLOGÍA

El objetivo principal de la generación de modelos, fue la de realizar la clasificación correcta de los individuos desmovilizados que pueden abandonar el proceso de reintegración, identificándose como atributo de clase la situación final frente al proceso. Los pasos que se siguieron en el desarrollo de esta investigación están relacionados con los tradicionales en un proceso de ML (Telikani, Tahmassebi, Banzhaf, & Gandomi, 2021), (Suresh, & Guttag, 2019) dividiéndolo en la validación, preprocesamiento de los datos, posteriormente signe las tareas de construcción de entrenamiento, construcción y evaluación del modelo, y por último el despliegue (Lakshmanan, Robinson, & Munn, 2020), (Djukanovic, Berger, Raidl, & Blum, 2019), en esta investigación la etapa de despliegue no se aplicó. Para el trabajo con *Auto ViML*, solo es necesario establecer cuál es el conjunto de datos que se utilizara, identificando el nombre del atributo que corresponde a la clase y la especificación de unos hiper-parámetros, relacionados con el conjunto de datos y la distribución de las clases.

3. RESULTADOS

El conjunto de datos utilizado corresponde a las estadísticas de las personas desmovilizadas de los grupos armados de Colombia que han ingresado al proceso de reintegración; cada uno de los registros corresponde a los atributos personales, familiares y la situación frente al proceso de desmovilización, para un total de 55517 registros 34 atributos, que se encuentran disponibles en la plataforma de datos abiertos de Colombia.

3.1 Preprocesamiento de datos e Ingeniería de características

Inicialmente se realizó un análisis exploratorio de la información en donde se encontraron valores anómalos en algunos de los atributos, como por ejemplo en el atributo sexo se tenían valores de masculino y femenino, en mayúscula y minúscula. Otros atributos tenían el valor de menos uno (-1), por lo tanto, fue necesario hacer imputación según su vecino más próximo. Para los valores categóricos fue aplicada la utilidad de

Python label encoder (Lakshmanan, Robinson, & Munn, 2020), que permite hacer la discretización automática de estos valores. Como clase para el conjunto de datos se tomó el atributo situación final frente al proceso, los valores de este atributo se distribuyeron entre Culminado y En Proceso, como las personas que no abandonan el proceso (identificado con uno y 31130 instancias) y Ausente del proceso y Fuera del Proceso, son las personas que abandonan en el proceso (identificado con cero y 21728 instancias). Existe otro valor asignado con No ha ingresado, para identificar a los desmovilizados que no se acogieron al proceso de reintegración, por lo cual fueron eliminados del conjunto de datos reduciendo el número de registros a 52858.

Luego de realizar la discretización se procedió a determinar la correlación que existía, encontrado 15 atributos con correlación mayor del 70%, por lo tanto, fueron descartados del conjunto de datos, reduciendo el número de características a 19 (Fu, Yan, & Huang, 2008). Para seleccionar los atributos más significativos del conjunto de datos, se utilizó el análisis de componentes principales, seleccionando 16 atributos (Aburomman & Bin Ibne Reaz, 2016).

3.2 Selección de modelo y optimización de hiper-parámetros

Para la selección del modelo con mayor exactitud, se dividió el conjunto de datos en dos subgrupos; uno para entrenamiento correspondiente al 80% y otro para pruebas con el 20%, con clases estratificadas en los dos conjuntos. Como lo muestra la figura 1, se utilizaron algoritmos de clasificación como máquinas de vectores de soporte, métodos de ensamble, regresión lineal, clasificador bayesiano, vecino más próximo, análisis discriminante, árboles y arboles de decisión impulsados por gradiente. También, la imagen muestra que la mayor exactitud la tuvieron los algoritmos *AdaBoostClassifier* con 96.8% y *XGBClassifier* 96.9%.

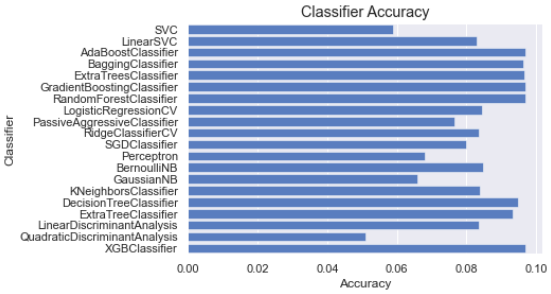

Fig. 1. Selección del algoritmo

Se selecciono el algoritmo *XGBClassifier*, para desarrollar el modelo, posteriormente se realizó la optimización de hiper-parámetros quedando los mejores parámetros *eval_metric* igual a error, *gamma* con 0.1 y *subsample* con 1.0, para un mejor resultado de clasificación del 97%, este valor permaneció luego de aplicar al modelo el conjunto de datos de prueba generando los valores de exactitud 97%, precisión 96%, sensibilidad 99% y *F1 Score* de 97%.

3.3 Trabajando con Auto ViML

Para hacer la automatización del ML, se tomó el conjunto de datos original y solo fue necesario etiquetar el atributo de clase 'Situación Final frente al proceso' y lo relacionado con la duplicidad de género que tenía el conjunto de datos. Luego de esos cambios el conjunto de datos fue procesado con la herramienta *Auto ViML*, inicialmente para los valores que se utilizaron, hace la discretización mediante procesamiento de lenguaje natural, reemplazando los valores originales por características de los textos que las contienen. Para la ingeniería de características encontró 12 variables altamente correlacionadas, y selecciono 15 atributos como las características más importantes. El algoritmo que fue seleccionado de forma automática fue *XGBClassifier*, y la configuración de los hiper-parámetros fue *eval_metric* igual a error, *gamma* con 1 y *subsample* con 0.7, como se muestra en la tabla 1, los resultados para el modelo generado con AutoML, tienen mejor rendimiento que el modelo generado de forma tradicional, esto puede deberse a como lo expresan Waringa, Charlotta y Renato (2020), al hecho que en AutoML se hace la transformación de características lo que mejora el aprendizaje del algoritmo. Particularmente *Auto ViML*, se apoya en el procesamiento del lenguaje natural, para los atributos categóricos de tipo cadena de caracteres, utilizando el tamaño de la cadena como valor del atributo (He, Zhao, & Chu, 2020), (Varshney, & Alemzadeh, 2017).

Tabla 1: Rendimiento del modelo

	Tradicional	**AutoML**
Accuracy	96.9%	98.8%
Precision	96.0%	98.0%
Recall	98.8%	99.9%
F1	97.4%	99.0%

En la tabla 2, se muestran los atributos que fueron seleccionados como los mejores predictores con las dos formas utilizadas para generar el modelo de clasificación; aunque en los dos métodos se utilizaron conceptos diferentes para determinar cuáles eran los mejores atributos, existe una coincidencia en cinco de los quince atributos seleccionados por cada técnica.

Tabla 2: Atributos comunes

Atributo	Descripción
BeneficioTRV	Recibió beneficio por Actividades Transversales en el año anterior
BeneficioFA	Recibió beneficio por Formación Académica en el año anterior
Nivel Educativo	Máximo ciclo aprobado en Formación Académica
Máximo Nivel FpT Reportado	Máximo nivel de formación para el trabajo
OcupacionEconomica	Ocupación económica para la fecha de corte
DesagregadoDesembolsoBIE	Situación en el Beneficio de Inserción Económica
Régimen de salud	Establece el régimen de salud
Clasificación Componente Específico	Posee Ruta Condicional

4. DISCUSION

Los resultados muestran algunas coincidencias en cuanto al algoritmo que fue seleccionado para generar el modelo, también existen atributos que fueron seleccionados por los dos procesos como los que mejor representan el conjunto de datos. Como se mostró en la tabla 1, el rendimiento para el caso mostrado fue el de AutoML, esto se produce porque existen procesamientos internos de las características que optimizan sus valores para mejorar la exactitud del modelo. Esto propone una nueva visión de trabajo que consiste en utilizar el AutoML, como una herramienta de extracción de características y de búsqueda del algoritmo que mejor exactitud tenga con los datos suministrados y preprocesados (Xin, Wu, Lee, Salehi, & Parameswaran, 2021).

5. CONCLUSIONES

La forma tradicional de construcción de modelos de ML, demanda un conocimiento técnico y especifico en el manejo de datos, lo cual supone que personas con poca experiencia o un nivel de dominio limitado en herramientas tecnológicas lo pueda ver complicado. Dado que las herramientas que permiten la construcción de los modelos, generalmente necesitan de una configuración en el más sencillo de los casos o deben dominar un lenguaje de programación que les permita generar las instrucciones necesarias para producir el modelo. Este nuevo enfoque de trabajo con ML, sugiere un incremento de las personas que pueden hacer ciencia de datos.

AutoML, propone una nueva forma de hacer ML, en donde el conocimiento que se debe tener se basa en la forma en que el algoritmo recibe los datos; el proceso posterior de esta especificación es completamente automático, la herramienta se encarga desde preprocesamiento de los datos, hasta la selección del algoritmo que generen el mejor modelo, embebiendo el proceso tradicional que debe seguir el desarrollador de solución de ML, esto permite que personas sin experiencia en estadística o programación puedan construir sus propios modelos sin necesidad de seguir el flujo de trabajo tradicional.

REFERENCIAS

Aburomman, A., & Bin Ibne Reaz, M. (2016). Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection. *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (págs. 636-640). Xi'an: IEEE.

Borana, J. (2016). Applications of Artificial Intelligence & Associated Technologies. *International Conference on Emerging Technologies in Engineering, Biomedical, Management and Science* (págs. 64-67). Jodhpur: SD-Technocrates.

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1-8.

Djukanovic, M., Berger, C., Raidl, G., & Blum, C. (2020). An A∗ search algorithm for the constrained longest common subsequence problem. *Information Processing Letters*, 1-12.

Fu, Y., Yan, S., & Huang, T. (2008). Correlation Metric for Generalized Feature Extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2229 - 2235.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Tercera ed.). Waltham: Morgan Kaufmann.

He, X., Zhao, K., & Chu, X. (2020). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 1-36.

Hernández Royett, J., Hernández, Y. F., Gil, M. de los A., & Cárdenas Barboza, E. (2018). Evaluación del modelo integrado de planeación y gestión (MIPG) en las entidades territoriales del estado colombiano. Aglala, 9(1), 444-463. http://revistas.curnvirtual.edu.co/index.php/aglala/article/view/1255

Jiménez-Carvelo, A., González-Casado, A., Bagur-González, G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, 25–39.

Lakshmanan, V., Robinson, S., & Munn, M. (2020). *Machine Learning Design Patterns.* Sebastopol: O'Reilly Media.

Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, 22071–22080.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López, Á., Heredia, I., . . . Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, 77–124.

Suresh, H., & Guttag, J. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arxiv*, 1-10.

Telikani, A., Tahmassebi, A., Banzhaf, W., & Gandomi, A. (2021). Evolutionary Machine Learning: A Survey. *ACM Computing Surveys*, 1-35.

Varshney, K., & Alemzadeh, H. (2017). On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. *Big Data*, 246-255.

Waringa, J., Lindvallc, C., & Umetona, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence In Medicine*, 1-12.

Xin, D., Wu, E., Lee, D., Salehi, N., & Parameswaran, A. (2021). Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. *CHI Conference on Human Factors in Computing Systems* (págs. 8-13). Yokohama: ACM.