08

Fecha de presentación: octubre, 2018
Fecha de aceptación: diciembre, 2018
Fecha de publicación: febrero, 2019

CAUSAS QUE AFECTAN LA PROMOCIÓN DE ESTUDIANTES

QUE CURSAN NIVELACIÓN EN LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO: UN ESTUDIO APLICANDO MINERÍA DE DATOS

CAUSES AFFECTING THE PROMOTION OF STUDENTS OF PRE-UNIVERSITY COURSES AT THE STATE TECHNICAL UNIVERSITY OF QUEVEDO: A STUDY APPLYING DATA MINING

Harold Elbert Escobar Terán[1]
E-mail: hescobar@uteq.edu.ec
ORCID: https://orcid.org/0000-0001-9165-6627
Amilkar Puris[1]
Pavel Novoa-Hernández[1]
E-mail: pnovoa@uteq.edu.ec
ORCID: https://orcid.org/0000-0003-3267-6753
[1]Universidad Técnica Estatal de Quevedo. Los Ríos. Ecuador.

Cita sugerida (APA, sexta edición)

Escobar Terán, H. E., Puris, A., & Novoa Hernández, P. (2019). Causas que afectan la promoción de estudiantes que cursan nivelación en la Universidad Técnica Estatal de Quevedo: un estudio aplicando minería de datos. *Universidad y Sociedad*, 11(2), 61-65. Recuperado de http://rus.ucf.edu.cu/index.php/rus

RESUMEN

La presente investigación tiene por objetivo analizar las causas de la promoción en estudiantes que cursan nivelación en la Universidad Técnica Estatal de Quevedo UTEQ aplicando técnicas de minería de datos. Se pretende modelar este problema como uno clasificación, que prediga con antelación si un estudiante necesita ser atendido, y lograr así que su desempeño mejore. El presente trabajo estudia por tanto, modelos de minería de datos que expliquen qué factores socio-económicos, médicos y psicológicos afectan la promoción de los estudiantes que cursan la nivelación en la Universidad Técnica Estatal de Quevedo (UTEQ). Para lograrlo, se siguió una metodología conocida en el ámbito del a minería de datos: CRISP-DM. Con ayuda del software WEKA y algoritmos de clasificación basados en reglas, se obtuvieron dos modelos que permiten predecir si el estudiante aprobará o no el curso de nivelación (basado en un examen final).

Palabras clave: Educación, Minería de datos, Reglas de asociación, Promoción, Análisis de datos, Ciencia de Datos.

ABSTRACT

The objective of this research is to analyze the causes of the promotion in students who are studying levelling in the State Technical University of Quevedo UTEQ applying data mining techniques. It is intended to model this problem as a classification, which predicts in advance if a student needs to be attended, and thus achieve better performance. The present work therefore studies data mining models that explain what socioeconomic, medical and psychological factors that affect the promotion of the students who attend levelling at the State Technical University of Quevedo (UTEQ). To achieve this, a well-known methodology was followed in the field of data mining: CRISP-DM. With the help of WEKA software and rule-based classification algorithms, two models were obtained that allow us to predict whether or not the student will pass the levelling course (based on a final exam).

Keywords: Education, Data Mining, Association Rules, Promotion, Data Analysis, Data Science.

INTRODUCCIÓN

La Universidad Técnica Estatal de Quevedo (UTEQ) contribuye al desarrollo y crecimiento de las habilidades de las personas, posibilitando en ellas el desarrollo de las competencias necesarias para que asuman con mejor preparación los desafíos de la sociedad actual. Además de reconocer que el único camino posible para aumentar la competencia laboral capacitada es la educación.

Uno de los problemas a enfrentar en la UTEQ es la nota con la que los estudiantes aprueban el curso de nivelación. La nota final del examen con el que aprueban los estudiantes afecta tanto en los ámbitos personales como en los institucionales, sociales y económicos. Muchas veces, para los estudiantes implica una condición de fracaso que afecta emocionalmente por la discrepancia con las aspiraciones personales. En lo institucional, implica una disminución del rendimiento académico de la universidad. En lo social, la calidad estudiantil contribuye a generar inequidad y desequilibrios sociales, además de entorpecer los objetivos y responsabilidades que la sociedad le ha entregado a la educación superior. En lo económico, se debería considerar el costo que esto implica para el sistema de educación nacional al invertir dinero en brindar una educación apropiada a la comunidad.

El rendimiento académico de los estudiantes se basa en diversos factores como los factores personales, socioeconómicos, psicológicos, médicos entre otros. Para un ser humano resulta difícil encontrar patrones que expliquen las causas del rendimiento académico en un estudiante. Son tantos los datos y variables asociadas que solo con ayuda de técnicas estadísticas, y más concretamente de la minería de datos (Maimon & Rokach, 2010), es posible realizar dicha tarea.

La minería de datos, es una colección de métodos de las estadísticas, las ciencias de la computación, la ingeniería y la inteligencia artificial para identificar patrones de comportamiento. En la minería de datos se hace énfasis en la identificación de patrones en los grandes volúmenes de datos, por lo que se considera esta técnica con el objeto de determinar patrones de comportamiento de las variables que tienen un impacto en la calidad estudiantil universitaria.

Una de las actividades más importantes en minería de datos es la clasificación. Esta se encuentra estrechamente relacionada con la minería de datos predictiva, ya que realiza la predicción a partir resultados conocidos que se encuentran en diferentes tipos datos.

La minería de datos, y más concretamente el aprendizaje automatizado ha impactado otros escenarios de la sociedad moderna actual. Ejemplos de ello son las investigaciones desarrolladas por Ranginkaman, Kazemi Kordestani, Rezvanian & Meybodi (2014); y Sikora, Krzystanek, Bojko & Spiechowicz (2011).

En la primera, se realiza un diagnóstico del drenaje en zonas costeras de Irán mediante el uso de métodos geoestadísticos, máquinas de soporte vectorial (SVM) (Campbell & Ying, 2011) en conjunto con un sistema de inferencia adaptativa neuro-difusa (ANFIS) (Abraham, 2005). En la segunda investigación, se aplica modelos de aprendizaje automatizado para describir y estimar en línea, el riesgo de contaminación por metano en minas. En el ámbito médico y biofarmacéutico, sobresale la investigación de Chazard, Preda, Merlin, Ficheur & Beuscart (2009), en el que se emplean árboles de decisión, para identificar situaciones que puedan derivar en riesgo de reacciones adversas de medicamentos. En la obtención del modelo se tuvieron en cuenta 10500 registros de pacientes de Dinamarca y Francia, para generar 500 reglas.

En Hathout & Metwally (2016)analysing data and finally extracting correlations and meaningful outcomes. In this context, binding energies could be used to model and predict the mass of loaded drugs in solid lipid nanoparticles after molecular docking of literature-gathered drugs using MOE® software package on molecularly simulated tripalmitin matrices using GROMACS®. Consequently, Gaussian processes as a supervised machine learning artificial intelligence technique were used to correlate the drugs' descriptors (e.g. M.W., xLogP, TPSA and fragment complexity, los autores emplean un algoritmo de aprendizaje automatizado para modelizar la carga de fármacos en nano-partículas de lípidos sólidos. Concretamente, aplicaron un proceso Gaussiano (Seeger, 2004) para correlacionar descriptores de los medicamentos. La investigación desarrollada por Revuelta-Zamorano, et al. (2016), permitió obtener un modelo predictivo para identificar las infecciones asociadas a la asistencia sanitaria (HAIS) en una unidad de cuidados intensivos (UCI). Variables como la edad, la duración de la estadía, la cama donde se alojó el paciente y el mes de ingreso constituyeron los factores de riesgo más relevantes para predecir HAIS en la UCI que fue objeto de estudio.

Dado que el nivel de promoción de los estudiantes que cursan nivelación en la UTEQ es aún deficiente, resulta necesario estudiar las causas de esta situación. Es por eso que la presente investigación tiene por objetivo: analizar las causas de la promoción en estudiantes que cursan nivelación en la UTEQ aplicando técnicas de minería de datos.

Se pretende modelar este problema como uno clasificación, que prediga con antelación si un estudiante necesita ser atendido, y lograr así que su desempeño mejore.

DESARROLLO

Para la consecución del objetivo trazado en la presente investigación, se siguió la metodología CRISP-DM1 (Wirth, 2000). Esta metodología involucra los pasos que se describen a continuación. Se ha hecho énfasis en los resultados obtenidos durante la presente investigación.

1. **Análisis del problema:** En esta etapa se procedió con entrevistas a diferentes expertos en educación y psicología. Se delinearon los objetivos preliminares y definieron un total de 72 variables, agrupadas en tres grandes categorías: socio-económicas, psicológicas y médicas. Asimismo, se logró firmar un compromiso de confidencialidad de los datos para que los estudiantes pudieran aportar información con la seguridad de anonimato. La principal conclusión en esta etapa fue el reconocimiento institucional sobre la necesidad de realizar un estudio sobre los factores que inciden en la promoción de los estudiantes, esto es, empleando la minería de datos (Bramer, 2013). Específicamente, este estudio se enfocó como una tarea de clasificación.
2. **Comprensión de los datos.** En esta etapa, se analizaron preliminarmente los datos provenientes de todos los estudiantes que asistieron a la modalidad de nivelación en el período lectivo curso 2015-2016. Sin embargo, no todas las variables pudieron ser medidas con éxito a los 3040 estudiantes participantes. De manera que se tuvo que realizar una selección de aquellas variables con menos del 50% de las mediciones (casos). En consecuencia, se obtuvo una versión preliminar del conjunto de datos con: 32 variables socio-económicas, 16 de tipo psicológicas, y 12 de tipo médicas.

Tabla 1. Reglas más importantes generadas con el modelo JRip de WEKA.

Antecedente	Consecuente	Número de casos cubiertos
Horas_de_sueño=0	1	108
Horas_de_sueño = 1 and actividades_saludables=0	1	44
Viaja_de_regreso_a_casa=4	1	15
Actividades_saludables=1 and constante=4	1	22
Fatiga_o_agotamiento=2 and paga_alquiler = 0	1	7
Horas_de_sueño=1 and edad_del_hermano=4 and bono_solidario=1	1	9
{ }	0	265

3. **Preparación de los datos.** El objetivo principal de esta etapa es pulir el conjunto de datos obtenido en la fase anterior de manera que sirva de entrada a la etapa 4 de Modelado. Aquí, se aplican varias actividades de transformación que incluyen: limpieza, balanceo de casos, entre otros. En el caso particular de esta investigación, se realizó una discretización (Liu, Hussain, Tan & Dash, 2002; Ramírez Gallego, et al., 2015) de las variables debido a que muchas de las mismas fueron de tipo respuesta abierta. Aquí es importante mencionar que, dado que nuestro interés consistió en analizar las causas de la promoción de los estudiantes, la variable clase (denominada EXAMEN) fue discretizada de forma binaria (0 si no aprueba, y 1 si aprueba).

Tabla 2. Reglas más importantes generadas con el modelo PART de WEKA.

Antecedente	Consecuente	Número de casos cubiertos
Horas_de_sueño=2 and viaja_de_regreso_a_casa=3 and edad_del_hermano=4	0	11
Horas_de_sueño=2 and edad_del_hermano=5	0	10
Horas_de_sueño=2 and edad_del_hermano=2	0	8
Horas_de_sueño=0 and tiempo_de_siestas=0 and pesadillas=2	1	29
Incontinencia=1 and horas_de_sueño=0	1	11

Posteriormente, debido a la gran diferencia de casos aprobados (2805) y desaprobados (235), se realizó un proceso de balanceado de datos con una técnica de submuestreo (undersampling) (López, Fernández,

[1] Siglas en inglés para Cross Industry Standard Process for Data Mining

García, Palade & Herrera, 2013). De manera que de los 3040 estudiantes (casos) con los que se comenzó el estudio, quedaron 470. Específicamente, 235 casos con EXAMEN = 0, y otros 235 casos con EXAMEN = 1.

Otra tarea importante, y en ocasiones necesaria para reducir la complejidad del análisis y contribuir a la precisión del estudio, es la selección de atributos (variables). Es de notar que el conjunto de datos antes de proceder con esta actividad, constaba de 60 variables. El objetivo aquí es identificar aquellas variables más relevantes (que más influyen y se relacionan con la variable clase EXAMEN) y reducir el número de variables que intervendrán en la generación del modelo. Particularmente, se empleó la herramienta WEKA (Witten & Frank, 2000) con el método de búsqueda GreedyStepwise y el evaluador de atributos CfsSubsetEval. Como resultado quedaron solo 15 variables:

1. NACIONALIDAD
2. TELEFONOFIJO
3. GENERO
4. VIAJADEREGRESOACASA
5. PAGAALQUILER
6. BONOSOLIDARIO
7. EDADDELHERMANO
8. CONSTANTE
9. FATIGAOAGOTAMIENTO
10. PESADILLAS
11. INCONTINENCIA
12. DISCAPACIDAD
13. ACTIVIDADESSALUDABLES
14. HORASDESUEÑO
15. TIEMPODESIESTAS

Lo anterior dejó al conjunto de datos listo para el proceso de obtención del modelo. Nótese que ahora el conjunto de datos, viéndolo de forma tabular, cuenta con 15 columnas (variables) y 470 filas (casos).

4. **Modelado.** El objetivo en esta etapa es obtener un modelo que permita realizar una de las tareas de la minería de datos. En este caso nos hemos centrado en la clasificación, y más concretamente en un modelo de clasificación basado en reglas. En nuestra opinión, este tipo de modelo permitirá explicar las causas de los resultados en la promoción empleando un lenguaje fácil de comprender por parte de los directivos de la UTEQ. Empleando la nuevamente la herramienta Weka y los algoritmos JRip y PART, se obtuvieron sendos modelos. En las Tablas 1 y 2 se ilustran las reglas más importantes generadas en ambos modelos predictivos. Como información adicional, en las tablas se incluye el número de casos cubiertos por cada regla. En particular, en la Tabla 1, la regla denotada por el símbolo {} corresponde al complemento de las otras reglas que la preceden. En otras palabras, si un estudiante no cumple con las reglas anteriores a {}, entonces se puede clasificar con consecuente EXAMEN = 0. En el caso del modelo JRip (Tabla 1) todas las reglas cuentan con una confianza del 85 %, mientras que las generadas por PART poseen un 90%
5. **Evaluación.** Una vez obtenido los modelos y resultados asociados, se procedió a evaluarlos. Para ello, se confrontaron las reglas con los expertos que participaron en la etapa 1. Con el empleo de entrevistas, se pudo constatar que los resultados (reglas) generadas por los modelos de minería de datos empleados, explican la promoción de los estudiantes de nivelación de manera satisfactoria. Sin embargo, resulta sorprendente que la variable más importante en el estudio fuera la relacionada con el tiempo de sueño.
6. **Explotación.** Los resultados obtenidos con la investigación fueron divulgados a las autoridades de la UTEQ para asistir futuros procesos de toma de decisiones. El objetivo es reaccionar con anticipación ante aquellos estudiantes propensos a suspender los cursos de nivelación, esto es, a través de un monitoreo frecuente de las variables más influyentes.

CONCLUSIONES

Los resultados del proceso de minería de datos llevado a cabo permiten concluir que: si las horas de sueño son menores de 6 horas diarias (valor nominal 0) el estudiante aprueba el examen final del curso (promueve). El 23% de los casos (108) cumplen con este patrón. Se puede observar que esta condición de las horas de sueño se presenta en la mayoría de las reglas de clasificación generadas por los modelos generados a partir de los algoritmos JRIP y PART. Esto se puede comprobar en las Tablas 1 y 2. Por ejemplo, vea que según el modelo de PART, si el estudiante duerme más de 8 diarias (horas_de_sueño = 2) resultará en que el estudiante desaprueba el curso.

Aunque estos resultados permiten aproximarse a esta problemática desde una arista cuantitativa, se requieren de estrategias educativas adicionales para lograr que el índice de promoción mejore en la UTEQ. Adicionalmente, existen otros tipos de análisis de la minería de datos que pueden ayudar a comprender mejor las causas de este problema aún presente en la UTEQ. Las investigaciones futuras estarán orientadas a estas y otras cuestiones relacionadas.

REFERENCIAS BIBLIOGRÁFICAS

Abraham, A. (2005). *Adaptation of Fuzzy Inference System Using Neural Learning BT - Fuzzy Systems Engineering: Theory and Practice*. En, N. Nedjah & L., de Macedo Mourelle (Eds.). (pp. 53–83). Berlin: Springer.

Bramer, M. (2013). *Introduction to Data Mining*. En, Principles of Data Mining. (pp. 1–8). London: Springer London.

Campbell, C., & Ying, Y. (2011). *Learning with Support Vector Machines*. Synthesis Lectures on Artificial Intelligence and Machine Learning, 5(1). Recuperado de https://www.morganclaypool.com/doi/abs/10.2200/S00324ED1V01Y201102AIM010

Chazard, E., Preda, C., Merlin, B., Ficheur, G., & Beuscart, R. (2009). *Data-Mining-based Detection of Adverse Drug Events*. Studies in Health Technology and Informatics, 150, 552-556. Recuperado de https://www.ncbi.nlm.nih.gov/pubmed/19745372

Hathout, R. M., & Metwally, A. A. (2016). *Towards better modelling of drug-loading in solid lipid nanoparticles: Molecular dynamics, docking experiments and Gaussian Processes machine learning*. European Journal of Pharmaceutics and Biopharmaceutics, 108, 262–268. Recuperado de https://www.ncbi.nlm.nih.gov/pubmed/27449631

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). *Discretization: An Enabling Technique*. Data Mining and Knowledge Discovery, 6(4), 393–423. Recuperado de https://link.springer.com/article/10.1023/A:1016304305535

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*. Information Sciences, 250, 113–141. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.1919&rep=rep1&type=pdf

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Berlín: Springer.

Ramírez Gallego, S, et al. (2015). *Data discretization: taxonomy and big data challenge*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(1), 5–21. Recuperado de https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1173

Ranginkaman, A. E., Kazemi Kordestani, J., Rezvanian, A., & Meybodi, M. R. (2014). *A note on the paper "A multi-population harmony search algorithm with external archive for dynamic optimization problems" by Turky and Abdullah*. Information Sciences, 288, 12–14. Recuperado de http://www.academia.edu/28852214/A_note_on_the_paper_A_multi-population_harmony_search_algorithm_with_external_archive_for_dynamic_optimization_problems_by_Turky_and_Abdullah

Revuelta-Zamorano, P., Sánchez, A., Rojo-Álvarez, J. L., Álvarez-Rodríguez, J., Ramos-López, J., & Soguero-Ruiz, C. (2016). *Prediction of Healthcare Associated Infections in an Intensive Care Unit Using Machine Learning and Big Data Tools BT - XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. En, E. Kyriacou, S. Christofides, & C. S. Pattichis (Eds.), (pp. 840–845). Cham: Springer International Publishing.

Seeger, M. (2004). *Gaussian Processes for Machine Learning*. International Journal of Neural Systems, 14(02), 69–106. Recuperado de https://www.worldscientific.com/doi/abs/10.1142/S0129065704001899

Sikora, M., Krzystanek, Z., Bojko, B., & Spiechowicz, K. (2011). *Application of a hybrid method of machine learning for description and on-line estimation of methane hazard in mine workings*. Journal of Mining Science, 47(4), 493–505. Recuperado de https://link.springer.com/article/10.1134/S1062739147040125

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*. Recuperado de https://www.semanticscholar.org/paper/CRISP-DM-%3A-Towards-a-Standard-Process-Model-for-Wirth/48b9293cfd4297f855867ca278f7069abc6a9c24

Witten, I. H., & Frank, E. (2000). Data Mining: *Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann.