

Modelo para predecir el rendimiento académico basado en redes neuronales y analítica de aprendizaje

Nelson Salgado Reyes¹, Jéfferson Beltrán Morales², Javier Guaña Moya³, Charles Escobar Teran⁴, Damian Nicolalde Rodriguez⁵, Gustavo Chafla Altamirano⁶

nesalgado@puce.edu.ec, jtbeltran@uce.edu.ec, eguana953@puce.edu.ec, cescobar637@puce.edu.ec, danicolalde@puce.edu.ec, gxchafla@puce.edu.ec

^{1,3,4,5,6} Docente Facultad de Ingeniería, Av. 12 de octubre 1076 y Roca, 17012184, Quito, Ecuador.

² Docente Facultad de Ingeniería Ciencias Físicas y Matemáticas, Av. Universitaria y Bolivia, 170129, Quito, Ecuador.

Pages: 258–266

Resumen: Un elemento de gran importancia para las instituciones educativas universitarias, educadores y estudiantes lo constituye el rendimiento académico de los mismos en el tránsito de su formación profesional. La minería de datos educativos desarrolla modelos y métodos para explorar los datos recopilados de los entornos de aprendizajes educativos mediante analíticas de aprendizajes con el fin de detectar patrones que permitan predecir variables de interés. La presente investigación describe un modelo predicativo de rendimiento académico usando técnicas de redes neuronales sobre un conjunto de datos reales de 300 estudiantes de la carrera de Sistemas de la Universidad Central del Ecuador. Este registro fue suministrado por el entorno virtual de aprendizaje <https://uvirtual.uce.edu.ec/> desarrollado en Moodle y usado en dicha Universidad.

Palabras-clave: rendimiento académico; redes neuronales, learning analytics; big data

Model to predict academic performance based on neural networks and learning analytics

Abstract: An element of great importance for university educational institutions, educators and students is the academic performance of them in the transition of their professional training. The mining of educational data develops models and methods to explore the data collected from the educational learning environments through learning analytics in order to detect patterns that allow predicting variables of interest. The present research describes a predictive model of academic performance using neural network techniques on a set of real data of 300 students of the Systems career of the Central University of Ecuador. This registration was provided by the virtual learning environment <https://uvirtual.uce.edu.ec/> developed in Moodle and used in said University.

Keywords: academic performance; neural networks, learning analytics; big data.

1. Introducción

Durante el tránsito de los estudiantes en la adquisición de competencias que les permita un buen desarrollo futuro de su profesión, afrontan el desafío constante de superar materias académicas. Según la teoría de aprendizaje, la probabilidad de éxito del durante sus estudios es un problema multifactorial, siendo la interacción aprendizaje-enseñanza un elemento transcendental. (Acevedo & Rocha, 2011)

A pesar de que el rendimiento académico es una variable multifactorial, muchos de los estudios alrededor de la misma incluyen solo factores personales y socioeconómicos; sin embargo, el surgimiento y la aplicación de las nuevas tecnologías de enseñanza sobre todo el uso de las plataformas virtuales, permiten a las universidades recolectar una gran cantidad de información en tiempo real. Estos cuantiosos datos electrónicos generados proporcionan un abordaje multivariante en el estudio del rendimiento académico. (Ye & Biswas, 2014)

La analítica de aprendizaje se puede definir como el proceso de determinar, evaluar e interpretar el significado de grandes volúmenes de datos educacionales; utilizando para ello algoritmos matemáticos. (Baker, 2014; Jaramillo & Paz, 2015) Aunque existen diversas técnicas para el análisis de este tipo de información, existen cada vez más trabajos enfocados al uso de algoritmos basados en redes neuronales artificiales. (Isljamovic & Suknovic, 2014)

Las redes neuronales artificiales (RNA) representan esquemas simplificados de la estructura física y del funcionamiento de una neurona biológica; por lo que sus algoritmos poseen capacidad de predecir resultados mediante la clasificación de patrones.

En los últimos años se han desarrollado múltiples tipologías de redes neuronales en función del problema a resolver. Una de las arquitecturas de mayor uso en problemas de clasificación y predicción, responde a redes perceptrón con múltiples capas a las que se asocian una cantidad de neuronas. De esta forma, el aprendizaje se obtiene mediante “retro propagación del error”, que se basa en obtener el mínimo de una función de error calculada a través de la salida del sistema y el modelo obtenido por el entrenamiento de la red. (Del Brío, Sanz, & Garcia, 2006)

El uso de analítica de aprendizaje y de estos modelos novedosos, aplicados a la educación no son todavía suficientemente utilizados en Ecuador. A pesar de que muchas Universidades en el país utilizan plataformas con entornos virtuales en los procesos de formación, los datos obtenidos no son aprovechados correctamente para realizar predicciones que permitan tomar decisiones oportunas. De aquí que el objetivo del presente trabajo se centra en elaborar un modelo de predictivo del rendimiento académico usando técnicas de redes neuronales sobre la información obtenida mediante analíticas de aprendizajes.

Dado la problemática, los resultados predictivos del modelo sirven a la Universidad para evitar casos de deserción de estudiantes, porque en general, la eficiencia y efectividad del sistema educativo en formar profesionales está actualmente poco trabajada por las universidades, y organizaciones gubernamentales competentes, lo que genera efectos negativos importantes para el país, sobre todo si se considera que el gobierno proyecta una fuerte inyección de inversión por la reforma educacional y la intención de utilizar este indicador como parte de las acreditaciones de las instituciones educativas a nivel superior.

Las variables motivacionales contribuyen a mejorar el rendimiento académico, pero no son las de más peso en la Universidad Central del Ecuador. Se destaca la motivación docente ligada a la heterogeneidad por facultad, la incidencia de la presión familiar, y la necesidad de ambientación al sistema previo al ingreso.

2. Revisión de la literatura

La mayoría de los estudios sobre “Learning Analytics” se centran en predecir el rendimiento académico del alumno cuantificando el grado final o por la aprobación o no de un determinado curso, dificultando la comparación de las diferentes investigaciones respecto a los predictores más estables de rendimiento académico. (Altujjar, Altamimi, Al-Turaik, & Al-Razgan, 2016; Il-Hyun, Yeonjeong, Jeonghyun, & Jongwoo, 2014)

A continuación, se detallan las que se han considerado más importantes:

Romero, et al (2014) recolectaron datos de estudiantes de un foro de discusión en línea, los clasificaron en varios subconjuntos de datos y luego evaluaron la precisión predictiva de cada conjunto a través de métodos de minería de datos. Para ello utilizaron un algoritmo de clasificación y datos de interacción de los alumnos antes de un examen para predecir el rendimiento en el aprendizaje de los mismos.

Huang&Fang (2013) aplicaron cuatro modelos matemáticos para predecir el rendimiento académico de estudiantes de un curso de ingeniería utilizando para ello las calificaciones finales de los mismos. Los resultados finales mostraron que los puntajes de los exámenes finales de los estudiantes eran predecibles con un 88% de precisión en base a ocho variables recopiladas de un sistema de gestión de aprendizaje (LMS).

Usman&Adenub (2013), desarrollaron un modelo basado en redes neuronales artificiales para predecir la calificación final de estudiantes universitarios antes de degradarse. Se trabajó con 30 estudiantes del Departamento de Ciencias de la Computación de la Universidad de Educación “Tai Solarin”. Los resultados mostraron la capacidad de las redes neuronales para predecir correctamente la calificación final de los estudiantes con un 92.7% de precisión.

Hu&Shih (2014) desarrollaron un sistema de alerta temprana basado en árboles de decisión para predecir si los estudiantes aprobarían o no. El modelo fue construido utilizando datos de 300 estudiantes y 13 variables recogidos mediante analítica en línea. Los resultados revelaron un 95% de precisión.

Isljamovic&Suknovic (2014), utilizan diferentes algoritmos de redes neuronales artificiales para encontrar el modelo más adecuado para la predicción de éxito académico de estudiantes graduados en la facultad de ciencias de la organización de la Universidad de Belgrado. El modelo permite identificar aquellos factores de mayor influencia en el rendimiento académico de los estudiantes.

Zacharis (2016), construyó un modelo basado en redes neuronales multi capas para predecir el rendimiento estudiantil en un entorno de aprendizaje mixto. Para ello utilizó datos de los estudiantes almacenados en un servidor Moodle y predijo el éxito del estudiante en curso, basado en cuatro actividades de aprendizaje: comunicación por correo electrónico, creación de contenido colaborativo con wiki, interacción de contenido

medida por archivos vistos y autoevaluación mediante cuestionarios en línea. El modelo predijo el rendimiento de los estudiantes con una tasa de clasificación correcta, del 98,3%.

3. Marco teórico

3.1. Analítica de aprendizaje en la educación

La disponibilidad de grandes conjuntos de datos obtenidos por la interacción de los estudiantes en los “Entornos Virtuales de Aprendizaje” en las instituciones educativas proporcionan grandes oportunidades para mejorar el proceso de aprendizaje de los mismos; dando lugar a un enfoque novedoso en el ámbito educativo conocido como “analítica de aprendizaje” (Learning Analytics). (Abu, 2016; Rojas, 2017)

El “learning analytics” aprovecha la gran cantidad de información que generan los estudiantes y sus contextos durante el proceso educativo con la finalidad de entender, mejorar y optimizar el aprendizaje y el rendimiento académico. (Conijn, Snijders, Kleingeld, & Matzat, 2017). Según Daud (2017), los datos obtenidos pueden ser útiles en diferentes contextos: en el caso de los docentes estos pueden monitorear el progreso de los estudiantes, los mismos que pueden retroalimentarse. De igual forma los directores y administradores esta información les permite tomar decisiones respecto a incrementos de aulas, tasas de graduación, ect.

3.2. Redes neuronales

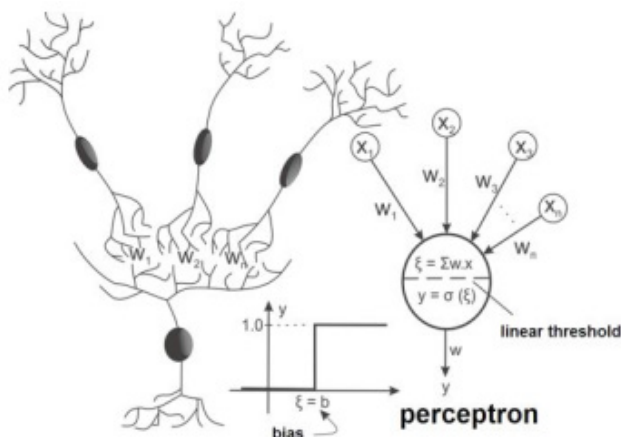


Figura 1 – Esquema de un modelo neuronal perceptrón

Fuente: (Del Brio, Sanz, & García, 2006)

Las redes neuronales artificiales (RNA) constituyen algoritmos computacionales inspirados en las redes neuronales biológicas, comportándose de forma similar a las mismas en sus funciones más comunes. Las RNA son entrenadas para aprender, generalizando resultados y extrayendo patrones de un conjunto de datos. (Del Brio, Sanz, & García, 2006)

Las redes neuronales artificiales reciben los datos de entrada del entorno y los agregan de una manera especial para formar una entrada de red (ξ), que pasan a través de un umbral lineal y la salida, “y” se transfiere a la neurona vecina o al entorno (Figura 1). El disparo neural se produce cuando ξ supera el umbral del borde denominado sesgo, “b”. La dinámica de la neurona lineal se asume comúnmente durante el cálculo de ξ y la entrada de la red se calcula como un producto escalar de las señales de entrada “x” independientemente de la neurona y su peso “w”.

Para n señales, la operación de la neurona perceptrón se expresa mediante:

$$y = \begin{cases} 1, \sum_{i=1}^n w_i x_i \geq b \\ 0, \sum_{i=1}^n w_i x_i < b \end{cases} \quad (1)$$

Un peso positivo ($w_i > 1$) aumenta la neurona, mientras que un peso negativo reduce e inhibe la actividad neural. Este tipo de neuronas artificiales se denomina perceptrón porque establece un mapeo entre la actividad de entrada (estímulo) y la señal de salida. El umbral neuronal puede considerarse como un nodo de entrada adicional cuyo valor es siempre 1, y su peso es igual a “b”.

En problemas no lineales las capas adicionales de neuronas se insertan entre la capa de entrada y la neurona de salida, por lo que se obtiene una arquitectura de red neuronal perceptrón multicapa (MLP). Como esta capa no puede interactuar con el entorno, se denomina capa oculta y los nodos de esta capa se denominan nodos ocultos. Las capas ocultas adicionales aumentan la capacidad del perceptrón para resolver un problema de clasificación no lineal.

Las redes neuronales con propagación hacia atrás representa es un tipo de MLP entrenado. Sin embargo, el proceso de aprendizaje es una actualización del algoritmo de perceptrón simple durante el cual los pesos se están ajustando en las capas ocultas.

La función de activación de las neuronas en la capa oculta desempeña un papel para proporcionar el aprendizaje de la red neuronal. Las neuronas de las capas ocultas no tienen una capacidad superior a la red de perceptrón ordinaria en el caso de problemas lineales. Las funciones de activación no lineales a menudo se usan para las neuronas en la capa de salida, lo que influye en que la red con múltiples capas sea particularmente poderosa. Las funciones de activación más frecuentemente usadas en el algoritmo de propagación hacia atrás son:

- Linear: $x; (-\infty, +\infty)$
- Sigmoidea logística: $\frac{1}{1+e^{-x}}; (0, +1)$
- Tangente hiperbólica: $\frac{e^x - e^{-x}}{e^x + e^{-x}}; (-1, +1)$
- Exponencial: $e^{-x}; (0, +\infty)$
- Escalón: $\begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}; [0, +1]$

4. Metodología

A continuación, se describe la metodología utilizada y aplicada a un conjunto de 300 estudiantes de la “Carrera de Ingeniería Informática” de la Universidad Central del Ecuador.

4.1. Recolección de datos

La información de los estudiantes se obtiene a través del intercambio de los mismos con el “Entorno Virtual de Aprendizaje” (<https://uvirtual.uce.edu.ec/>), desarrollado en Moodle. Los registros de datos de suministrados por el sistema fueron extraídos y sometidos a una limpieza utilizando los paquetes “DataClean” y “DataCombine” del software de uso libre “R”. De esta forma detectamos las variables disponibles que formaron parte de la investigación y que mostramos en la tabla 1

Variable predictora	
Número total de clicks	Número de consultas al profesor
Cantidad de sesiones en línea	Participación en los foros
Tiempo total en línea (min)	Participación en chat
Número de visitas a la página del curso	Participación en video-colaboración
Número de comentarios	Número de recursos vistos
Irregularidad del tiempo de estudio	Número de enlaces vistos
Irregularidad del intervalo de estudio	Número de páginas de contenido vistas
Mayor periodo de inactividad	Número de publicaciones de discusión vistas
Tiempo hasta la primera actividad (min)	Número total de publicaciones de discusión
Tiempo promedio por sesión (min)	Numero de cuestionarios iniciados
Número de intentos por prueba	Número de asignaciones enviadas
Número de cuestionarios vistos	Número de presentación vistas
Número de cuestionarios pasados	Número de ediciones en la wiki
Número de wiki vistas	Grado promedio de evaluación

Tabla 1 – Variables predictoras usadas

4.2. Análisis de datos

Las variables obtenidas fueron procesadas, utilizando el paquete “nnet” de R, que permite la creación, aplicación, visualización e implementación de redes neuronales de alimentación directa con una sola capa oculta, para modelos multinomiales log-lineales.

El diseño de la red neuronal utilizada se basó en una topología de red neuronal multicapa (MLP), la cual fue entrenada de forma independiente para cada asignatura obteniendo predicciones únicas de forma correcta.

La arquitectura de la red estuvo determinada por una capa de entrada y una capa intermedia u oculta con funciones de activación sigmoideal tangente hiperbólica y una capa de salida con función de activación lineal.

5. Resultados

La tabla 2, describe la precisión de clasificación y recuperación para cada curso y las diferencias en el rendimiento predictivo de la red neuronal entrenada individualmente en los datos de cada curso.

Num. curso	Curso	Exact.	Exact. base	Recordado	Grados asignación
1	Programación I	0.621	0.533	0.521	-
2	Programación II	0.772	0.685	0.570	-
3	Estructura de Datos	0.588	0.610	0.621	-
4	Programación Distribuida	0.866	0.810	0.857	Si
5	Aplicaciones de Minería de Datos	0.917	0.895	0.906	Si
	Promedio	0.7528		0.6950	-

Tabla 2 – Rendimiento de la red neuronal por cursos

Según los resultados obtenidos, el curso de *Aplicaciones de Minería de Datos* presenta el mayor valor de precisión y recuerdo con un 91,7% y 89,5% respectivamente. Al observar los resultados restantes, las precisiones de predicción van del 58.8% para el curso de *Estructura de Datos* al 86,6% para el curso de *Programación Distribuida*. Al promediar los puntajes obtenidos, encontramos una precisión del 75,28%. De igual forma podemos encontrar que los puntajes de recuerdo tienen una alta variabilidad que va del 52,1% en *Programación I* hasta el 90,6% para el curso de *Aplicaciones de Minería de Datos*; con valor promedio de recuperación de 69,50%.

Junto a los resultados anteriores se extrajeron las características de importancia para cada uno de los cursos, los cuales se exponen en la tabla 3. En el caso del curso de *Minería de Datos*, que como habíamos mencionado presenta el mayor valor de precisión, tenemos que el grado del cuestionario fue la característica más relevante con un 12,4%; la regularidad del estudio y el promedio de clics por sesión fueron los siguientes mejores predictores con un 12,4% y 10,9% respectivamente. En contraste a estos resultados tenemos que el grado promedio de las evaluaciones representa una característica de poca importancia en este curso, pues solo obtuvo un puntaje de 2,4%; resultado que puede explicarse ya que algunas calificaciones son un registro de una determinada tarea mas no refleja que tan bien se realizó la misma.

El otro curso que, según los resultados, tiene calificaciones y asignación como predictor es el de *Estructura de Datos*. Para el mismo, el promedio de las evaluaciones fue la característica más importante representando un 25,3%; seguido del promedio de clics por sesión con un 15,6%. Las características menos relevantes fueron: los mensajes enviados, publicaciones en foros/chat/videos, el grado promedio del examen y el número de cuestionarios realizados con un puntaje de 0.0%.

5. Conclusiones

El objetivo de esta investigación es determinar la efectividad de un modelo basado en redes neuronales para predecir el rendimiento del estudiante. La revisión de la literatura indicó que estos enfoques superan a todos los otros clasificadores, con respecto a la precisión de la predicción. Una red neuronal de perceptrón multicapa fue entrenada por un algoritmo de propagación inversa, para predecir la capacidad de pasar exitosamente la carrera.

Características	Curso 4	Curso 5
Regularidad de estudio	12.4	11.6
Grado promedio de examen	19.5	0.0
Número de cuestionarios realizados	9.4	0.0
Máximo periodo de inactividad	9.7	10.0
Longitud total de sesión	10.0	10.3
Total número de sesiones	8.3	9.6
Total número de clics	7.4	11.9
Promedio de clics por sesión	10.9	13.5
Promedio de longitud de sesión	8.1	8.9
Número de asignaciones realizadas	1.9	1.9
Mensajes enviados	0.0	0.0
Publicaciones en foro, chat, video	0.0	0.0
Grado promedio de evaluación	2.4	22.4

Tabla 3 – Características importantes por cada curso

La tasa de precisión de la clasificación es alta, con un promedio de 75,28% para todos los cursos; lo que muestran la efectividad de los predictores obtenidos para la predicción del rendimiento académico.

Referencias

- Abu, A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 212–220.
- Acevedo, C., & Rocha, F. (2011). Estilos de aprendizaje, género y rendimiento. *Journal of Learning Styles*, 71-84.
- Altujjar, Y., Altamimi, W., Al-Turaik, & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science*, 65–71.
- Baepler, P., & Murdoch, C. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, 1–9.
- Baker, R. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent Systems*, 78–82.
- Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, 1–15.
- Daud, A., Radi, N., Ayaz, R., Lytras, M., & Abbas, F. (2017). Predicting Student Performance using Advanced Learning Analytics. *Proceedings of the 26th International Conference on World Wide Web Companion* (págs. 415-421). Australia: WWW '17 Companion.

- Del Brio, B. M., Sanz, A., & García, A. (2006). *Redes neuronales y sistemas borrosos*. EUU: RA-MA EDITORIAL.
- González, J. C., Ramos, S., & Hernández, S. (2017). Modelo Difuso del Rendimiento Académico Bi-explicado. *Revista de Sistemas y Gestión Educativa*, 55–64.
- Hamasa, H., Indiradevi, S., & Kizhakkethottam, J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*, 326–332.
- Hu, Y., Lo, C., & Shih, S. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 469–478.
- Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 133–145.
- Il-Hyun, J., Yeonjeong, P., Jeonghyun, K., & Jongwoo, S. (2014). Analysis of Online Behavior and Prediction of Learning Performance in Blended Learning Environments. *Educational Technology International*, 71–88.
- Isljamovic, S., & Suknovic, M. (2014). Predicting students academic performance using artificial neural network : A case study. *The Eurasia Proceedings of Educational & Social Sciences (EPESS)*, 68–72.
- Jaramillo, A., & Paz, H. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. *Revista Tecnológica ESPOL*, 64–90.
- Lu, O., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Journal of Educational Technology & Society*, 220–232.
- Rojas, P. (2017). Learning Analytics: A Literature Review. *Educ.Educ*, 106–128.
- Schalk, P., Wick, D., Turner, P., & Ramsdell, M. (2011). Predictive assessment of student performance for early strategic guidance. In *Frontiers in Education Conference (FIE)*. Rapid City, Estados Unidos de América.
- Usman, O., & Adenubi, A. (2013). Artificial Neural Network (ANN) Model for Predicting Students' Academic Performance. *Journal of Science and Information Technology*, 23–37.
- Ye, C., & Biswas, G. (2014). Early prediction of student dropout and performance in moocs using higher granularity temporal information. *Journal of Learning Analytics*, 169–172.
- Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 17–29.

© 2019. This work is published under
<https://creativecommons.org/licenses/by-nc-nd/4.0/>(the
“License”). Notwithstanding the ProQuest Terms and
Conditions, you may use this content in accordance with the
terms of the License.