www.concienciadigital.o

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

Predicción del nivel de riesgo de reprobación estudiantes de educación superior usando un modelo de red neuronal artificial.

Prediction of the risk level of failure of higher education students using an artificial neural network model.

Gisel Katerine Bastidas Guacho. [1], Patricio Xavier Moreno Vallejo. [2] & María Elena Vallejo Sanaguano. [3]

Recibido: 14-06-2021 / Revisado: 24-06-2021 /Aceptado: 12-07-2021/ Publicado: 05-08-2021

Abstract.

DOI: https://doi.org/10.33262/concienciadigital.v4i3.1.1816

The desertion of undergraduate students and high academic failure rates is a problem in Ecuador's higher education institutions. If the failure rate in a subject is high, then the number of students who must retake the subject is also high. Therefore, it limits the available resources and makes the educational institutions' authorities constantly restructure physical spaces and teachers. On the other hand, educational data mining uses machine learning and deep learning techniques to analyze and model educational data to predict students' academic performance. Previous studies propose the use of different models of artificial neural networks to predict academic performance; however, these models focus on using only academic data and some students' sociodemographic data. On the contrary, in the present study, educational, sociodemographic, and economic data are considered, which were gathered through digital surveys and educational systems of a

[1] Escuela Superior Politécnica de Chimborazo, Facultad de Informática y Electrónica, Carrera de Software, Riobamba, Chimborazo, Ecuador. gis.bastidas@espoch.edu.ec, https://orcid.org/0000-0002-6070-7193

[2] Escuela Superior Politécnica de Chimborazo, Facultad de Administración de Empresas. Carrera de Gestión del Transporte, Riobamba, Chimborazo, Ecuador. xavier.moreno@espoch.edu.ec, https://orcid.org/0000-0002-9317-9884

[3] Escuela Superior Politécnica de Chimborazo, Facultad de Recursos Naturales. Carrera de Forestal, Riobamba, Chimborazo, Ecuador. mvallejo@espoch.edu.ec, https://orcid.org/0000-0003-0026-5917

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

higher education institution, and a multi-layer perceptron network is proposed to predict the risk of failure of a student, which will allow students, teachers and authorities to know the risk of loss in a subject so that the corresponding actions can be taken to lower the failure rate. The proposed model reached an accuracy of approximately 88%, demonstrating good performance. Additionally, we compare the proposed model's performance with a decision tree's performance and a logistic regression model; these models obtained approximately 85% and 82% accuracy, respectively.

Keywords: Prediction, Academic Performance, Failure, MLP.

Resumen.

La deserción de los estudiantes universitarios de las carreras y las altas tasas de reprobación es un problema en las instituciones de educación superior en el Ecuador y mientras más alta es la tasa de reprobación en una asignatura, mayor es el número de estudiantes que deben cursar nuevamente dicha asignatura lo cual limita los recursos disponibles y hace que las autoridades de las instituciones educativas realicen una constante reestructuración de espacios físicos y de docentes. Por lo tanto, el objetivo del presente estudio es usar la minería de datos educativa con técnicas de aprendizaje de máquina y aprendizaje profundo para analizar y modelar datos educativos de tal forma que se puede predecir el rendimiento académico de un estudiante. El diseño de la investigación fue mixta y longitudinal debido a que se analizó información obtenida durante 6 periodos académicos. A diferencia de estudios previos, en el presente estudio se consideran datos académicos, sociodemográficos y socioeconómicos los cuales fueron obtenidos mediante encuestas digitales y sistemas informáticos académicos de una institución de educación superior y se propone un modelo de red neuronal artificial MLP para predecir el nivel de riesgo de reprobación de los estudiantes, el cual permitirá a estudiantes, docentes y autoridades conocer el riesgo de reprobación en una asignatura de la forma que se pueda tomar las acciones correspondientes con el fin de menorar la tasa de reprobación. El modelo propuesto alcanzó una certeza de aproximadamente el 88% demostrando un buen desempeño. Adicionalmente, se comparó el rendimiento del modelo propuesto con el rendimiento de un modelo de árbol de decisión y de un modelo de regresión logística aplicados al mismo conjunto de datos, estos modelos obtuvieron una certeza de aproximadamente 85% y 82%, respectivamente.

Palabras claves: Predicción, Rendimiento Académico, Reprobación, MLP.

Introducción.

La minería de datos, también conocida como el descubrimiento de conocimiento en bases de datos, se enfoca en obtener información novedosa y potencialmente útil a partir de conjuntos de datos extensos (Baker, 2010). Desde su creación, la minería de datos se ha aplicado en varias áreas incluyendo la educación a partir de la necesidad de predecir el

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

comportamiento de los estudiantes para poder asistirlos de forma oportuna para que se puedan graduar sin inconvenientes. Se han realizado avances importantes en la educación superior al utilizar la minería de datos para predecir hasta con un 85% de certeza que estudiantes se graduarán y quienes no lo lograrán (Davis et al., 2007). Comúnmente, las estimaciones se basan en ciertas características como la nota promedio al final de un periodo académico o el nivel de ingresos. Con el pasar de los años, algunos enfoques basados en probabilidad, estadística, aprendizaje de máquina, programación dinámica, entre otros, se han aplicado en la minería de datos educacional (MDE). Siendo los más populares, los métodos basados en probabilidades, seguido por los métodos de aprendizaje de maquina y estadística, que en total abarcan un 88% de los métodos propuestos a lo largo de la historia de la MDE. Un 90% de los enfoques aplican tareas relacionadas a clasificación, agrupamiento, regresión y reglas de asociación (Peña-Ayala, 2014). La clasificación se aplica cuando se tiene una variable categórica que puede ser binaria o multiclase, como, por ejemplo, una variable binaria que indica si un estudiante aprueba o no una materia, o una variable multiclase que indica si el rendimiento del estudiante es bajo, medio o alto. En cualquiera de los casos los modelos de clasificación buscan predecir el valor de la variable para nuevas observaciones que, en este caso, cada observación corresponde a un estudiante. Los modelos que utilizan el enfoque de aprendizaje de máquina se optimizan utilizando ciertos algoritmos especializados que se basan en información histórica de los estudiantes. Por ejemplo, algunos algoritmos de clasificación incluyen la máquina de soporte de vectores (SVM) que están basados en kernels (Cristianini & Shawe-Taylor, 2000), arboles de decisión (Rokach & Maimon, 2005), bosques aleatorios (Zhang & Ma, 2012), regresión logística (Bonaccorso, 2017), k vecinos cercanos (Deng et al., 2016), y redes neuronales (Goodfellow et al., 2016). Por otra parte, el agrupamiento se utiliza cuando se desea crear grupos de estudiantes que compartan características similares de forma no supervisada. Con relación al desempeño de los estudiantes, se puede buscar dos grupos de estudiantes en donde uno contenga a los estudiantes con probabilidad de aprobar un nivel y el otro con los estudiantes con baja probabilidad de aprobar un nivel. Por lo general, en los enfoques que utilizan agrupamiento, es necesario indicar el número de grupos k que se desean descubrir previo a la ejecución del algoritmo. La regresión se utiliza para modelar los datos X en base a una función de regresión que permita obtener los valores futuros resolviendo la función obtenida para nuevos valores de X. Generalmente, se usa modelos de regresión lineal que son optimizados con el principio de los mínimos cuadrados. Sin embargo, debido a que los datos no siempre tienen una correlación lineal, también se aplican modelos de aprendizaje de máquina que pueden aprender funciones polinómicas más complejas. Por último, las reglas de asociación se utilizan para crear ciertas condiciones al estilo si-entonces, en donde, el cumplimiento de ciertas reglas da como resultado el valor de la variable objetivo. Los modelos basados en reglas de decisión utilizan la entropía para optimizar la generación de reglas.

(Roblyer & Davis, 2008) proponen un modelo basado en regresión logística para predecir la probabilidad de aprobación y de esa forma dar soporte a los estudiantes con baja probabilidad de aprobación para que se pueda prevenir el fracaso. En dicho estudio se

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

indica que la regresión logística es útil para predecir de forma acertada los estudiantes que aprueban un curso, pero al momento de predecir los estudiantes que fallan un curso, el modelo tiene un rendimiento pobre que alcanza una certeza de apenas el 30%. Por otra parte, (Chang & Kim, 2021) aplican regresión logística para obtener la probabilidad de que un estudiante apruebe o falle un curso en línea en base a tres conjuntos de variables que incluyen los antecedentes del estudiante, actividades de aprendizaje realizadas por el estudiante y las características individuales del curso tomado por el estudiante. El estudio realizado por (Chang & Kim, 2021) considera información referente al curso para realizar las predicciones, como son variables que indican si el curso tiene un examen final acumulativo, si el curso fue dado en la primavera, la tasa histórica promedio de aprobación del curso, y la tasa histórica promedio de aprobación con el profesor encargado del curso. A diferencia de los artículos previamente revisados, el presente artículo aplica aprendizaje profundo con el perceptrón multicapa para la predicción del nivel de riesgo de reprobación de un estudiante universitario tomando en cuenta algunas variables de tipo sociodemográficos, socioeconómicos y académicos.

El artículo se ha organizado en las siguientes secciones: Metodología, en donde se describe los métodos y técnicas utilizadas para la obtención de los datos y la selección del modelo de predicción. En la sección de Resultados se presenta el rendimiento del modelo basado en métricas como la exactitud y la sensibilidad. En la sección Discusión se analiza desde un punto crítico los resultados obtenidos. En la última sección se presentan las conclusiones del presente estudio.

Metodología.

En esta sección se presenta la metodología utilizada en la implementación de un modelo predictivo del nivel de riesgo de reprobación de estudiantes de educación superior. Primeramente, se realizó la recolección de los datos a ser usados en el entrenamiento del modelo propuesto, estos datos fueron preprocesados usando diferentes técnicas de preprocesamiento de datos como reducción de dimensionalidad, eliminación de datos vacíos, codificación one-hot para datos categóricos y normalización de datos. Posteriormente, se procedió a entrenar el modelo propuesto y validar el rendimiento del mismo.

Datos

El conjunto de datos utilizado consta de datos sociodemográficos, socioeconómicos y académicos de estudiantes pertenecientes a una institución de educación superior como se muestra en la Tabla 1. Los datos académicos fueron extraídos de las actas de calificaciones de 6 periodos académicos. Estas actas se encuentraban en formato Excel por cada semestre y asignatura. Por otro lado, los datos sociodemográficos y socioeconómicos de los estudiantes se obtuvieron mediante encuestas digitales realizadas a los estudiantes. Dado que los datos utilizados proceden de diversas fuentes primero se realizó una integración de los datos obteniendo un conjunto de datos de 2974 registros de los cuales se realizó una limpieza de datos eliminando instancias con valores perdidos y removiendo datos irrelevantes. Adicionalmente, se realizó la conversión de datos

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

categóricos a una codificación One-Hot y se aplicó normalización MinMax. Esto resultó en un conjunto de datos con 227 variables por lo que se procedió a realizar una reducción de dimensionalidad utilizando el análisis de componentes principales (PCA) y se tomó los 100 primeros componentes principales. Dentro del conjunto de datos, cada observación se encuentra etiquetada con el nivel de riesgo de reprobación que puede tomar los valores de alto, medio, o bajo. Dado que los datos se encontraban desbalanceados, se utilizó las técnicas de oversampling y undersamplig aleatorio que permitieron balancear el conjunto de datos obteniendo 2400 registros para el entrenamiento del modelo y 480 registros de prueba. Como resultado el conjunto de datos final tiene un tamaño de 2400 observaciones con 100 variables para entrenamiento y 480 observaciones con 100 variables para pruebas.

Dato	Variable
Datos Sociodemográficos	Número_miembros_familia Sector Nivel_instrucción_padre Nivel_instrucción_madre Ocupación_madre Ocupación_padre Con_quién_vive
Datos Académicos	Periodo_ académico Código Asignatura Número_créditos Horas_semanales Nivel Paralelo Nota_Parcial1 Nota_Parcial2 Nota_Parcial3 Evaluación_Acumulativa Requiere_evaluación_final Evaluación_Final Requiere_evaluación_recuperación Evaluación_Recuperación Porcentaje_asistencia Aprobación
Datos Socioeconómicos	Trabaja Manutención_hogar Tipo_vivienda Internet_fijo Dispositivo_electrónico_casa Dispositivo_electrónico_compartido Tiempo_uso_dispositivo_electrónico

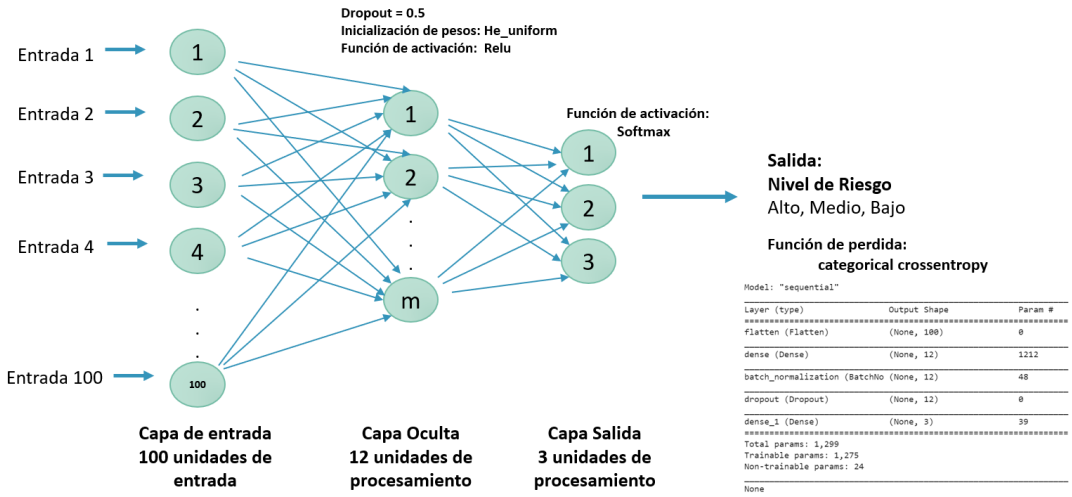
Tabla 1: Conjunto de datos
Fuente: Elaboración propia.

Modelo

En el presente estudio se propone una red neuronal artificial basada en el perceptrón multicapa conocido como MLP por su nombre en inglés Multi-Layer Perceptron, el cuál

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

es un modelo de aprendizaje profundo. La red propuesta predice el nivel de riesgo de reprobación en una asignatura de un estudiante universitario mediante la evaluación de datos académicos, sociodemográficos y socioeconómicos. La arquitectura tiene una profundidad de 3 capas: la primera capa corresponde a la de entrada, la segunda capa es oculta y la última capa es de salida como se muestra en la Ilustración 1. La capa de entrada contiene 100 unidades de entrada por lo que ingresan vectores de dimensión 1x100 que corresponde a los datos de cada estudiante. La capa oculta contiene 12 unidades de procesamiento, esta cantidad de unidades de procesamiento se consideró en base a lo propuesto en (Altaf et al., 2019), la función de activación de esta capa es la unidad lineal rectificada, conocida como Relu, los pesos fueron inicializados con He (He et al., 2015). La optimización se realiza mediante la propagación de la raíz cuadrada de la media (RMSprop - Root Mean Squared Propagation) (Tieleman & Hinton, 2012). Por otro lado, la capa de salida contiene 3 unidades de procesamiento las cuales corresponden a los 3 niveles de riesgo de reprobación: Alto, Medio, Bajo y la función de activación es Softmax.



Model: "sequential"

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 100)	0
dense (Dense)	(None, 12)	1212
batch_normalization (BatchNo	(None, 12)	48
dropout (Dropout)	(None, 12)	0
dense_1 (Dense)	(None, 3)	39

Total params: 1,299
Trainable params: 1,275
Non-trainable params: 24

None

Ilustración 1: Arquitectura de la MLP propuesta
Fuente: Elaboración propia.

Resultados.

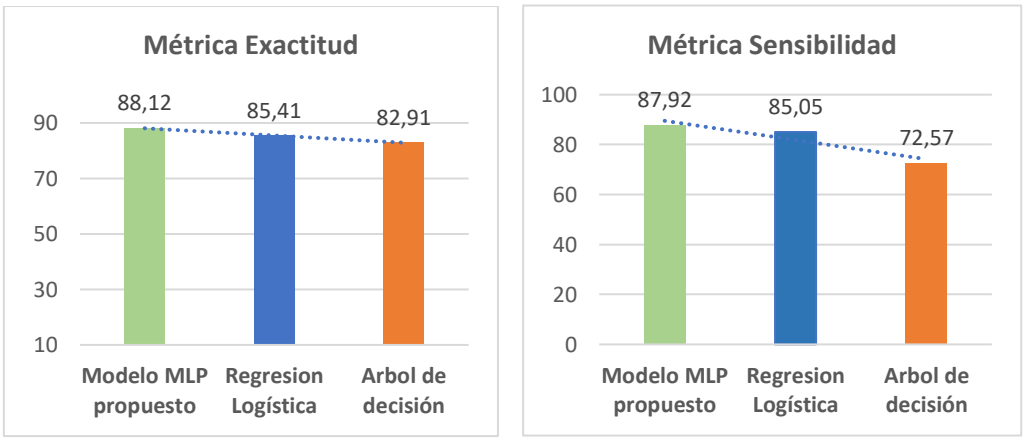
En esta sección se presentan los resultados obtenidos de la evaluación del rendimiento del modelo propuesto. La implementación del modelo se realizó en Python 3.6 usando la librería Keras. Para la definición de la arquitectura de la red se realizaron pruebas con distintas configuraciones e hiperparámetros que se muestran en la Tabla 2**¡Error! No se encuentra el origen de la referencia.**. Mediante la prueba de diferentes combinaciones de configuraciones de red se pudo mejorar los resultados desde aproximadamente 82% hasta 88% de exactitud. Por lo que la arquitectura final del modelo propuesto se basa en el mejor resultado obtenido de esta experimentación: 12 unidades de procesamiento, dropout de 0.2, inicialización de pesos con he_uniform, optimizador RMSProp, 1 capa oculta, batch size de 400 y 200 epochs. Las funciones de activación utilizadas son Relu y Softmax en la capa de oculta y salida, respectivamente.

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

Configuración/Hiperparámetro	Valores
Dropout	[0.2, 0.3, 0.4,0.5]
Número de capas ocultas	1,2,3
Batch size	[20,40,60,80, 100, … ,400]
Número de épocas	10,50,75,100,150,200,300,400
Algoritmo de inicialización de pesos	He_uniform y random_normal
Optimizadores	RMSProp, SGD, Adagrad, Adamax

Tabla 2: Configuraciones e hiperparámetros para experimentar y definir la red propuesta.
Fuente: Elaboración propia

Se evaluó el desempeño del modelo propuesto usando las métricas de exactitud y sensibilidad. El modelo alcanzó una exactitud de 88.12% en la predicción del nivel de riesgo de reprobación y una sensibilidad de 87.92%. Adicionalmente, se realizó un comparativo del rendimiento del modelo propuesto con el rendimiento de un modelo de regresión logística y un árbol de decisión entrenados con el mismo conjunto de datos. Los resultados de este comparativo se muestran en la Ilustración 2. El modelo de regresión logística obtuvo 85.41% y 85.05% de exactitud y sensibilidad, respectivamente. Mientras que el árbol de decisión obtuvo una exactitud de 82.91% y una sensibilidad de 72.57%. En base a estos resultados, se puede evidenciar que el modelo propuesto tiene un mejor rendimiento comparado a los otros modelos que han sido utilizados en estudios previos. Finalmente, el modelo propuesto fue integrado en un sistema de escritorio el cual permite realizar la predicción mediante el modelo MLP entrenado.



a) b)

Ilustración 2: Resultados del rendimiento de modelo propuesto respecto a un modelo de regresión logística y un árbol de decisión. En la gráfica a) se visualiza el comparativo de la exactitud y en b) se muestra los resultados de la sensibilidad obtenida por cada modelo evaluado.
Fuente: Elaboración propia

Discusión.

Hace algunos años era poco factible utilizar modelos de aprendizaje profundo por las limitaciones computacionales, sin embargo, hoy en día, se puede aprovechar las capacidades computacionales de los nuevos equipos para entrenar este tipo de modelos

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

hasta en computadores portátiles de última generación. Esto hace posible que en la actualidad se pueda minar datos educacionales con el fin de extraer conocimiento para mejorar los procesos educativos. Durante el desarrollo del presente estudio, se analizaron diferentes técnicas de aprendizaje de máquina que han sido utilizadas para la predicción de forma oportuna del nivel de rendimiento estudiantil con el fin de dar soporte a los estudiantes con bajo rendimiento para disminuir las tasas de repitencia estudiantil y así mejorar la calidad en la educación superior. Estudios previos han utilizado en su mayoría enfoques basados en regresión logística como (Chang & Kim, 2021; Roblyer & Davis, 2008) que solo permiten realizar una clasificación binaria. Por lo tanto, el aporte del presente estudio es ir más allá de una clasificación binaria, teniendo una clasificación multiclase, para lo cual se definieron 3 categorías para el nivel de riesgo de reprobación de una asignatura: Alto, Medio y Bajo. La definición del número de clases puede ser fácilmente extendido en el modelo propuesto con el fin de tener un mayor detalle del nivel de riesgo de reprobación. Para realizar el incremento del número de clases simplemente se debe añadir más unidades de procesamiento a la capa de salida del modelo. Adicionalmente, debido a que se ha demostrado la eficiencia del modelo propuesto, a futuro se puede extender este estudio utilizando la misma arquitectura pero incluyendo nuevas variables en el conjunto de datos referentes a las características de la asignatura y de los profesores como se propone en (Chang & Kim, 2021).

Conclusiones.

En el presente artículo se propone un modelo de predicción del nivel de riesgo de reprobación basado en un modelo de aprendizaje profundo multicapa perceptrón. Este modelo tiene como objetivo permitir a estudiantes, docentes y autoridades de educación superior conocer de manera temprana el nivel de riesgo de reprobación en una asignatura, de tal forma que se pueda tomar acciones inmediatas al respecto. Adicionalmente, el conjunto de datos construido en el presente estudio incluye datos sociodemográficos, socioeconómicos y académicos los cuales sirvieron para entrenar el modelo propuesto. La utilización de este conjunto de datos en el modelo propuesto lo diferencia de los modelos existentes que solamente incluyen datos académicos y en algunos de los casos datos sociodemográficos para hacer la predicción del rendimiento académico. El modelo propuesto además predice el nivel de riesgo definido en tres categorías alto, medio y bajo mientras que la mayoría de los modelos existentes realizan una clasificación binaria de las clases aprobado-reprobado. Finalmente, el modelo propuesto muestra un buen rendiendo comparado con los modelos de regresión logística y árboles de decisión, logrando obtener una exactitud y sensibilidad de 88.12% y 87.92%, respectivamente. Por lo que el modelo propuesto supera al modelo de regresión logística en aproximadamente 3% de exactitud y 6% de exactitud respecto al árbol de decisión.

Referencias bibliográficas.

Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A case study on educational data mining. *ACM International Conference Proceeding Series*, 59–64.

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

https://doi.org/10.1145/3325917.3325919

Baker, R. S. J. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112–118.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Chang, H. M., & Kim, H. J. (2021). Predicting the pass probability of secondary school students taking online classes. *Computers and Education*, *164*(December 2020), 104110. https://doi.org/10.1016/j.compedu.2020.104110

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Davis, C. M., Hardin, J. M., Bohannon, T., & Oglesby, J. (2007). Data mining applications in higher education. *Data Mining Methods and Applications*, 123–148. https://doi.org/10.1201/b15783

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143–148.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4 PART 1), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Roblyer, M. D., & Davis, L. (2008). Predicting success for virtual school students: Putting research-based models into practice. *Online Journal of Distance Learning Administration*, *11*(4).

Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165–192). Springer.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, *4*(2), 26–31.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

ISSN: 2600-5859
Vol. 4, N°3.1, p. 95-104, agosto, 2021

PARA CITAR EL ARTÍCULO INDEXADO.

Bastidas Guacho, G. K., Moreno Vallejo, P. X., & María Elena Vallejo Sanaguano. (2021). Predicción del nivel de riesgo de reprobación estudiantes de educación superior usando un modelo de red neuronal artificial. ConcienciaDigital, 4(3.1), 95-104. https://doi.org/10.33262/concienciadigital.v4i3.1.1816