

MODELO BASADO EN MACHINE LEARNING PARA EL NEURORRENDIMIENTO ACADÉMICO DE ESTUDIANTES UNIVERSITARIOS

Francisco Gamarra Gómez ^{1,a}

RESUMEN

Objetivo. Establecer el sobresaliente modelo predictivo para las métricas de Machine Learning, entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda en la Universidad José Carlos Mariátegui Filial Tacna, 2018-I. **Materiales y métodos.** Se utilizó el diseño no experimental y la correspondencia del coeficiente de determinación R^2 , en un enfoque cuantitativo. La población estuvo compuesta por 650 estudiantes, con una muestra de 194 estudiantes, el nivel de confianza fue de 95%, el intervalo de error es de 5%. Los instrumentos empleados fueron la media de evaluación por estudiante, la nómina de características individuales y encuestas acerca de la longitud de onda. **Resultados.** La regresión de decisión forestal con 0,985251 fue el mejor modelo predictivo; además, se obtuvo un buen índice Kappa, determinado por medio del software Azure Machine Learning Studio. **Conclusiones.** Para cada escuela se determinó el mejor modelo predictivo, a medida que el entrenamiento se realice con mayor población, aumentara el índice Kappa.

Palabras clave: Big data; Machine Learning; Modelo y regresión; Neurorendimiento académico.

MACHINE LEARNING BASED MODEL FOR THE ACADEMIC NEURORENDMENT OF UNIVERSITY STUDENTS

ABSTRACT

Objective. Stablishing the outstanding predictive model for Machine Learning metrics, between Academic Neuro-Performance, Neuro-Learning, Student Characteristics and Wavelength at the José Carlos Mariátegui Filial Tacna University, 2018-I. **Methodology.** Non-experimental design and the correspondence of the coefficient of determination R^2 were specified, in a quantitative approach. Population was composed of 650 students, with a sample of 194 students, the confidence level was 95%, the error interval is 5%. Instruments, the average evaluation per student, payroll of individual characteristics and surveys about wavelength. **Results.** The regression of forest decision with 0.9885251 was the best predictive model and a very good Kappa index was obtained determined through the Azure Machine Learning Studio software. **Conclusion.** For each school the best predictive model was determined, a measure that the training is carried out with a larger population, increases the Kappa index.

Keywords: Big data; Machine learning; Model and regression; Academic Neuro-Performance.

¹ Universidad José Carlos Mariátegui, Tacna, Perú.

^a Investigador principal. Maestro en Ciencias con mención en Gestión Empresarial.

INTRODUCCIÓN

En las universidades de todo el planeta emerge en los últimos años el interés por la calidad de la tarea educativa que prestan. En ese sentido, se ha efectuado la tesis doctoral que utilizó el promedio de las calificaciones como medida del rendimiento académico aplicando la estadística descriptiva, la estadística inferencial, y el análisis multivariado, cuyo estudio predictivo concluye que los análisis de regresión múltiple les permite afirmar que existen factores diferenciales del rendimiento académico en las 18 carreras existentes ⁽¹⁾.

Asimismo, el modelo general alcanza un coeficiente de determinación ajustado de 59,6%; en la presente investigación se utilizó el Azure Machine Learning Studio y se obtuvo una determinación cuyo valor es de 98,53%.

Como consecuencia de ello, surgen los procesos de autoevaluación a fin de determinar las fortalezas y debilidades de las instituciones, y propiciar actividades correctivas de las imperfecciones halladas. Razón por la cual las instituciones de enseñanza superior públicas y particulares han iniciado metodologías de autoestudio organizacional; es aquí donde el neurorendimiento universitario se toma en cuenta como un destacado indicativo en la formación profesional de un educando en un periodo determinado ⁽²⁾.

Este estudio ayuda a determinar el mejor modelo predictivo relacional de coeficiente de determinación basado en métricas de Machine Learning entre el neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda ⁽³⁾.

La utilización del adecuado color, así como la apropiada combinación y colocación, posibilita la influencia en las emociones, la atención y la conducta de los individuos cuando se está realizando el neuroaprendizaje, de tal forma que los estudiantes recuerdan las imágenes a color con más facilidad que los de negro y blanco ⁽⁴⁾.

Error absoluto medio (MAE)

Es el valor medio de la desigualdad absoluta de la estimación real y la predicha ⁽⁵⁾.

Error cuadrático medio (RMSE)

Es el valor medio de la diferencia al cuadrado entre el valor real y el valor predicho para todos los puntos de datos ⁽⁶⁾.

Error absoluto relativo (RAE)

Es la suma de todos los errores absolutos dividida entre la suma de las distancias entre los valores de etiqueta correctos y el promedio de todos los valores de etiqueta adecuadas ⁽⁷⁾.

Error cuadrático relativo (RSE)

Es la suma de todos los errores absolutos cuadráticos dividida entre la suma de las distancias cuadráticas existente entre los valores de etiqueta correctos y el promedio de todos los valores de etiqueta adecuadas.

Coeficiente de determinación

Determina la capacidad de un modelo para predecir futuros resultados, representa la dispersión alrededor de la línea de regresión, si a cuyo valor lo multiplicamos por 100 nos indica el porcentaje de cambio de la variable dependiente ⁽⁸⁾.

Objetivo general

Determinar el mejor modelo predictivo relacional de coeficiente de determinación basado en métricas de Machine Learning entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda en la Universidad José Carlos Mariátegui Filial Tacna, 2018-I.

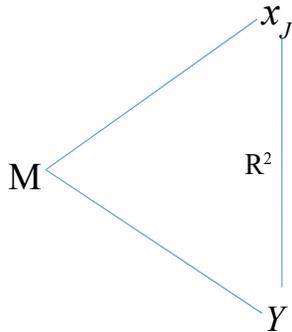
Hipótesis de investigación

Existe un modelo predictivo relacional de coeficiente de determinación basado en métricas de Machine Learning entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda en la Universidad José Carlos Mariátegui Filial Tacna, 2018-I.

Para la justificación de la hipótesis, se manipuló el chi cuadrado junto a la concordancia de las variables; además, se esgrimió las métricas de machine learning obteniendo el coeficiente de determinación y el índice Kappa.

MATERIALES Y MÉTODOS

Investigación aplicada⁽⁹⁾, de diseño relacional, utilizando métrica de Machine Learning de coeficiente de determinación⁽¹⁰⁾.



Donde:
 M = Muestra de estudiantes por ciclo y carreras profesionales.
 X_j = Neuroaprendizaje, características de los estudiantes, longitud de onda.
 Y = Neurorendimiento académico.
 R^2 = Coeficiente de determinación (relación entre (X, Y)).

Población

La población(650 estudiantes) estuvo conformada por los educandos de pregrado, en las carreras profesionales de Derecho, Contabilidad, Educación Inicial, Educación Especialidad, Educación Física, Ingeniería Comercial, Ciencias Administrativas y Marketing Estratégico, y Psicología.

$$n = \frac{Z^2 * N * p * q}{E^2 * (N - 1) + Z^2 * p * q}$$

Tabla 1. Población y muestra. Estudiantes de la Universidad José Carlos Mariátegui, Filial Tacna, 2018-I

Escuela profesional	Población y muestra	
	N	n
Ciencias Administrativas y Marketing Estratégico	19	64
Contabilidad	32	106
Derecho	103	344

Escuela profesional	Población y muestra	
	N	n
Educación Especialidad - Educación Física	3	11
Educación Inicial	17	56
Ingeniería Comercial	15	51
Psicología	5	18
Total	194	650

Fuente: Oficina de Servicios Académicos, Evaluación y Registro Central.

Tabla 2. Variables de investigación

Dimensiones	Nivel	Indicador
		Medida lograda
X1: Neuroaprendizaje	Elevado	20-18 excelente
		17-16 muy bueno
	Intermedio	15-13 bueno
X2: Características de los estudiantes	Intermedio	12-11 deficiente
		00-10 muy deficiente
	Inferior	Edades (años)
X3: Longitud de onda	Joven	15-20 adolescencia
		20-39 adultos jóvenes
	Adulto	40-49 adultos medios
		50-59 adultos maduros
Adulto mayor	60-69 vejez inicial	
X3: Longitud de onda	Longitud de Onda	(1) Amarillo 570-590nm
		(2) Verde 495-590nm
		(3) Rojo 645-700nm
		(4) Azul 450-475nm

Fuente: elaboración propia

Modelo predictivo

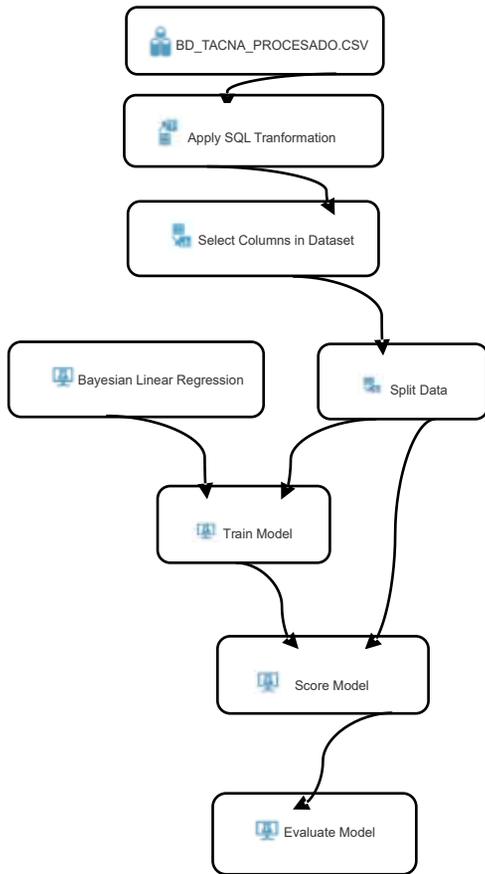


Figura 1. Plantilla predictiva Machine Learning

En la Figura 1, parte superior del modelo, se encuentra la base de datos BD-Tacna Procesado en formato CSV que contiene la información de todos los educandos de las carreras de la Universidad José Carlos Mariátegui, Filial Tacna 2018-I.

Para seleccionar a las carreras profesionales, con el fin de realizar las predicciones, así como a toda la Filial Tacna, se seleccionan las tres variables X1, X2, X3 haciendo uso de Apply SQL Transformación.

Luego pasa por el bloque de tratamiento de datos Select Columns In Data Set para seleccionar las columnas que se desea procesar (promedio acumulado, nota-curso, promedio ERP, edad, sexo, longitud de onda, ciclo, programa, estudio) en la parte inferior se hallan los algoritmos regresión lineal bayesiano, regresión de decisión forestal, regresión lineal, *Neural Network Regression*, y regresión del árbol de decisión impulsada, en el que se basa la

predicción.

Seguidamente, se efectúa el adiestramiento del modelo que radica en la actuación del algoritmo sobre una porción de los datos disponibles. Para hacer uso de los valores necesarios para el adiestramiento, se usa el bloque «división de datos», que divide el total de los valores en dos subclases de muestras aleatorias que comprenden proporcionalmente el 70% y el 30% del total.

El 70% de los valores se usa ligado al algoritmo. El egreso de este bloque suministra el modelo ya instruido, que a su momento sirve como ingreso para efectuar una valoración del mismo, con el bloque modelo de puntuación que señala cuál es la predicción. Este bloque tiene como entrada el restante 30% separados anteriormente.

Por último, y para comprobar el rendimiento de los diferentes modelos, se incorpora el bloque «evaluar modelo» que suministra las métricas ya explicadas de exactitud del modelo.

Combinamos y comparamos todas las métricas de rendimiento de regresión usando los módulos *execute R script* y *addrows*. El módulo *evaluate model* produce una tabla con una sola fila que contiene varias métricas.

En el módulo *execute R script*, extraemos las métricas de rendimiento de regresión y agregamos el nombre del modelo correspondiente, tener en consideración ignorar la probabilidad de registro negativo y extraer las cinco métricas restantes. Todos los módulos *execute R script* producen una tabla con una sola fila que contiene el nombre del modelo y varias métricas. Finalmente, combinar todas las métricas de rendimiento de regresión juntas.

El *software* utilizado es el Microsoft Azure Machine Learning Studio, el cual se encuentra en la nube.

RESULTADOS

Tabla 3. Resultados de métricas de Machine Learning - Psicología

<i>Model</i>	<i>Mean Absolute Error</i>	<i>Root Mean Squared Error</i>	<i>Relative Absolute Error</i>	<i>Relative Squared Error</i>	<i>Coefficient of Determination</i>
Regresión lineal bayesiano	0,181	0,206	0,126	0,017	0,983
Regresión de decisión forestal	0,998	1,316	0,693	0,676	0,324
Regresión lineal	0,160	0,175	0,111	0,012	0,988
Regresión de la red neuronal	0,821	0,997	0,570	0,388	0,612
Regresión del árbol de decisión impulsada	1,538	1,832	1,068	1,311	-0,311

En la Tabla 3, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,988 y el menor error cuadrático medio es 0,175.

Tabla 4. Resultados de métricas de Machine Learning - Derecho

<i>Model</i>	<i>Mean Absolute Error</i>	<i>Root Mean Squared Error</i>	<i>Relative Absolute Error</i>	<i>Relative Squared Error</i>	<i>Coefficient of Determination</i>
Regresión lineal bayesiano	0,253	0,292	0,313	0,065	0,935
Regresión de decisión forestal	0,033	0,131	0,041	0,013	0,987
Regresión lineal	0,252	0,292	0,313	0,064	0,936
Regresión de la red neuronal	0,249	0,288	0,308	0,063	0,937
Regresión del árbol de decisión impulsada	0,112	0,270	0,139	0,055	0,945

En la Tabla 4, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,987 y el menor error cuadrático medio es 0,270.

Tabla 5. Resultados de métricas de Machine Learning - Ingeniería Comercial

<i>Model</i>	<i>Mean Absolute Error</i>	<i>Root Mean Squared Error</i>	<i>Relative Absolute Error</i>	<i>Relative Squared Error</i>	<i>Coefficient of Determination</i>
Regresión lineal bayesiano	0,288	0,324	0,360	0,088	0,912
Regresión de decisión forestal	0,276	0,456	0,345	0,173	0,827
Regresión lineal	0,284	0,324	0,356	0,087	0,913
Regresión de la red neuronal	0,535	0,616	0,668	0,316	0,684
Regresión del árbol de decisión impulsada	0,616	0,895	0,770	0,667	0,333

En la Tabla 5, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,913 y el menor error cuadrático medio es 0,324.

Tabla 6. Resultados de métricas de Machine Learning - Ciencias Administrativas y Marketing Estratégico

<i>Model</i>	<i>Mean Absolute Error</i>	<i>Root Mean Squared Error</i>	<i>Relative Absolute Error</i>	<i>Relative Squared Error</i>	<i>Coefficient of Determination</i>
Regresión lineal bayesiano	0,254	0,296	0,335	0,099	0,901
Regresión de decisión forestal	0,261	0,454	0,343	0,234	0,766
Regresión lineal	0,253	0,296	0,333	0,099	0,901
Regresión de la red neuronal	0,386	0,441	0,508	0,221	0,779
Regresión del árbol de decisión impulsada	0,411	0,589	0,541	0,394	0,606

En la Tabla 6, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,901 y el menor error cuadrático medio es 0,296.

Tabla 7. Resultados de métricas de Machine Learning - Contabilidad

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Regresión lineal bayesiano	0,263	0,295	0,383	0,107	0,893
Regresión de decisión forestal	0,148	0,313	0,215	0,120	0,880
Regresión lineal	0,261	0,292	0,380	0,105	0,895
Regresión de la red neuronal	0,522	0,626	0,760	0,482	0,518
Regresión del árbol de decisión impulsada	0,126	0,286	0,183	0,101	0,899

En la Tabla 7, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,899 y el menor error cuadrático medio es 0,286.

Tabla 8. Resultados de métricas de Machine Learning - Educación Inicial

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Regresión lineal bayesiano	0,268	0,321	0,569	0,175	0,825
Regresión de decisión forestal	0,230	0,409	0,488	0,284	0,716
Regresión lineal	0,280	0,340	0,594	0,196	0,804
Regresión de la red neuronal	0,410	0,525	0,871	0,469	0,531
Regresión del árbol de decisión impulsada	0,434	0,544	0,923	0,503	0,497

En la Tabla 8, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,825 y el menor error cuadrático medio es 0,321

Tabla 9. Resultados de métricas de Machine Learning - Educación Física

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Regresión lineal bayesiano	0,309	0,400	0,696	0,720	0,280
Regresión de decisión forestal	0,347	0,361	0,781	0,586	0,414
Regresión lineal	0,416	0,444	0,935	0,866	0,114
Regresión de la red neuronal	0,754	0,874	1,697	3,439	-2,439
Regresión del árbol de decisión impulsada	1,042	1,143	2,344	5,883	-4,883

En la Tabla 9, los resultados de las métricas de Machine Learning son: el mejor coeficiente de determinación es 0,414 y el menor error cuadrático medio es 0,361.

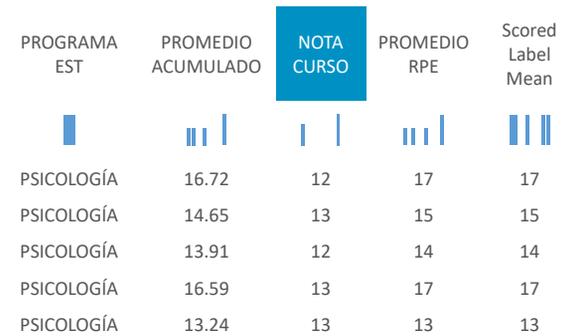


Figura 2. Resultados de predicción de la Carrera de Psicología



Figura 3. Resultados de predicción de la Carrera de Derecho



Figura 4. Resultados de predicción de la Carrera Comercial



Figura 7. Resultados de predicción de la Carrera Educación inicial



Figura 5. Resultados de predicción de la Carrera Ciencias Administrativas y Marketing Estratégico

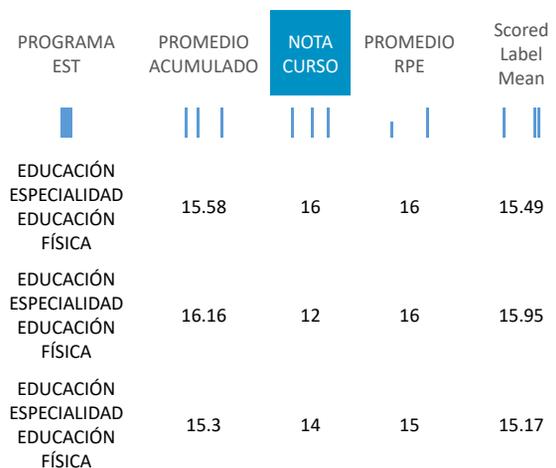


Figura 8. Resultados de predicción de la Carrera Educación Física



Figura 6. Resultados de predicción de la Carrera de Contabilidad

DISCUSIÓN

Entre los educandos de la escuela Profesional de Psicología se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de un elevado poder de predicción (98,8%), como se indica en la Tabla 3; además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la escuela profesional de Derecho se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de una elevada poder de predicción 98,7% (Tabla 4); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la escuela profesional de Ingeniería Comercial se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de un elevado poder de predicción 91,3% (Tabla 5); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la Escuela Profesional de Ciencias Administrativas y Marketing Estratégico se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de un elevado poder de predicción 90,1% (Tabla 6); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la Escuela Profesional de Contabilidad se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue Regresión Lineal de un elevado poder de predicción 89,9% (Tabla 7); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la Escuela Profesional de Educación Inicial se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de un elevado poder de predicción 82,5% (Tabla 8); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Entre los educandos de la Escuela Profesional de Educación Física se obtuvo un modelo sobresaliente relacional basado en métricas de Machine Learning que fue regresión lineal de un elevado poder de predicción 41,4% (Tabla 9); además, contiene un buen índice Kappa entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda.

Hipótesis

- a) Formulación de hipótesis
- b) Nivel de significancia
 $\alpha = 0,05$ ó 5% de error

c) Estadístico de prueba

$$X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} + e_{ij})^2}{e_{ij}}$$

d) Prueba de chi cuadrado para el modelo predictivo de regresión de decisión forestal.

Tabla 10. Resultados para la prueba de hipótesis y relación multivariado

Valor X^2	gl	Significación Asintónica (p)	R2	Índice Kappa
31,44	12,00	0,02	0,99	Muy buena

Fuente: elaboración propia

e) Regla de decisión

Se rechaza H_0 si:

$$\sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} + e_{ij})^2}{e_{ij}} \geq X^2_{0,05, gl=12}$$

El valor de chi cuadrado calculado es igual a 31,44 y el chi cuadrado tabla es igual a 21,03, por lo que se desestima la hipótesis nula y se acepta la hipótesis alterna, confirmado por el valor de significación asintótica p menor a 0,05. En consecuencia, hay bastante certeza estadística para aseverar que existe una relación muy buena del índice de Kappa con coeficiente de determinación igual a 0,99 entre el modelo predictivo basado en métricas de Machine Learning y neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda de estudiantes de la Universidad José Carlos Mariátegui, filial Tacna, 2018-I a un grado significativo de 0,05 para el modelo predictivo regresión de decisión forestal.

CONCLUSIONES

El mejor modelo predictivo, que explica la relación entre las variables, es la regresión lineal con una métrica de coeficiente de determinación de 0,988 y con un índice Kappa muy bueno. El 98,8% de la varianza total de la variable de respuesta explica su relación con sus variables predictivas de la

regresión entre neurorendimiento académico, neuroaprendizaje, características de los estudiantes y la longitud de onda en la Universidad José Carlos Mariátegui Filial Tacna, 2018-I, con un nivel de significancia asintótica P menor a 0,05.

RECOMENDACIONES

Para favorecer la administración académica con el propósito de que los educandos desarrollen su neurorendimiento académico, se sugiere que también se tenga la información socioeconómica

(situación laboral, financiación de estudios) y actitudinal (causa de elección de estudios, gusto, antelación al porvenir Académico y profesional) a partir de la incorporación del estudiante al centro superior de estudio, con esto se podría incrementar la exactitud y el nivel de concordancia.

Realizar investigaciones para el neurorendimiento académico en las universidades públicas utilizando el modelo basado en Machine Learning Studio con los algoritmos adaptados a la realidad de cada institución universitaria.

REFERENCIAS BIBLIOGRÁFICAS

1. Guzmán M. Modelo predictivo y explicativos del rendimiento académico universitario: caso de una institución privada en México. [Tesis doctoral]. [Madrid 516]: Universidad Compu- tence de Madrid; 2012.
2. Campos A. Uniendo las neurociencias y la educación en la búsqueda del desarrollo humano. Revista digital la Educación, 143 [Internet]. 2010; Disponible en: http://educocoea.org/portal/La_Educacion_Digital/laeducacion_143/articulos/neuroeducacion.pdf.
3. Guzmán M. Modelo predictivo y explicativos del rendimiento académico universitario: caso de una institución privada en México. [Tesis doctoral]. [Madrid 516]: Universidad Compu- tence de Madrid; 2012.
4. Brown M. Learning Analytics: The Coming Third Wave. EDUCAUSE Learning Initiative. abril de 2011;4.
5. García D. Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en Learning Analytic: (análisis del aprendizaje), mediante uso de técnicas multivariantes. [Tesis doctoral]. [Sevilla, 176]: Departamento de estadística e investigación operativa, Universidad de Sevilla.; 2015.
6. Shin Y, Raudenbusch S. The causal effect of class size on academic achievement multivariate instrumental variable estimators with data missing at random. Journal of Educational and Behavioral Statistics. 2011;154-85.
7. Goldstein H. Multilevel Statistical Models. 4th Edition. UK: Wiley; 2011. 384 p.
8. Michie D, Spiegelhalter DJ, Taylor CC. Machine Learning, Neural and Statistical Classification. 1994; Disponible en: <http://www1.maths.leeds.ac.uk/~charles/statlog/>.
9. Hernández Sampieri R, Fernández Collado C, Baptista Lucio P. Metodología de la investigación. 6a. ed. México D.F.: McGraw-Hil; 2014. 634 p.
10. Carrasco Diaz S. Metodología de la investigación científica. Lima: San Marcos. Lima; 2008.

Correspondencia

Francisco Gamarra Gómez
Universidad José Carlos Mariátegui - Filial Tacna - Urb. Los Olivos A-16, Tacna 23000 - Perú
fgamara@hotmail.com