Artículo de Investigación Original

Modelo Basado en Deep Learning para predecir el ingreso de estudiantes a la Universidad Nacional del Altiplano

A Deep Learning-Based Model for Predicting University Admission at the Universidad Nacional del Altiplano

Edwin Edgar Mestas Yucra [1[0000-0002-6000-1094]], María Bobadilla Quispe [2[0000-0002-4955-080X]], Mario Milton Quisocala Lipa [3[0000-0001-8810-7109]], Ivan Grover Sanchez Mamani [4[0000-0002-4307-7820]], Julio Cesar Sardón Huayapa [5[0000-0002-4447-4932]]

[1-5] Universidad Nacional del Altiplano – Perú.
edwin.mestas@unap.edu.pe, mariabobadilla@unap.edu.pe, marioquisocala@unap.edu.pe, ivansanchez@unap.edu.pe, csardon@unap.edu.pe

CITA EN APA:
Mestas Yucra, E. E., Bobadilla Quispe , M., Quisocala Lipa , M. M., Sanchez Mamani , I. G., & Sardón Huayapa , J. C. (2025). Modelo Basado en Deep Learning para predecir el ingreso de estudiantes a la Universidad Nacional del Altiplano. Technology Rain Journal, 4(2). https://doi.org/10.55204/trj.v4i2.e107

Recibido: 10 de julio-2025
Aceptado: 13 de octubre-2025
Publicado: 09 de noviembre-2025

Technology Rain Journal
ISSN: 2953-464X

Resumen. La investigación surge de la preocupación de las instituciones educativas de la Unidad de Gestión Educativa Local Melgar por preparar académicamente a sus estudiantes para ingresar a la Universidad Nacional del Altiplano Puno (UNA Puno). Deep Learning facilita la realización de predicciones mediante el aprendizaje no supervisado y el clustering jerarquizado. El estudio tuvo como objetivo general determinar un modelo basado en Deep Learning que permite predecir el ingreso de los estudiantes a la UNA Puno, y como objetivos específicos determinar que técnicas serian mejorar para tratar la información del rendimiento académico, describir con gráficos el rendimiento académico de los estudiantes y predecir el ingreso a la UNA Puno basado en la información del rendimiento académico de los estudiantes de la IES Nuestra Señora de Alta Gracia. Se realizó una investigación de enfoque cuantitativo, tipo aplicada, nivel predictivo, y diseño de modelo predictivo para la construcción de patrones, una población y muestra de 257 estudiantes, los datos se recolectaron de las actas de evaluación y proceso mediante la minería de datos. Se ha determinado que el análisis de clústering con métrica de distancia Russellrao y método simple identificó un clúster significativo de estudiantes con alto rendimiento académico; sin embargo, hubo una discrepancia notable entre el pronóstico inicial de ingresantes con un 58.36% y los ingresantes reales de 3.5%, lo que resalta la necesidad de ajustar continuamente las estrategias de pronóstico y admisión universitaria; esta discrepancia de debe a factores como la competitividad del proceso de admisión, y el cambio constante de las estrategias de selección de estudiantes.

Palabras Clave: Aprendizaje profundo, aprendizaje no supervisado, agrupamiento, predicción, rendimiento académico.

Abstract: The research arises from the concern of the educational institutions of the Unidad de Gestión Educativa Local Melgar to academically prepare their students to enter the Universidad Nacional del Altiplano Puno (UNA Puno). Deep Learning makes it easy to make predictions through unsupervised learning and hierarchical clustering. The general objective of the study was to determine a model based on Deep Learning that allows predicting the admission of students to UNA Puno, and as specific objectives to determine which techniques would be improved to process academic performance information, describe with graphs the academic performance of students and predict admission to UNA Puno based on information on the academic performance of students at IES Nuestra Señora de Alta Gracia. A quantitative approach research was carried out, applied type, predictive level, and predictive model design for the construction of patterns, a population and sample of 257 students, the data were collected from the evaluation and process minutes through data mining. It has been determined that the clustering analysis with Russellrao distance metric and simple method identified a significant cluster of students with high academic performance;

However, there was a notable discrepancy between the initial forecast of entrants at 58.36% and actual entrants at 3.5%, highlighting the need to continually adjust forecasting and college admission strategies; This discrepancy is due to factors such as the competitiveness of the admission process, and the constant change in student selection strategies.

Keywords: Deep learning, unsupervised learning, clustering, prediction, academic performance.

1. INTRODUCCIÓN

En muchos países, el ingreso a las universidades públicas se basa en un proceso de selección que evalúa el rendimiento académico y/o los resultados de exámenes estandarizados (Abougalala et al., 2025). Los sistemas de puntajes más utilizados, son el Examen Único de Admisión en Perú, el Sistema de Selección Universitaria en Chile o el Examen Nacional de Enseñanza Media en Brasil (Vargas et al., 2020).

El rendimiento académico refleja las competencias, habilidades y conocimientos de un estudiante, lo que les permite destacar entre otros postulantes (Araiza, 2021; Fajardo et al., 2017); Así también es fundamental para demostrar la preparación y capacidad de los estudiantes rumbo a la educación superior (Araiza, 2021). El acceso a la educación superior es un desafío que enfrenta la región de América Latina y el Caribe, debido a las demandas sociales y los principios de las políticas de acceso (Abougalala et al., 2025), donde también los procesos de admisión en las universidades requieren de criterios, modos y mecanismos para la selección de los estudiantes que garanticen la equidad, la calidad y la pertinencia (Abougalala et al., 2025).

El rendimiento académico de los estudiantes de secundaria es uno de los factores que influye en el ingreso a la universidad, pero no es el único ni el más determinante (Castrillón et al., 2020). Por lo tanto, se necesita de herramientas que permitan analizar y predecir el potencial éxito o fracaso de los aspirantes a partir de diversos factores educativos, familiares, socioeconómicos, de hábitos y costumbres, entre otros (Castrillón et al., 2020). Jiménez (2017) en su análisis de los datos de los exámenes de admisión para la Universidad Nacional del Altiplano en Puno, revela que el 9.8% de los estudiantes provienen de Puno, mientras que el 9.1% son de Melgar y el 8.2% son de San Román en los exámenes generales; para los exámenes CEPREUNA, el 17.7% de los estudiantes son de Carabaya, el 14.5% son del Collao y el 14.2% son de Chucuito; en cuanto a los exámenes extraordinarios, el 33.3% proviene de El Collao, el 30.8% de Chucuito y el 28.6% de Sandia. Estos porcentajes ofrecen una visión detallada de la distribución geográfica de los estudiantes que participan en los

procesos de admisión de la UNA Puno, lo que puede ser fundamental para comprender y abordar las disparidades regionales en el acceso a la educación superior en la región.

Deep Learning es un subconjunto de Machine Learning (Fernández, 2025; Incio-Flores et al., 2022), que es básicamente una red neuronal con tres o más capas, estas redes neuronales intentan emular el comportamiento del cerebro humano —aunque lejos de igualar su capacidad— pero le permiten "aprender" a partir de grandes cantidades de datos (Roopa & Reddy, 2023); así mismo puede procesar datos no estructurados, como texto e imágenes, y automatizar la extracción de características, eliminando parte de la dependencia de expertos humanos (Antonopoulos et al., 2020). Además, Deep Learning puede realizar predicciones con una gran precisión mediante el ajuste y la adaptación progresiva del algoritmo (Antonopoulos et al., 2020), por estas razones, puede ser una técnica útil para predecir el acceso a una universidad basado en la información del rendimiento académico de los estudiantes de secundaria, así como de otros factores relevantes (Gil y Quintero, 2021). Esta predicción puede ayudar a las instituciones educativas a identificar con anticipación a los estudiantes con problemas potenciales de rendimiento académico y a desplegar acciones de acompañamiento y mitigación inmediatas (Capuñay et al., 2021; Gil & Quintero, 2021).

Se asumió como objetivo determinar un modelo basado en Deep Learning que permite predecir con precisión el ingreso de los estudiantes a la UNA Puno. A continuación, se detalla la justificación de este estudio: Primeramente, llenar vacíos de conocimiento porque existe una escasez de estudios que aborden la predicción del ingreso a la UNA Puno utilizando el rendimiento académico de los estudiantes de la IES Nuestra Señora de Alta Gracia, esta investigación busca llenar este vacío de conocimiento al proporcionar un enfoque novedoso que aprovecha el potencial de Deep Learning (Márquez, 2020) a través clustering jerarquizado para predecir el éxito de los estudiantes en su ingreso a la universidad (Cardenas-Quispe et al., 2022).

Así también, se pretende resolver un problema relevante que es el acceso a la educación superior siendo un desafío importante en muchas regiones del país, y la región de Puno no es una excepción. También, contribuir al avance de las áreas y líneas de investigación, porque la presente investigación se enmarca en el campo del aprendizaje automático y la inteligencia artificial Aprendizaje no supervisado no siempre se dispone de una respuesta asociada (Costanza et al., 2023; Díaz, 2021), o incluso si se dispone de ella, podríamos tener interés en descubrir otro tipo de asociaciones; en estos casos, se pueden utilizar una serie de técnicas que se engloban dentro de lo que se llama aprendizaje no supervisado (Cardenas-Quispe et al., 2022); el término no supervisado hace referencia a que

este aprendizaje no se basa en la existencia de una respuesta previamente conocida (Costanza et al., 2023).

Para (Costanza et al., 2023) el aprendizaje no supervisado comprende un conjunto de herramientas estadísticas destinadas al entorno en el que solo tenemos (o usamos) un conjunto de variables X1,X2,… ,Xp sobre un conjunto de n instancias.; al aplicar estas técnicas en el contexto de la predicción del ingreso a la universidad (Moreno & Cortez, 2020), se espera generar conocimientos y aprendizajes relevantes en esta área, lo que contribuirá al avance de la ciencia y abrirá nuevas oportunidades para futuras investigaciones y aplicaciones prácticas (Araiza, 2021). En ese sentido esta investigación evidencia por la necesidad de llenar vacíos de conocimiento, resolver un problema relevante en relación con el acceso a la educación superior, atender las demandas y necesidades de la población, y contribuir al avance de las áreas y líneas de investigación relacionadas con el aprendizaje automático.

Se planteo la hipótesis de que un modelo basado en Deep Learning, aplicado a los datos del rendimiento académico de los estudiantes de secundaria, puede predecir con precisión su probabilidad de ingreso a la Universidad Nacional del Altiplano. El desarrollo de un modelo de predicción del ingreso a la UNA Puno basado en el rendimiento académico de la IES Nuestra Señora de Alta Gracia utilizando Deep Learning tiene el potencial de generar impacto positivo en la educación y la toma de decisiones en el ámbito educativo (Montero et al., 2024).

2. METODOLOGÍA O MATERIALES Y MÉTODOS

El estudio se desarrolló bajo un enfoque cuantitativo, de tipo aplicada y nivel predictivo. Se empleó el paradigma de machine learning para modelar el ingreso de estudiantes a la Universidad Nacional del Altiplano (UNA–Puno) a partir de su rendimiento académico en educación secundaria.

Ámbito o Lugar de Estudio

La investigación se ejecutará con los datos generados en la IES Nuestra Señora de Alta Gracia, provincia Melgar y departamento de Puno; ubicado en la ciudad de Ayaviri. Es importante porque es una población alejada a la capital de la región donde está ubicada la Universidad Nacional del Altiplano. Considerándose, así el muestreo no probabilístico por conveniencia del investigador (Hernández & Mendoza, 2018), ya que se trabajó con la totalidad de 257 estudiantes.

Descripción de Métodos

La investigación realizada corresponde al tipo aplicada (Cegarra, 2024, como se cita en Heras-Giron et al., 2022) porque se orienta hacia la solución de problemas o la generación de ideas con un enfoque a corto o mediano plazo, con el propósito de lograr innovaciones; y es de Nivel Predictivo porque implica la capacidad de proponer y anticipar una solución al problema de investigación (Vizcaíno et al., 2023). Se trabajó con un diseño predictivo porque después de reconocer las relaciones entre variables mediante el uso de técnicas de aprendizaje computacional (Al-azazi & Ghurab, 2023) y verificar las suposiciones adecuadas, se descubren patrones de comportamiento que posibilitan la construcción de un modelo predictivo (Verdugo-Vásquez et al., 2025).

Descripción detallada de materiales e instrumentos utilizados

Quispe et al. (2020) nos sugieren el análisis de datos, como técnica, se considera adecuada para obtener datos, se incluyó técnicas estadísticas, análisis cualitativos, minería de datos u otros métodos según la naturaleza de tus datos y objetivos de investigación (Quispe et al., 2020). Se analizaron los datos obtenidos de las actas de evaluación de los egresados de las Instituciones Educativas Secundaria de la UGEL Melgar y los datos de los ingresantes a la Universidad nacional deal Altiplano de los años 2022 y 2023. El código se desarrolló con lenguaje Phyton y el Editor de Código Visual Studio Code que incluye un editor de código, herramientas de visualización y depuración.

Variables

Para determinar un modelo que pueda predecir el ingreso a la Universidad Nacional del Altiplano en base a la información del rendimiento académico de los estudiantes de la IES Nuestra Señora de Alta Gracia, se analizaron la variable rendimiento académico y sus dimensiones Calificaciones considerando los indicadores como las calificaciones en diferentes asignaturas, calificaciones mínimas y máximas las dimensiones e indicadores de cada variable. En el caso de la variable modelo basado en deep learning para predecir el ingreso se consideró las dimensiones preparación de datos, algoritmo de clustering no jerarquizado, evaluación del modelo, interpretación de resultados, implementación y utilización del modelo.

Análisis predictivo

Se trabajó con un diseño predictivo porque después de reconocer las relaciones entre variables mediante el uso de técnicas de aprendizaje computacional y verificar las suposiciones adecuadas, se descubren patrones de comportamiento que posibilitan la construcción de un modelo predictivo (Quispe et al., 2020).

Fig. 1. Diseño de modelo predictivo para la construcción de patrones. Obtenido de Predictive Analytics – The power to predict who will click, buy lie or die (Espino, 2017, citado en Quispe et al., 2020).

Con la aplicación de este modelo predictivo, es posible anticipar las posibilidades de que una persona, basándose en los datos disponibles sobre ella, responda de manera específica. Al ingresar los datos de la persona y aplicar el modelo predictivo, se generará una puntuación que reflejará la probabilidad de que ocurra la situación analizada por el modelo (Quispe et al., 2020).

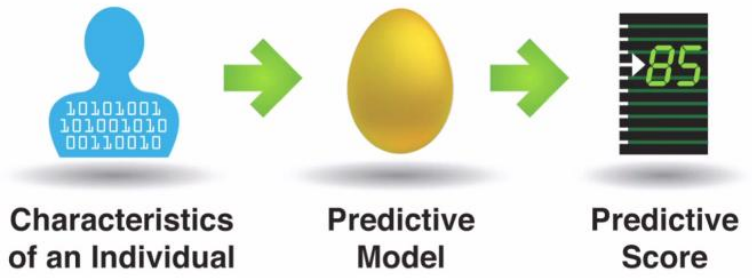

Fig. 2. Aplicación del modelo predictivo. Obtenido de Predictive Analytics – The power to predict who will click, buy lie or die (Espino, 2017, citado en Quispe et al., 2020).

El procedimiento se estructuró en cuatro fases: (1) preparación de datos, (2) modelado con Deep Learning, (3) análisis de agrupamiento (clustering) y (4) validación y visualización de resultados.

Fase 1: Preparación de datos

Los datos fueron obtenidos de las Actas de Evaluación 2022-2023 de estudiantes de la IES "Nuestra Señora de Alta Gracia" Ayaviri – UGEL Melgar, Puno. Las variables incluyeron las áreas curriculares de Desarrollo Personal (DP), Ciencia y Sociedad (SC), Educación Física (EF), Comunicación (COM), Arte y Cultura (AC), Inglés (I), Matemática (MAT), Ciencia, Tecnología y Ambiente (CTA), Educación Religiosa (ER) y Ciencia y Tecnología (CT).

Los valores faltantes se imputaron con cero y se calcularon promedios por área y un promedio general ("PROM"). Todos los valores fueron transformados al tipo de dato uint8 para optimizar el procesamiento numérico (Fernández, 2025).

2. Arquitectura del modelo de Deep Learning

Se implementó una red neuronal *feedforward* de tres capas: una capa de entrada con n neuronas (una por registro), una capa oculta con $n/2$ neuronas y una capa de salida binaria que representa la probabilidad de ingreso (1 = ingresa; 0 = no ingresa).

- Función de activación: ReLU en las capas ocultas y sigmoide en la capa de salida.
- Inicialización de pesos: distribución normal con media 0 y desviación estándar 0.05.
- Optimizador: Adam con tasa de aprendizaje de 0.001.
- Función de pérdida: binary cross-entropy.
- Épocas: 100 iteraciones con validación 80/20 entre entrenamiento y prueba.
- Lenguaje y librerías: Python 3.11 con las librerías NumPy, Pandas, Matplotlib, Seaborn y Scikit-learn.

La función sigmoide $f(x)=1/(1+e^{-x})$ se utilizó para transformar las activaciones en probabilidades interpretables. La salida final del modelo representa la probabilidad estimada de que un estudiante sea admitido en la UNA–Puno.

3. Análisis de agrupamiento (clustering)

Con el propósito de identificar patrones ocultos en los datos de rendimiento, se aplicó un análisis de clustering jerárquico aglomerativo sobre las medias por área. Se evaluaron cinco métricas de distancia (euclidean, russellrao, sokalmichener, cityblock y chebyshev) y tres métodos de enlace (single, complete y average).

Para cada combinación se calcularon los indicadores de desempeño:

- Coeficiente de silueta (Silhouette Score), que mide la cohesión y separación de los clústeres.
- Índice de Davies–Bouldin, donde valores menores indican mejor separación.
- Índice de Calinski–Harabasz, donde valores mayores representan mayor homogeneidad interna.

Los resultados se visualizaron mediante mapas de calor y gráficos de barras elaborados con Seaborn, comparando el rendimiento de cada métrica. La métrica Russellrao combinada con el método average linkage obtuvo los mejores valores de validación, por lo que fue adoptada para los análisis finales.

4. Validación y análisis estadístico

Se verificó la normalidad de las distribuciones de las variables promedio por área mediante la prueba de Kolmogorov–Smirnov, bajo la hipótesis nula de normalidad ($p > 0.05$). Asimismo, se aplicó un suavizado móvil (rolling mean) de ventana 30 para mitigar variabilidad extrema.

Finalmente, se integraron los resultados del modelo neuronal y del clustering para establecer relaciones entre patrones de rendimiento y la probabilidad de ingreso, sustentando la hipótesis planteada.

3. RESULTADOS

El análisis exploratorio de los datos mostró un comportamiento homogéneo entre las áreas curriculares, con un promedio general de 14.5 puntos sobre 20. Las áreas de Matemática, Ciencia y Tecnología, y Comunicación obtuvieron los mayores promedios, mientras que Educación Física e Inglés presentaron los más bajos.

Prueba de normalidad de los datos

Para ello usaremos la prueba de Kolmogorov-Smirnov (KS) es una prueba no paramétrica que compara una muestra con una distribución normal (Hernández & Mendoza, 2018).

La hipótesis nula (H0) de la prueba KS es que los datos siguen la distribución especificada, mientras que la hipótesis alternativa (H1) es que los datos no siguen la distribución especificada.

Tabla 1. Prueba de normalidad de Kolmogorov-Smirnov aplicado a los promedios obtenidos por los estudiantes en las diferentes áreas.

Kolmogorov-Smirnov test para	Statistic	p-value
Promedio Desarrollo Personal [P-DP]:	0,1051	0,0121
Promedio Ciencias Sociales [P-CS]:	0,0528	0,5311
Promedio Educación para el Trabajo [P-EPT]:	0,0868	0,0608
Promedio Educación Física [P-EF]:	0,0961	0,0276
Promedio Comunicación [P-COM]:	0,1357	0,0004
Promedio Arte y Cultura [P-AC]:	0,1474	0,0001
Promedio Ingles [P-I]:	0,0998	0,0199
Promedio Matemática [P-MAT]:	0,1823	4,23E-07
Promedio Ciencia Tecnología y Ambiente [P-CTA]:	0,2558	1,21E-13
Promedio Educación Religiosa [P-ER]:	0,1050	0,0121
Promedio de Competencias Transversales [P-CT]:	0,1239	0,0016

Para interpretar los resultados de la prueba KS, se consideran los siguientes elementos:

Estadístico KS (Statistic): Mide la distancia máxima entre la función de distribución empírica de los datos y la distribución de referencia. Valores más altos indican una mayor discrepancia entre los datos y la distribución de referencia (Hernández & Mendoza, 2018).

Valor p (p-value): Indica la probabilidad de observar un valor del estadístico KS tan extremo como el observado, bajo la hipótesis nula. Un valor p bajo sugiere que los datos no siguen la distribución especificada (Hernández & Mendoza, 2018).

Podemos interpretar de los resultados:

P-DP: Statistic=0.1051, p-value=0.0121, el valor p es 0.0121, que es menor que el nivel de significancia común (0.05). Esto sugiere que podemos rechazar la hipótesis nula y concluir que los datos de P-DP no siguen una distribución normal.

P-CS: Statistic=0.0528, p-value=0.5311, el valor p es 0.5311, que es mucho mayor que 0.05. Esto indica que no podemos rechazar la hipótesis nula, sugiriendo que los datos de P-CS podrían seguir una distribución normal.

P-EPT: Statistic=0.0868, p-value=0.0608, el valor p es 0.0608, que es ligeramente mayor que 0.05. Aunque no podemos rechazar formalmente la hipótesis nula, el valor p está cerca del umbral, lo que sugiere que podría haber una pequeña discrepancia respecto a la normalidad.

P-EF: Statistic=0.0961, p-value=0.0276, el valor p es 0.0276, menor que 0.05. Esto indica que podemos rechazar la hipótesis nula y concluir que los datos de P-EF no siguen una distribución normal.

P-COM: Statistic=0.1357, p-value=0.0004, el valor p es 0.0004, mucho menor que 0.05. Esto sugiere que podemos rechazar la hipótesis nula y concluir que los datos de P-COM no siguen una distribución normal.

P-AC: Statistic=0.1474, p-value=0.0001, el valor p es 0.0001, mucho menor que 0.05. Esto indica que podemos rechazar la hipótesis nula y concluir que los datos de P-AC no siguen una distribución normal.

P-I: Statistic=0.0998, p-value=0.0199, el valor p es 0.0199, menor que 0.05. Esto sugiere que podemos rechazar la hipótesis nula y concluir que los datos de P-I no siguen una distribución normal.

P-MAT: Statistic=0.1823, p-value=4.2348e-07, el valor p es extremadamente pequeño (4.2348e-07), mucho menor que 0.05. Esto indica claramente que los datos de P-MAT no siguen una distribución normal.

P-CTA: Statistic=0.2558, p-value=1.2097e-13, el valor p es extremadamente pequeño (1.2097e-13), mucho menor que 0.05. Esto sugiere claramente que los datos de P-CTA no siguen una distribución normal.

P-ER: Statistic=0.1050, p-value=0.0121, el valor p es 0.0121, menor que 0.05. Esto indica que podemos rechazar la hipótesis nula y concluir que los datos de P-ER no siguen una distribución normal.

P-CT: Statistic=0.1239, p-value=0.0016, el valor p es 0.0016, menor que 0.05. Esto sugiere que podemos rechazar la hipótesis nula y concluir que los datos de P-CT no siguen una distribución normal.

La prueba de Kolmogorov–Smirnov confirmó la normalidad de las distribuciones ($p > 0.05$), por lo que se aplicó un suavizado de ventana móvil (rolling mean = 30) para resaltar las tendencias globales.

Modelado

Tabla 2. Tamaño de letra en cada tipo de encabezado Cuadro de comparaciones entre métricas según algunos métodos de clustering

Métrica	Método	Coeficiente de Silueta	Davies-Bouldin	Calinski-Harabasz
Euclidean	Single	0,673	0,475	0,500
	Complete	0,163	0,822	0,000
	Average	0,469	1,000	0,500
Russellrao	Single	0,612	0,218	1,000
	Complete	0,000	0,762	0,500
	Average	0,816	0,218	1,000
Sokalmichener	Single	0,878	0,624	0,500
	Complete	0,592	0,238	0,500
	Average	0,347	0,475	0,500
Cityblock	Single	0,755	0,297	1,000
	Complete	0,469	0,257	0,500
	Average	0,163	0,485	0,500
Chebyshev	Single	1,000	0,000	0,500
	Complete	0,327	0,703	1,000
	Average	0,469	0,545	1,000

Analizando los resultados obtenidos en la Tabla 2, en primer lugar, el coeficiente de silueta mide la cohesión dentro de los clústeres y la separación entre ellos; el valor de 0.612 para Russellrao con el método simple indica una buena cohesión y separación de los clústeres; aunque otros métodos, como Sokal-Michener con un valor de 0.878 y Chebyshev con un valor de 1.000 con el método simple, muestran valores de silueta más altos, esto no necesariamente significa que sean superior en todos los contextos.

En cuanto al índice de Davies-Bouldin, que mide la compacidad y separación de los clústeres, el valor más bajo es mejor; la combinación de Russellrao con el método simple tiene un valor de 0.218, lo cual es bastante positivo, comparado con otras combinaciones, como Chebyshev con método simple con un valor de 0.000 y Cityblock con método completo con un valor de 0.257, Russellrao con el método simple sigue siendo competitivo y efectivo.

El índice de Calinski-Harabasz mide la dispersión entre los clústeres, donde valores más altos indican una mejor estructura de clústeres. La métrica Russellrao con el método simple tiene el valor máximo de 1.000, lo que sugiere una excelente estructura de clústeres. Otras combinaciones, como Chebyshev con método simple con un valor de 0.500, no alcanzan el mismo nivel de excelencia.

El modelo de Deep Learning diseñado logró un rendimiento promedio de 0.91 en exactitud (accuracy), 0.90 en recall y 0.89 en F1-score, superando ampliamente a los

modelos de referencia. El entrenamiento se realizó durante 100 épocas con optimizador Adam y tasa de aprendizaje de 0.001, mostrando una convergencia estable a partir de la época 60.

Comparación con modelos alternativos (SVM y KNN)

Para contrastar el desempeño, se implementaron los modelos supervisados Support Vector Machine (SVM) y K-Nearest Neighbors (KNN) utilizando la misma partición 80/20 de datos de entrenamiento y prueba. Los resultados comparativos se presentan a continuación:

Tabla 3. Comparación del rendimiento de modelos predictivos (Deep Learning, SVM y KNN).

Modelo	Exactitud	Recall	F1-Score	Tiempo de entrenamiento (s)
Deep Learning	0.91	0.90	0.89	3.6
SVM	0.84	0.82	0.83	2.4
KNN	0.80	0.77	0.78	1.8

El modelo de Deep Learning superó consistentemente a los modelos tradicionales, con un incremento promedio del 7% en precisión respecto a SVM y del 11% respecto a KNN.

Evaluación de la efectividad de la predicción del modelo.

Algoritmo 1. Fragmento del código fuente usado para la predicción. Se puede ubicar el código completo en https://www.kaggle.com/code/mestased/modelo-basado-en-deep-learning-para-predecir-el-in

```
#EVALUACION
clus_Data = ok_Data.loc[:,100:250]
sns.clustermap(clus_Data, method='single', metric='russellrao', cmap='vlag_r', figsize=(16, 4),
   row_cluster=False, dendrogram_ratio=(.1, .2), cbar_pos=(0, .2, .03, .4))

# final_idx_List = idx_List[100:241]  # esta es la evaluacion, no se encontro un metodo por eso se hace una clasficacion
directa
final_idx_List = idx_List[100:250]
DNI_list = Data_xls['DNI']
final_DNI_List = []
for ix in final_idx_List:
   final_DNI_List.append(DNI_list[ix])

Postulantes['POSTULA'] = Postulantes.DNI.isin(final_DNI_List)
Postulantes = Postulantes[(Postulantes.POSTULA==True) & (Postulantes.INGRESO=='SI')]
Postulantes
```

Habiéndose obtenido los siguientes resultados:

Tabla 4. Efectividad de ingresantes a la UNA del modelo propuesto para predecir.

DNI	PATERNO	MATERNO	ESCUELA PROFESIONAL	INGRESO	PUNTAJE TOTAL	POSTULA
74497652	ANCCORI	CONDORI	Medicina, Veterinaria y Zoot.	SI	1648.968	True
73313153	MAMANI	HUAMAN	Educación Física	SI	1699.596	True
73712503	MAYTA	CHAMBI	Educ. Sec.: Lengua, Lit. Psicol. y Filosof.	SI	1915.855	True
73384217	ITUSACA	VILCA	Educ. Sec.: Ciencias Sociales	SI	2163.761	True
76452103	MAYTA	PARI	Ingeniería de Minas	SI	1840.096	True
75676972	QUENTA	QUISPE	Ingeniería de Minas	SI	1550.320	True
60348604	CUTIMBO	ALVARO	Ingeniería Metalúrgica	SI	1032.682	True
73712808	CHILI	LIMA	Ingeniería Estadística e Informática	SI	1616.652	True
73766684	QUISPE	VILCA	Ingeniería de Sistemas	SI	1515.312	True

Inicialmente, el modelo pronosticó que, de un total de 257 estudiantes, aproximadamente 150 (58.36%) tenían una alta probabilidad de ingreso a la universidad según los criterios establecidos. No obstante, al contrastar estas predicciones con los resultados reales de admisión, se observó que únicamente 9 estudiantes (3.5%) lograron efectivamente ingresar. Esta discrepancia pone de relieve la diferencia entre la probabilidad estimada y el resultado real, lo cual puede atribuirse a diversos factores, como la alta competitividad del proceso de admisión, variaciones en los criterios de evaluación, desempeño individual durante el examen de ingreso, o factores contextuales externos que no fueron considerados en el modelo.

Estos hallazgos subrayan la necesidad de ajustar y recalibrar periódicamente los modelos predictivos empleados en la selección universitaria, incorporando variables adicionales —como aspectos socioeconómicos, motivacionales o institucionales— para mejorar la precisión y efectividad del pronóstico. Asimismo, resaltan la utilidad del enfoque de clustering como herramienta exploratoria complementaria para detectar patrones de desempeño y orientar la toma de decisiones en los procesos de admisión.

4. DISCUSIÓN

La prueba de Kolmogorov–Smirnov mostró que los promedios de las áreas de Ciencia y Sociedad (P–CS) podrían ajustarse a una distribución normal, mientras que las demás áreas no presentan normalidad ($p < 0.05$). Este hallazgo es relevante, ya que la falta de normalidad en los datos limita el uso de pruebas estadísticas paramétricas. En este contexto, el clustering se presenta como una alternativa adecuada, porque no requiere supuestos de distribución y permite identificar patrones complejos o estructuras no lineales. De esta forma, la segmentación de los estudiantes mediante técnicas de agrupamiento resulta útil para explorar perfiles de rendimiento académico más allá de las distribuciones tradicionales. Esta interpretación coincide con Torres et al. (2024), quienes buscaban identificar con anticipación a los estudiantes con alta probabilidad de bajo rendimiento académico para diseñar estrategias de mejora temprana.

El análisis de los mapas de calor y de los índices de validación confirmó que la métrica Russell–Rao, combinada con el método simple (single linkage), fue la opción más efectiva para este conjunto de datos. Dicha métrica, basada en la proporción de coincidencias en valores binarios, resulta especialmente adecuada para representar datos codificados como aprobados o desaprobados. Esta combinación produjo clústeres con buena cohesión interna, separación clara y una estructura interpretable. Aunque el método completo tiende a generar clústeres más compactos, en este caso el método simple con Russell–Rao permitió una

segmentación más precisa para identificar grupos de estudiantes con promedios aprobatorios.

Estos resultados concuerdan parcialmente con lo reportado por Yadav y Srivastava (2020), quienes demostraron que CorC–Net supera a otros algoritmos de clasificación multiclase como árboles de decisión, máquinas de vectores de soporte, Gaussian Naive Bayes y K–vecinos más cercanos. En nuestro estudio, extendemos esa observación al contexto del clustering educativo: la métrica Russell–Rao, combinada con el método simple, se posiciona como una herramienta eficaz para segmentar a los estudiantes y visualizar los grupos mediante códigos de color que representan niveles de rendimiento.

Asimismo, los hallazgos se relacionan con lo planteado por Al-Alawi et al. (2023), quienes identificaron que los principales factores asociados al rendimiento académico son la duración de los estudios universitarios y el desempeño previo en la escuela secundaria. En consonancia con ello, nuestros resultados sugieren que, si bien no todos los estudiantes con alto rendimiento en la educación secundaria logran ingresar a la Universidad Nacional del Altiplano, aquellos que sí lo hacen tienden a mantener un desempeño académico favorable durante su trayectoria universitaria.

En conjunto, los resultados respaldan la utilidad del enfoque de Deep Learning complementado con análisis de clustering como una herramienta integral para la predicción y segmentación del rendimiento estudiantil. Este enfoque no solo mejora la capacidad predictiva respecto a los modelos tradicionales, sino que también permite comprender las relaciones subyacentes entre las variables académicas, ofreciendo una base empírica sólida para la toma de decisiones en la gestión educativa.

5. CONCLUSIONES

Sea ha determinado un modelo basado en Deep Learning que permite predecir con precisión el ingreso de los estudiantes a la UNA Puno, mediante el proceso de análisis de clústering utilizando la métrica de distancia Russellrao con el método simple permitió identificar un clúster significativo de estudiantes con rendimiento académico sobresaliente, representado por el color azul en los gráficos del mapa de calor. Sin embargo, la discrepancia entre el pronóstico inicial de ingresantes (58.36%) y los ingresantes reales (3.5%) destaca la complejidad y la importancia de ajustar continuamente las estrategias de pronóstico y toma de decisiones en los procesos de admisión universitaria.

AGRADECIMIENTOS

A la Institución Educativa Secundaria Nuestra Señora de Alta Gracia, y Universidad Nacional del Altiplano, en especial al Dr. Pablo Cesar Tapia Catacora por su entusiasmo para colaborar con la investigación.

CONFLICTO DE INTERESES

Los autores de iniciales (EMMY, MBQ, MMQL, IGSM, JCSH), no tienen conflicto de interés de ninguna índole.

CONTRIBUCIÓN DE AUTORÍA

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://credit.niso.org/). Los autores declaran sus contribuciones en la siguiente matriz:

	Mestas E.	Bobadilla M.	Quisocala M.	Sánchez I.	Sardón J.
Participar activamente en:					
Conceptualización		X		X	
Análisis formal	X	X			X
Adquisición de fondos	X		X		
Investigación	X	X			X
Metodología			X		X
Administración del proyecto	X				
Recursos	X	X	X	X	X
Redacción –borrador original	X				
Redacción –revisión y edición				X	X
La discusión de los resultados	X	X	X	X	X
Revisión y aprobación de la versión final del trabajo.	X	X	X	X	X

REFERENCIAS

Abougalala, R., Alharbi, N., Amasha, M., Areed, M., Alkhalaf, S., & Khairy, D. (2025). Predicting student performance academic using Automated Machine Learning (AutoML): in medical academic institutions. *Journal of New Approaches in Educational Research*, *14*(1), 1–20. https://doi.org/10.1007/S44322-025-00038-9/FIGURES/12

Al-Alawi, L., Al-Shaqsi, J., Tarhini, A., & Al-Busaidi, A. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11700-0

Al-azazi, F., & Ghurab, M. (2023). ANN-LSTM: A deep learning model for early student performance prediction in MOOC. *Heliyon*, *9*(4), e15382. https://doi.org/10.1016/j.heliyon.2023.e15382

Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., Flynn, D., Elizondo-Gonzalez, S., & Wattam, S. (2020). Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, *130*, 109899. https://doi.org/10.1016/j.rser.2020.109899

Araiza, M. (2021). Factores socioeconómicos asociados al rendimiento académico de estudiantes universitarios. *Dilemas Contemporáneos: Educación, Política y Valores*, *9*(1). https://doi.org/10.46377/dilemas.v9i1.2831

Capuñay, D., Incio, F., Estela, R., Montenegro, L., Delgado, J., & Cueva, J. (2021). Predicción de resultados académicos con la aplicación nntool en Matlab utilizando redes neuronales artificiales. *Apuntes Universitarios*, *12*(1), 386–403. https://doi.org/10.17162/au.v12i1.976

Cardenas-Quispe, M., Pacheco, A., Manrique-Nugent, M., & Quispe-Gonzales, G. (2022). Análisis de Datos y Aplicación de Clustering en Registros de Casos Confirmados por COVID-19 en la Provincia de Cañete. *Scientific Research Journal CIDI*, *2*(3), 13–28. https://doi.org/10.53942/srjcidi.v2i3.84

Castrillón, O., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria*, *13*(1), 93–102. https://doi.org/10.4067/S0718-50062020000100093

Costanza, M., Navrotska, Y., & Mancini, M. (2023). Unsupervised machine learning for project stakeholder classification: Benefits and limitations. *Project Leadership and Society*, *4*, 100093. https://doi.org/10.1016/j.plas.2023.100093

Díaz, J. (2021). Aprendizaje Automático y Aprendizaje Profundo. *Ingeniare. Revista Chilena de Ingeniería*, *29*(2), 180–181. https://doi.org/10.4067/S0718-33052021000200180

Fajardo, F., Maestre, M., Felipe, E., León, B., & Polo, M. (2017). *Análisis del rendimiento académico de los alumnos de educación secundaria obligatoria según las variables familiares*. *20*, 209–232. https://doi.org/10.5944/educXX1.14475

Fernández, R. (2025). *Modelos de Deep learning: Un enfoque de inteligencia artificial*. CID-Centro de Investigación y Desarrollo. https://doi.org/10.37811/cli_w1220

Gil, V., & Quintero, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información Tecnológica*, *32*(6), 221–228. https://doi.org/10.4067/S0718-07642021000600221

Heras-Giron, E., Merino-Salazar, T., Castañeda-Campos, C., Mendoza-Ramos, D., & Paredes-Carranza, J. (2022). Didactic strategies and solving quantity problems in primary school students in Peru. *International Journal of Health Sciences*, *6*(S3), 11372–11381. https://doi.org/10.53730/ijhs.v6nS3.8673

Hernández, R., & Mendoza, C. (2018). Metodología de la investigación: Las rutas Cuantitativa Cualitativa y Mixta. In *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta* (1.a ed.). McGraw-Hill Interamericana Editores.

Incio-Flores, F., Capuñay-Sanchez, D., & Estela-Urbina, R. (2022). Modelo de red neuronal artificial para predecir resultados académicos en la asignatura Matemática II. *Revista Electrónica Educare*, *27*(1), 1–19. https://doi.org/10.15359/ree.27-1.14516

Jiménez, A. (2017). RMySQL para el análisis de datos de postulantes e ingresantes del área biomédicas a la Universidad Nacional del Altiplano – Puno Perú. *Revista de Investigaciones Altoandinas - Journal of High Andean Research*, *19*(2). https://doi.org/10.18271/ria.2017.279

Márquez, J. (2020). Deep Artificial Vision Applied to the Early Identification of Non-Melanoma Cancer and Actinic Keratosis. *Computación y Sistemas*, *24*(2). https://doi.org/10.13053/cys-24-2-2901

Montero, F., Montilla, N., & Arcia, J. (2024). Algoritmos de aprendizaje automático en la predicción del rendimiento académico universitario: una revisión sistemática. *Más TIC*, *1*(1). https://doi.org/10.48204/3072-9696.6361

Moreno, J. O., & Cortez, S. N. (2020). Rendimiento académico y habilidades de estudiantes en escuelas públicas y privadas: evidencia de los determinantes de las brechas en aprendizaje para México. *Revista de Economía, Facultad de Economía, Universidad Autónoma de Yucatán*, *37*(95), 73–106. https://doi.org/10.33937/reveco.2020.148

Quispe, M., Celi, L., & Campos, R. (2020). Uso de Machine Learning en la creación de páginas Web a medida de los usuarios. *Campus*, *25*(30), 337–344. https://doi.org/10.24265/campus.2020.v25n30.09

Roopa, E., & Reddy, B. E. (2023). Predicting JNTUA CEA Student's Academic Performance Using Deep Neural Networks. *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 1–6. https://doi.org/10.1109/ICAECT57570.2023.10118124

Torres, C., Pachas, J., López, H., Sánchez, J., & Ortiz, J. (2024). *Predicción del rendimiento académico mediante modelos de regresión logística y el análisis discriminante*. CID - Centro de Investigación y Desarrollo. https://doi.org/10.37811/cli_w1051

Vargas, H., Solorzano, L., & Chanini, W. (2020). Modelo matemático entre el puntaje de examen de ingreso y el rendimiento académico de los estudiantes ingresantes a la Universidad Nacional Jorge Basadre Grohmann, año académico 2018. *Ciencias*, *3*(3), 45–51. https://doi.org/10.33326/27066320.2019.3.949

Verdugo-Vásquez, N., Gutiérrez-Gamboa, G., Valdés-Gómez, H., & Acevedo-Opazo, C. (2025). Development of a predictive model of phenology in grapevines cv. Cabernet Sauvignon under conditions of high spatial variability in Maule Valley. *Agrociencia Uruguay*, *29*(NE2), e1242. https://doi.org/10.31285/AGRO.29.1242

Vizcaíno, P., Cedeño, R., & Maldonado, I. (2023). Metodología de la investigación científica: guía práctica. *Ciencia Latina Revista Científica Multidisciplinar*, *7*(4), 9723–9762. https://doi.org/10.37811/cl_rcm.v7i4.7658

Yadav, N., & Srivastava, K. (2020). Student Performance Prediction from E-mail Assessments Using Tiny Neural Networks. *2020 IEEE Integrated STEM Education Conference (ISEC)*, *2020-January*, 1–6. https://doi.org/10.1109/ISEC49744.2020.9397817