UNIVERSIDAD NACIONAL DE SAN MARTÍN

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Categorización a estudiantes universitarios en niveles de riesgo de deserción en base al algoritmo de aprendizaje no supervisado basado en densidad

Tesis para obtener el título profesional de Ingeniero de Sistemas e Informática

AUTOR:
Luis Gerardo Salazar Ramírez

ASESOR:
Ing. Dr. Miguel Ángel Valles Coral

Tarapoto – Perú

2022

UNIVERSIDAD NACIONAL DE SAN MARTÍN

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Categorización a estudiantes universitarios en niveles de riesgo de deserción en base al algoritmo de aprendizaje no supervisado basado en densidad

AUTOR:
Luis Gerardo Salazar Ramírez

Sustentada y aprobada el 09 de diciembre del 2022, por los siguientes jurados:

Ing. Mg. Richard Enrique Injante Oré
Presidente

Lic. M. Sc. Edwin Augusto Hernández Torres
secretario

Ing. Carlos Francois Hidalgo Reátegui
Vocal

Universidad Nacional de San Martín
Facultad de Ingeniería de Sistema e Informática
Jr. Vía Universitaria S/Nº - Ciudad Universitaria - Morales
Telef. (042) 525688 - 524074 - Anexo 109

ACTA DE SUSTENTACIÓN
PARA OPTAR EL TÍTULO DE INGENIERO DE SISTEMAS E INFORMÁTICA

En la Universidad Nacional de San Martín, Facultad de Ingeniería de Sistemas e Informática, bajo la Modalidad Virtual, en el Marco de la Emergencia Nacional por el COVID-19; a las 12:05 horas del día viernes 9 de diciembre del año 2022 mediante ZOOM por https://usnm-edu-pe.zoom.us/j, se reunieron los **miembros del Jurado Calificador,** integrado por:

Presidente	:	**ING. MG. RICHARD ENRIQUE INJANTE ORE**
Secretario	:	**LIC. EDWIN AUGUSTO HERNANDEZ TORRES**
Vocal	:	**ING. CARLOS FRANCOIS HIDALGO REATEGUI**

Para evaluar la Tesis "CATEGORIZACIÓN A ESTUDIANTES UNIVERSITARIOS EN NIVELES DE RIESGO DE DESERCIÓN EN BASE AL ALGORITMO DE APRENDIZAJE NO SUPERVISADO BASADO EN DENSIDAD" presentada por el Bachiller LUIS GERARDO SALAZAR RAMÍREZ, participando en calidad de asesor el Ing. Dr. Miguel Ángel Valles Coral.

Los señores miembros del Jurado, después de haber atendido la sustentación y evaluada las respuestas a las preguntas formuladas y terminada la réplica; luego de debatir entre sí, reservada y libremente lo declaran APROBADO por UNANIMIDAD con el calificativo de 19 (DIECINUEVE), equivalente a EXCELENTE en fe de lo cual firmamos la presente acta, siendo las 13:20 horas del mismo día, con lo que se dio por terminado el Acto de Sustentación.

.....
ING. M.G. RICHARD ENRIQUE INJANTE ORE
Presidente

.....
LIC. EDWIN AUGUSTO HERNANDEZ TORRES
Secretario

.....
ING. CARLOS FRANCOIS HIDALGO REATEGUI
Vocal

Constancia de asesoramiento

El que suscribe el presente documento, Ing. Dr. Miguel Ángel Valles Coral.

Hace constar:

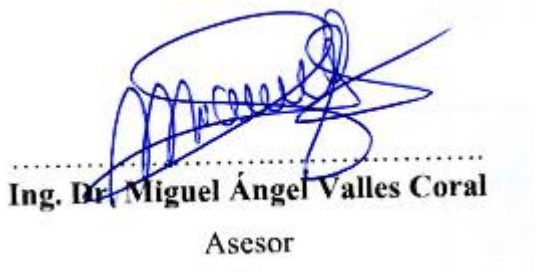
Que, he revisado la tesis titulada: **Categorización a estudiantes universitarios en niveles de riesgo de deserción en base al algoritmo de aprendizaje no supervisado basado en densidad,** en fechas del cronograma a fin de optimizar y agilizar la investigación, elaborada por el señor:

Bachiller en Ingeniería de Sistemas e Informática: **Luis Gerardo Salazar Ramírez**

La que encuentro conforme en estructura y en contenido. Por lo que doy conformidad para los fines que estime conveniente, y para que conste, firmo en la ciudad de Tarapoto.

Tarapoto, 09 de diciembre del 2022.

Atentamente:



Ing. Dr. Miguel Ángel Valles Coral

Asesor

Declaración de autenticidad

Luis Gerardo Salazar Ramírez, con DNI N° 70119118**,** bachiller de la Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de San Martín, autor de la tesis titulada: **Categorización a estudiantes universitarios en niveles de riesgo de deserción en base al algoritmo de aprendizaje no supervisado basado en densidad.**

Declaro bajo juramento que:

1. La tesis presentada es de mi autoría.
2. La redacción fue realizada respetando las citas y referencias de las fuentes bibliográficas consultadas.
3. Toda la información que contiene la tesis no ha sido auto plagiada;
4. Los datos presentados en los resultados son reales, no han sido alterados ni copiados, por tanto, la información de esta investigación debe considerarse como aporte a la realidad investigada.

Por lo antes mencionado, asumo bajo responsabilidad las consecuencias que deriven de mi accionar, sometiéndome a las leyes de nuestro país y normas vigentes de la Universidad Nacional de San Martín.

Tarapoto, 09 de diciembre del 2022.

Luis Gerardo Salazar Ramírez
DNI N° 70119118

Dedicatoria

El presente trabajo está dedicado con un inmenso y especial cariño a mi mamá, Berenice Ramírez Panduro, quien con todo el valor y la tenacidad que la caracteriza, inculcó en mí los valores y principios que me llevaron hasta este momento de mi vida, quien con su amor y aliento siempre me impulsó a seguir mis sueños, convirtiéndolos incluso en sus propios sueños y quien, con toda su fe, jamás dejó ni por un momento de creer en mí. Por su esfuerzo y amor incondicional, estaré eternamente agradecido.

Agradecimientos

Manifiesto mi más profundo agradecimiento a mi asesor, el Dr. Miguel Ángel Valles Coral, quien me animó a comenzar esta investigación y me brindó los conocimiento y herramientas necesarias para poder culminarla.

Agradezco también a mi familia y amigos, quienes me apoyaron durante toda la travesía que representó el llegar hasta este momento. Por último, agradecer a los docentes de la Universidad Nacional de San Martín, quienes depositaron en mí las enseñanzas necesarias para llevar a cabo esta investigación.

Índice general

Índice de tablas

Índice de figuras

Resumen

La etapa de formación superior universitaria, es un proceso que expone a los estudiantes a un estrés físico y mental prolongado, así como también a la autoexigencia con el fin de superar los retos que supone una carrera universitaria. Estos estímulos, de manera prolongada, van calando en su salud y estabilidad, tanto física como mental. Dicho desgaste, los expone a un determinado riesgo de deserción, el cual, posteriormente resulta crucial en la consecución o finalización de sus estudios. Es por esto que se planteó como objetivo categorizar a los estudiantes de la Universidad Nacional de San Martín en función al riesgo de deserción. Por lo cual, se realizó un estudio aplicado y de nivel descriptivo, bajo un diseño no experimental, utilizando una muestra de 670 estudiantes a los cuales se les proporcionó un conjunto de preguntas mediante una interfaz de chatbot sobre una plataforma web. Tras recopilar sus respuestas, estas fueron organizadas en una tabla de datos dentro de un archivo portable para que mediante un conjunto de técnicas de limpieza y normalización, fueran preprocesadas, esto con el fin de ser posteriormente sometidas al algoritmo DBSCAN. Luego de este procedimiento, los resultados fueron redimensionados para su visualización mediante PCA, así como también, fueron sometidos a técnicas de validación: a través de la proyección de los clústeres en la nube de puntos, y el análisis del coeficiente de Silhouette. Como resultado se obtuvieron 5 clústeres, con un coeficiente de Silhouette de 0.478, siendo finalmente etiquetadas y jerarquizadas con ayuda de un experto. Es así como se concluye que, se logró categorizar exitosamente a los estudiantes de la Universidad Nacional de San Martín en función al riego de deserción.

Palabras clave: Clusterización, DBSCAN, algoritmo de aprendizaje no supervisado, riesgo de deserción, PCA

Abstract

Higher education stage is a process that exposes students to prolonged physical and mental stress, as well as self-demanding in order to overcome the challenges of a university career. These stimuli, over a prolonged period, affect both physical and mental health and stability. Such attrition exposes students to a certain risk of dropping out, which subsequently proves to be crucial in the achievement or completion of their studies. The objective of this study was to categorize the students of the National University of San Martín according to their dropout risk. An applied and descriptive study was conducted under a non-experimental design, using a sample of 670 students who were provided with a set of questions through a chatbot interface on a web platform. Their responses were then organized into a data table in a portable file and preprocessed through a set of cleaning and normalization techniques in order to be subsequently subjected to the DBSCAN algorithm. After this procedure, the results were resized for visualization through PCA, as well as subjected to validation techniques, through cluster projection in the point cloud and Silhouette coefficient analysis. As a result, 5 clusters were obtained, with a Silhouette coefficient of 0.478, being finally labeled and hierarchized with the help of an expert. In conclusion, it was possible to successfully categorize the students of the National University of San Martín according to the risk of desertion.

Keywords: Clustering, DBSCAN, unsupervised learning algorithm, dropout risk, PCA

Introducción

La universidad se presenta dentro de la sociedad como una institución destinada a la enseñanza, investigación de nuevos conocimientos. El estudiante es la piedra angular sobre la cual se rigen los principales propósitos de dicha institución (Díaz-Méndez et al., 2019). Es por ello por lo que resulta fundamental contar con estrategias y mecanismos que aseguren su cuidado y permanencia, así como el cumplimiento de las condiciones adecuadas para que puedan llevar a cabo el correcto proceso de aprendizaje es fundamental. No asegurar dichas condiciones trae consigo no sólo el empobrecimiento en la calidad educativa, sino también en la deserción de estudiantes y la fuga de talentos (Noboa et al., 2018).

La etapa universitaria representa uno de los retos más importantes dentro de la vida de cada persona que la experimenta; en el Perú, en la mayoría de los casos, es una etapa que transcurre al término de la educación secundaria, momento por el cual, el estudiante experimenta una serie de cambios que van desde lo social hasta lo emocional. Durante el tiempo que dura esta etapa, el individuo es expuesto a un conjunto de nuevas experiencias y responsabilidades que requieren un alto esfuerzo físico y mental (Barreto Osama & Salazar Blanco, 2020).

Todo esto suele llevar al desgaste y la autoexigencia, una combinación que genera ansiedad, una respuesta normal e involuntaria que al incrementarse produce un conjunto de síntomas de carácter físicos y psicológicos que no sólo afectan notablemente el proceso de aprendizaje y desenvolvimiento social, sino que puede llegar a poner en riesgo su salud y hasta su propia vida (Vargas et al., 2018).

Durante la pandemia de COVID-19 muchos gobiernos tomaron medidas de aislamiento social, lo que generó un cambio repentino en el estilo de vida de la población, y que a su vez afectó negativamente la salud mental de las personas (Kumar & Nayar, 2020). Entre ellos, los estudiantes universitarios vieron limitados no sólo la forma de interacción entre su círculo social, sino también, vieron detenidas sus actividades académicas y en muchos casos también laborales, sumado a esto se encuentra toda la carga emocional que generó la crisis, dando como resultado el aumento de síntomas vinculados a la ansiedad y la depresión (Fruehwirth et al., 2021).

Aquellas medidas de aislamiento, además, representan una limitación y/o imposibilitan el poder llevar a cabo con normalidad los programas de seguimiento y acompañamiento psicológico de manera presencial, afectando no sólo el alcance, sino también los recursos de las universidades. Por eso, se hace urgente crear maneras más eficientes de realizar dichos programas (Feng & Zhang, 2021).

Actualmente en la Universidad Nacional de San Martín, existe un proyecto relacionado a la digitalización del proceso de tutoría a través de Chatbots, sin embargo, no existía el conocimiento necesario para proporcionar sustitutos a los expertos que evalúan la condición psicológica de los alumnos.

Esto debido a múltiples factores: en primer lugar, no existía una base de conocimiento y/o modelos que estén debidamente articulados y desde los cuales se pudiera desarrollar una propuesta de solución, además, no existían propuestas que integren los avances en las tecnologías de procesamiento de datos y tampoco un esfuerzo coordinado entre los profesionales y expertos de las TICs y la salud mental.

Son varias las consecuencias vinculadas a este problema: La baja o nula atención a los problemas de los estudiantes tiende a genera un aumento en la insatisfacción de los servicios brindados por la institución, lo que a su vez genera desinterés y baja motivación en el estudiante, crea una percepción negativa hacia la institución e incremente las manifestaciones de molestia de la población universitaria (Zulu & Mutereko, 2020).

No atender las necesidades psicológicas de los estudiantes, tiende a generar un conjunto de desórdenes mentales que impiden el pleno desempeño académico de los estudiantes (Brandão et al., 2017). El bajo rendimiento académico puede ser entendido también como el resultado de no aplicar las estrategias adecuadas y específicas al estilo de aprendizaje de los estudiantes (Freiberg-Hoffmann et al., 2017).

Es importante señalar ciertas limitantes a las que está sujeta la resolución de dicha problemática. Los cuales son: El desinterés o poco convencimiento de cierta parte de la comunidad universitaria hacia la idea del uso de las TICs para la resolución de los problemas expuestos (Huanca-Arohuanca et al., 2020). El limitado acceso a las TICs en los hogares de una parte de la población universitaria (Benites, 2021). Desinterés por parte de alumnos y tutores para participar en programas de seguimiento psicológicos y académicos (Rochin Berumen, 2021).

Bajo el contexto planteado, se consideró necesario encontrar una solución que brinde sustitutos a los mecanismos convencionales dedicados al seguimiento académico y emocional de los estudiantes para categorizar a dicha población en base al riesgo de deserción. Solución que integre los avances de las TICs como es el caso de la minería de datos y el machine learning, que planteen la base de conocimiento para próximos proyectos y que a su vez sirva como precedente del trabajo y esfuerzo conjunto entre profesionales de salud mental y de las TICs.

A través de la presente investigación, se pudo desarrollar un modelo de agrupamiento que pudo categorizar exitosamente a los alumnos de pregrado de la UNSM a través de la utilización de técnicas modernas de aprendizaje de máquina, así como también instrumentos de evaluación psicológica. Dicho modelo permitirá entender mejor el estado actual de los estudiantes de pregrado la UNSM, facilitando así el planteamiento de más y mejores alternativas de solución.

El objetivo general planteado fue categorizar a los estudiantes universitarios en función al riesgo de deserción en la Universidad Nacional de San Martín. Y los objetivos específicos fueron: I) Plantear una propuesta de solución que integre los nuevos avances de las TICs; II) Desarrollar una base de conocimiento desde la cual desarrollar una solución; III) Generar una solución integral con el esfuerzo conjunto entre profesionales de la salud mental y de las TICs.

El presente estudio está estructurado en tres capítulos principales: Capítulo I) en el que se muestra los antecedentes de la investigación, las bases teóricas y definición de términos básicos; Capítulo II) en el cual se describe el tipo, nivel y diseño de la investigación, la población y muestra, técnicas y validación de los instrumentos de recolección de datos y los métodos de procesamiento y análisis; y el capítulo III) donde se presentan los resultados y discusión de la investigación en torno a los objetivos planteados. Por último, se presentan las conclusiones, recomendaciones, referencias bibliográficas y anexos derivados.

CAPÍTULO I
REVISIÓN BIBLIOGRÁFICA

1.1 Antecedentes de la investigación

Según Masud et al. (2020) en su investigación "Monitoreo no intrusivo de comportamiento y patrones de movimiento para detectar el nivel de severidad de depresión clínica a través del smartphone", introduce un método cuyo objetivo es evaluar el nivel de depresión de un individuo usando el smartphone para el monitoreo de sus actividades diarias. Para ello, hace uso del smartphone del individuo, así como el cuestionario de salud del paciente de nueve ítems (PHQ-9). Las características del tiempo del sensor de aceleración del smartphone son tomadas junto a una máquina de soporte vectorial (SVM) para clasificar las actividades físicas. Además, la información de ubicación geográfica se agrupa mediante un sensor GPS de teléfono inteligente para simplificar los patrones de movimiento. Se extrajeron un total de 12 características de la actividad física y los patrones de movimiento de las personas y se analizaron junto con sus puntajes semanales de depresión utilizando el Cuestionario de salud del paciente de nueve ítems. Utilizando un método de selección de características de envoltura (wrapper method), se seleccionó un subconjunto de características y se aplicó a un modelo de regresión lineal para estimar la puntuación de depresión. Después se utiliza el algoritmo de máquina de soporte vectorial para clasificar el nivel de severidad de depresión entre los individuos. Como resultado este método tiene una precisión del 87.2% en los casos de depresión severa, por lo que supere a otros modelos de clasificación, incluyendo el k-nearest neighbor (vecinos más cercanos) y redes neuronales artificiales. En conclusión, este método es una solución costo-efectiva de largo plazo para la identificación de depresión en individuos y puede lograrlo sin invadir su espacio personal o generar molestias en su día a día.

Según Piorecký et al. (2019) en su investigación "Un método robusto para el diagnóstico temprano trastorno del espectro autista a partir de señales en el EEG basado en la selección de características y el método DBSCAN" tiene como objetivo presentar un método robusto para el diagnóstico temprano de ASD a partir de señales en el EEG. La población del estudio consiste en 34 niños de entre 3 a 12 años de edad con ASD y 11 niños sanos en el mismo rango de edad. El método propuesto utiliza características lineales y no lineales tales como espectro de poder, transformada de Wavelet, transformada rápida de fourier (FFT), dimensión fractal, dimensión de correlación, exponente de Lyapunov, entropía,

análisis de fluctuación sin tendencia y probabilidad de sincronización para la descripción de las señales del EEG, así como la utilización de agrupamiento basado en densidad para la eliminación de artefactos y brindar robustez. Además, la selección de características se aplica en base a diferentes criterios tales como Información Mutua (MI), Ganancia de Información (IG), Mínima Redundancia Máxima Relevancia (mRmR) y Algoritmo Genético (GA). Por último, para la decisión final se utilizan los clasificadores K-Nearest-Neighbor (KNN) y Support Vector Machines (SVM). Como resultado, la investigación indica que la precisión de clasificación del enfoque que utiliza SVM es del 90,57%, mientras que para KNN es del 72,77%. Además, la sensibilidad del método propuesto es del 99,91% para SVM y del 91,96% para KNN.

Ahuja & Banga (2019) en su investigación tiene como objetivo analizar el estrés en los estudiantes universitarios en diferentes momentos de su vida, para ello calcula el estrés mental de los estudiantes una semana antes de un examen y durante el uso de Internet. El conjunto de datos se tomó del Instituto Jaypee de Tecnología de la Información y consistió en datos de 206 estudiantes. Se aplican cuatro algoritmos de clasificación: Regresión lineal, Naïve Bayes, Random Forest y SVM, y la sensibilidad, la especificidad y la precisión se utilizan como parámetro de rendimiento. La precisión y el rendimiento de los datos se mejoran aún más mediante la aplicación de la validación cruzada de diez fases. La precisión más alta registrada fue por algoritmo de máquina de soporte vectorial (Support Vector Machine) (85,71%).

Xie et al. (2021) en su investigación "Predicción de enfermedades cardiovasculares usando información del peso para el aprendizaje basado en densidad" busca resolver el problema de la distribución desigual de los puntos de muestra en los conjuntos de datos (datasets) médicos. Para ello, utilizando el algoritmo de agrupamiento espacial de aplicaciones con ruido (DBSCAN) basado en la densidad, se propone un enfoque de aprendizaje ponderado para utilizar la información de densidad de conjuntos de datos para la predicción precisa de enfermedades cardiovasculares (ECV). Selecciona características importantes mediante el algoritmo de bosque aleatorio (Random Forest), divide los puntos de muestra en tres tipos y los pondera utilizando diferentes valores por aprendizaje de peso según la densidad. Sus resultados muestran que, en comparación con los modelos de aprendizaje automático convencionales, el enfoque de validación cruzada mostró que el rendimiento de los modelos de aprendizaje automático con aprendizaje por peso podría lograr una precisión mejorada en 3 puntos porcentuales con el conjunto de datos de Stroke y más de 10 puntos

porcentuales con los de la Universidad de California, Irvine (UCI). Concluye en que, los modelos de aprendizaje automático construidos que combinan las características originales y la característica de peso pueden aprender información de densidad, identificar de manera más efectiva los límites de decisión y lograr un mejor rendimiento.

Li et al. (2021) en su trabajo titulado "Un enfoque de agrupamiento de conjuntos no supervisado para el análisis de patrones de comportamiento de los estudiantes", buscan poder crear reglas específicas basándose en los patrones de comportamiento de los estudiantes para poder ser aplicados especialmente con patrones inesperados. Para llevarlo a cabo, proponen un framework de clustering no supervisado con el fin de utilizar datos del comportamiento de los estudiantes para descubrir patrones. El procedimiento propuesto extrae características de comportamiento de las dos perspectivas de la estadística y la entropía y luego combina el algoritmo DBSCAN y algoritmos k-means para descubrir patrones de comportamiento. Para evaluar el desempeño del framework propuesto, se llevan a cabo experimentos con seis tipos de datos de comportamiento producidos por estudiantes universitarios en una universidad de Beijing y se analizan las relaciones entre los diferentes patrones de comportamiento y los promedios de calificaciones de los estudiantes. Los resultados muestran que el framework no solo puede detectar patrones de comportamiento anómalos, sino también encontrar patrones convencionales. Se concluye que los departamentos de estudiantes pueden adoptar medidas más específicas para la intervención y los servicios especializados a través de los resultados del clustering.

1.2 Bases teóricas

1.2.1 Clustering (Agrupamiento)

Es una metodología por lo general de aprendizaje no supervisado, en el cual los objetos en un conjunto de datos son acumulados de acuerdo con sus características similares tales como distancia espacial y es comúnmente usado tengo una gran cantidad de campos de estudio.

Existen diferentes enfoques para clasificar lo que caracteriza distintos grupos en los datos.

Desde un punto de vista procedimental, muchos métodos de agrupamiento intentan encontrar una partición de los datos en k grupos, de modo que las diferencias dentro del grupo se minimicen mientras que las diferencias entre grupos se maximizan. Las nociones de disimilitud dentro del conglomerado y disimilitud entre conglomerados se definen utilizando la función de distancia d dada.

Desde un punto de vista estadístico, tales métodos corresponden a un enfoque paramétrico, donde se supone que la densidad desconocida $p(x)$ de los datos es una mezcla de k densidades $pi(x)$, cada una de las cuales corresponde a uno de los k grupos en los datos; se supone que pi (x) proviene de alguna familia paramétrica (por ejemplo, distribuciones gaussianas) con parámetros desconocidos, que luego se estiman a partir de los datos.

- Por el contrario, la agrupación basada en densidad es un enfoque no paramétrico, donde los grupos en los datos se consideran las áreas de alta densidad de una densidad $p(x)$. Los métodos de agrupación basados en densidad no requieren el número de agrupaciones como parámetros de entrada, ni hacen suposiciones sobre la densidad subyacente $p(x)$ o la varianza dentro de los grupos que pueden existir en los datos. En consecuencia, los clústeres basados en densidad no son necesariamente grupos de puntos con alta similitud dentro del conglomerado como lo mide la función de distancia d, pero pueden tener una “forma arbitraria” en el espacio de características; a veces también se les llama "agrupaciones naturales". (Webb et al., 2011)

1.2.2 Clustering basado en densidad

La agrupación en clústeres basada en la densidad de una nube de puntos hace referencia a métodos de aprendizaje no supervisados cuya función es la identificación de grupos (clústeres) específicos en los datos, basados en que un clúster es una región dentro de un espacio de datos contigua de alta densidad de elementos, disociada de otros clústeres similares por zonas contiguas de punto bajo. densidad. Los datos ubicados en las regiones de separación de baja densidad de puntos generalmente se consideran ruido o valores atípicos (Webb et al., 2011).

La densidad de datos puede evaluarse analizando la vecindad de cada objeto de datos. Existen dos formas posibles de definir la vecindad de un objeto.
Primero, cuando el radio de vecindad de un objeto se expresa como aquella distancia euclidiana al k-ésimo vecino más próximo, el tamaño de vecindad se define dinámicamente dependiendo de la densidad de datos. Los vecindarios de los objetos son relativamente pequeños en regiones densas del espacio de datos y son considerablemente más grandes en regiones menos densas del espacio de datos. La segunda opción es asumir el mismo radio de vecindad para todos los objetos de datos mientras se agrupan los datos como se hace en el método DBSCAN (Daszykowski & Walczak, 2009).

1.2.3 DBSCAN

Agrupamiento espacial basado en densidad de aplicación con ruido (DBSCAN), es un algoritmo de agrupamiento basado en densidad, en la que la densidad de datos en una vecindad de un radio predefinido se evalúa para cada objeto y se expresa como el número de objetos en esa vecindad. Por lo tanto, se pueden identificar tres tipos de objetos de datos en la agrupación de DBSCAN, estos son: Objetos centrales (core), objetos fronterizos (border) y objetos periféricos (outlying). (Ester et al., 1996)

- El objeto central contiene un número predefinido de objetos, ***k***, en su vecindad de radio ***r***.
- El objeto de datos se denomina objeto de borde si hay menos de ***k*** objetos en su vecindad de radio ***r***, pero al menos uno de ellos es un objeto central
- El llamado objeto de datos periférico es un objeto con menos de ***k*** objetos en su vecindad de radio ***r***, y ninguno de ellos es un objeto central

Un grupo de objetos es una región en el espacio de datos, donde la densidad de datos está por encima de un valor de umbral predefinido. Las vecindades de objetos que satisfacen esa condición se fusionan, de modo que los clústeres contienen objetos centrales y fronterizos (los objetos periféricos no pertenecen a ningún grupo). Los pasos principales del algoritmo DBSCAN se pueden resumir de la siguiente manera:

1. Se define el radio de la vecindad (r), y el número mínimo de objetos considerados como un grupo (k) dentro de un radio r.

2. Se determina un objeto central entre los objetos aún no procesados y se inicia una nueva clase *g*:
   a. Se marca este objeto como ya procesado y recupera todos sus vecinos en la vecindad de radio r.
   b. Se asigna los vecinos al grupo *g* y se les agrega a la llamada lista de semillas (*seeds list*).

3. Hasta que la lista de semillas no esté vacía:
   a. Se selecciona cualquier objeto de la lista de semillas, se marca como procesado y este es eliminado de la lista de semillas.
   b. Se obtiene todos sus vecinos dentro de la vecindad de radio r, se agrega a la lista de semillas aquellos vecinos que aún no están incluidos en la lista de semillas y se asigna a la clase g.

4. Una vez que la lista de semillas esté vacía, se configura $g \leftarrow g+1$ y se regresa al paso 2.
5. Si no hay objetos centrales entre los objetos sin procesar, estos no pertenecerán a ningún grupo y, por lo tanto, son considerados objetos periféricos (outlying objects).

1.2.4 Técnicas de validación de clústeres

Uno de los más importantes problemas en el ámbito del análisis de clústeres es la validación o evaluación de los resultados para encontrar la cantidad de grupos o clústeres que mejor se adapte a los datos proporcionados. Este es el principal reto de la validez de los clústeres.

A continuación, discutiremos los conceptos fundamentales de esta área mientras presentamos los diversos enfoques de validez de clústeres propuestos en la literatura.

Los métodos de agrupamiento tienen el objetivo de descubrir grupos característicos presentes en un universo de datos. Por lo general, se tiende a buscar agrupaciones cuyos miembros estén estrechamente próximos entre sí (es decir, tengan un alto grado de similitud) y que estos estén bien separados de los demás clústeres (Draszawka & Szymański, 2011).

Un reto al que enfrentar es la decisión del número adecuado de clústeres que se ajusten a un conjunto de datos dados. Para aquellos casos en los que los conjuntos de datos sólo tienen dos dimensiones, el espectador puede comprobar con relativa facilidad de manera visual la validez de los resultados (es decir, qué tan bien el algoritmo descubrió las agrupaciones del conjunto de datos). Para este tipo de situaciones, claramente la verificación visual de los resultados de la agrupación dentro del espacio de datos es crucial. Por otro lado, en aquellas situaciones en las que el conjunto de datos tiene carácter multidimensional (tres o más dimensiones), tanto visualizar de manera eficaz el conjunto de datos, como percibir los clústeres a través de las diferentes herramientas visuales de verificación disponibles es una tarea extremadamente complicada para las limitadas capacidades de los seres humanos, por lo que no es posible percibir ni asimilar espacios de una gran cantidad de dimensiones. (Theodoridis & Koutroumbas, 2008)

El término: validez de agrupamiento, hace referencia al proceso de comprobación de los resultados de un algoritmo de clusterización. A grandes rasgos y dependiendo de la

disponibilidad de información externa de los conjuntos de datos, existen tres enfoques para verificar la validez de los clústeres (Haouas et al., 2017)

1.2.4.1. Enfoque externo

La validación externa se puede utilizar cuando la partición real de los datos agrupados se conoce a priori. Al conocer las categorías (o clases) de los objetos de datos, podemos compararlos con los clústeres creados por un algoritmo. Es sabido que la validación externa a comparación de los demás tipos de validación (interna y relativa), es más precisa. Este tipo de validación es especialmente importante, al momento que uno intenta encontrar el mejor método de agrupamiento para un objetivo específico y usualmente usa una variedad de algoritmos en un determinado conjunto de datos con una estructura de clases bien conocida.(Brun et al., 2007)

Cuando la división correcta está disponible, los resultados del algoritmo de agrupamiento se comparan con ella basándose en los índices de validez externos, los cuales incluyen: Índice de Rand (Pfitzner et al., 2009), índice de Rand ajustado (Rand, 1971), índice de Jaccard (Bakshi et al., 2014), etc.

Campo et al. (2016) propuso un índice de validez externa que podría aplicarse a clústeres superpuestos sobre la base del enfoque probabilístico intuitivo, con la que se podría medir correctamente la similitud entre conjuntos de datos con diferentes niveles de superposición. Lee et al. (2018) propuso un índice de validez basado en la calificación de datos de vectores de soporte (SVDD), en el que la compacidad de un grupo se mide en el espacio del núcleo. Los resultados de su experimento muestran que SVDD se puede aplicar a conjuntos de datos con grandes superposiciones y valores atípicos, especialmente para formas arbitrarias y superpuestas.

1.2.4.2. Enfoque interno

Las técnicas de validación interna emplean el hecho de que los clústeres son conjuntos de objetos compactos y bien separados.

Cuando la información externa del conjunto de datos no está disponible, los resultados de la partición son utilizados para la validación de la calidad de la partición, estos índices de validación se denominan “índices de validación internos”, los cuales se basan en la estrechez, la separabilidad, la conectividad y el grado de superposición, así como de otros aspectos que describen la información de la estructura geométrica de los datos. Desafortunadamente, la conclusión que se repite a menudo de estos estudios es que no

existe un mejor índice de validación interna, porque todos dependen de los datos.(Draszawka & Szymański, 2011)

Algunos de estos índices solo consideran la información de la estructura geométrica del conjunto de datos, como: Índice de Dunn (Dunn, 1973), índice DB (Davies & Bouldin, 1979), índice CH (Caliñski & Harabasz, 1974) y el índice de Silhouette (Rousseeuw, 1987). Algunos de estos indicadores solo consideran la estrechez, y algunos solo consideran la separabilidad, mientras que algunos consideran tanto la estrechez como la separabilidad, y utilizan la suma o el cociente de estas dos medidas para formar nuevos índices de validez. (Hartigan et al., 1981) presenta el índice de Hartigan y lo aplica al algoritmo de agrupación de K-means para determinar el número óptimo de agrupaciones; (Krzanowski & Lai, 1988) propusieron un índice de "modelo de codo" llamado KL. Cuando el valor del índice aumenta o disminuye con el valor equivalente del número de clústeres k, el valor que toma k es correspondientemente la cantidad óptima de clústeres. (Kashyap & Bhattacharya, 2017) propusieron el índice de densidad espacial (SDI) basado en la medición de la densidad, que puede identificar eficazmente el número de conglomerados. (Haouas et al., 2017) con base en la distancia dentro del clúster y la distancia entre clústeres, propuso el índice HF, que toma en consideración las medidas de estrechez y separabilidad. El número de clústeres correspondiente al valor mínimo de HF es el número óptimo de clústeres.

1.2.4.3. Enfoque relativo

El enfoque relativo para la validación de clústeres se basa en repetir el mismo algoritmo de agrupación varias veces utilizando diferentes parámetros y eligiendo los resultados más estables. Por ejemplo, si el número de clústeres es uno de los parámetros de entrada, se intenta agrupar utilizando varios números de clústeres y se elige aquel para el que los índices internos son mejores. Si el número de clústeres no es el parámetro, la validación relativa permite elegir los valores de los parámetros que se encuentran en el medio del rango más amplio para el que el número de clústeres es constante. (Halkidi et al., 2001)

Se establece un objetivo de decisión antes del agrupamiento, luego se ejecuta el algoritmo de clustering varias veces bajo diferentes parámetros y se evalúa el resultado de acuerdo con los criterios preestablecidos de evaluación de clústeres, por último, se determina el número óptimo de clústeres y el resultado de división óptimo. (Rojas-Thomas et al., 2017)

Debido a que el clustering es un mecanismo de aprendizaje no supervisado y la información externa generalmente no está disponible, los índices de validez internos son los más utilizados. (Li et al., 2020)

1.2.5 Deserción universitaria

El comprender la deserción de los estudiantes de educación de nivel superior se ha enmarcado en un vasto conjunto de perspectivas y teorías de análisis, que por lo general han tenido las características de los estudiantes tradicionales como punto de inicio (Carvajal & Cervantes, 2018).

Bean & Metzner (1985), llevaron a cabo el desarrollo de un modelo conceptual dirigido a estudiantes no tradicionales considerando la escasez de fundamentos teóricos para orientar la investigación sobre la deserción universitaria entre este tipo de población estudiantil. La estructura de dicho modelo indica que la decisión de continuar o salir en la universidad está directamente relacionada con un grupo de variables: antecedentes y características individuales del estudiante, variables psicológicas, académicas, institucionales y ambientales (Khuong, 2014).

Una vez completado un proceso de categorización y agrupamiento (clustering), (Carvajal & Cervantes, 2018) sugieren cuatro categorías de factores: El primero se refiere a las condiciones y características personales y situacionales de los estudiantes que intervienen en su permanencia. El segundo examina los factores académicos que impactan sobre la decisión de abandonar la escuela. El tercero investiga factores circunstanciales, que ocurren inesperadamente influyendo en la permanencia de los estudiantes. Por último, el cuarto hace referencia a la oferta universitaria y las experiencias que los estudiantes tuvieron con ella durante el tiempo en el que se llevaron estudios en la institución.

Cuando se trata de variables de entrada para predecir la deserción, existen diferentes modelos y características de entrada que se encuentran en la literatura, algunos enfocados en analizar un solo tipo de datos, mientras que otros mezclan diferentes grupos de datos (Heredia-Jimenez et al., 2020).

Con respecto a un solo tipo de datos, por ejemplo, en el estudio según Ting (2001), los autores utilizaron variables psicosociales con un modelo estadístico en 690 estudiantes, y los resultados oscilaron entre el 11,8% y el 22% de la varianza. El estudio de Viloria & Parody (2016), se centró en analizar datos académicos como el primer grado parcial y el

porcentaje de inasistencia acumulado de 171 alumnos. Desafortunadamente, no se menciona la tasa de predicción. El estudio de Kappe & Van Der Flier (2012) combinó la inteligencia, la personalidad y los predictores de motivación en 137 estudiantes y los resultados mostraron un 33% de la varianza en el GPA y un 30% de la varianza en el tiempo hasta la graduación.

En términos de mezclar diferentes grupos de datos, en Aulck et al. (2016), los autores utilizaron información demográfica, de ingreso a la universidad y registros completos de calificaciones de 69,116 estudiantes, generando un total de más de 700 características y tuvieron una precisión del 66%. En Ameri et al. (2016), los autores utilizaron el modelo de riesgos proporcionales de Cox con distintos grupos de variables, como demografía, antecedentes familiares, información financiera, preparatoria, matrícula universitaria y créditos semestrales de 11,121 estudiantes. La precisión de su modelo osciló entre el 71% y el 82%. En Jiménez et al. (2019), los atributos utilizados fueron sobre el comportamiento del estudiante en la primera parte de su programa de pregrado de 498 estudiantes, ignorando los atributos clásicos. La precisión de sus modelos osciló entre el 74% y el 78%. Otro modelo predictivo, así como un EWS, es presentado por Baneres et al. (2019), que solo utiliza el rendimiento académico y los datos recopilados de LMS, los rangos de precisión van del 79% al 92%.

Si bien todos los modelos anteriores son precisos en sus contextos, no son escalables porque usan muestras pequeñas, solo usan los primeros años académicos o se enfocan en campos específicos o programas de pregrado. Además, una combinación de diferentes variables, como psicológicas, demográficas y académicas, podría provocar resultados sesgados o predicciones menos precisas. Además, ciertos modelos no se pueden replicar porque las variables no se especifican, sino que se mencionan de manera general. Por lo tanto, se necesita un modelo que pueda ser escalable y replicable para comprender el camino del estudiante de principio a fin y considerar las variables que causan un impacto en el rendimiento del aprendizaje (Heredia-Jimenez et al., 2020).

1.3 Definición de términos básicos

Inteligencia artificial:

Es un subcampo de la informática, dedicado a proporcionar a las computadoras capacidades para la resolución inteligente de problemas, es decir, resolver problemas complejos de una manera que consideraríamos inteligente (Hügle et al., 2021).

Aprendizaje máquina (ML):

El machine learning o aprendizaje automático, un subcampo de la inteligencia artificial, proporciona algoritmos (secuencias de instrucciones informáticas bien definidas que resuelven un problema específico) que construyen modelos matemáticos basados en datos muestreados (Hügle et al., 2021).

Aprendizaje supervisado:

En el aprendizaje supervisado, los modelos de aprendizaje automático se entrenan en base a ejemplos dados, que consisten en entradas y salidas deseadas proporcionadas por un experto (por ejemplo, el médico). Los resultados pueden ser un conjunto de categorías (por ejemplo, enfermedad activa, moderada o en remisión) o pueden ser números (por ejemplo, la puntuación absoluta DAS28). Los modelos entrenados para generar una selección de categorías (por ejemplo, nivel de enfermedad bajo, moderado o activo) se denominan modelos de clasificación. Los modelos entrenados para generar números reales se denominan modelos de regresión. La diferencia entre clasificación y regresión se muestra en la Figura 3 para el ejemplo de predicción de niveles de enfermedad (Heredia-Jimenez et al., 2020).

Aprendizaje no supervisado:

El aprendizaje no supervisado tiene como objetivo encontrar patrones o estructuras subyacentes aún desconocidos en datos sin etiquetar. Los métodos de aprendizaje no supervisados se utilizan a menudo para procesar grandes bases de datos, como grandes grupos de pacientes. También pueden agrupar a los pacientes (subdividiéndolos en grupos) y caracterizar valores atípicos u otras características importantes. Pueden también usarse para reducir la dimensionalidad de los datos dados (Hügle et al., 2021).

Dataset:

Un dataset o conjunto de datos es una colección de datos que se utiliza para algún propósito específico de aprendizaje automático (Sammut & Webb, 2010).

Máquinas de vectores de soporte (SVM):

Las máquinas de vectores de soporte (SVM) son una clase de algoritmos lineales que se pueden utilizar para tareas de clasificación, regresión, estimación de densidad, detección de nuevas características en los datos y otras aplicaciones. En el caso de clasificar en dos clases, esta clase de algoritmos encuentran un hiperplano que divide las dos categorías de

datos con el mayor margen posible. Todo esto supone una gran precisión para generalizar en datos no descubiertos y permite la utilización de métodos de optimización específicos que permiten que las SVM aprendan de una gran cantidad de datos (Sammut & Webb, 2010).

Random forest (Bosque aleatorio):

Es un algoritmo de aprendizaje basado en conjuntos de árboles de decisión que es utilizado para la regresión y la clasificación concernientes al pronóstico de una variable continua. Es un híbrido del algoritmo Bagging con el método del subespacio aleatorio (random subspace), y hace uso de árboles de decisión como base de su clasificador. Cada árbol se crea a partir de un conjunto muestra aleatorio del conjunto de datos original (Sammut & Webb, 2010).

CAPÍTULO II
MATERIALES Y MÉTODOS

2.1 Tipo y nivel de investigación

La presente investigación fue del tipo aplicada, dado a que los resultados generados pudieron ser aplicados de manera inmediata para la solución de un problema real.

El nivel fue descriptivo, pues se llevó a cabo una recopilación de datos desde la cual se llevó a cabo la observación y posterior procesamiento para obtener una solución.

2.2 Diseño de investigación

El diseño de la presente investigación fue no experimental de desarrollo tecnológico, pues no se manipuló ninguna variable y sólo se limitó al estudio y análisis de los datos preexistentes para desarrollar una solución que pudiera mejorar las técnicas actuales.

2.3 Población y muestra

2.3.1 Población:

El presente proyecto presenta una población de estudio conformada por los estudiantes de pregrado matriculados durante el semestre 2021-II, siendo un total de 5575 individuos (fuente: Oficina de Asuntos Académicos, semestre 2021-I).

2.3.2 Muestra:

Para el cálculo de la muestra se utilizó la fórmula de población finita:

$$n = \frac{z^2(p * q)}{e^2 + \frac{(z^2(p * q))}{N}}$$

Siendo:

- n = Tamaño de la muestra.
- z = Nivel de confianza deseado, en este caso, 0.95
- p = Proporción de la población con la característica deseada, en este caso, 0.5
- q = Proporción de la población sin la característica deseada, en este caso, 0.5
- e = Nivel de error dispuesto a cometer, en este caso, 0.017
- N = Tamaño de la población, en este caso, 5575

Como resultado, se obtuvo un valor de 670.88, lo que se tradujo a un valor de la muestra de 670 estudiantes de pregrado matriculados en el semestre académico 2021-II. Tras la obtención de la muestra, se dispuso la selección bajo el criterio de selección aleatoria simple por escuela.

2.4 Técnicas e instrumento de recolección de datos

Para el proceso de recopilación de datos, se desarrolló una aplicación de chatbot sobre una plataforma web, en la que los alumnos seleccionados respondieron a un conjunto de preguntas basadas en un conjunto de instrumentos de evaluación psicológica.

Para acceder a la plataforma, se generaron enlaces específicos para cada estudiante con el fin de identificar correctamente al individuo, dicho enlace era enviado de manera periódica al correo electrónico de los participantes. Este proceso se llevó a cabo durante el transcurso del semestre académico 2021-II.

La interfaz de chatbot brindaba una serie de preguntas a los participantes, los cuales respondían de acuerdo con su propio criterio. Al finalizar una ronda de preguntas, el estudiante podía decidir si pasar a la siguiente ronda, o continuar en otro momento. Este proceso se repitió hasta conseguir completar los instrumentos seleccionados.

Tras la recopilación, los datos fueron almacenados en una base de datos relacional para su posterior procesamiento.

2.5 Técnicas de procesamiento y análisis de datos

Durante la ejecución del proyecto, se llevaron a cabo un conjunto de técnicas, tanto de preprocesamiento, procesamiento y visualización de los datos para su posterior análisis. La metodología empleada puede ser explicada en el siguiente diagrama:

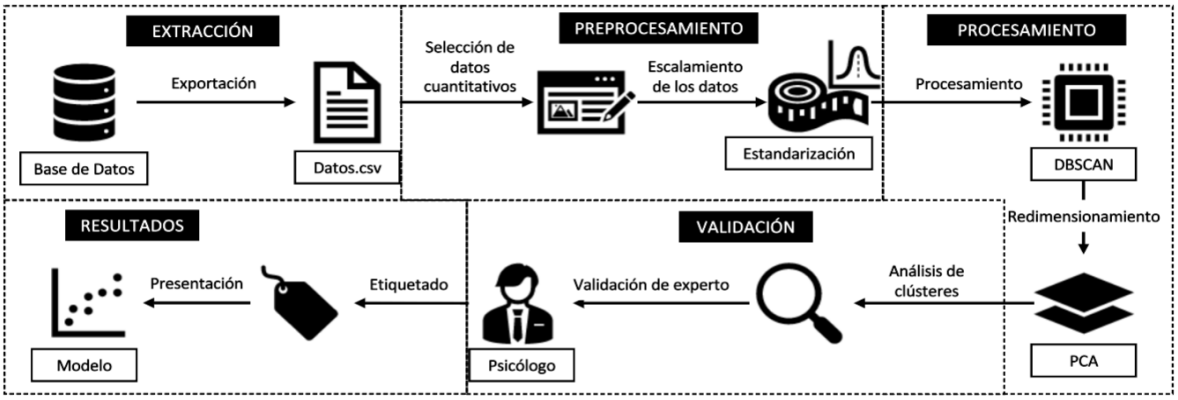


Figura 1. Representación gráfica de flujo de procesamiento de los datos
Fuente: Elaboración propia

2.5.1 Extracción:

Una vez que los datos de los estudiantes intervenidos fueron almacenados en una base de datos, estos fueron exportados a un esquema de datos resumido que permitieran su procesamiento con mayor simplicidad, facilidad y rapidez.

2.5.2 Preprocesamiento:

A partir de este paso, el tratamiento de los datos se llevó a cabo bajo entorno de desarrollo integrado (IDE) de código abierto para programación científica en el lenguaje Python, denominado Spyder V5.2.2.

2.5.2.1. Selección de datos cuantitativos:

Con el propósito de poder llevar a cabo el procesamiento de los datos a través del algoritmo de aprendizaje no supervisado DBSCAN, se tuvieron que remover aquellos datos que no aportaran valores cuantitativos al modelo.

Para ello, se comenzó importando los datos desde el archivo "Datos.csv" a la variable "data":

```
data = pd.read_csv("Datos.csv")
```

La variable "data" contaba con las siguientes columnas:

Tabla 1
Columnas de datos de la variable "data"

Columna	Tipo
CODIGO	String
HABITOS DE ESTUDIO	Int
ADAPTACION Y CONVIVENCIA	Int
DEPRESION	Int
ANSIEDAD	Int

Fuente: Elaboración propia

Debido a que los datos almacenados en la columna "CODIGO" tienen la finalidad de identificar al estudiante y no aportan valores relevantes para el modelo, se optó por su eliminación:

```
data = data.drop("CODIGO", axis=1)
```

Como resultado, tenemos las siguientes columnas con sus respectivos tipos de datos:

Tabla 2
Columnas de datos en la variable "data" tras aplicarse los cambios

Columna	Tipo
HABITOS DE ESTUDIO	Int
ADAPTACION Y CONVIVENCIA	Int
DEPRESION	Int
ANSIEDAD	Int

Fuente: Elaboración propia

2.5.2.2. Escalamiento de los datos

Una vez realizado el proceso de selección y limpieza de datos, se analizaron los datos estadísticos descriptivos del conjunto de datos resultante:

Index	HABITOS DE ESTUDIO	ADAPTACION Y CONVIVENCIA	DEPRESION	ANSIEDAD
count	670	670	670	670
mean	3.57	0.696	1.09	0.275
std	0.967	0.467	0.317	0.697
min	1	0	0	0
25%	3	0	1	0
50%	4	1	1	0
75%	4	1	1	0
max	5	2	3	3

Figura 2. Datos estadísticos de las columnas de datos en la variable "data"
Fuente: Elaboración propia

A partir de dichos datos, se pudo reconocer una discordancia entre las distintas columnas respecto a la escala o rango en el que sus datos fluctuaban (máximo y mínimos).

Esto, debido a que las puntuaciones realizadas en los procesos de recopilación de datos son representadas en escalas o rangos de datos diferentes, como se puede ver a continuación:

Tabla 3
Escala de posibles valores en las columnas de datos de la variable "data"

Columna	Escala
HABITOS DE ESTUDIO	0 - 5
ADAPTACION Y CONVIVENCIA	0 - 2
DEPRESION	0 - 3
ANSIEDAD	0 - 3

Fuente: Elaboración propia

Con el fin de brindarle al algoritmo de aprendizaje no supervisado DBSCAN, datos en un mismo formato y escala, se procedió a escalar los datos a través de métodos de normalización:

```
X = preprocessing.StandardScaler().fit_transform(data)
```

Este proceso, denominado “estandarización”, escala los datos en base a una distribución normal, ajustando la media a 0 y la varianza a 1.

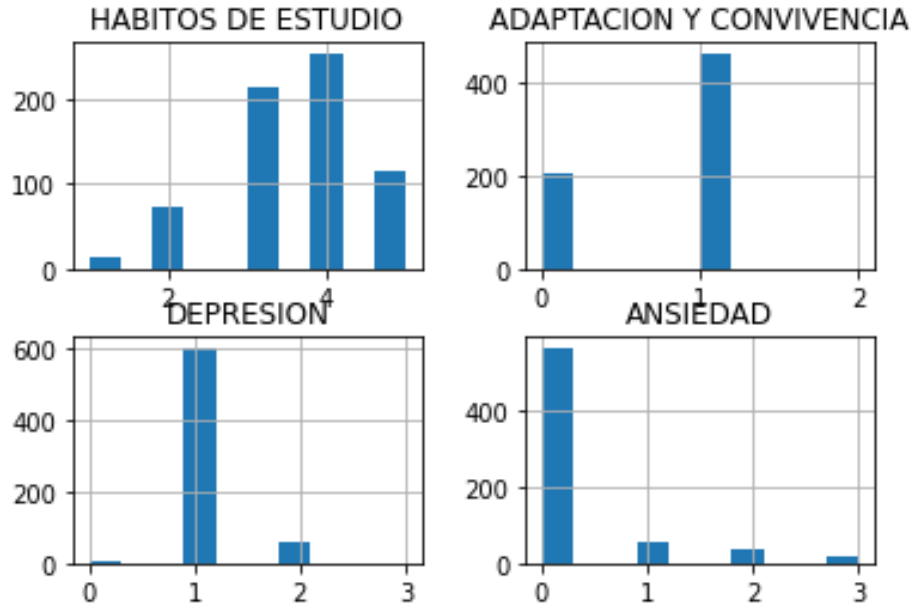


Figura 3. Distribución de los datos antes de la normalización
Fuente: Elaboración propia

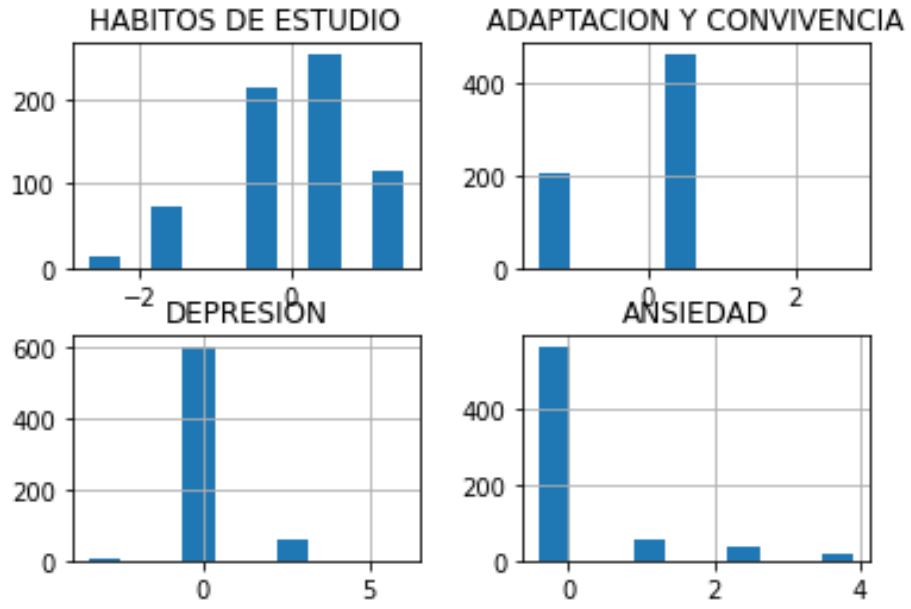


Figura 4. Distribución de los datos después de la normalización
Fuente: Elaboración propia

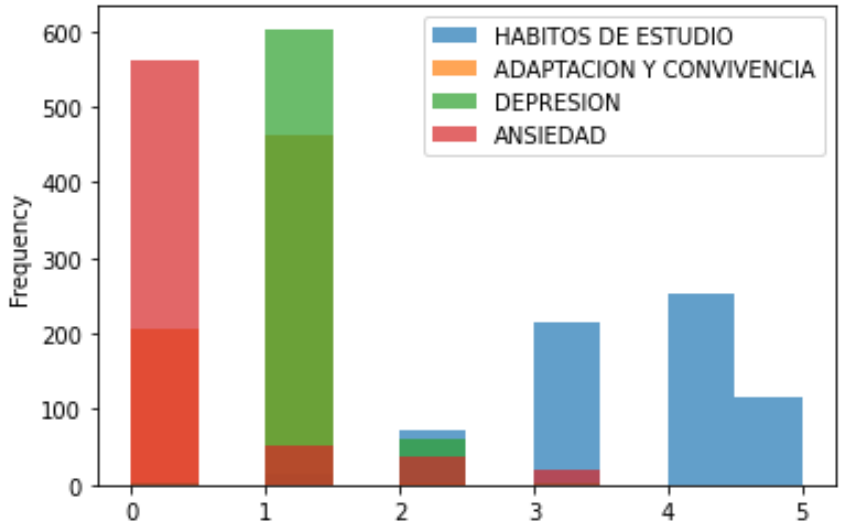


Figura 5. Contraste de las distribuciones de datos antes de la normalización
Fuente: Elaboración propia

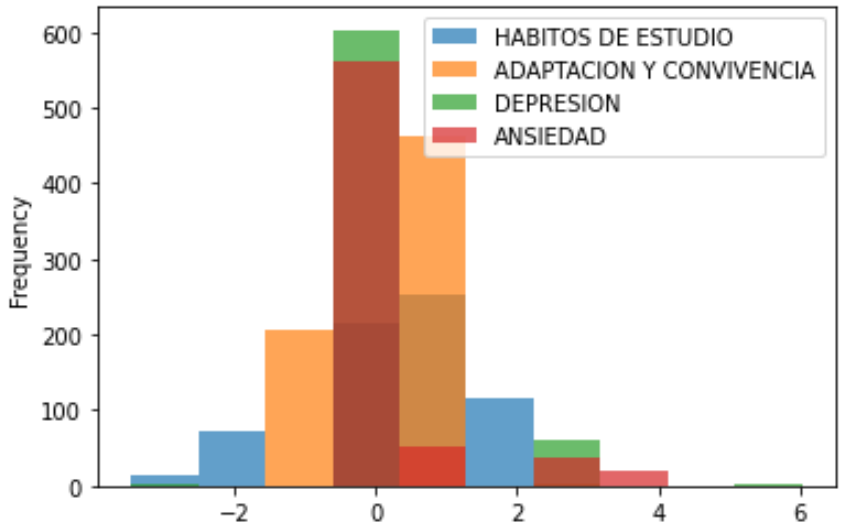


Figura 6. Contraste de las distribuciones de datos después de la normalización
Fuente: Elaboración propia

2.5.3 Procesamiento:

El procesamiento de los datos se realizó a través del algoritmo de aprendizaje no supervisado DBSCAN.

El mecanismo detrás de DBSCAN se puede explicar de la siguiente manera: Los datos son presentados en una matriz multidimensional, los cuales en su conjunto serán considerados como el grupo universo. Dado un radio (Eps) para cada uno de los puntos en un espacio euclidiano y a través de un número mínimo de puntos (MinPts), la vecindad de un punto se define como:

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$$

Dados los parámetros Eps y MinPts, DBSCAN elige aleatoriamente un punto núcleo como semilla y recupera todas las muestras de densidad alcanzable (dentro del radio Eps) desde la semilla para formar un clúster, aquellos puntos que no pertenecen a un grupo son considerados como ruido.

2.5.3.1. Selección de los parámetros Eps y Min_pts

Para comenzar con el procesamiento, fue crucial contar con los parámetros que el algoritmo requiere para su ejecución, Eps y MinPts. Dichos parámetros se calcularon mediante la ejecución iterativa del propio algoritmo sobre un rango de valores Eps y MinPts, esto con el fin de recopilar sus resultados y contrastarlos con los objetivos del proyecto.

El método y el criterio empleado para la selección de los parámetros del algoritmo se basan en el análisis del coeficiente de Silhouette, el cual se calcula de la siguiente manera:

$$S = \frac{b - a}{max(a, b)}$$

- Siendo “a” la distancia media entre los elementos en el interior de los clústeres.
- Siendo “b”, la suma de disimilitudes de los elementos dentro de un clúster “A”, respecto a los elementos dentro de un clúster “B” (Fig. 6.).

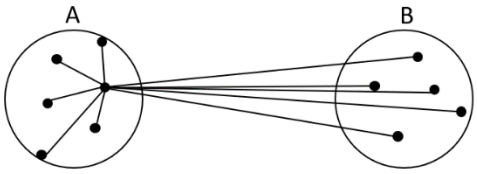


Figura 7. Ilustración de los elementos involucrados en el cómputo del coeficiente de Silhouette
Fuente: (Rousseeuw, 1987)

Se comenzó importando los recursos necesarios:

- Numpy para los cálculos numéricos.
- Pandas para la manipulación de datos en esquemas llamados dataframes.
- El módulo “preprocessing” de la librería sci-kit learn.
- El algoritmo Nearest Neighbors para el cálculo de distancias entre los elementos del set de datos.
- El método product del paquete itertools para generar combinaciones en base a los elementos de dos o más listas de datos.

- El algoritmo de aprendizaje no supervisado DBSCAN para ser ejecutado en base a las combinaciones de parámetros generados por el método product.

```
import numpy as np
import pandas as pd
from sklearn import preprocessing
from sklearn.neighbors import NearestNeighbors
from itertools import product
from sklearn.cluster import DBSCAN
```

Se llevó a cabo el proceso de selección y escalamiento de datos.

```
data = pd.read_csv("Datos.csv")
data = data.drop("CODIGO", axis=1)
X = preprocessing.StandardScaler().fit_transform(data)
```

A continuación, se calcularon las distancias entre los vecinos más cercanos a cada punto en el ser de datos:

```
cercanos = NearestNeighbors()
vecinos = cercanos.fit(X)
distancias, indices = vecinos.kneighbors(X)
```

Luego de calcular las distancias, se extrajeron los datos estadísticos del cuadro de distancias con el fin de verificar sus valores máximos y mínimos:

```
est_distancias = pd.DataFrame(distancias).describe()
```

Index	0	1	2	3	4
count	670	670	670	670	670
mean	0	0.0512857	0.0930871	0.139095	0.166415
std	0	0.295199	0.401938	0.464692	0.504116
min	0	0	0	0	0
25%	0	0	0	0	0
50%	0	0	0	0	0
75%	0	0	0	0	0
max	0	3.32552	3.32552	3.47091	3.47091

Figura 8. Datos estadísticos de la matriz de distancias obtenidas por el algoritmo Nearest Neighbors
Fuente: Elaboración propia

Siendo 0 y 3.47 ≈ 3.5, los valores mínimos y máximos respectivamente.

Con el rango de datos obtenido para los valores Eps, se optó por asignar un rango de Min_pts arbitrario de 5 a 15, basado en los objetivos del proyecto:

```
valores_eps = np.arange(0.1, 3.4, 0.1)
min_pts = np.arange(5,16)
```

Se generaron las combinaciones de parámetros en base a las listas de rangos de datos generadas anteriormente, y se inicializaron variables para el almacenamiento de los datos resultantes a la ejecución iterativa del algoritmo con cada uno de los parámetros propuestos:

```
params_dbscan = list(product(valores_eps, min_pts))
nro_clusters = []
pts_sil = []
_valores_eps = []
_pts_min = []
ruido = []

for p in params_dbscan:
    dbscan_cluster          =          DBSCAN(eps=p[0],
min_samples=p[1]).fit(X)
    _valores_eps.append(p[0])
    _pts_min.append(p[1])

nro_clusters.append(len(np.unique(dbscan_cluster.labels_
))-1)
    ruido.append(list(dbscan_cluster.labels_).count(-1))
    pts_sil.append(metrics.silhouette_score(X,
dbscan_cluster.labels_))
```

Por último y tras la ejecución de todos los casos de interés, se obtuvieron 362 resultados, de entre los cuales, se seleccionaron los parámetros 1.7 y 6 para los valores Eps y Min_pts respectivamente:

Index	Número de Clústers	Coeficiente de Silhoueti	Valores épsilon	Mínimo de putos	Ruido
177	5	0.484959	1.7	6	9
156	5	0.484065	1.5	7	10
167	5	0.484065	1.6	7	10
178	5	0.484065	1.7	7	10

Figura 9. Resultados obtenidos tras la ejecución de DBSCAN
Fuente: Elaboración propia

Esto, debido a los siguientes criterios:

- Coeficiente de sillhouette mayor a 0.
- Baja cantidad de puntos de tipo ruido.
- Número de clústeres mayor a 3.

2.5.3.2. Aplicación de DBSCAN

Retomando desde el final de los procesos de preprocesamiento, se aplicó DBSCAN con los parámetros obtenidos anteriormente:

```
db = DBSCAN(eps=1.7, min_samples=6).fit(X)
etiquetas = db.labels_
```

El resultado del cómputo del algoritmo sobre el set de datos, se almacenó en la variable etiquetas, siendo este, una lista de 670x1 cuya única columna contiene las etiquetas de los clústeres generados, siendo su índice en Y, el índice correspondiente al alumno dentro del set de datos inicial:

Ind	Cl
0	0
1	1
2	0
3	0
4	0
..	..
665	1
666	1
667	1
668	1
669	2

Sumado a este resultado, se extrajeron los siguientes datos:

- Número de clústeres.
- Número de puntos ruido.
- Coeficiente de Silhouette.

```
n_clusters_ = len(set(etiquetas)) - (1 if -1 in etiquetas else 0)
n_ruido_ = list(etiquetas).count(-1)
c_silhouette = metrics.silhouette_score(X, etiquetas);
```

Resultado:

```
NÚMERO ESTIMADO DE CLÚSTERES: 5
NÚMERO ESTIMADOS DE PUNTOS RUIDO: 9
COEFICIENTE DE SILHOUETTE:  0.48495862439988846
```

Estos datos fueron posteriormente de mayor relevancia durante el proceso de validación de clústeres.

2.5.3.3. Redimensionamiento para la visualización

Un paso previo al proceso de validación de los clústeres generados fue la composición tridimensional de la nube de puntos y sus respectivos clústeres.

Esta nube de puntos, sin embargo, se genera a partir de cuatro características o columnas de datos, es decir, cuatro dimensiones. Por lo que a priori, resulta imposible representar dicha nube de puntos sobre un plano tridimensional.

Es por esta razón que resulta necesario redimensionar dicho set de datos mediante alguna técnica procesamiento o compresión de datos. PCA o Análisis de Componentes Principales, es una de las más populares técnicas que responden a este inconveniente. Se importó entonces desde el módulo “decomposition”, el método PCA:

```
from sklearn.decomposition import PCA
```

Mediante esta técnica, se pudo reducir exitosamente el set de datos de cuatro dimensiones, a las tres necesarias para su representación gráfica:

```
pca = PCA(n_components=3)
pca.fit(pd.DataFrame(X))
xyz = pca.transform(pd.DataFrame(X))
```

CAPÍTULO III
RESULTADOS Y DISCUSIÓN

Con la ejecución del algoritmo sobre el set de datos propuestos, se obtuvieron un conjunto de resultados que posteriormente fueron utilizados para aplicar distintas técnicas de validación con el fin de verificar la exactitud del modelo planteado.

3.1 Validación:

3.1.1 Validación visual

Con el set de datos correctamente redimensionado, se procedió a importar las librerías de generación de gráficos:

```
import matplotlib.pyplot as plt
```

Se inicializó y configuró el lienzo, colores y ejes para llevar a cabo la representación:

```
fig = plt.figure()
ax = fig.add_subplot(111, projection="3d")
unique_labels = set(etiquetas)
colors = [plt.cm.Spectral(each) for each in np.linspace(0, 1,
len(unique_labels))]
colors[0]=[0, 0, 0, 1]
patron_colores = [(colors[etiquetas[i]+1]) for i in range(len(X))]
```

Y finalmente se procedió a dibujar sobre el lienzo dicho set de datos:

```
ax.scatter(xyz[:, 0], xyz[:, 1], xyz[:, 2], s=20, c=patron_colores, marker="o")
plt.title("NÚMERO ESTIMADO DE CLÚSTERES: %d" % n_clusters_)
ax.set_xlabel('Eje X')
ax.set_ylabel('Eje Y')
ax.set_zlabel('Eje Z')
plt.show()
```

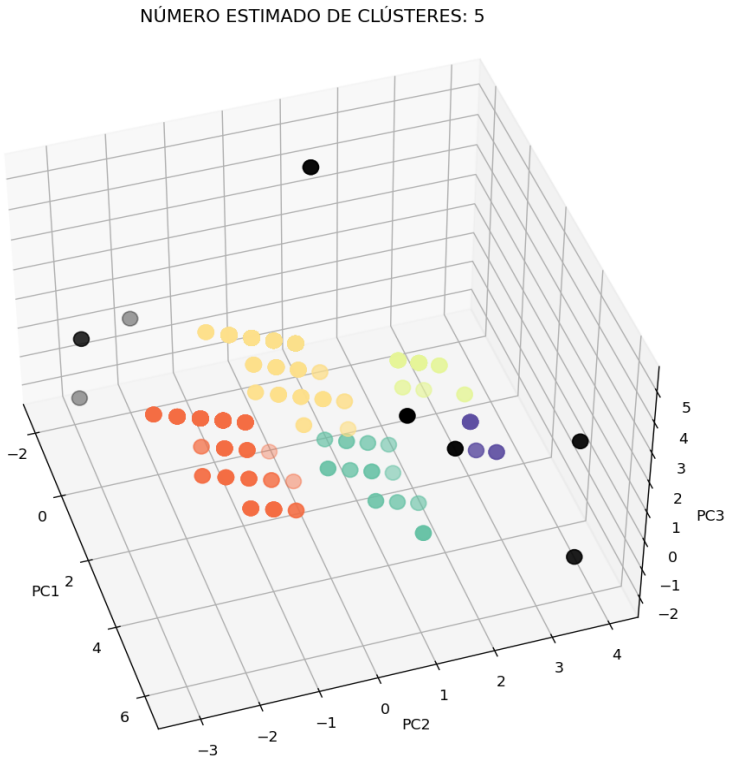


Figura 10. Representación visual de la nube de puntos clusterizada en tres dimensiones
Fuente: Elaboración propia

Tras obtener la representación gráfica de la nube de puntos del set de datos y sus respectivos clústeres identificados por colores (Fig. 7), es posible diferenciar estructuras fuertemente definidas. Cabe destacar que dicha representación tridimensional, es el resultado de un proceso de redimensionamiento del set de datos de cuatro dimensiones, a través del PCA. Sin embargo, a pesar de no contar con la representación de la nube de puntos en su estado inicial, la técnica de reducción de características logra ofrecer una estructura en la que es posible reconocer visualmente los clústeres, teniendo cercanos a aquellos elementos de su propio clúster y lejanos a aquellos no pertenecientes o ajenos a su grupo. Este resultado es coherente con otras investigaciones realizadas anteriormente como la de Khojastehnazhand & Roostaei (2022) en la que utilizaron la técnica de PCA para evaluar los resultados obtenidos con distintas técnicas de clasificación de características, por su parte Lopes et al. (2022) hacen uso de PCA durante el preprocesamiento de los datos, con el que logró reducir el número de características de 103 a 58, manteniendo el 95% de la varianza de los datos iniciales.

3.1.2 Enfoque de validación interna:

Dada la inexistencia de modelos previos con los cuales sea posible comparar el presente modelo, se optó por emplear un enfoque de validación interna, siendo el coeficiente de Silhouette el índice de rendimiento a evaluar.

El coeficiente de Silhouette ha sido utilizado ampliamente como un índice de validación para la selección de un número adecuado de clústeres (Ozaki et al., 2016). Denotando un alto grado de separabilidad entre clústeres y similitud entre miembros de un mismo clúster, así como un alto grado de disimilitud entre miembros de distintos clústeres, siendo los más deseados aquellos cuyos valores son mayores a cero (Rousseeuw, 1987).

Tras la ejecución de DBSCAN, se extrajeron un conjunto de resultados, entre los cuales se destaca el ya mencionado coeficiente de Silhouette:

```
NÚMERO ESTIMADO DE CLÚSTERES: 5
NÚMERO ESTIMADOS DE PUNTOS RUIDO: 9
COEFICIENTE DE SILHOUETTE:  0.48495862439988846
```

Con lo obtenido, se puede entender que el modelo generado durante la ejecución del proyecto, proporciona un coeficiente de Silhouette válido en términos del análisis de índices internos. Así mismo, Zhao et al. (2020) en su investigación, propone un modelo de clasificación de imágenes, en la que, a través del análisis de índices internos, obtiene un coeficiente de Silhouette de 0.68 puntos, lo que resulta coherente la teoría planteada por Rousseeuw (1987). De igual manera, Valarmathy & Krishnaveni (2020) plantea un nuevo método clusterización basado en DBSCAN, obteniendo un coeficiente de Silhouette de 0.78, y a su vez, utilizando este parámetro para evaluar el rendimiento de su modelo.

3.1.3 Validación de resultados por el experto:

En colaboración con el Psic. Dr. Juan Rafael Juárez Díaz, docente en educación básica especial incorporado al 5to nivel de la ley de reforma magisterial, con 25 años de servicio. Lic. en Educación con mención en ciencias sociales y Lic. en Psicología; Maestro en Educación con mención en Investigación, Dr. en Administración y Dr. en Ciencias de la educación; así mismo, egresado del Doctorado en Psicología Educacional y Tutorial, y siendo uno de los autores del proyecto “Caracterización del proceso de tutoría a estudiantes de la UNSM aplicando un modelo de atención virtual basado en chatbots”, y en calidad de experto de la salud mental, se dio a la tarea de validar e identificar los patrones reconocidos

en cada clúster de datos. Tras la generación de los clústeres, se pudo observar los siguientes patrones en la distribución de los resultados:

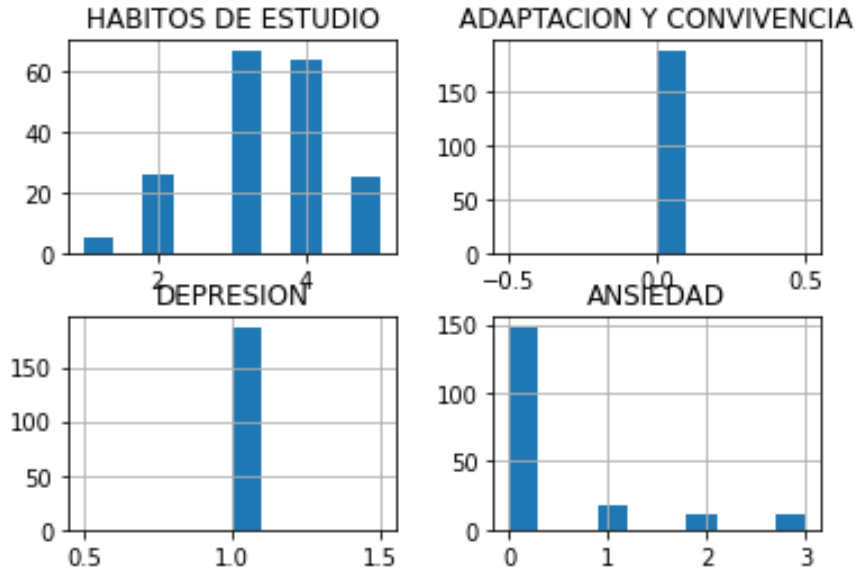


Figura 11. Distribución de resultados obtenidos en cada prueba sobre los integrantes del clúster 0
Fuente: Elaboración propia

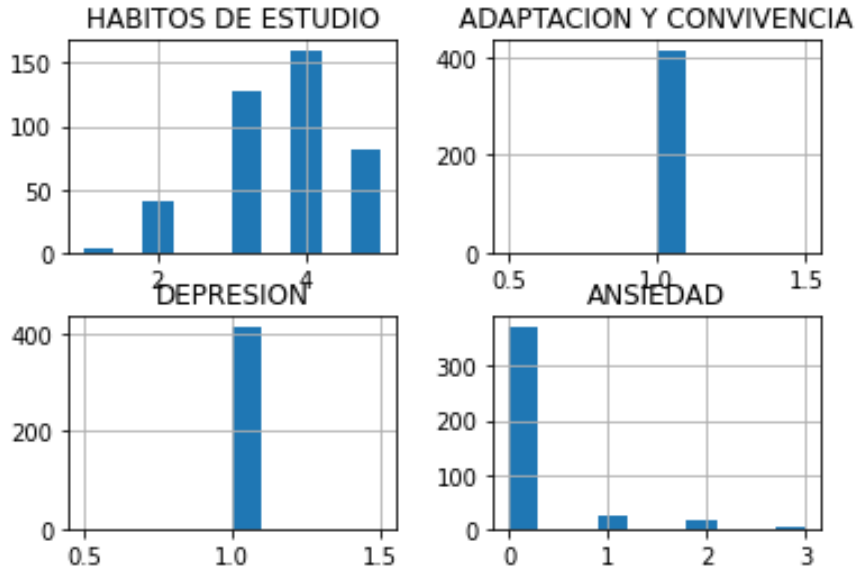


Figura 12. Distribución de resultados obtenidos en cada prueba sobre los integrantes del clúster 1
Fuente: Elaboración propia

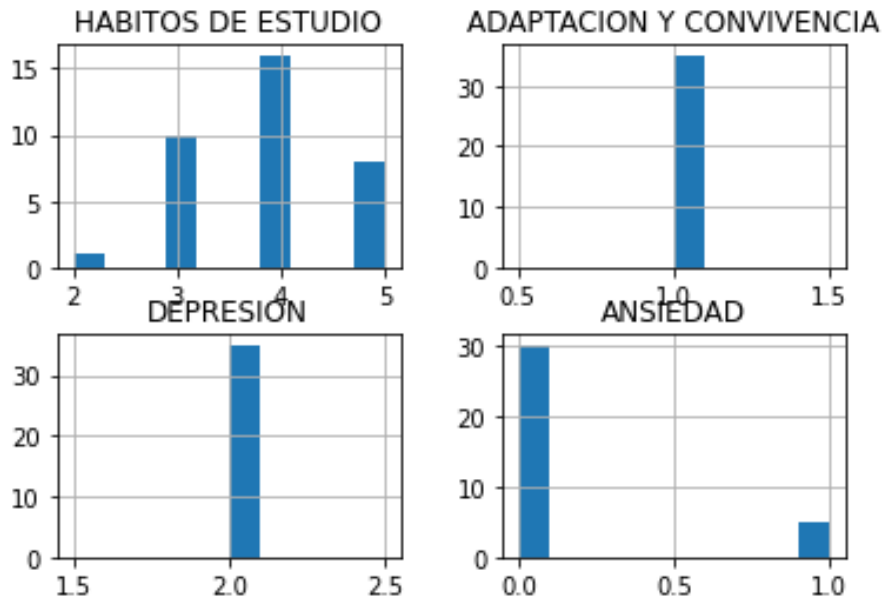


Figura 13. Distribución de resultados obtenidos en cada prueba sobre los integrantes del clúster 2
Fuente: Elaboración propia

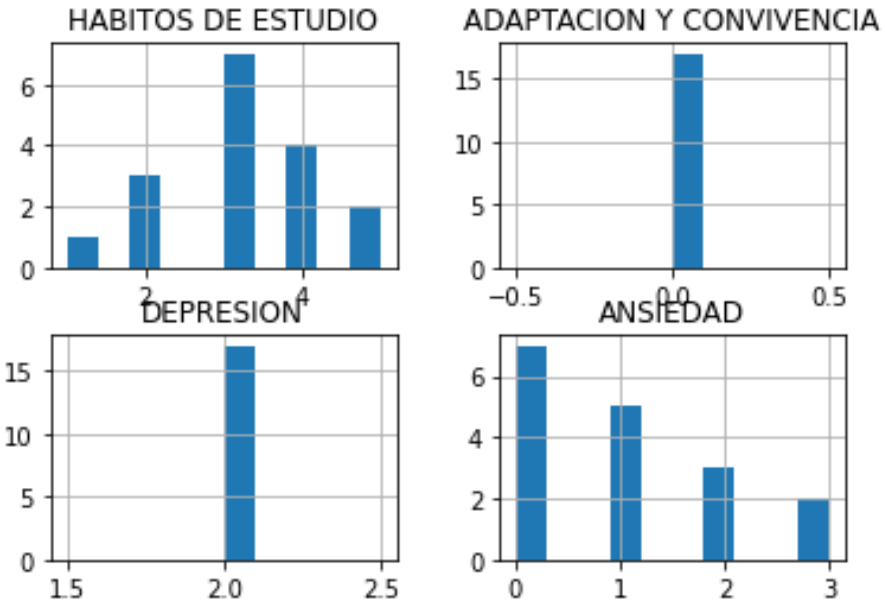


Figura 14. Distribución de resultados obtenidos en cada prueba sobre los integrantes del clúster 3
Fuente: Elaboración propia

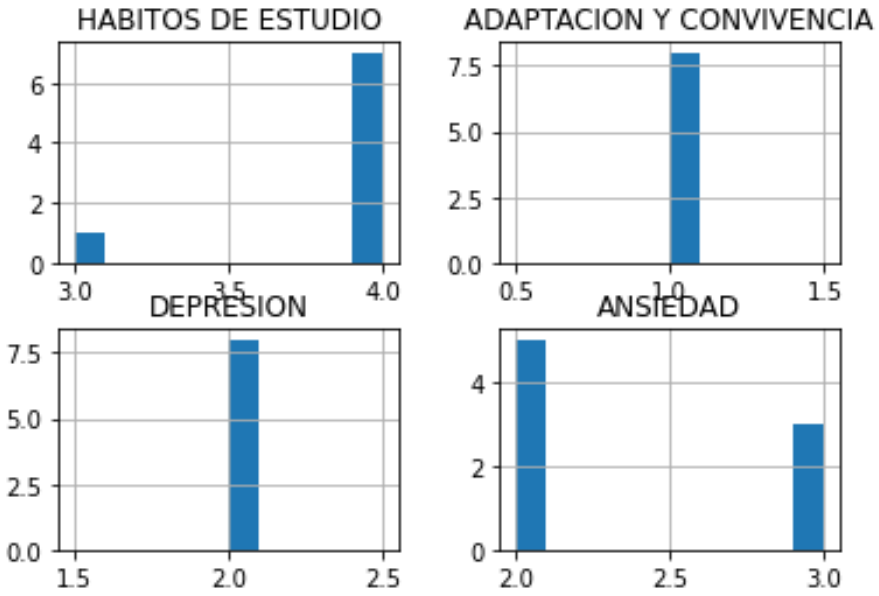


Figura 15. Distribución de resultados obtenidos en cada prueba sobre los integrantes del clúster 4
Fuente: Elaboración propia

Con la media de los resultados obtenidos en las distintas distribuciones y el criterio del profesional anteriormente mencionado, se procedió al etiquetado de los resultados:

Tabla 4.
Valores de la media de la distribución de los datos según clústeres

Clúster	Hábitos de estudio	Adaptación y convivencia	Depresión	Ansiedad	Nivel de riesgo
0	3.417112	0	1	0.385027	2 = Bajo
1	3.659420	1	1	0.144928	1 = Muy bajo
2	3.885714	1	2	0.142857	3 = Medio
3	3.176471	0	2	1	5 = Muy alto
4	3.875000	1	2	2.375000	4 = Alto

Fuente: Elaboración propia

CONCLUSIONES

1. Se logró unir los esfuerzos conjuntos entre profesionales de la salud mental y técnicas de aprendizaje no supervisado para generar una solución integral ante la inexistencia de un modelo que permitiera categorizar a los estudiantes universitarios de pregrado en función del riesgo de deserción en la Universidad Nacional de San Martín.
2. Se desarrolló un modelo de agrupamiento que sirve como base de conocimiento del panorama actual de los estudiantes de pregrado de la Universidad Nacional de San Martín con el cual es posible implementar futuras soluciones.
3. Se planteó una propuesta de solución que integra técnicas, modelos y metodologías actualmente estudiadas en el ámbito de las TICs, específicamente, del campo del aprendizaje de máquina.
4. Se logró categorizar exitosamente a los estudiantes de pregrado de la Universidad Nacional de San Martín en cinco niveles en función al riesgo de deserción a través un modelo de agrupamiento basado en DBSCAN.

RECOMENDACIONES

1. Para próximos trabajos, se recomienda seguir integrando las distintas disciplinas del conocimiento humano para generar más y mejores modelos que permitan entender las distintas problemáticas en la realidad actual dentro de la población universitaria de la Universidad Nacional de San Martín.
2. Con el fin de entender mejor el estado y la realidad de los estudiantes de la Universidad Nacional de San Martín, se deben seguir generando nuevos modelos que sirvan como base de conocimiento para futuros trabajos y soluciones.
3. Los avances en las TICs permiten un mejor procesamiento de los datos a evaluar, por lo que se recomienda la utilización de técnicas, modelos, metodologías e instrumentos modernos y actualizados.
4. Se recomienda seguir desarrollando nuevos modelos de categorización de los estudiantes de la Universidad Nacional de San Martín en función al riesgo de deserción con la finalidad de brindar una mejor y más amplia perspectiva de la realidad en cuestión.

REFERENCIAS BIBLIOGRÁFICAS

Ahuja, R., & Banga, A. (2019). Mental stress detection in university students using machine learning algorithms. *Procedia Computer Science*, *152*, 349–353. https://doi.org/10.1016/j.procs.2019.05.007

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. *International Conference on Information and Knowledge Management, Proceedings*, *24-28-Octo*, 903–912. https://doi.org/10.1145/2983323.2983351

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting Student Dropout in Higher Education*.

Bakshi, S., Jagadev, A. K., Dehuri, S., & Wang, G. N. (2014). Enhancing scalability and accuracy of recommendation systems using unsupervised learning and particle swarm optimization. *Applied Soft Computing Journal*, *15*, 21–29. https://doi.org/10.1016/j.asoc.2013.10.018

Baneres, D., Rodríguez-Gonzalez, M. E., & Serra, M. (2019). An Early Feedback Prediction System for Learners At-Risk within a First-Year Higher Education Course. *IEEE Transactions on Learning Technologies*, *12*(2), 249–263. https://doi.org/10.1109/TLT.2019.2912167

Barreto Osama, D., & Salazar Blanco, H. A. (2020). Agotamiento Emocional en estudiantes universitarios del área de la salud. *Universidad y Salud*, *23*(1), 30–39. https://doi.org/10.22267/rus.212301.211

Bean, J. P., & Metzner, B. S. (1985). A Conceptual Model of Nontraditional Undergraduate Student Attrition. *Review of Educational Research*, *55*(4), 485. https://doi.org/10.2307/1170245

Benites, R. (2021). La Educación Superior Universitaria en el Perú post-pandemia. *Políticas y Debates Públicos*, *1*(1), 1–11.

Brandão, A. S., Bolsoni-Silva, A. T., & Loureiro, S. R. (2017). The predictors of graduation: Social skills, mental health, academic characteristics. *Paideia*, *27*(66), 117–125. https://doi.org/10.1590/1982-43272766201714

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, *40*(3), 807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Caliñski, T., & Harabasz, J. (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Campo, D. N., Stegmayer, G., & Milone, D. H. (2016). A new index for clustering validation with overlapped clusters. *Expert Systems with Applications*, *64*, 549–556. https://doi.org/10.1016/j.eswa.2016.08.021

Carvajal, R. A., & Cervantes, C. T. (2018). Approaches to college dropout in Chile. *Educacao e Pesquisa*, *44*(1).

Daszykowski, M., & Walczak, B. (2009). Density-Based Clustering Methods. *Comprehensive Chemometrics*, *2*, 635–654. https://doi.org/10.1016/B978-044452701-1.00067-3

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Díaz-Méndez, M., Paredes, M. R., & Saren, M. (2019). Improving society by improving education through service-dominant logic: Reframing the role of students in higher education. *Sustainability (Switzerland)*, *11*(19). https://doi.org/10.3390/su11195292

Draszawka, K., & Szymański, J. (2011). External validation measures for nested clustering of text documents. *Studies in Computational Intelligence*, *369*, 207–225. https://doi.org/10.1007/978-3-642-22732-5_18

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231.

Feng, Y., & Zhang, Y. (2021). Evaluation and Analysis of Mental Health Level of College Students With Financial Difficulties Under the Background of COVID-19. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.649195

Freiberg-Hoffmann, A., Stover, J. B., & Donis, N. (2017). Influence of learning strategies on learning styles: Their impact on academic achievement of college students from Buenos aires. *Problems of Education in the 21st Century*, *75*(1), 6–18.

Fruehwirth, J. C., Biswas, S., & Perreira, K. M. (2021). The Covid-19 pandemic and mental health of first-year college students: Examining the effect of Covid-19 stressors using longitudinal data. *PLoS ONE*, *16*(3 March 2021). https://doi.org/10.1371/journal.pone.0247999

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems 2001 17:2*, *17*(2), 107–145. https://doi.org/10.1023/A:1012801612483

Haouas, F., Ben Dhiaf, Z., Hammouda, A., & Solaiman, B. (2017, August 23). A new efficient fuzzy cluster validity index: Application to images clustering. *IEEE*

International Conference on Fuzzy Systems. https://doi.org/10.1109/FUZZ-IEEE.2017.8015651

Hartigan, J. A., Spath, H., & Ryzin, J. Van. (1981). Clustering Algorithms. In *Journal of Marketing Research* (99th ed., Vol. 18, Issue 4). John Wiley & Sons, Inc. https://doi.org/10.2307/3151350

Heredia-Jimenez, V., Jimenez, A., Ortiz-Rojas, M., Marín, J. I., Moreno-Marcos, P. M., Muñoz-Merino, P. J., & Kloos, C. D. (2020). An early warning dropout model in higher education degree programs: A case study in Ecuador. In T. Y.-S. G. D. V. K. P.-S. M. P.-S. M. H. I. Z.-P. M. A. O.-R. M. S. E. Munoz-Merino P.J. Kloos C.D. (Ed.), *CEUR Workshop Proceedings* (Vol. 2704, pp. 58–67). CEUR-WS. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095968225&partnerID=40&md5=1cbf6f7fa11f03822ed5fdbad20a9171

Huanca-Arohuanca, J. W., Supo-Condori, F., Sucari Leon, R., Supo Quispe, L. A., Huanca-Arohuanca, J. W., Supo-Condori, F., Sucari Leon, R., & Supo Quispe, L. A. (2020). El problema social de la educación virtual universitaria en tiempos de pandemia, Perú. *Revista Innovaciones Educativas*, *22*(Especial), 115–128. https://doi.org/10.22458/IE.V22IESPECIAL.3218

Hügle, M., Omoumi, P., van Laar, J. M., Boedecker, J., & Hügle, T. (2021). Applied machine learning and artificial intelligence in rheumatology. *Rheumatology Advances in Practice*, *4*(1). https://doi.org/10.1093/rap/rkaa005

Jiménez, F., Paoletti, A., Sánchez, G., & Sciavicco, G. (2019). Predicting the Risk of Academic Dropout with Temporal Multi-Objective Optimization. *IEEE Transactions on Learning Technologies*, *12*(2), 225–236. https://doi.org/10.1109/TLT.2019.2911070

Kappe, R., & Van Der Flier, H. (2012). Predicting academic success in higher education: What's more important than being smart? *European Journal of Psychology of Education*, *27*(4), 605–619. https://doi.org/10.1007/s10212-011-0099-9

Kashyap, M., & Bhattacharya, M. (2017). A density invariant approach to clustering. *Neural Computing and Applications*, *28*(7), 1695–1713. https://doi.org/10.1007/s00521-015-2145-z

Khojastehnazhand, M., & Roostaei, M. (2022). Classification of seven Iranian wheat varieties using texture features. *Expert Systems with Applications*, *199*(May 2021), 117014. https://doi.org/10.1016/j.eswa.2022.117014

Khuong, H. (2014). Evaluation of a Conceptual Model of Student Retention at a Public Urban Commuter University. *Dissertations*. https://ecommons.luc.edu/luc_diss/1092

Krzanowski, W. J., & Lai, Y. T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics*, *44*(1), 23. https://doi.org/10.2307/2531893

Kumar, A., & Nayar, K. R. (2020). COVID 19 and its mental health consequences. In *Journal of Mental Health* (pp. 1–2). https://doi.org/10.1080/09638237.2020.1757052

Lee, S. H., Jeong, Y. S., Kim, J. Y., & Jeong, M. K. (2018). A new clustering validity index for arbitrary shape of clusters. *Pattern Recognition Letters*, *112*, 263–269. https://doi.org/10.1016/j.patrec.2018.08.005

Li, X., Liang, W., Zhang, X., Qing, S., & Chang, P. C. (2020). A cluster validity evaluation method for dynamically determining the near-optimal number of clusters. *Soft Computing*, *24*(12), 9227–9241. https://doi.org/10.1007/s00500-019-04449-7

Li, X., Zhang, Y., Cheng, H., Zhou, F., & Yin, B. (2021). An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns. *IEEE Access*, *9*, 7076–7091. https://doi.org/10.1109/ACCESS.2021.3049157

Lopes, M. L. B., Barbosa, R. de M., & Fernandes, M. A. C. (2022). Unsupervised Learning Applied to the Stratification of Preterm Birth Risk in Brazil with Socioeconomic Data. *International Journal of Environmental Research and Public Health*, *19*(9). https://doi.org/10.3390/ijerph19095596

Masud, M. T., Mamun, M. A., Thapa, K., Lee, D. H., Griffiths, M. D., & Yang, S. H. (2020). Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. *Journal of Biomedical Informatics*, *103*, 103371. https://doi.org/10.1016/j.jbi.2019.103371

Noboa, C., Ordóñez, M., & Magallanes, J. (2018). Statistical learning to detect potential dropouts in higher education:Aa public university case study. In O. X. del Mar Perez Sanagustin M. (Ed.), *CEUR Workshop Proceedings* (Vol. 2231). CEUR-WS.

Ozaki, R., Hamasuna, Y., & Endo, Y. (2016). A Method of Two-Stage Clustering Based on Cluster Validity Measures. *Proceedings - 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2016*, 410–415. https://doi.org/10.1109/SCIS-ISIS.2016.0093

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*(3), 361–394. https://doi.org/10.1007/s10115-008-0150-6

Piorecký, M., Štrobl, J., & Krajca, V. (2019). Automatic EEG classification using density based algorithms DBSCAN and DENCLUE. *Acta Polytechnica*, *59*(5), 498–509. https://doi.org/10.14311/AP.2019.59.0498

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846. https://doi.org/10.2307/2284239

Rochin Berumen, F. L. (2021). Deserción escolar en la educación superior en México: revisión de literatura. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *11*(22). https://doi.org/10.23913/ride.v11i22.821

Rojas-Thomas, J. C., Santos, M., & Mora, M. (2017). New internal index for clustering validation based on graphs. *Expert Systems with Applications*, *86*, 334–349. https://doi.org/10.1016/j.eswa.2017.06.003

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sammut, C., & Webb, G. I. (2010). Encyclopedia of Machine Learning. In *Encyclopedia of Machine Learning*. Springer US. https://doi.org/10.1007/978-0-387-30164-8

Theodoridis, S., & Koutroumbas, K. (2008). Pattern Recognition. In *Pattern Recognition* (4th Editio). Elsevier Inc.

Ting, S. M. R. (2001). Predicting Academic Success of First-Year Engineering Students from Standardized Test Scores and Psychosocial Variables. *International Journal of Engineering Education*, *17*(1), 75–80.

Valarmathy, N., & Krishnaveni, S. (2020). A novel method to enhance the performance evaluation of DBSCAN clustering algorithm using different distinguished metrics. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.09.623

Vargas, M., Talledo-Ulfe, L., Heredia, P., Quispe-Colquepisco, S., & Mejia, C. R. (2018). Influence of Habits on Depression in the Peruvian Medical Student: Study in Seven Administrative Regions. *Revista Colombiana de Psiquiatria*, *47*(1), 56–64. https://doi.org/10.1016/j.rcp.2017.01.008

Viloria, A., & Parody, A. (2016). Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence. *Indian Journal of Science and Technology*, *9*(46). https://doi.org/10.17485/ijst/2016/v9i46/107369

Webb, G. I., Fürnkranz, J., Fürnkranz, J., Fürnkranz, J., Hinton, G., Sammut, C., Sander, J., Vlachos, M., Teh, Y. W., Yang, Y., Mladeni, D., Brank, J., Grobelnik, M., Zhao, Y., Karypis, G., Craw, S., Puterman, M. L., & Patrick, J. (2011). Density-Based Clustering. In *Encyclopedia of Machine Learning* (pp. 270–273). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_211

Xie, J., Wu, R., Wang, H., Chen, H., Xu, X., Kong, Y., & Zhang, W. (2021). Prediction of cardiovascular diseases using weight learning based on density information. *Neurocomputing*, *452*, 566–575. https://doi.org/10.1016/j.neucom.2020.10.114

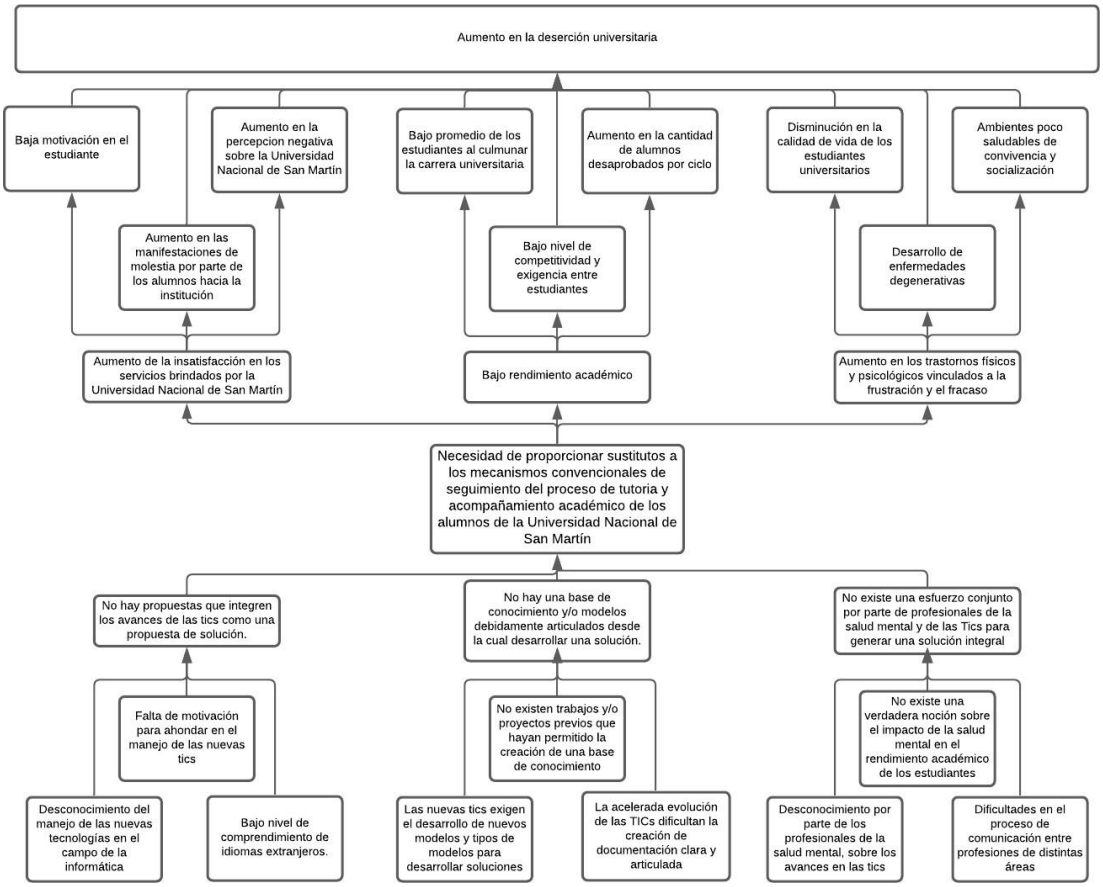
Zhao, Z., Zhao, J., Song, K., Hussain, A., Du, Q., Dong, Y., Liu, J., & Yang, X. (2020). Joint DBN and Fuzzy C-Means unsupervised deep clustering for lung cancer patient

stratification. *Engineering Applications of Artificial Intelligence*, *91*(December 2019), 103571. https://doi.org/10.1016/j.engappai.2020.103571

Zulu, W. V, & Mutereko, S. (2020). Exploring the Causes of Student Attrition in South African TVET Colleges: A Case of One KwaZulu-Natal Technical and Vocational Education and Training College. *Interchange*, *51*(4), 385–407. https://doi.org/10.1007/s10780-019-09384-y

ANEXOS

Anexo 1. Árbol de problema



Anexo 2. Árbol de problema

<table>
<tr><th>Formulación del problema</th><th>Objetivos</th><th>Hipótesis</th><th>Técnicas e instrumentos</th></tr>
<tr>
<td>¿Cuáles son las categorías de estudiantes en función al riesgo de deserción descubierta en base al algoritmo de agrupamiento basado en densidad?</td>
<td>Objetivo general
Categorizar a los estudiantes universitarios en función al riesgo de deserción en la Universidad Nacional de San Martín.
Objetivos específicos
- Plantear una propuesta de solución que integre los nuevos avances de las TICs.
- Desarrollar una base de conocimiento desde la cual desarrollar una solución.
- Generar una solución integral con el esfuerzo conjunto entre profesionales de la salud mental y de las TICs.</td>
<td>Hipótesis general
El algoritmo de agrupamiento basado en densidad DBSCAN categoriza a estudiantes universitarios en niveles de riesgo de deserción.

Hipótesis alterna:
H_1: El algoritmo de agrupamiento basado en densidad DBSCAN sí categoriza a estudiantes universitarios en niveles de riesgo de deserción.

Hipótesis nula:
H_0: El algoritmo de agrupamiento basado en densidad DBSCAN no categoriza a estudiantes universitarios en niveles de riesgo de deserción.</td>
<td rowspan="2">Técnicas:
Análisis exploratorio de datos.
Técnicas de selección de características.
Preprocesamiento de datos.
DBSCAN.

Instrumentos:
Sistema de información</td>
</tr>
<tr>
<td>Diseño de investigación</td>
<td>Población y muestra</td>
<td>Variables, dimensiones e indicadores</td>
</tr>
<tr>
<td>Tipo: Aplicada
Enfoque: Cualitativo
Diseño: No experimental</td>
<td>La población del presente proyecto consta de datos recopilados en el proyecto de Tutoría Virtual basado en Chatbot de los estudiantes de la Universidad Nacional de San Martín.
La muestra consta de datos recopilados en el proyecto de Tutoría Virtual basado en Chatbot de los estudiantes de la Universidad Nacional de San Martín.</td>
<td>
<table>
<tr><td rowspan="16">Categorización estudiantes universitarios en niveles de riesgo de deserción</td><td rowspan="4">Índices internos</td><td>Cohesión</td></tr>
<tr><td>Separación</td></tr>
<tr><td>Coeficiente de Silhouette</td></tr>
<tr><td>Correlación</td></tr>
<tr><td rowspan="12">Índices estadísticos</td><td>Desviación estándar</td></tr>
<tr><td>Media</td></tr>
<tr><td>Varianza</td></tr>
<tr><td>Asimetría</td></tr>
<tr><td>Error estándar de asimetría</td></tr>
<tr><td>Curtosis</td></tr>
<tr><td>Rango</td></tr>
<tr><td>Mínimo</td></tr>
<tr><td>Máximo</td></tr>
<tr><td>Suma</td></tr>
<tr><td>Percentiles</td></tr>
<tr><td>Mediana</td></tr>
<tr><td>Moda</td></tr>
<tr><td>Error estándar de la media</td></tr>
</table>
</td>
</tr>
</table>

UNIVERSIDAD NACIONAL DE SAN MARTÍN

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

UNIDAD DE INVESTIGACIÓN

CERTIFICADO DE PORCENTAJE DE SIMILITUD

Yo, Edwin Augusto Hernández Torres, en mi condición de Director de la Unidad de Investigación de la Facultad de Ingeniería de Sistemas e Informática he realizado la verificación de similitud del informe final titulado: **CATEGORIZACIÓN A ESTUDIANTES UNIVERSITARIOS EN NIVELES DE RIESGO DE DESERCIÓN EN BASE AL ALGORITMO DE APRENDIZAJE NO SUPERVISADO BASADO EN DENSIDAD,** presentado por el bachiller: LUIS GERARDO SALAZAR RAMÍREZ, para optar el título de Ingeniero de Sistemas e Informática.

Que, habiendo cumplido con lo establecido en el reglamento de similitud y originalidad y considerando la revisión, evaluación y análisis realizado, utilizando el reporte del software de similitud textual, cuyo porcentaje es **18 %**.

Certifico que la similitud del documento está en nivel **ACEPTABLE**.

Se emite el presente certificado con fines de continuar con el trámite respectivo.

Morales, agosto del 2023.

.....

Lic. M. Sc. EDWIN AUGUSTO HERNANDEZ TORRES
Director de la Unidad de Investigación de la FISI-UNSM