UNIVERSIDAD CATÓLICA SANTO TORIBIO DE MOGROVEJO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Aplicación predictiva para apoyar el diagnóstico de trastornos psicológicos en alumnos de la Institución Educativa Privada Santo Toribio de Mogrovejo

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO DE SISTEMAS Y COMPUTACIÓN

AUTOR

Miguel Esteban Torres Becerra

ASESOR

Mariana Chavarry Chankay

https://orcid.org/0000-0001-5136-7177

Chiclayo, 2025

Aplicación predictiva para apoyar el diagnóstico de trastornos psicológicos en alumnos de la Institución Educativa Privada Santo Toribio de Mogrovejo

PRESENTADA POR

Miguel Esteban Torres Becerra

A la Facultad de Ingeniería de la
Universidad Católica Santo Toribio de Mogrovejo
para optar el título de

INGENIERO DE SISTEMAS Y COMPUTACIÓN

APROBADA POR

Noblecilla Vinces William Alfredo
PRESIDENTE

Nicho Cordova Ernesto Ludwin
SECRETARIO

Mariana Chavarry Chankay
VOCAL

Dedicatoria

A mi familia, por su paciencia, amor y apoyo incondicional a lo largo de todos estos años, brindándome siempre la fuerza para continuar en cada paso de este proceso. También a todos los compañeros que he tenido el honor de llamar amigos, con quienes compartí este camino y que siempre estuvieron allí cuando más los necesité.

Agradecimientos

A mi asesora, la ingeniera Mariana Chavarry, por su inestimable ayuda, el valioso aporte de sus conocimientos y su constante apoyo durante la realización de esta investigación.
A la ingeniera María Arangurí, por su dedicación y apoyo, siempre dispuesta a brindar su orientación tanto en clases como fuera de ellas.
A la psicóloga Danitsa Rocillo, por su valiosa contribución profesional y sus conocimientos, que fueron esenciales para hacer posible este proyecto.

Aplicación predictiva para apoyar el diagnóstico de trastornos psicológicos en alumnos de la Institución Educativa Privada Santo Toribio de Mogrovejo

INFORME DE ORIGINALIDAD

8%	6%	1%	5%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	tesis.usat.edu.pe Fuente de Internet	1%
2	www.coursehero.com Fuente de Internet	1%
3	hdl.handle.net Fuente de Internet	1%
4	idoc.pub Fuente de Internet	1%
5	Submitted to ITESM: Instituto Tecnologico y de Estudios Superiores de Monterrey Trabajo del estudiante	1%
6	Submitted to Universidad Católica Santo Toribio de Mogrovejo Trabajo del estudiante	1%
7	Óscar Claret González Ortiz. "Fundamentos de ingeniería industrial", Ecoe Ediciones S. A. S., 2023	<1%

Índice

Resumen

La detección temprana de trastornos mentales en la infancia es esencial, ya que afectan la capacidad de los menores para realizar actividades diarias y su desarrollo educativo. Esta problemática se refleja en la Institución Educativa Privada Santo Toribio de Mogrovejo. En esta investigación, se desarrolló una aplicación predictiva para la identificación rápida de trastornos psicológicos en alumnos, facilitando su tratamiento. Utilizando la metodología CRISP-DM, se revisaron 13 estudios previos para seleccionar el algoritmo de clasificación adecuado, en los cuales los que más resaltaban fueron red neuronal artificial, Random Forest y KNN, siendo este último el elegido para el modelo predictivo tras haber sido evaluado con Weka, una herramienta de software para aprendizaje automático. KNN logró una precisión del 86%, exactitud del 79%, recall del 54%, especificidad del 92% y puntaje F1 del 69%. Este modelo predictivo fue implementado en una aplicación web, la cual permite a los alumnos realizar un cuestionario para obtener un diagnóstico presuntivo para autismo, TDAH y trastornos de lenguaje, y al personal de la institución gestionar los resultados. La aplicación proporciona una solución eficiente para identificar y tratar trastornos psicológicos en alumnos, mitigando los desafíos asociados con la escasez de recursos y personal especializado en entornos educativos. Su implementación puede mejorar significativamente la atención y el bienestar emocional de los estudiantes, evitando consecuencias negativas a largo plazo como dificultades académicas, problemas de conducta, baja autoestima y dificultades en las relaciones interpersonales.

Palabras clave: KNN, trastorno psicológico, sistema predictivo, machine learning, aprendizaje supervisado.

Abstract

Early detection of mental disorders in childhood is essential, since they affect the ability of children to perform daily activities and their educational development. This problem is reflected in the Santo Toribio de Mogrovejo Private Educational Institution. In this research, a predictive application was developed for the rapid identification of psychological disorders in students, facilitating their treatment. Using the CRISP-DM methodology, 13 previous studies were reviewed to select the appropriate classification algorithm, in which the most outstanding ones were artificial neural network, Random Forest and KNN, the latter being the one chosen for the predictive model after being evaluated with Weka, a machine learning software tool. KNN achieved a precision of 86%, accuracy of 79%, recall of 54%, specificity of 92% and F1 score of 69%. This predictive model was implemented in a web application, which allows students to take a questionnaire to obtain a presumptive diagnosis for autism, ADHD and language disorders, and the institution's staff to manage the results. The app provides an efficient solution to identify and treat psychological disorders in students, mitigating the challenges associated with the scarcity of resources and specialized personnel in educational environments. Its implementation can significantly improve students' attention and emotional well-being, avoiding long-term negative consequences such as academic difficulties, behavioral problems, low self-esteem and difficulties in interpersonal relationships.

Keywords: KNN, psychological disorder, predictive system, machine learning, supervised learning.

Introducción

En la infancia, los trastornos mentales son frecuentes, y debido a la pandemia y al confinamiento sufrido para evitar contagios del virus COVID-19, se presentan con más frecuencia. Los menores que han vivido una situación de aislamiento y la experimentan de forma negativa tienden a presentar más cuadros relacionados con el estrés o la ansiedad, por lo que su detección y atención temprana es de suma importancia, según Erades et al. [1]. Estos trastornos impiden que los niños realicen sus actividades diarias con normalidad e interfieren en su comportamiento. Como se menciona en el artículo [2], su identificación es vital para manejar la situación y evitar problemas a largo plazo. Debido a esto, los padres o responsables de los menores deben estar atentos a señales y ayudarlos lo antes posible, siendo estos trastornos la razón principal de problemas psicológicos en menores de 4 a 15 años, como se expresa en el artículo [3].

La Organización Mundial de la Salud [4] advierte que los trastornos mentales son extremadamente frecuentes de manera global, afectando aproximadamente a una de cada ocho personas en el mundo. Entre estos trastornos, los de ansiedad y depresión son los más comunes tanto en hombres como en mujeres de diferentes edades. Además, el suicidio es una de las principales causas de muerte entre los jóvenes, y puede haber hasta 20 intentos por cada fallecimiento. Los trastornos mentales graves, como la esquizofrenia, también tienen un impacto significativo, ya que quienes los padecen suelen vivir entre 10 y 20 años menos, en gran parte debido a enfermedades físicas prevenibles. En términos económicos, los trastornos mentales generan enormes costos para la sociedad, principalmente debido a la pérdida de productividad, siendo la esquizofrenia uno de los trastornos más costosos. Aunque los trastornos de ansiedad y depresión son menos costosos por persona, su alta prevalencia contribuye sustancialmente al gasto total. A pesar de la magnitud de este problema, los sistemas de salud mental en muchos países enfrentan grandes deficiencias, dedicándose menos del 2 % de los presupuestos de salud a esta área, y predominando la atención en hospitales psiquiátricos sobre los servicios comunitarios. En muchos lugares, las personas afectadas carecen de acceso a atención de calidad o tienen miedo de buscar ayuda debido a la estigmatización, lo cual agrava las brechas en cobertura y tratamiento.

En España, según un estudio realizado en septiembre de 2020, debido a la pandemia, los jóvenes fueron uno de los grupos más vulnerables, identificándose un aumento en ansiedad, estrés y depresión. Asimismo, los estudios centrados en la infancia arrojan resultados sobre la

situación emocional de los menores. Aunque algunos se sienten contentos por pasar más tiempo en casa y con la familia, existe un alto porcentaje que llora más (55.54%), se siente más nervioso (70.17%), se enfada más (74.66%) y se siente más triste, según Gómez et al. [5]. En el artículo de Rusca et al. [6] encontramos una encuesta realizada en 546 personas, proviniendo el 59% de Lima, 25,9% de Arequipa y 6,1% de La Libertad; siendo el resto en departamentos como Junín, Callao e Ica; el 69,2% (371) de los encuestados reportaron cambios en la conducta y emociones en sus menores hijos en la situación actual; siendo los síntomas más comunes y preocupantes: desgano, miedo, ansiedad, oposicionismo, ansiedad de separación, tartamudez, mayor sensibilidad, inquietud motora y tendencia al llanto. Así mismo, se manifiesta en la Consulta Regional para la Agenda Adolescente y Joven 2021-2026 [7], que en la región Lambayeque uno de los principales problemas de salud adolescentes de 15 a 18 años, son los problemas de salud mental, además de que existen muy pocos servicios de salud mental comunitaria en la región Lambayeque.

La Institución Educativa Privada Santo Toribio de Mogrovejo cuenta con más de 1000 alumnos, como se observa en el anexo 1, y dispone de solo tres psicólogos para atender a los estudiantes. Dado el alto número de alumnos, el personal disponible resulta insuficiente para cubrir adecuadamente todas las necesidades psicológicas, lo que puede llevar a que algunos estudiantes no reciban el diagnóstico y la ayuda necesaria. Mientras estos estudiantes no sean atendidos, es probable que sus problemas psicológicos o emocionales se agraven con el tiempo, afectando su desempeño escolar y su bienestar personal en el futuro. Frente a esta realidad, surge la siguiente pregunta de investigación: ¿Cómo puede una aplicación predictiva apoyar en el diagnóstico temprano y el tratamiento de trastornos psicológicos en alumnos de la Institución Educativa Privada Santo Toribio de Mogrovejo?

En el artículo de Castillo et al. [8] se nos explica que los factores importantes en los trastornos mentales de los niños y jóvenes son la familia y la escuela. En la familia los factores a tomar en cuenta son su convivencia y estilos parentales, en general separaciones matrimoniales, pérdidas en la familia o problemas en esta, como es el caso de una conducta parental negligente y autoritaria, en la que los padres dan órdenes a su hijo, sin atender a razones pudiendo llegar a agredirlo si el menor no obedece; y en la escuela, el trato con el profesorado o la interacción con los demás estudiantes, en el cual el menor puede desarrollar estos trastornos por abusos, ya sea por un mal trato de parte de un docente, o ser víctima de acoso por parte de algún o algunos compañeros. Los padres y el profesorado están expuestos a un mayor grado de ansiedad y estrés; si no saben afrontar correctamente estas dificultades, pudiendo influir negativamente en los menores. Estas dificultades se pueden extender en el

contexto escolar, ya que cualquier logro que consigan, podría ser desacreditado; deteriorando su autoestima y autoconcepto. Esto puede mantenerse o empeorarse en la adultez, teniendo consecuencias en su salud personal y pública igualmente.

Este trabajo se justifica desde varias perspectivas: científica, económica, social y tecnológica, y tiene como propósito abordar la identificación temprana de trastornos psicológicos en alumnos mediante el uso de tecnologías avanzadas como la Inteligencia Artificial (IA).

Desde el punto de vista científico, esta investigación busca fomentar el desarrollo futuro en el campo de la intersección entre psicología y ciencias de la computación, con el objetivo de crear herramientas que mejoren tanto el diagnóstico como el tratamiento de trastornos psicológicos en los estudiantes. Los problemas emocionales y psicológicos pueden afectar significativamente el desempeño académico de los alumnos. En muchos casos, aquellos que enfrentan dificultades psicológicas pueden repetir el año escolar y, en consecuencia, ser expulsados, lo que se traduce en una pérdida de estudiantes para las instituciones educativas. Este aspecto conecta con la justificación económica de la investigación, que está orientada a reducir la deserción escolar, evitando la disminución de ingresos para las instituciones educativas.

Desde una perspectiva social, el desarrollo de esta investigación permitirá brindar un tratamiento adecuado a los estudiantes que padezcan algún tipo de trastorno psicológico. Esto no solo ayudará a prevenir que sus problemas persistan o empeoren con el tiempo, sino que también contribuirá a su bienestar a largo plazo, reduciendo el riesgo de consecuencias como la baja autoestima, la habituación al fracaso o problemas que puedan afectar su vida personal, su rendimiento académico e incluso su salud mental en la adultez.

Finalmente, desde el punto de vista tecnológico, la justificación se basa en la aplicación de técnicas de Inteligencia Artificial para optimizar el proceso de apoyo al diagnóstico de estos trastornos. Esto permitirá acelerar la identificación de problemas psicológicos, minimizando el tiempo de diagnóstico y ayudando a evitar complicaciones en el futuro.

El objetivo principal de esta investigación es el desarrollo de una aplicación predictiva de apoyo para la identificación de trastornos psicológicos en alumnos rápidamente para su pronto tratamiento. Para esto, se determinó el algoritmo de predicción más apropiado para el apoyo del diagnóstico de trastornos psicológicos, para posteriormente utilizarlo para desarrollar un modelo predictivo. Posteriormente este fue evaluado para determinar su precisión como un instrumento de apoyo para la obtención de un diagnóstico y finalmente se desplegó el modelo

predictivo en una aplicación web, permitiendo facilitar el diagnóstico de trastornos psicológicos.

Con esto, se espera poder apoyar al diagnóstico de trastornos psicológicos de los estudiantes de la I.E.P. Santo Toribio de Mogrovejo, evitando así consecuencias negativas que puedan repercutir en su vida estudiantil, personal y, a largo plazo, profesional.

Revisión de literatura

Antecedentes

En el trabajo realizado por Almadhor et al. [9], se explora la utilización de técnicas de aprendizaje activo y machine learning para la predicción de ansiedad en estudiantes. Ante el aumento en la prevalencia de la ansiedad en el alumnado y sus consecuencias en el rendimiento académico, este estudio examina cómo estas tecnologías pueden emplearse para comprender y predecir los niveles de ansiedad en los alumnos. Para ello, se emplean estrategias de aprendizaje activo que incrementan la efectividad de los modelos de machine learning en esta predicción, destacando la utilidad de estas metodologías en la precisión de los modelos de ansiedad. Se utilizan dos conjuntos de datos con información sobre el comportamiento estudiantil, empleando modelos como K-Nearest Neighbors (KNN), Regresión Logística (LR), XGBoost (XGB), Naive Bayes (NB) y Random Forest (RF) para construir modelos predictivos. Los experimentos revelaron que el modelo de Regresión Logística basado en aprendizaje activo obtuvo una puntuación de 0.61, mientras que Random Forest alcanzó una precisión promedio de 0.60 en el primer dataset. En el segundo, Random Forest se destacó como el modelo más efectivo, logrando una precisión del 83%. Estos resultados brindan valiosa información sobre el rendimiento de los modelos en métricas clave y destacan el potencial del aprendizaje automático y el aprendizaje activo para predecir y gestionar la ansiedad en los estudiantes.

La investigación de Nash et al. [10] propone un innovador modelo utilizando machine learning con el objetivo de detectar síntomas de TDAH en adultos, utilizando un conjunto de datos multimodal que integra video RGB con datos sobre puntos faciales, postura corporal y movimientos de las manos. Este enfoque busca abordar la problemática del diagnóstico de TDAH en adultos, ya que las herramientas tradicionales, como el DSM-V, requieren que los síntomas se manifiesten en la niñez. Sin un diagnóstico en la infancia, la prevalencia de TDAH en adultos es mayor y este trastorno no diagnosticado conlleva consecuencias graves, como altas tasas de desempleo, productividad reducida y un riesgo significativo de problemas

legales y de salud mental. El modelo propuesto fue validado mediante técnicas de validación cruzada aleatoria y de dejar uno fuera, logrando una precisión del 98.67%, una precisión de 98.01% y una tasa de recall de 98.88%. Estos resultados destacan el potencial del modelo para ser aplicado en entornos clínicos, mejorando significativamente la detección de síntomas de TDAH sin requerir el historial de infancia del paciente, lo que representa un avance importante en la salud conductual y abre nuevas posibilidades no solo para el diagnóstico, sino también para el tratamiento del TDAH.

La investigación de Hasan et al. [11] aborda la dificultad de la detección temprana de autismo, un trastorno del desarrollo neurológico que afecta a pacientes en su día a día. Aunque el diagnóstico temprano es clave para mitigar la gravedad del trastorno, aún persisten desafíos en la identificación precisa del TEA. El estudio propone un marco para evaluar diversas técnicas de Machine Learning (ML) en este contexto, empleando cuatro estrategias de escalado de características y clasificadores como Support Vector Machine, Ada Boost, Random Forest y K-Nearest Neighbors, entre otros. Los experimentos se realizaron con cuatro conjuntos de datos estándar (para Niños pequeños, Niños, Adolescentes y Adultos), y los resultados mostraron que Ada Boost alcanzó una precisión del 99.25% para los Niños pequeños y 97.95% para los Niños, mientras que Linear Discriminant Analysis obtuvo una precisión del 97.12% para los Adolescentes y 99.03% para los Adultos. Además, los métodos de escalado Normalizer y Quantile Transformer demostraron ser los más efectivos para los conjuntos de datos de Niños pequeños y Adolescentes, respectivamente. La investigación también destaca la importancia de las características mediante varias técnicas de selección, lo que ayuda a tomar decisiones en la detección temprana del TEA, ofreciendo un enfoque prometedor para la mejora de los métodos actuales de diagnóstico.

En el artículo de Alghamdi et al. [12], se expresó que en las plataformas de redes sociales muchos usuarios emplearon estos medios para reflejar sus vidas personales. Estos usuarios diferían en cuanto a antecedentes, idioma, edad y nivel educativo. La estrecha relación entre estas plataformas y sus usuarios generó una cantidad enorme de información que podía ser explotada por los investigadores. Sus publicaciones fueron analizadas utilizando procesamiento de lenguaje natural (PLN) para predecir rasgos psicológicos como la depresión. Sin embargo, hasta donde se sabe, ningún estudio había utilizado las redes sociales para predecir trastornos de salud mental en publicaciones en árabe, especialmente la depresión. Por lo tanto, en dicho estudio se investigó la aplicación de procesamiento de lenguaje natural y aprendizaje automático en texto en árabe para la predicción de la depresión, y se evaluó y comparó el rendimiento de los métodos. La metodología de investigación se

basó en la recopilación de texto en árabe de foros en línea y en la aplicación de un enfoque basado en léxicos o de un enfoque basado en machine learning. En el primer enfoque, se creó el léxico ArabDep y se utilizó un algoritmo basado en reglas para predecir síntomas de depresión utilizando dicho léxico; sin embargo, en el segundo enfoque, los datos fueron anotados con la ayuda de un psicólogo, se extrajeron características de texto de publicaciones en árabe y finalmente se aplicaron algoritmos de machine learning, con los cuales se lograron predecir síntomas de depresión. Los resultados mostraron que los enfoques aplicados tuvieron un rendimiento prometedor en la predicción de síntomas de depresión, con una precisión de más del 80%, una tasa de recuperación del 82% y una precisión del 79%.

En el trabajo realizado por Zhao et al. [13], comentan que el diagnóstico de autismo se basa en revisar el comportamiento del paciente, pero es un proceso lento y laborioso. Para mejorar esto, se utilizó: análisis discriminante lineal, máquinas de vectores de soporte, árbol de decisiones, bosque aleatorio y k vecinos más próximos (clasificadores de Machine Learning) para probar que se puede diagnosticar autismo utilizando cadenas cinemáticas restringidas. Veinte niños con TEA y veintitrés con desarrollo típico realizaron tareas motoras, y se usaron cinco algoritmos de AA para analizar sus movimientos. Como resultado, se obtuvo que el algoritmo KNN mostró la mayor precisión (88.37%) en la identificación del TEA, demostrando que este y otros algoritmos como máquina de vectores, análisis discriminatorio lineal, árbol de decisiones, bosque aleatorio, son de gran ayuda para clasificar la precisión de herramientas como características cinemáticas restringidas para la identificación de autismo.

En la investigación de Meng y Zhang [14], se explica que los jóvenes universitarios son el grupo más activo, sensible y vulnerable a diversos problemas psicológicos en la sociedad actual, observándose un incremento en la incidencia de ansiedad, depresión y tasas de suicidio. Con el fin de prestar una atención más efectiva al desarrollo psicológico de los estudiantes universitarios, este estudio propone un método para identificar automáticamente la ansiedad en este grupo utilizando un sistema difuso Takagi-Sugeno-Kang (TSK) y características profundas. Primero, se realiza un preprocesamiento de los EEG recopilados de los estudiantes. Luego, se emplea una red neuronal convolucional (CNN) para extraer características profundas de los datos ingresados. Para finalizar, se aplica el sistema difuso TSK para clasificar estas características y obtener el resultado final del reconocimiento. A través de experimentos realizados en conjuntos de datos estándar y conjuntos de datos creados específicamente para este estudio, se ha confirmado la eficacia del método propuesto para identificar la ansiedad. Además, se ha demostrado que las características profundas proporcionan una información más completa que las características tradicionales. La

capacidad del sistema difuso TSK para lidiar con el ruido garantiza un rendimiento sólido en la clasificación y generalización. Los resultados obtenidos permiten identificar rápidamente a los estudiantes con trastornos de ansiedad y facilitan la investigación sobre los problemas psicológicos en este grupo utilizando el sistema difuso TSK, lo que permite aumentar la eficacia de las escuelas y profesores al investigar casos de problemas psicológicos en estudiantes

Bases teóricas:

Inteligencia Artificial:

Rouhiainen [15] explica que la inteligencia artificial es la habilidad de las máquinas de lograr un aprendizaje de datos y utilizarlo para tomar decisiones igual a un ser humano, con la ventaja de que estos dispositivos no necesitan reposo y son capaces de analizar volúmenes grandes de datos al mismo tiempo. De esta manera, se logra una menor cantidad de errores, ahorrar tiempo y dinero; y automatizar procesos. La inteligencia artificial ya ayuda en distintos ámbitos de la vida de los humanos, tales como:

- Ciberseguridad: importante para bancos y sistemas de pagos.
- Clasificación de objetos: visto principalmente en la industria de vehículos.
- Contenido en redes sociales: Se utiliza como herramienta de marketing o para concientizar sin fines de lucro a distintas organizaciones.
- Procesamiento de datos de pacientes: Mejora la eficacia y eficiencia de la atención médica.
- Reconocimiento de imágenes: importante para múltiples ámbitos de la industria.

También define machine learning como una disciplina principal de la inteligencia artificial. Permite a las máquinas u ordenadores, sin estar programadas para ello, aprender, permitiendo predicciones de alguna situación en particular. Utiliza algoritmos con la capacidad de aprender a partir de patrones de datos. Son tres los tipos de aprendizaje:

- Aprendizaje supervisado: en el cual, los algoritmos utilizan datos, los cuales con anterioridad han sido organizados, utilizando estos para categorizar la nueva información. La intervención humana es necesaria para la retroalimentación.

- Aprendizaje no supervisado: los algoritmos no utilizan datos que hayan sido organizados con anterioridad; deben de categorizar la nueva información por sus propios medios sin la intervención humana.
- Aprendizaje de refuerzo: cada vez que un algoritmo acierte, debemos indicar que ha realizado su tarea correctamente, de esta manera el algoritmo va a prendiendo a medida que obtiene experiencia.

Machine Learning tiene 7 tipos principales de algoritmos:

- Algoritmos de aprendizaje profundo: Estos se ejecutan a lo largo de varias capas de algoritmos de redes neuronales, pasando a medida que avanzan una representación cada vez más simplificada.
- Algoritmos de reducción de dimensión: Reducen la cantidad de variables a tener en cuenta para hallar la información necesaria.
- Algoritmos de redes neuronales: Existen varias unidades en distintas capas, conectándose unas con otras. Utilizadas para trabajar con datos de alta dimensión que tengan relaciones no lineales, con una relación entre variables que sea difícil de comprender.
- Algoritmos de árbol de decisión: Se utiliza un método de bifurcación para analizar cada posible resultado dependiendo de las decisiones que se tomen. Cada nodo es una variable específica y, cada rama, un resultado.
- Algoritmos de agrupación: Se trabaja con datos no etiquetados, por lo que se utilizan en el aprendizaje no supervisado. Busca grupos en los datos, utilizando la variable K para contabilizar los grupos y así asignar iterativamente a los K grupos cada punto de datos
- Algoritmos bayesianos: Es un tipo de algoritmo de clasificación. Permite predecir una categoría a partir de un conjunto dado de características basándose en la probabilidad.
- Algoritmos de regresión: El programa comprende las relaciones entre las variables. Se enfoca en una dependiente y varias cambiantes, volviéndose muy útil para el pronóstico.

En este trabajo se empleó el algoritmo K-Nearest Neighbors (KNN), un método de aprendizaje supervisado ampliamente utilizado en tareas de clasificación y regresión. KNN clasifica las nuevas instancias comparándolas con los vecinos más cercanos en el conjunto de datos, utilizando métricas de distancia para identificar patrones y asignar etiquetas. Este enfoque fue seleccionado debido a su simplicidad y efectividad en problemas donde los

datos presentan relaciones complejas que pueden identificarse mediante proximidad en un espacio multidimensional.

K Vecino más cercano:

Según IBM [16], KNN es un método para la clasificación de casos a partir de su parecido con otros casos, denominados "vecinos". A los nuevos casos, denominados "reservas", se les calcula la distancia con los otros casos del modelo y se clasifican en categorías según el mayor número de vecinos más próximos. Este método se utiliza para reconocer patrones en los datos sin requerir coincidencias exactas. En el presente trabajo, se emplea KNN para identificar patrones en los datos de los estudiantes, lo cual permitirá apoyar en el diagnóstico temprano de posibles trastornos psicológicos, facilitando así una atención más efectiva y oportuna.

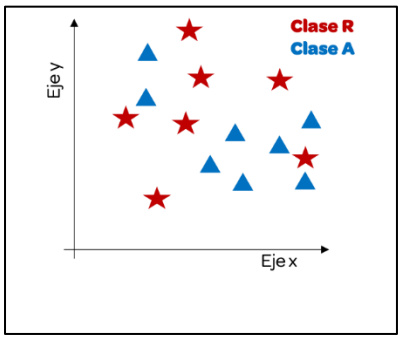

Fig. 1 Método K vecinos más próximos [17]

CRISP-DM:

Cross-Industry Standard Process for Data Mining, es definido según IBM [18] como una metodología utilizada para minería de datos, contiene descripciones de sus respectivas fases y tareas, ofreciendo un resumen del ciclo vital de la minería de datos. Es flexible, permitiendo en el proyecto avanzar y retroceder entre las distintas fases. Estas son:

- Comprensión del negocio: Se busca alinear los objetivos de negocio con los del aprendizaje automático.
- Comprensión de los datos: Conocer los datos, estructura, distribución y calidad de estos.
- Preparación de los datos: Obtener datos finales, mediante limpieza de datos para una posterior integración.
- Modelado: Construcción de un algoritmo para los objetivos del proyecto, seleccionando técnicas, estrategias y evaluar su fiabilidad.

- Evaluación: Se evalúa la integración del algoritmo con el objetivo de negocio.
- Despliegue: Desplegar los resultados a los usuarios finales, con posteriores mantenimientos.

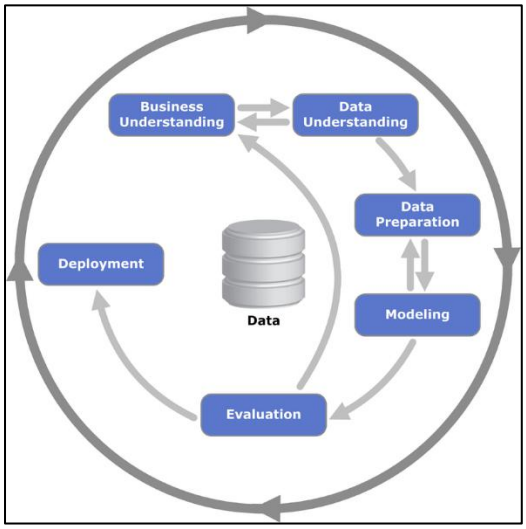

Fig. 2 Fases de metodología CRISP-DM [18]

Trastornos psicológicos:

En el artículo presentado en MedlinePlus [19], se menciona que los trastornos psicológicos son padecimientos que tienen impacto en el sentimiento, pensamiento, comportamiento y estado de ánimo del paciente. Se dividen entre ocasionales o crónicas, afectando la capacidad de relacionarse con otras personas y el funcionamiento cada día del afectado.

Los trastornos psicológicos abarcan una variedad de categorías, tales como:

- Trastornos de ansiedad
- Trastornos psicóticos
- Trastornos de estado de ánimo
- Trastornos de la alimentación
- Trastornos de la personalidad
- Trastorno de estrés postraumático
- Trastorno del neurodesarrollo

En esta investigación, se abordarán específicamente tres trastornos del neuro-desarrollo: trastorno del espectro autista, trastorno por déficit de atención e hiperactividad (TDAH), y el trastorno del lenguaje y fonológico.

Los factores que contribuyen a la probabilidad de sufrir un trastorno psicológicos son los siguientes:

- Exposición a un virus o producto químicos durante el embarazo

- Consumo de drogas o alcohol
- Tener una enfermedad seria y/o terminal
- Sentirse solitario o aislado
- Genética
- Su experiencia de vida
- Factores biológicos
- Lesión cerebral traumática

Materiales y métodos

La presente investigación fue aplicada según OECD [20], puesto que, con la adquisición de nuevos conocimientos, se planea cumplir con un objetivo y propósito práctico específico, siendo en este caso, el desarrollar una aplicación predictiva de apoyo para el diagnóstico de trastornos psicológicos en alumnos, con el fin de comenzar su tratamiento lo más pronto posible.

Los métodos de investigación que se empleó basándonos en el libro de Bernal [21] se presentan en la tabla I:

TABLA I

MÉTODOS DE INVESTIGACIÓN

Método	Sustento por el cual será empleado en la investigación
Analítico	Se estudiarán de forma particular cada trastorno psicológico que puedan sufrir los estudiantes de que se encuentren en primaria y secundaria.
Deductivo	Se obtendrá información validada a partir de la investigación realizada para su aplicación.
Implementación	Se implementará un sistema predictivo mediante una aplicación web.

Para el desarrollo de este proyecto, se utilizó la Metodología CRISP-DM para el apartado de machine learning, el cual comprende las siguientes fases definidas por IBM [22]:

- **Comprensión del negocio.**
  - Fijar los objetivos de negocio.
  - Evaluación de la situación actual.
  - Establecer los objetivos a nivel de aprendizaje automático.
  - Elaboración de un plan de proyecto.
- **Compresión de datos.**
  - Realizar procesos de captura de datos.

    - Brindar una descripción del juego de datos.
    - Ejecutar tareas para la exploración de datos.
    - Gestión de la calidad de datos para la identificación de problemas y proponer soluciones.
- **Preparación de datos.**
    - Establecer el universo de datos.
    - Limpieza de datos.
    - Elaboración de un juego de datos idóneo para ser usado en modelos de aprendizaje automático.
    - Integrar datos de fuentes hírbridas.
- **Modelado.**
    - Elegir las técnicas de aprendizaje automático que más se adecúen a el dataset y los objetivos.
    - Determinar una estrategia para verificar la calidad del algoritmo.
    - Construcción de un algoritmo utilizando las técnicas seleccionadas sobre el juego de datos.
    - Adecuar el modelo a los objetivos establecidos anteriormente.
- **Evaluación.**
    - Evaluación del algoritmo.
    - Revisar todo el proceso de aprendizaje automático hasta ese momento
    - Repetir fases anteriores de ser necesario.
- **Despliegue.**
    - Diseño de un procedimiento de despliegue del algoritmo.
    - Seguimiento y mantenimiento de la parte más operativa de esta fase.

En la última fase, el modelo predictivo se desplegó en una aplicación web, la cual fue desarrollada con la metodología XP, con las siguientes fases según Beck y Andres [23]:

- **Fase de exploración.**
    - Evaluar la situación actual.

- **Fase de planificación.**
    - Establecer las historias de usuario.
- **Fase de diseño.**
    - Diseñar la arquitectura del sistema.
    - Diseñar la infraestructura del sistema.
    - Diseñar la base de datos.
    - Realizar tarjetas de Clase-Responsabilidad-Clase.
    - Diseñar interfaces
- **Fase de codificación.**
    - Crear relaciones en los modelos.
    - Crear controladores de los modelos.
    - Codificar interfaces.
- **Fase de pruebas.**
    - Realizar pruebas de caja blanca.
    - Realizar pruebas de caja negra.

Resultados y discusión

Determinar el algoritmo de predicción más apropiado para el apoyo del diagnóstico de trastornos psicológicos.

Existen cuatro tipos de aprendizaje en machine learning según SAP [24]: supervisado, no supervisado, de refuerzo y semisupervisado. El aprendizaje supervisado adquiere conocimiento utilizando datos etiquetados, a diferencia de el no supervisado, el cual trabaja con datos no etiquetados. El aprendizaje de refuerzo entrena modelos mediante retroalimentación de recompensas y castigos, y el semisupervisado combina datos no estructurados con una pequeña cantidad de datos etiquetados para mejorar el aprendizaje.

Para este proyecto, dado que se utilizarán datos etiquetados y un dataset previo, el aprendizaje será supervisado. Según Bismart [25], este se divide en clasificación, que asigna objetos a categorías basándose en datos etiquetados, y regresión, que predice valores numéricos continuos. Se usará clasificación, ya que el modelo determinará la

categoría (un trastorno psicológico) a la que pertenece una nueva variable basada en las respuestas del cuestionario.

Proporcionar una descripción del juego de datos:

Las variables que se tendrán en cuenta se muestran en la tabla II:

TABLA II

	Variable	Tipo de variable	Valores posibles
Independiente	Sociabilidad	Numérica discreta	0-42
	Empatía	Numérica discreta	0-24
	Fijación	Numérica discreta	0-39
	Creatividad	Numérica discreta	0-18
	Percepción	Numérica discreta	0-27
	Desatención	Numérica discreta	0-9
	Hiperactividad	Numérica discreta	0-9
	Lenguaje	Numérica discreta	0-5
	Fonética	Numérica discreta	0-6
	Edad	Numérica discreta	5-17

Dependiente	Diagnóstico	Categórica nominal	• Sano. • Autismo. • Lenguaje. • Fonologico. • Lenguaje_fonologico. • TDAH. • TH. • TDA. • Autismo_TDA. • TH_lenguaje. • Autismo_TDAH_lenguajefonologico. • Autismo_TDA_lenguajefonologico. • Autismo_TH_lenguaje. • Autismo_TDA_lenguaje. • Autismo_lenguaje. • Autismo_TH. • Autismo_TDAH. • TH_lenguajefonologico. • TDAH_lenguajefonologico. • Autismo_TDAH_lenguaje. • Autismo_lenguajefonologico. • Autismo_TH_lenguajefonologico. • TDA_lenguajefonologico. • TDA_lenguaje. • TDAH_lenguaje.

Seleccionar las técnicas de aprendizaje automático más adecuadas para nuestro juego de datos y nuestros objetivos:

Posteriormente, se utilizaron 13 antecedentes para encontrar los algoritmos más utilizados en investigaciones de machine learning relacionadas con la salud mental, lo cual se ve en la tabla III:

TABLA III

Antecedente	Algoritmo	EXACTITUD	RECALL	PRECISIÓN	VALOR-F	CURVA ROC
[26]	KNN	87.23%	-	-	-	-
[27]	Naive Bayes	88%	88%	87%	87%	-
	KNN	91%	91%	85%	87%	-
	Máquinas de soporte vectorial	92%	92%	84%	84%	-
[28]	KNN	80.42%	-	-	-	87%
	Árbol de decisión	80.69%	-	-	-	86%
	Random Forest	81.22%	-	-	-	90%
	Stacking	81.75%	-	-	-	86%
[29]	Máquinas de soporte vectorial	89.39%	89.39%	89.4%	89.39%	-
	Multinominal Naive Bayes	89.07%	89.08%	89.08%	89.07%	-
	Descenso de gradiente estocástico	89.43%	89.43%	89.45%	89.42%	-
	Random Forest	87.69%	87.7%	87.69%	87.69%	-
[30]	Red neuronal artificial	75.5%	-	75.65%	-	-
	Random Forest	72.9%	-	75.65%	-	-
[31]	AdaBoost	93.8%	95.15%	92.5%	-	-
	Máquinas de soporte	94.5%	91.3%	93.75%	-	-

	vectorial					
	XGB	94.5%	91.3%	94.05%	-	-
	Random Forest	95.05%	91.6%	94.65%	-	-
	Red neuronal artificial	95.9%	98.75%	94.9%	-	-
[32]	Random Forest	-	-	60.73%	-	-
	Red neuronal artificial	-	-	76.8%	-	-
[33]	Máquinas de soporte vectorial	85%	85.6%	85.4%	85%	-
	Multinominal Nave Bayes	78%	78.6%	77%	78%	-
[34]	Red neuronal convolusional	61%	61%	69%	63%	-
[35]	Random Forest	-	63.15%	-	-	-
	Red Neuronal Artificial	-	75.2%	-	-	-
[36][37]	KNN	100%	100%	100%	100%	-
	Máquinas de soporte vectorial	99.45%	98.59%	100%	99.28%	-
	Random Forest	98.54%	96.45%	99.84%	98.06%	-
[38]	Máquinas de soporte vectorial	-	69.65%	62.55%	61.1%	69.7%
	Naive Bayes	-	70.35%	63.5%	63.45%	77.6%
	Random Forest	-	68.35%	65.95%	66.9%	77.6%

Luego de ver la tabla anterior, podemos ver en la tabla IV que los algoritmos que mejor clasificaban en estos antecedentes son:

TABLA IV

Algoritmo	Número de veces usado
Red Neuronal Artificial	4
KNN	3
Random Forest	2
Descenso de gradiente ectocástico	1
Red Neuronal Convolusional	1
Máquinas de soporte vectorial	1

Estos algoritmos se probarán con el dataset a utilizar en esta investigación en Weka, un software con el cual podremos utilizar herramientas para medir su desempeño. Esto se visualizará en las figuras 3 a 5:

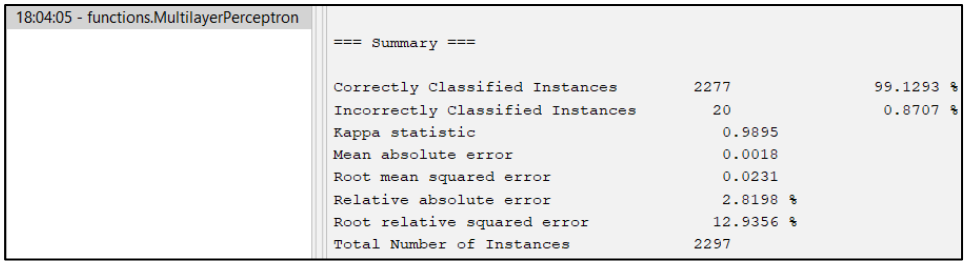

Fig. 3 Probando el dataset con Redes neuronales

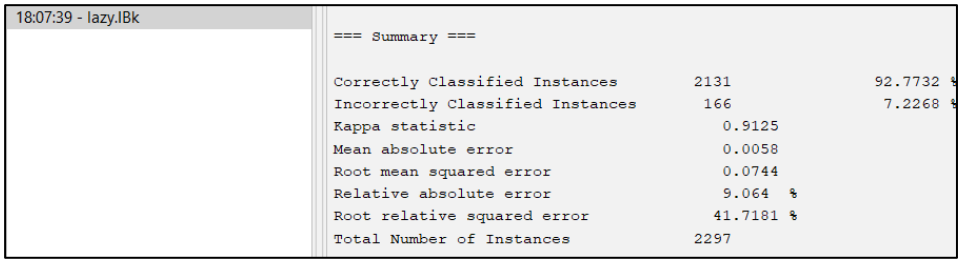

Fig. 4 Probando el dataset con KNN

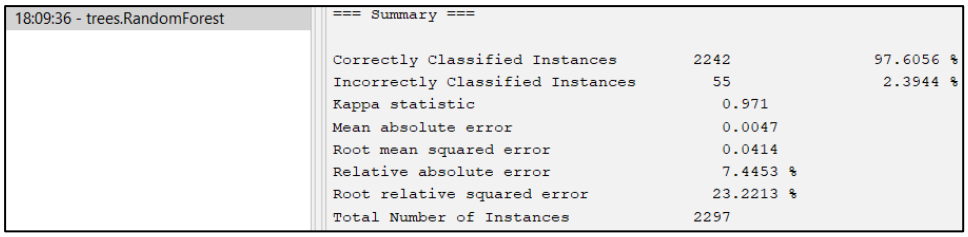

Fig. 5 Probando el dataset con Random Forest

Cada algoritmo de clasificación planteado en el anterior paso fue evaluado en los siguientes factores:

- Instancias correctamente clasificadas (ICC)
- Instancias incorrectamente clasificadas (IIC)
- Coeficiente Kappa (CK)
- Error absoluto medio (EAM)
- Distancia media cuadrática mínima (DMCM)
- Error relativo (ER)
- Error cuadrático relativo de la raíz (ECRR)

A continuación, La tabla V con las métricas utilizadas para cada algoritmo:

TABLA V

Algoritmo	ICC	IIC	CK	EAM	DMCM	ER	ECRR
Redes Neuronales	99.13%	0.87%	0.989	0.002	0.0231	2.82%	12.94%
KNN	92.77%	7.23%	0.913	0.006	0.0744	9.06%	41.72%
Random Forest	97.61%	2.39%	0.971	0.005	0.0414	7.45%	23.22%

Como se puede verificar en la tabla anterior, estos tres algoritmos otorgan muy buenos resultados, teniendo un porcentaje de instancias correctamente clasificadas mayor al 90%. Betolaza [39] describe las redes neuronales como un algoritmo capaz de trabajar con datos incompletos luego de haber sido entrenado, posee tolerancia a fallos y almacenamiento del conocimiento en la propia red en lugar de una base de datos exterior, no obstante, es muy exigente a nivel de hardware, requiriendo procesadores con procesamiento en paralelo, dificulta la introducción de datos a la red y con el paso del tiempo estas tienden a perder su eficiencia; IBM [40] se menciona que el algoritmo de

Random Forest es flexible con respecto a la falta de datos, logrando estimarlos en caso falten una parte de estos, sin embargo es muy lento para el procesamiento de datos, debido a que procesa para cada árbol individual, requiriendo muchos más recursos, y, IBM [41] nuevamente, describe al KNN como un algoritmo que a pesar de que al no tener un buen escalado y ocupe más memoria de almacenamiento frente a otros algoritmos de clasificación, es muy fácil de implementar, se adapta fácilmente a nuevas muestras de entrenamiento y requiere de pocos parámetros para su funcionamiento.

Debido a esto, para ofrecer una solución simple y efectiva, que se mantenga trabajando correctamente con el paso del tiempo y sea fácil de actualizar el conjunto de datos, además de no consumir tantos recursos, se utilizará el algoritmo de KNN para la clasificación de trastornos psicológicos en alumnos de la institución educativa.

Desarrollar un modelo predictivo empleando el algoritmo seleccionado para el apoyo del diagnóstico de trastornos psicológicos.

Primero tenemos la función “manhattanDistance”, que calcula la distancia entre dos puntos en un espacio de características. La distancia de Manhattan se define como la suma de las diferencias absolutas entre las coordenadas de los puntos en cada dimensión.

IBM [41] explica que la fórmula de la distancia de Manhattan mide el valor absoluto entre dos puntos:

Distancia de Manhattan (M) = |x1 - x2| + |y1 - y2|.

La función toma dos parámetros: $point1 y $point2, que representan los dos puntos que se van a comparar. Ambos son representados como arrays, donde cada elemento de estos pertenece a una coordenada en el espacio de características. Luego, realiza un bucle que recorre todas las coordenadas de los puntos. En cada iteración, calcula la diferencia absoluta entre las coordenadas correspondientes de los dos puntos y las suma al valor acumulado de la distancia. Esto se realiza utilizando la función abs() para obtener el valor absoluto de la diferencia. Después de recorrer todas las coordenadas, la función devuelve el valor total de la distancia de Manhattan entre los dos puntos, como se puede ver en la figura 6:

```php
public function manhattanDistance($point1, $point2)
{
    $distance = 0;
    $numFeatures = count($point1);

    for ($i = 0; $i < $numFeatures; $i++) {
        $distance += abs($point1[$i] - $point2[$i]);
    }

    return $distance;
}
```

Fig. 6 función “manhattanDistance”

Continuando, tenemos la función “normalizeFeatures” que se encarga de normalizar las características de los datos. La normalización es un proceso en el cual se realiza un ajuste a los valores de las características para que se encuentren dentro de un rango específico, en este caso [0, 1]. Esto es útil para garantizar que las características tengan la misma escala y no tengan más relevancia unas sobre otras al realizar cálculos o análisis.

La función toma un parámetro $data, que representa los datos a normalizar. Los datos se representan como una matriz, donde cada fila corresponde a una instancia y cada columna corresponde a una característica.

La función inicializa con dos arreglos: $minValues y $maxValues, ambos con una longitud igual al número de características. Estos arreglos se utilizan para almacenar los valores mínimos y máximos de cada característica.

Luego, se realiza un bucle para encontrar los valores mínimos y máximos de cada característica recorriendo todas las instancias de datos. En cada iteración, se compara el valor actual de la instancia con el valor mínimo y máximo almacenados en los arreglos correspondientes, actualizando de cumplirse que es mayor o menor que los que poseen esos puestos.

Después de obtener los valores mínimos y máximos de cada característica, se realiza otro bucle para normalizar las características en el rango [0, 1]. En cada iteración, se calcula el rango de la característica restando el valor mínimo al valor máximo. Luego, se verifica si el rango es diferente de cero para evitar divisiones por cero. Si el rango es distinto de cero, se normaliza el valor de la característica restando el valor mínimo y dividiendo por el rango. Si el rango es cero, se establece el valor de la característica en cero como valor predeterminado.

La fórmula matemática se observa en el artículo de Li et al. [42], la cual es:

$$\mathbf{X(normalizada)=x\text{-}min(x)/(max(X)-min(X))}$$

Al finalizar, se devuelve un arreglo con los datos normalizados. Todo esto se ve en la figura 7:

```php
public function normalizeFeatures($data)
{
    $numFeatures = count($data[0]);
    $minValues = array_fill(0, $numFeatures, PHP_FLOAT_MAX);
    $maxValues = array_fill(0, $numFeatures, PHP_FLOAT_MIN);

    // Encontrar los valores mínimos y máximos de cada característica
    foreach ($data as $instance) {
        for ($i = 0; $i < $numFeatures; $i++) {
            $minValues[$i] = min($minValues[$i], $instance[$i]);
            $maxValues[$i] = max($maxValues[$i], $instance[$i]);
        }
    }

    // Normalizar las características en el rango [0, 1]
    foreach ($data as &$instance) {
        for ($i = 0; $i < $numFeatures; $i++) {
            $range = $maxValues[$i] - $minValues[$i];
            if ($range != 0) {
                $instance[$i] = ($instance[$i] - $minValues[$i]) / $range;
            } else {
                $instance[$i] = 0; // Valor predeterminado cuando el rango es cero
            }
        }
    }

    return $data;
}
```

Fig. 7 función “normalizeFeatures”

Se sigue con la función “predict, encargada de realizar una predicción de clase para un punto de prueba utilizando el algoritmo de K-Nearest Neighbors (K-NN).

El primer paso es leer los datos de entrenamiento desde un archivo CSV. Se inicializan dos arreglos, $trainingData y $labels, para almacenar las variables independientes y la variable dependiente de cada instancia (fila) de entrenamiento, respectivamente.

A continuación, se aplica la normalización de características a los datos de entrenamiento y al punto de prueba con la función ya definida previamente “normalizeFeatures”, que normaliza los arreglos $trainingData y [$testPoint] (el punto de prueba , que sería el nuevo vecino a ser comparado con los anteriores, se convierte en un arreglo de un solo elemento para ser compatible con la función “normalizeFeatures”). Esto garantiza que las características estén en el mismo rango para un cálculo de distancia más preciso.

Después de la normalización, se calculan las distancias entre el punto de prueba y todas las instancias de entrenamiento utilizando la función “manhattanDistance”. Las distancias se almacenan en el arreglo $distances.

A continuación, se ordenan las distancias de menor a mayor utilizando asort(), lo que nos permite obtener los primeros K vecinos más cercanos. Se utiliza la función array_slice para seleccionar los primeros K elementos del arreglo $distances, y se almacenan en el arreglo $neighbors.

A continuación, se cuenta la frecuencia de cada clase entre los vecinos. Se inicializa el arreglo $classCount para almacenar el recuento de cada clase. Se recorren los vecinos y se incrementa el contador correspondiente a la clase en $classCount.

Luego, se encuentra la variable dependiente con mayor frecuencia entre los vecinos. Se inicializa la variable $prediction para almacenar la predicción de la variable dependiente y $maxCount para almacenar el recuento máximo. Se recorre el arreglo $classCount y se compara el recuento de cada clase con $maxCount. Si el recuento es mayor, se actualiza la variable $prediction con la clase correspondiente y se actualiza $maxCount.

Finalmente, se devuelve la predicción. Esto se ve en la figura 8:

```php
public function predict($k, $trainingDataFile, $labelColumn, $testPoint)
{
    // Leer los datos de entrenamiento desde un archivo CSV
    $trainingData = array();
    $labels = array();

    $csvFile = fopen($trainingDataFile, 'r');
    if ($csvFile) {
        // Leer los encabezados y descartarlos
        $headers = fgetcsv($csvFile);

        while (($line = fgetcsv($csvFile)) !== false) {
            $features = array_slice($line, 0, count($line) - 1);
            $class = end($line);
            $trainingData[] = $features;
            $labels[] = $class;
        }
        fclose($csvFile);
    }

    // Normalizar características
    $trainingData = $this->normalizeFeatures($trainingData);
    $testPoint = $this->normalizeFeatures([$testPoint])[0];

    // Calcular distancias
    $distances = array();
    $numTrainingInstances = count($trainingData);
    for ($i = 0; $i < $numTrainingInstances; $i++) {
        $distances[$i] = $this->manhattanDistance($trainingData[$i], $testPoint);
    }


    // Ordena las distancias de menor a mayor
    asort($distances);

    // Obtiene los primeros K vecinos más cercanos
    $neighbors = array_slice($distances, 0, $k, true);

    // Cuenta las clases de los vecinos
    $classCount = array();
    foreach ($neighbors as $index => $distances) {
        $class = $labels[$index];
        if (isset($classCount[$class])) {
            $classCount[$class]++;
        } else {
            $classCount[$class] = 1;
        }
    }

    // Encuentra la clase con mayor frecuencia entre los vecinos
    $prediction = '';
    $maxCount = 0;
    foreach ($classCount as $class => $count) {
        if ($count > $maxCount) {
            $lastCommaPosition = strrpos($class, ";");
            $prediction = substr($class, $lastCommaPosition + 1);
            $maxCount = $count;
        }
    }

    // Leer las etiquetas reales del archivo CSV de datos de prueba
    $actualLabels = array();
    $csvTestFile = fopen($testPoint, 'r');
    if ($csvTestFile) {
        // Leer los encabezados y descartarlos
        $headers = fgetcsv($csvTestFile);

        while (($line = fgetcsv($csvTestFile)) !== false) {
            $actualClass = $line[$labelColumn];
            $actualLabels[] = $actualClass;
        }
        fclose($csvTestFile);
    }
```

Fig. 8 función “prediction”

Este algoritmo se ejecuta posteriormente en la función “saveFinalInstace” en la cual la función “predict” recibe como parámetros los puntajes obtenidos al responder el cuestionario en la aplicación web para cada una de las variables independientes y la edad del alumno, fijándose que se comparará con 10 vecinos cercanos. Posteriormente a esto, la predicción es registrada en la base de datos como un diagnóstico para ser revisado por el experto. Esto se refleja en la figura 9:

```
$trainingDataFile = 'D:/Mio/U/Trabajos/Ciclo X/Seminario de tesis 1/AppDiagnostico/dataset_final.csv';

$records = TestVariableRecord::where('test_instance_id', '=', $rowTestInstance->id)->orderBy('test_variable_type_id')->get();
$scores = $records->pluck('sum')->all();

$testPoint = array_merge($scores, [$rowTestInstance->age]);
$k = 10;
$prediction = $this->predict($k, $trainingDataFile, $testPoint);

$rowTestDiagnose = new TestDiagnose();
$rowTestDiagnose->diagnose = $prediction;
$rowTestDiagnose->test_instance_id = $rowTestInstance->id;
$rowTestDiagnose->save();
```

Fig. 9 Realizando una predicción

Evaluar la precisión del modelo predictivo para determinar su precisión como herramienta para el apoyo del diagnóstico de trastornos psicológicos.

Se evaluó el modelo predictivo utilizando una matriz de confusión, con las siguientes métricas:

- Exactitud: 0.86 de un valor máximo de 1.
- Precisión: 0.79 de un valor máximo de 1.
- Recall: 0.54 de un valor máximo de 1.
- Especificidad: 0.92 de un valor máximo de 1.
- Puntaje F1: 0.69 de un valor máximo de 1.

Como se puede ver, una de las métricas es precisión, que obtuvo un valor de 0.79, demostrando una precisión sólida. De esta forma, se evaluó la precisión del modelo predictivo para determinar su precisión como herramienta para el apoyo del diagnóstico de trastornos psicológicos

Desplegar el modelo predictivo en una aplicación web que permita facilitar el diagnóstico de trastornos psicológicos.

Agregamos tablas a la base de datos del sistema de la I.E.P necesarias para crear y gestionar los cuestionarios psicológicos. Posteriormente a ello, se crearon las interfaces para responder los cuestionarios psicológicos. Al momento de responderse, es cuando se utiliza el algoritmo para realizar la clasificación y almacenarse en la base de datos. También podemos buscar los cuestionarios realizados por nombre, aula o DNI; para que finalmente se decidiera que estos fueran confirmados o desestimados. De esta forma, se desplegó el modelo predictivo en una aplicación web que permita facilitar el diagnóstico de trastornos psicológicos.

Discusiones

En el trabajo de Almadhor et al. [9], se centraron en la identificación de ansiedad en estudiantes, empleando modelos como Random Forest, Naive Bayes, XGBoost, Regresión Logística y K-Nearest Neighbors, siendo este último el algoritmo utilizado igualmente en este trabajo de investigación para el apoyo al diagnóstico de autismo, TDAH y trastorno de lenguaje y/o fonológico.

En el artículo de Nash et al. [10], buscaban detectar síntomas de TDAH en adultos utilizando aprendizaje automático, dado que las herramientas tradicionales requieren de síntomas en la niñez. En esta investigación, además de TDAH y trastorno de lenguaje y/o fonológico, también se enfocó en apoyar el diagnóstico de autismo en alumnos, logrando así evitar consecuencias negativas futuras para estos alumnos.

Hasan et al. [11], expresó la dificultad de la detección temprana de autismo, por lo que utilizaron cuatro estrategias de escalado de características y clasificadores como Ada Boost, Random Forest, Support Vector Machine y K-Nearest Neighbors, los cuales mostraron resultados significativos para la detección del trastorno neurológico. En esta investigación igualmente se utilizó K-Nearest Neighbors para apoyar al diagnóstico de autismo, además de también TDAH y trastorno de lenguaje fonológico.

En la investigación de Alghamdi et al. [12], se busca identificar la depresión en textos en idioma árabe. En el presente trabajo, se pueden realizar cuatro diagnósticos: autismo, TDAH y trastorno del lenguaje y/o fonológico, además de generar reportes para el psicólogo de la institución educativa, que le servirán de apoyo en la gestión de casos para dicha institución.

Con respecto al trabajo realizado por Zhao et al. [13], probaron múltiples algoritmos de predicción para el diagnóstico del trastorno del espectro autista (TEA), siendo el algoritmo KNN el que obtuvo mayor precisión. En el producto acreditable de esta investigación también se seleccionó KNN como el algoritmo más factible, habiéndose evaluado previamente junto con redes neuronales, Random Forest, máquinas de soporte vectorial y descenso de gradiente estocástico. No obstante, a diferencia de la referencia mencionada, este proyecto también diagnostica TDAH y trastornos del lenguaje y/o fonológico, además de autismo.

En el trabajo realizado por Meng y Zhang [14], se diagnostica únicamente ansiedad en jóvenes universitarios utilizando un sistema difuso Takagi-Sugeno-Kang (TSK), a diferencia del producto acreditable de esta investigación, que identifica cuatro trastornos diferentes: autismo, TDAH y trastorno del lenguaje y/o fonológico, abarcando a niños desde los 4 hasta adolescentes de 17 años.

Conclusiones

- Se logró determinar que el algoritmo de predicción más apropiado para el apoyo del diagnóstico de trastornos psicológicos fue el KNN, utilizando el aprendizaje supervisado. Este tipo de aprendizaje utiliza datos previamente etiquetados para entrenar el algoritmo de predicción, en este caso, KNN, el cual compara la proximidad de un punto de datos individual con los anteriores datos previamente etiquetados.
- Se logró desarrollar un modelo predictivo empleando el algoritmo seleccionado para el apoyo del diagnóstico de trastornos psicológicos, con el cual, se podrá diagnosticar entre los distintos trastornos abarcados: autismo, TDAH, trastorno del lenguaje y/o fonológico.
- Se logró evaluar la precisión del modelo predictivo para determinar su precisión como herramienta para el apoyo del diagnóstico de trastornos psicológicos, para esto se utilizaron los siguientes factores con sus respectivos resultados: exactitud (86%), precisión (79%), recall (54%), especificidad (92%) y puntaje F1 (69%).
- Se logró desplegar el modelo predictivo en una aplicación web que permita facilitar el diagnóstico de trastornos psicológicos, el cual podrá ser utilizado tanto por los alumnos o postulantes para realizar el test psicológico y obtener el diagnóstico, como por el psicólogo de la institución educativa para gestionar estos resultados.
-

Recomendaciones

- Se propone utilizar otros algoritmos de predicción para futuros proyectos, como pueden ser Redes Neuronales y Random Forest, los cuales fueron considerados al momento de seleccionar un algoritmo y también dieron buenos resultados.
- Se sugiere incluir análisis de texto mediante inteligencia artificial para lograr el diagnóstico mediante textos elaborados por el alumno.

Referencias

[1] N. Erades y A. Morales, «Impacto psicológico del confinamiento por la COVID-19 en niños españoles: un estudio transversal», rpcna, vol. 7, n.o no 3, pp. 27-34, 2020, doi: 10.21134/rpcna.2020.mon.2041. Accedido el 28 de abril de 2022. [En línea]. Disponible: https://www.revistapcna.com/sites/default/files/006_0.pdf

[2] Anonymous "Learn how to identify if your child suffers from anxiety," CE Noticias Financieras, 2021. Available: http://usat.lookproxy.com/wire-feeds/learn-how-identify-if-your-child-suffers-anxiety/docview/2572131478/se-2.

[3] EFE News Service. "Los transtornos mentales, entre las principales enfermedades en la infancia: CIENCIA ENFERMEDADES," EFE News Service, 2018. Available: https://www.proquest.com/docview/2024024413/F1A154BD59E446AFPQ/1?accountid=37610

[4] OMS. "Informe mundial sobre salud mental", OMS, 2022. Accedido el 31/10/2024. [En línea]. Disponible: https://www.who.int/es/publications/i/item/9789240050860

[5] I. Gómez-Becerra, J. M. Flujas, M. Andrés, P. Sánchez-López, y M. Fernández-Torres, «Evolución del estado psicológico y el miedo en la infancia y adolescencia durante el confinamiento por la COVID-19», rpcna, vol. 7, n.o no 3, pp. 11-18, 2020, doi: 10.21134/rpcna.2020.mon.2029. Accedido el 28 de abril de 2022. [En línea]. Disponible: https://www.revistapcna.com/sites/default/files/004_0.pdf

[6] F. Rusca-Jordan, C. Cortez-Vergara, B. C. Tirado-Hurtado, y M. Strobbe-Barbat, «Una aproximación a la salud mental de los niños, adolescentes y cuidadores en el contexto de la COVID-19 en el Perú», ACTA MEDICA PERUANA, vol. 37, n.o 4, dic. 2020, doi: 10.35663/amp.2020.374.1851. Accedido el 28 de abril de 2022. [En línea]. Disponible: http://www.scielo.org.pe/pdf/amp/v37n4/1728-5917-amp-37-04-556.pdf

[7] CONSULTA REGIONAL PARA LA AGENDA ADOLESCENTE Y JOVEN 2021-2026, MCLCP, Chiclayo, 2021. Accedido el 28 de abril de 2022. [En línea]. Disponible: https://www.mesadeconcertacion.org.pe/storage/documentos/2021-04-29/sistematizaciondeconsulta-agendaadolescencia-joven.pdf

[8] K. M. Castillo Barberán, P. G. Chávez Quimi, y M. J. Zoller Andina, «Factores familiares y escolares que influyen en los problemas de conducta y de aprendizaje en los niños», ACADEMO (Asunción), vol. 6, n.o 2, jul. 2019, doi: 10.30545/academo.2019.jul-dic.3, Accedido: 29 de abril de 2022. [En línea].

Disponible en: http://scielo.iics.una.py/pdf/academo/v6n2/2414-8938-academo-6-02-124.pdf

[9] A. Almadhor et al., "Multi-Class Adaptive Active Learning for Predicting Student Anxiety," in IEEE Access, vol. 12, pp. 58097-58105, 2024, doi: 10.1109/ACCESS.2024.3391418, Accedido: 10 de noviembre de 2024. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/10505253

[10] C. Nash, R. Nair and S. M. Naqvi, "Insights Into Detecting Adult ADHD Symptoms Through Advanced Dual-Stream Machine Learning," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 3378-3387, 2024, doi: 10.1109/TNSRE.2024.3450848. Accedido: 10 de noviembre de 2024. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/10654367

[11] S. M. Mahedy Hasan, M. P. Uddin, M. A. Mamun, M. I. Sharif, A. Ulhaq and G. Krishnamoorthy, "A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders," in *IEEE Access*, vol. 11, pp. 15038-15057, 2023, doi: 10.1109/ACCESS.2022.3232490, Accedido: 10 de noviembre de 2024. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9999443

[12] N. S. Alghamdi, H. A. Hosni Mahmoud, A. Abraham, S. A. Alanazi, y L. Garcia-Hernandez, «Predicting Depression Symptoms in an Arabic Psychological Forum», IEEE Access, vol. 8, pp. 57317-57334, 2020, doi: 10.1109/ACCESS.2020.2981834. Accedido: 13 de mayo de 2022. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9040556

[13] Z. Zhao et al., «Applying Machine Learning to Identify Autism With Restricted Kinematic Features», IEEE Access, vol. 7, pp. 157614-157622, 2019, doi: 10.1109/ACCESS.2019.2950030. Accedido: 13 de mayo de 2022. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/8886387

[14] X. Meng y J. Zhang, «Anxiety Recognition of College Students Using a Takagi-Sugeno-Kang Fuzzy System Modeling Method and Deep Features», IEEE Access, vol. 8, pp. 159897-159905, 2020, doi: 10.1109/ACCESS.2020.3021092. Accedido: 13 de mayo de 2022. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9184941

[15] L. Rouhiainen, Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro. Barcelona: Alienta Editorial, 2019. Accedido: 07 de junio de 2022. [En línea]. Disponible en: https://www.planetadelibros.com/libros_contenido_extra/40/39307_Inteligencia_artificial.pdf

[16] IBM. «Análisis vecino más cercano», IBM, 2021. Accedido: 14 de junio de 2022. [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=features-nearest-neighbor-analysis

[17] IBM. «Análisis vecino más cercano», IBM, 2021. Accedido: 14 de junio de 2022. [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=features-nearest-neighbor-analysis

[18] IBM. « Conceptos básicos de ayuda de CRISP-DM»,IBM, 2021. Accedido: 12 de junio de 2022. [En línea]. Disponible en: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

[19] MedlinePlus. « Enfermedades mentales», National Library of Medicine, 2021. Accedido: 9 de junio de 2022. [En línea]. Disponible en: https://medlineplus.gov/spanish/mentaldisorders.html

[20] OECD, Manual de Frascati 2015: Guía para la recopilación y presentación de información sobre la investigación y el desarrollo experimental. OECD, 2018. doi: 10.1787/9789264310681-es. Accedido: 21 de mayo de 2022. [En línea]. Disponible en: https://www.oecd-ilibrary.org/science-and-technology/manual-de-frascati-2015_9789264310681-es

[21] C. A. Bernal Torres, Metodología de la investigación. Distrito Federal: Pearson Educación, 2010. Accedido: 21 de mayo de 2022. [En línea]. Disponible en: https://abacoenred.com/wp-content/uploads/2019/02/El-proyecto-de-investigación-F.G.-Arias-2012-pdf.pdf

[22] IBM, "Guía de CRISP-DM de IBM SPSS Modeler," IBM Documentation, 18.4.0. Accedido el 1 de junlio de 2023. [En línea]. Disponible: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf.

[23] K. Beck y C. Andres, Extreme programming explained: embrace change, 2nd ed. Boston, MA: Addison-Wesley, 2005.

[24] SAP. "¿Qué es machine learning?" www.sap.com. Accedido el 1 de julio de 2023. [En línea]. Disponible: https://www.sap.com/latinamerica/products/artificial-intelligence/what-is-machine-learning.html

[25] Bismart, "Tipos de Análisis Predictivo: Clasificación y Regresión," Bismart Blog, Accedido el 1 de julio de 2023, [En línea]. Disponible: https://blog.bismart.com/tipos-analisis-predictivo-clasificacion-regresion.

[26] G. R, P. Prajwal Annigeri, N. Huddar, K. Kumar y K. KP, "STRESS PREDICTION IN WORKING ENVIRONMENT USING KNN", Int. Res. J. Modernization Eng. Technol. Sci., vol. 4, n. º 7, p. 1982–1986, julio de 2022. Accedido el 1 de julio de 2023. [En línea]. Disponible: https://www.irjmets.com/uploadedfiles/paper/issue_7_july_2022/28315/final/fin_irjmets1657987943.pdf

[27] A. Damayunita, R. S. Fuadi y C. Juliane, "Comparative Analysis of Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) Algorithms for Classification of Heart Disease Patients", Jurnal Online Informatika, vol. 7, n. º 2, p. 219–225, diciembre de 2022. Accedido el 1 de julio de 2023. [En línea]. Disponible: https://doi.org/10.15575/join.v7i2.919

[28] K. Vaishnavi, U. Nikhitha Kamath, B. Ashwath Rao y N. V. Subba Reddy, "Predicting Mental Health Illness using Machine Learning Algorithms", J. Physics: Conf. Ser., vol. 2161, n. º 1, pp. 012021, enero de 2022. Accedido el 1 de junio de 2023. [En línea]. Disponible: https://doi.org/10.1088/1742-6596/2161/1/012021

[29] O. Oyebode, F. Alqahtani and R. Orji, "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews," en IEEE Access, vol. 8, pp. 111141-111158, 2020, Accedido el 1 de julio de 2023. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9115602

[30] G. Tsang, S. -M. Zhou and X. Xie, "Modeling Large Sparse Data for Feature Selection: Hospital Admission Predictions of the Dementia Patients Using Primary Care Electronic Health Records," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1-13, 2021. Accedido el 1 de julio de 2023. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9268962

[31] S. A. Amin et al., "Data Driven Classification of Opioid Patients Using Machine Learning–An Investigation," in IEEE Access, vol. 11, pp. 396-409, 2023. Accedido el 1 de julio de 2023. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9991956

[32] M. A. Yanez Soffia, “Análisis sobre modelos predictores de depresión mediante la interpretación de lenguaje natural a partir de textos usando machine learning”, mayo de 2023. Accedido el 5 de setiembre de 2023. [En línea]. Disponible en: https://oa.upm.es/75140/

[33] D. G. Forero-Chaux y J. D. Vergara-Rojas, “Método automático para apoyar en la identificación de síntomas de depresión mediante el análisis en narrativa oral”, 2021, Accedido: 5 de setiembre de 2023. [En línea]. Disponible en: https://hdl.handle.net/10983/26291

[34] K. E. Charcopa Lajones y F. J. Rosemberg Fariño, «Use of machine learning algorithms for diagnosis and treatment of psychological pathologies», bachelorThesis, 2023. Accedido: 5 de setiembre de 2023. [En línea]. Disponible en: http://dspace.ups.edu.ec/handle/123456789/24405

[35] A. Calvo Mendoza, "Implementación de dos modelos predictivos para TeenSmart International que determinan la probabilidad de una persona de incurrir en intento de suicidio y en actividad sexual temprana", Tesis de maestría, Univ. Cenfotec, 2022. Accedido el 5 de septiembre de 2023. [En línea]. Disponible: https://repositorio.ucenfotec.ac.cr/handle/123456789/318

[36] F. Garcia Moreno, M. Bermudez Edo, J. Garrido y J. Perez Mármol, "Evaluación de emociones y salud emocional en mayores mediante wearables y Machine Learning", pp. 1–10, 2022. Accedido el 5 de septiembre de 2023. [En línea]. Disponible: https://www.researchgate.net/publication/365184505_Evaluacion_de_emociones_y_salud_emocional_en_mayores_mediante_wearables_y_Machine_Learning#read

[37] F. M. Garcia-Moreno, M. Bermudez-Edo, E. Rodríguez-García, J. M. Pérez-Mármol, J. L. Garrido, y M. J. Rodríguez-Fórtiz, "A machine learning approach for semi-automatic assessment of IADL dependence in older adults with wearable sensors", International Journal of Medical Informatics, vol. 157, p. 104625, 2022. Accedido el 5 de setiembre de 2023. [En línea]. Disponible en: https://doi.org/10.1016/j.ijmedinf.2021.104625

[38] L. Flesia et al., "Predicting Perceived Stress Related to the Covid-19 Outbreak through Stable Psychological Traits and Machine Learning Models," Journal of Clinical Medicine, vol. 9, no. 10, p. 3350, Oct. 2020. Accedido el 5 de setiembre de 2023. [En línea]. Disponible en: doi: 10.3390/jcm9103350

[39] X. Maestre Betolaza, "IMPLEMENTACIÓN DE REDES NEURONALES EN PLATAFORMAS HARDWARE PARA SU APLICACIÓN EN INGENIERÍA ELÉCTRICA", Tesis de maestría, BIZK. INGENIARITZA ESKOLA, Bilbao, 2021. Accedido el 5 de septiembre de 2023. [En línea]. Disponible en: https://addi.ehu.es/bitstream/handle/10810/54026/TFM_XabierMaestreBetolaza.pdf?sequence=1

[40] IBM. "¿Qué es un bosque aleatorio?". Accedido 5 de setiembre de 2023. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/random-forest

[41] IBM. "¿Qué es KNN?". Accedido 5 de setiembre de 2023. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/knn

[42] B. Li, P. Lu, . ”Normalización de datos: referencia de componente - Azure Machine Learning”, 1 de junio de 2023. ?”. Accedido 5 de setiembre de 2023. [En línea]. Disponible en: https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/normalize-data?view=azureml-api-2 (accedido 5 de octubre de 2023).

Anexos

ANEXO N° 01. CANTIDAD DE ESTUDIANTES EN LA I.E.P SANTO TORIBIO DE MOGROVEJO

INSTITUCIÓN EDUCATIVA PRIVADA
"SANTO TORIBIO DE MOGROVEJO"
CHICLAYO

ESTUDIANTES 2024

nivel	grado	seccion	TOTAL		total
			H	M	
INICIAL	3 años	A	7	3	10
		B	6	3	9
	4 años	A	10	9	19
		B	11	9	20
	5 años	A	8	9	17
		B	11	6	17
		C	11	7	18
	TOTAL INICIAL		**64**	**46**	110
PRIMARIA	1	A	20	11	31
		B	18	9	27
		C	16	12	28
	2	A	16	13	29
		B	15	11	26
		C	15	13	28
	3	A	28		28
		B	32		32
		C		32	32
	4	A	29		29
		B	29		29
		C		34	34
	5	A	34		34
		B	29		29
		C		31	31
	6	A	20		20
		B	20		20
		C		18	18
	TOTAL PRIM.		**321**	**184**	**505**
SECUNDARIA	1	A	31		31
		B	32		32
		C		35	35
	2	A	31		31
		B	32		32
		C		33	33
	3	A	28		28
		B	27		27
		C		33	33
	4	A	22		22
		B	18		18
		C		36	36
	5	A	26		26
		B	28		28
		C		34	34
	TOTAL SEC.		**275**	**171**	**446**
TOTAL ALUMNOS					**1061**