

UNIVERSIDAD CATÓLICA SANTO TORIBIO DE MOGROVEJO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN



**SISTEMA PREDICTIVO DE RENDIMIENTO ACADÉMICO EN BASE
A FACTORES INFLUYENTES EN ESTUDIANTES DEL 1º
SECUNDARIA EN UN COLEGIO DE LAMBAYEQUE**

**TESIS PARA OPTAR EL TÍTULO DE
INGENIERO DE SISTEMAS Y COMPUTACIÓN**

AUTOR

LUIS JESUS GARCIA PEREDO

ASESOR

MARLON EUGENIO VICLHEZ RIVAS

<https://orcid.org/0000-0003-2979-0731>

Chiclayo, 2021

**SISTEMA PREDICTIVO DE RENDIMIENTO ACADÉMICO EN BASE
A FACTORES INFLUYENTES EN ESTUDIANTES DEL 1º
SECUNDARIA EN UN COLEGIO DE LAMBAYEQUE**

PRESENTADA POR

LUIS JESUS GARCIA PEREDO

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

Jury Yesenia Aquino Trujillo
PRESIDENTE

Segundo Jose Castillo Zumaran
SECRETARIO

Marlon Eugenio Vilchez Rivas
VOCAL

Dedicatoria

A Dios por permitirme llegar hasta aquí con buena salud y sobre todo con vida, por dotarme de las capacidades que me permiten afrontar los retos de la vida diaria y aprender de ellos, para que nunca se aparte de mi lado.

A mis padres por ser mi motivación, por guiar mis pasos día a día, confiar en mi desde que tengo uso de razón y porque sé que su amor hacia mi será para toda la vida.

Agradecimientos

Primero a Dios por estar conmigo siempre y demostrarme que con él no hay camino imposible, por llenarme de esa Fe que no te deja rendirte y por permitirme tener a mi familia a mi lado.

A mi familia, sobre todo a mis padres por tanto amor, aliento y entrega.

Al Mgtr. Ing. Marlon Vílchez por no ser solo un asesor sino también un amigo que guía con la experiencia y las palabras correctas, a la Ing. Marya Aranguri por guiarme en diferentes momentos de mi carrera y a todos mis amigos los cuales sus palabras sirvieron de impulso.

Índice

Resumen	6
Abstract	7
Introducción	8
Revisión de literatura	10
Materiales y métodos	15
Resultados y discusión	20
Conclusiones	31
Recomendaciones	32
Referencias	33
Anexos	35

Resumen

Hoy en día los estudiantes de los primeros años de secundaria de la I.E. Sara Bullón no presentan un rendimiento académico estable, este es influenciado por factores psicológicos. La institución intenta tomar acciones para mejorar esta situación, pero no han dado resultado ya que no tienen información clave que las facilite. Frente a esta problemática se planteó como objetivo principal de la investigación, la implementación de un sistema de predicción de rendimiento académico en base a diversos factores de influencia que facilite información sobre el rendimiento académico de los estudiantes. Primero se aplicó minería de datos con el uso de la metodología CRISP-DM para extraer información de la data de los estudiantes y establecer el algoritmo base del sistema de predicción (Random Forest o Decision Tree). Luego se automatizó el modelo para que mediante una interfaz se ingresen datos del estudiante y devuelva una predicción en tiempo real, fue integrada en un sistema que abarca el proceso de autorización de los apoderados de los estudiantes y muestreo de reportes para las predicciones realizadas, este sistema se realizó mediante la metodología RUP. Se concluyó que el algoritmo con la medida más óptima de asertividad fue Random Forest con 0.95, el cual se usó como base en la automatización del modelo, también se obtuvo el grado de importancia de cada variable (factor de influencia) en el modelo y se comparó la correlación obtenida de cada dimensión psicológica con respecto al rendimiento académico en base a la escala de Pearson.

PALABRAS CLAVE: Rendimiento Académico, Minería de datos, Random Forest, Decision Tree, Automatización, Predicción en tiempo real.

Abstract

Today the students of the first years of secondary school of the I.E. Sara Bullón doesn't present stable academic performance, this is influenced by psychological factors. The institution tries to take actions to improve this situation, but they have not been successful since they don't have key information to facilitate them. Faced with this problem, the main objective of the research was to implement a prediction system for academic performance based on various influencing factors that provides information on the academic performance of students. First, data mining was applied with the use of the CRISP-DM methodology to extract information from the students' data and establish the base algorithm of the prediction system (Random Forest or Decision Tree). Then the model was automated so that through an interface student data is entered and a prediction is returned in real time, it was integrated into a system that encompasses the authorization process of the students' proxies and sampling of reports for the predictions made, this system was performed using the RUP methodology. It was concluded that the algorithm with the most optimal measure of assertiveness was Random Forest with 0.95, which was used as a basis in the automation of the model, the degree of importance of each variable (influence factor) in the model was also obtained and compared the correlation obtained from each psychological dimension with respect to academic performance based on the Pearson scale.

KEYWORDS: Academic Performance, Data Mining, Random Forest, Decision Tree, Automation, Real-time Prediction.

Introducción

Se pueden buscar y encontrar diferentes definiciones sobre el rendimiento académico, la más global es la que lo establece como una variable esencial que le permite al Sistema Educativo evolucionar, pero la definición que más se adapta a esta investigación es la que establece Requena [1] afirmando al rendimiento académico como una consecuencia de la relación entre el desempeño y la capacidad del estudiante. Aunque a simple vista esta relación parezca algo fácil de comprender, hay una serie de factores que influyen de manera directa, Oliver [2] manifiesta que a veces solo se piensa que el alumno presenta un bajo desempeño académico debido a su mala organización con respecto a sus tareas o a sus malos hábitos de estudio, pero no se evalúa más allá, no se toman en cuenta otros factores que también son influyentes, como el caso de los factores psicológicos que cumplen un rol vital para la determinación del rendimiento académico.

En una evaluación realizada a los países que pertenecen a la Organización para la Cooperación y el Desarrollo Económico (OCDE), se obtuvo que un 28% de los estudiantes no llegan al nivel de conocimientos básicos establecido en al menos tres de las asignaturas evaluadas por PISA, además un 12% de estudiantes presentan un rendimiento académico bajo en todos los cursos. Se identificó una tendencia en los estudiantes que se encuentran dentro de este porcentaje, la cual demuestra que su desempeño académico se ve afectado por la falta de motivación, perseverancia y confianza [3]. La influencia negativa de estos factores se da de manera global, afirmando la importancia que se le debe dar al seguimiento y la estabilidad del cociente emocional en todos los estudiantes. También cabe recalcar que algunos países de Latinoamérica presentaron un incremento en sus porcentajes de bajo desempeño académico, según las interpretaciones realizadas por BBC MUNDO sobre el informe de la OECD los países fueron: Argentina (56%), Colombia (60%) y Perú (68%) [4]. En México [5] se pudo realizar un estudio más profundo donde se obtuvo el promedio de la cantidad de horas semanales que un estudiante con buen desempeño les dedica a los deberes, siendo este un aproximado de 6.6 horas; en el otro lado de la situación se encuentran los estudiantes con bajo rendimiento con un promedio alarmante de 3.3 horas semanales dedicadas a los deberes.

La situación en el Perú con respecto al rendimiento académico es alarmante, PISA evaluó en diferentes instituciones a un aproximado de 6,035 estudiantes que redondean los 15 años y los resultados de estas evaluaciones nos colocaron como el país con el peor rendimiento académico de Latinoamérica. Pero Llorente afirma que los análisis de PISA no poseen ningún valor, debido a que no evalúan nada, solo examinan respuestas en base a un modelo reducido de ciertas áreas de tres materias y a la larga no aportan en nada a la educación, el bajo rendimiento académico no se combate con “exámenes”, porque se sabe que el alumno no es una calificación, en la etapa de la adolescencia se presentan notables alteraciones físicas y psicológicas que traen consigo factores que afectan el desempeño de un estudiante, esto se debe contrarrestar mediante estudios que contribuyan a identificar las razones que originan dicho rendimiento académico o identificar perfiles que presentan los estudiantes con bajo rendimiento, esa identificación permitirá tomar acciones de gran importancia para disminuir el porcentaje de fracasos en las evaluaciones que se realicen [6].

A nivel local la problemática sigue siendo evidente, ya que según evaluaciones psicológicas aplicadas en alumnos con bajo rendimiento académico de 1º año de secundaria en una institución educativa de Lambayeque (véase Anexo N.º 4) se pudo observar la inestabilidad emocional que estos presentan, además, las evaluaciones dieron una posible causa con respecto al problema de bajo rendimiento académico que se presenta en la institución. Se dice “posible

causa” porque, aunque se evalúen factores como: autoestima y motivación, estos aún siguen siendo insuficientes para determinar el causante del bajo desempeño de los estudiantes.

Asimismo, el intento de la institución por tratar de mejorar la situación fue casi nula, debido a que buscaban masificar una posible solución que a la larga terminaba siendo ineficaz, porque no a todos los estudiantes que obtienen un bajo rendimiento se les apoya de la misma forma. Estas supuestas “soluciones” en vez de generar una mejora, solo desencadenaban más problemáticas siendo una de ellas la incapacidad del área de psicología para abastecerse en la evaluación de todos los estudiantes y generar información clave que apoye la toma de decisiones, por si fuera poco, a todo esto, se le sumaba la ausencia del uso de tecnología en los procesos de la institución lo que no les permitía contar con información digitalizada para que pueda ser visualizada y utilizada de manera más eficaz en busca de soluciones.

Tomando el bajo desempeño académico influenciado por factores psicológicos como el problema general y la falta de información para manejarlo, se estableció como objetivo general: la implementación de un sistema de predicción de rendimiento académico en base a diversos factores de influencia que permita integrar no unos pocos factores sino diversos factores psicológicos junto al factor académico, para generar información clave que apoye la toma de decisiones en la mejora del desempeño académico de los estudiantes. Esta integración se lleva a cabo con evaluaciones del cociente emocional mediante el Test BarOn (Adolescentes), el formato de evaluación del test es a través de 30 preguntas que son respondidas mediante la escala de Likert, las dimensiones que evalúa son: intrapersonal, interpersonal, adaptabilidad, manejo de estrés y estado anímico, los resultados que se obtienen son porcentajes de cada dimensión, los cuales son procesados y relacionados con el promedio ponderado de los estudiantes para encontrar una tendencia mediante un algoritmo minería de datos y así posteriormente obtener pronósticos de rendimiento, por último el modelo será automatizado para que pueda generar predicciones en tiempo real mediante el ingreso de datos. De igual manera, se establecieron objetivos específicos como:

Determinar el algoritmo que permita obtener la medida más óptima de asertividad en la predicción del rendimiento académico basado en los diversos factores de influencia analizados. Determinar la importancia de cada variable (factores de influencia) dentro del modelo. Evaluar comparativamente la correlación que tienen las dimensiones psicológicas con el rendimiento académico del estudiante.

Los beneficios que obtiene la institución al utilizar el sistema forman parte del plan de mejora establecido por la OECD [5] los cuales son: identificar estudiantes que tengan tendencia a presentar un bajo rendimiento y ofrecer un apoyo adecuado a la situación que presente el estudiante. La implementación del sistema de predicción de rendimiento académico permitirá cubrir en cierta parte el vacío de información existente con respecto a la integración de diversos factores psicológicos e histórico de notas para el apoyo del desempeño académico.

La presente investigación se justifica científicamente porque se desarrolló un sistema de predicción de rendimiento académico en base a diversos factores influyentes, en este caso psicológicos y de desempeño (notas) para evaluar la proyección del desempeño del estudiante, todo esto bajo un enfoque científico. También contribuye en el área de educación e innovación tecnológica, brindando más conocimientos sobre el uso de algoritmos de aprendizaje automático para demostrar la importancia de la correlación que tiene el cociente emocional con respecto al desempeño académico.

Por otro lado, se justifica económicamente ya que se cuenta con disponibilidad de recursos económicos para desarrollarla, a su vez la institución educativa se puede beneficiar económicamente porque le permite reducir costos de pago a personas que se encargan de realizar reportes de pronósticos en Excel con la data histórica de notas y remuneraciones a

personal de apoyo en el área de psicología para cubrir evaluaciones de todos los estudiantes del grado. Es necesario que la institución educativa destine recursos económicos en capacitaciones de manejo de tecnologías de información para los docentes, las cuales ayudarían a que ellos desarrollen habilidades en el uso de estas.

Socialmente es justificada puesto que, según encuestas realizadas en el Perú, se demuestra que somos una sociedad muy preocupada por el tema de la educación y esta tesis sirve como una evidencia de que no somos indiferentes con respecto a la situación actual de la educación en nuestro país y que se continúa trabajando e investigando para poder mejorar esta situación [7]. También puede demostrar como algunos factores influyen en la variación del desempeño y rendimiento académico de los estudiantes; y a su vez facilita la toma acciones de mejora y prevención no solo en la institución, sino de manera general en el ámbito educativo.

Finalmente, la investigación se justifica tecnológicamente debido a que la institución educativa cuenta con los recursos tecnológicos para el desarrollo y la utilización del sistema. Se hizo uso de algoritmos de minería de datos para analizar datos seleccionados y luego pronosticar el rendimiento académico del estudiante, esto le permite a la institución mejorar la visualización y utilización de la información de los estudiantes con respecto a sus notas y cociente emocional.

Revisión de literatura

Antecedentes

Los modelos de predicción en base a técnicas de minería de datos han sido utilizados en el ámbito educativo para predecir el rendimiento académico en base a notas o algunos factores psicológicos. Lo que se pretende en este proyecto es integrar casi en su totalidad los factores psicológicos con las notas para generar una predicción y posteriormente automatizar el modelo, en base a eso no hay registros de que se haya realizado una investigación, pero sí existen algunas que presentan una cierta similitud:

Internacionales

Imran, Latif y Saqlain [8] intentan evidenciar los problemas que ocasiona una mala calidad en los datos y la importancia del procesamiento de estos. Presentaron un modelo de predicción de desempeño académico del estudiante basado en la técnica de Árboles de decisión y clasificándolo de diferentes maneras. Para ejecutar el modelo plantearon una metodología basada en 7 etapas, esperando en la última etapa un resultado de aprobado o desaprobado. Luego de ejecutar el modelo y aplicar el método de conjunto para procesar los datos de diferentes maneras, concluyen que la aplicación de este método mejoró el rendimiento de la clasificación, obteniendo un 95% de precisión. Se seleccionó esta investigación ya que sirve de orientación para un correcto procesamiento de datos y también evidencia como el uso de un método en la clasificación de los datos mejora el rendimiento de estos.

Mohammed [9] tomó como problemática la imprecisión de los resultados del algoritmo Decision Tree frente casos de EDM, su objetivo fue aplicar una técnica modificada de dicho algoritmo y comparar su mejora con otros algoritmos. No utilizó metodología ya que evaluó algoritmos como tal, hizo la modificación con el fin de comprender que tanto variaban en asertividad los datos de los estudiantes y posteriormente predecir el rendimiento académico

(Pasar o Fallar). Usó la herramienta Accord.Net en su predicción y concluyó que el algoritmo modificado superó el asertividad del algoritmo base, creando una nueva alternativa de uso. Se consideró este proyecto porque se realizó una comparativa de los resultados de los algoritmos que se están tomando en esta investigación, además fue de ayuda para poder analizar algunas limitaciones con respecto a otros algoritmos y también para evaluar el posible uso de la herramienta automatizada Accord.Net en la predicción.

Nacionales

Holgado [10] implementó una solución de minería de datos con la metodología CRISP-DM, para plantear una solución frente a la dificultad de detectar los patrones de bajo rendimiento en la institución educativa. Comparó los algoritmos C5.0 y Cart para encontrar que factores influyen más en el bajo rendimiento académico para la búsqueda se tomó en cuenta las variables de: número de cursos llevados, servicio del comedor y pago de pensiones; al final el autor logró identificar los patrones requeridos y concluyó que el algoritmo C5.0 le dio los mejores resultados con un 77% de efectividad. Esta investigación se tomó en cuenta para analizar la implementación de la metodología CRISP-DM y la manera en que fueron tratados los datos destinados a la minería.

Yamao [11] se encontró frente a la problemática de deserción en los alumnos del primer ciclo de la institución educativa, como solución a este problema, planteo la predicción del rendimiento académico tomando en cuenta indicadores sociales (relación con pares, motivación, género), económicos y académicos. Aplicó la metodología CRISP-DM, logró predecir el rendimiento académico mediante técnicas de minería de datos y concluyó que el mejor resultado fue de C5.0 con un 82.87% de asertividad. La razón por la que se consideró este antecedente fue porque trabajó con factores sociales en la predicción, sirviendo de guía para el procesamiento de mis datos, también utilizó grafica de puntos para representación de análisis.

Locales

Luego de haber realizado una búsqueda exhaustiva en diversos repositorios de investigaciones y bases de datos de tesis, no se encontraron antecedentes a nivel local con el requerimiento de antigüedad establecido (mayor a 2017) para esta investigación.

Bases teóricas científicas

Rendimiento Académico

Hay muchas teorías para definir el rendimiento académico ya que es estudiado desde diferentes enfoques y eso ciertas veces suele confundir al momento que uno se quiere referir a aptitud escolar o desempeño académico. Pero el planteamiento de esta investigación gira alrededor de la definición establecida por Requena [1] afirma que el rendimiento académico es resultado del desempeño y capacidad del estudiante, explicándolo mejor, es el resultado del desempeño que se ha visto reflejado desde el inicio hasta el final de un proceso de aprendizaje, este que también se complementa con otros factores como la capacidad intelectual, las notas obtenidas, etc.

Factores influyentes en el rendimiento académico de un adolescente

La adolescencia trae consigo una transición enorme en la estabilidad emocional de un estudiante, esta que en algunos casos trae consecuencias negativas a nivel académico. La nota es uno de los factores que generalmente es considerado en todo momento para tener una idea del rendimiento académico, pero esto es totalmente insuficiente, ahora es indispensable considerar otro tipo de factores como el desempeño académico que viene siendo influenciado por el cociente emocional. Esto hace que el rendimiento académico se vuelva producto del junte de aptitudes, capacidades, personalidad, estado anímico del adolescente.

Según un artículo redactado por Perú21 [12] no todo afecta negativamente, también se han evaluado factores que se pueden reforzar a esa edad para apoyar en la mejora del desempeño y por ende obtener un buen rendimiento académico: Desarrollo físico, Desarrollo emocional, Medio social, Relación familiar y Relación con docente.

Todos los adolescentes pasan por diversos cambios, ya sean emocionales, hormonales, etc., pero lo que se tiene que hacer es reforzar su cociente emocional para que este sepa dominar factores que influyen no solo en su rendimiento sino también en su vida diaria.

Cociente Emocional

Bar-On [13] lo define como la inteligencia para manejar un grupo de habilidades y conocimientos en el aspecto emocional que influyen en la capacidad de una persona para afrontar situaciones del entorno. Según el modelo de Bar-On, se pudo segmentar el cociente emocional en 5 capacidades: intrapersonal, interpersonal, manejo de estrés, estado ánimo y adaptabilidad. Dentro del ámbito escolar la inteligencia emocional es la que permite al alumno mantener una estabilidad en su desempeño académico, la diferencia es muy clara con respecto a la capacidad intelectual. Según investigaciones de Llibre [14] el IQ de un estudiante solo permite predecir un 20% de su éxito, mientras que el cociente emocional un 80%, estudios indican que el promedio de ponderación del cociente emocional como factor de influencia está muy por encima que el de cualquier otro.

Cociente emocional en adolescentes

Cuando llega la etapa de la adolescencia esta trae consigo un aumento en la intensidad de las emociones en los adolescentes, se encuentran dentro de un sube y baja emocional, se genera una mezcla de diferentes tipos de sensaciones como: angustia, tristeza, soledad, felicidad, madurez, etc.; si estas sensaciones no son manejadas correctamente llegan a repercutir de manera permanente en su vida [15]. A base de eso nace la necesidad de que las instituciones educativas planteen estrategias que faciliten la permanente evaluación de la inteligencia emocional de los estudiantes.

Goleman [16] afirma que el cociente emocional le permite al adolescente desenvolverse de manera correcta en su entorno, usar emociones para alcanzar objetivos y ser consciente de sus propias decisiones, bastan solo estos 3 puntos para suponer a qué nivel impactaría en su desempeño académico si es que no se evalúa el cociente emocional.

Evaluación del cociente emocional mediante Test BarOn

El test de BarOn tiene como fin evaluar únicamente el cociente emocional, esta evaluación consiste en responder aproximadamente 120 preguntas mediante la escala de Likert [17]. Su resultado es un porcentaje dentro del rango de 1-100 con respecto a cada una de las 5 capacidades que abarca el cociente emocional:

Intrapersonal: asertividad, autoconcepto, autorrealización, independencia.

Interpersonal: empatía, relaciones, responsabilidad social.

Adaptabilidad: solución de problemas, prueba de realidad, flexibilidad.

Manejo de estrés: tolerancia, control de impulsos.

Estado animo: felicidad, optimismo.

Para esta investigación se seleccionará el test de acuerdo a la edad (10-13años), será el test BarOn de 30 preguntas para adolescentes.

Sistemas Web

Se le puede llamar sistema web a las aplicaciones que permiten su utilización mediante la conexión con un servidor en la nube, son muy utilizados debido a la cantidad de beneficios que proporcionan, algunos de estos son: la autonomía de su sistema operativo, ahorro de costes en hardware, apoyo al trabajo colaborativo y seguridad de la información. Hoy en día la opción de implementar una solución a través de un sistema web es prácticamente segura, estos no limitan a los usuarios que solo cuentan con conocimientos básicos en tecnología, ya que son adaptables a ellos y además sus porcentajes de error son bajos, debido a que generalmente son corregidos de manera inmediata [18].

Lenguaje PHP

Sintes [19] define PHP como un lenguaje de programación abierto y apropiado para el desarrollo web, también se considera como un lenguaje de programación interpretado y no tipificado, teniendo una sintaxis parecida a la del lenguaje de programación C. Tiene la ventaja de que se puede insertar en HTML.

Metodología RUP

Metodoss [20] analizó la metodología RUP y afirma que esta metodología utiliza la orientación a objetos en su diseño, posee técnicas y prácticas que están abaladas comercialmente. Todas las fases de esta metodología generan artefactos que van siendo mejorados constantemente al pasar por cada una de ellas, las fases son:

Fase de diseño: contiene pautas de los objetivos, arquitectura y planificación para el acuerdo entre ambas partes.

Fase de elaboración: diseñar el sistema siendo complemento de la documentación de los casos de uso.

Fase de construcción: inicialización de desarrollo físico del Software códigos y pruebas.

Fase de transición: plan de despliegue y entrega, evaluación de calidad del Software para satisfacción del usuario.

Con respecto a la investigación se ha seleccionado la fase de diseño de esta metodología, debido a que utiliza las mejores prácticas para desarrollar software moderno y manejar de forma efectiva las reglas del negocio.

Minería de datos

Gonzales define la minería de datos como el proceso de extracción de información entendible y útil a base de medianas o grandes cantidades de datos, esta definición tiende a ser la más genérica y global. Pero Piatetsky-Shapiro le da un enfoque un poco más orientado, afirmando que la minería de datos es el conjunto de técnicas y herramientas que se aplican a la elaboración de conocimiento en base a datos recogidos o almacenados, todo esto con el fin de realizar predicciones de tendencias o conductas [21].

Jiménez y Álvarez [22] expresan que hoy por hoy el interés de utilizar la minería de datos en el ámbito educacional aumentó, enfocándose en predicciones y agrupamientos en base a los

datos de plataformas educativas, esto con el fin de conocer más al estudiante, desarrollar más el entorno en el que trabaja y sobre todo apoyar su desempeño académico.

Precisamente este caso en el que se pretende apoyar el desempeño de estudiantes es donde se adapta perfectamente uno de los usos de la minería de datos, tomando un enfoque educativo EDM (Educativo Data Mining).

Técnica de minería de datos

Las técnicas de la minería de datos tienen parte de su origen en la rama de la estadística y de la inteligencia artificial. Estas a su vez se pueden definir como la aplicación de algoritmos sobre un conjunto de datos para obtener el tipo de resultado esperado [23].

Analizando los fines de la investigación y leyendo sobre las diferentes técnicas de minería de datos, se pudo seleccionar el enfoque de Modelos de predicción + Técnicas de DM que permite entrenar y probar un modelo a base del algoritmo seleccionado [24].

Entre algunos algoritmos tenemos:

Random Forest: trabaja con la media de varios árboles de decisión que están elaborados con muestras de los datos. Utiliza la lógica de cada árbol separado es impreciso, pero la suma de estos alcanza mejores resultados.

Árboles de decisión: pretenden organizar datos mediante elecciones en ramas de influencia, utilizan condicionales de sí/no.

Metodología de desarrollo de minería de datos

En la guía modeladora de IBM [25] se define a CRISP-DM como una metodología utilizada para proyectos de análisis de datos, su proceso se contempla como un proyecto profesional estableciendo un contexto mucho más influyente en la elaboración de modelos. Esta abarca las fases de un proyecto, las tareas y relaciones entre estas; su ciclo de vida está dividido en 6 fases: Entendimiento del negocio: se definen las necesidades del cliente y objetivos del proyecto para solucionar problemática.

Entendimiento de datos: estudio y comprensión de los datos, se coleccionan los datos para descubrir algo sobre ellos.

Preparación de data: análisis completo de los datos y selección de sus características, incluye proceso de transformación y limpieza de datos.

Modelado: se eligen técnicas que se van a utilizar para el modelado.

Evaluación: se revisa a fondo lo realizado anteriormente y se evalúan aspectos que no hayan sido considerados.

Despliegue: se pone en producción el modelo, pero no se da por concluido hasta terminar prueba del cliente.

Cabe resaltar que esta metodología no presenta una secuencia rígida en sus fases, esto quiere decir que puede haber un desarrollo hacia atrás y hacia adelante en las fases.

Lenguaje R orientado a minería de datos

R es un lenguaje y a su vez un entorno para la programación que está orientada al análisis de datos, entre sus principales características están las diversas herramientas de estadística y la capacidad para generar gráficos. Pero no se le orienta a la minería de datos solo por el simple hecho de que maneja estadísticas, sino que hay muchos otros aportes que este brinda si es que le damos un enfoque de minería de datos. Aportes como el fácil manejo de matrices y vectores, paquetes diseñados especialmente para selección, recodificación y recuperación de datos mediante lenguaje R, lo que lo identifica con esta investigación es que se creó en el ámbito académico [26].

Algoritmo para predicción automática

Se le proporciona un conjunto de datos estudiados que contenga las entradas y salidas esperadas, el algoritmo debe hallar un método para poder llegar a ellas. Luego ya debe ser capaz de recibir y analizar datos de entrada para poder predecir los valores que se muestran en la salida, todo dentro de un rango de asertividad aceptable [27].

Python

Python es un lenguaje de programación orientado a objetos que está preparado para realizar programas, aplicaciones, servidores de red o hasta páginas web. Presenta una característica como lenguaje interpretado ya que tiene como finalidad el desarrollo de una sintaxis que priorice la legibilidad del código, es considerado también multiparadigma ya que soporta varias orientaciones, ofrece una ventaja en la velocidad de desarrollo [28]. Este lenguaje también facilita el desarrollo de algoritmos de rápido aprendizaje como los que son necesarios en proyectos de aprendizaje automático.

Flask

Es un Microframework escrito en Python que facilita el desarrollo de Aplicaciones Web. Su característica “micro” no significa que sea pequeño, sino que al instalarlo vamos a tener las herramientas necesarias para crear una aplicación web funcional, pero en algunos casos se necesitan nuevas funcionalidades y para eso existe un conjunto muy grande de extensiones que se pueden instalar con Flask, las cuales le van agregando dichas funcionalidades [29].

Materiales y métodos

Tipo de investigación

Se desarrolló una investigación de tipo tecnológica o también denominada investigación de desarrollo, puesto que Cegarra [30] establece como finalidad de esta investigación, la elaboración de un artefacto innovador que traiga consigo beneficios en la empresa o mercado.

Métodos de investigación

Los métodos de investigación empleados son los siguientes:

Tabla 1: Métodos de investigación

Método	Descripción
Análítico	Estudio y análisis de la problemática con respecto al desempeño académico de los estudiantes que presenta la institución educativa.
Deductivo	Obtención de una conclusión en base a premisas identificadas para el correcto planteamiento de la propuesta de solución al problema.
Implementación	Se pondrá en funcionamiento el sistema de predicción de rendimiento establecido como propuesta de solución, a través de metodologías seleccionadas.

Técnicas e instrumentos de recolección de datos

A continuación, en la siguiente tabla se muestra las técnicas e instrumentos que fueron útiles para la recolección de datos.

Tabla 2: Técnicas e instrumentos de recolección de datos

Técnicas	Instrumentos	Elementos de la población	Propósito
Entrevista	Guía de entrevista (ver anexo N.º 2)	Subdirector	Obtener información más detallada de la situación y conocer disponibilidad para realizar cambios en sus procesos.
Entrevista	Guía de entrevista (ver anexo N.º 3)	Área de Psicología	Conocer las limitaciones que presentan para ayudar a los estudiantes y analizar propuestas que refuercen el modelo.
Revisión documental	Recurso electrónico (ver anexo N.º 4)	Área de psicología	Obtener evidencias que demuestren la inestabilidad en la salud mental de los alumnos.

Procedimientos

Metodologías de desarrollo

A continuación, se mencionan las fases de las metodologías a seguir, en este caso se emplearon 2 metodologías: La metodología CRISP-DM, que se utilizará para la elaboración del modelo de minería de datos, en la última fase el modelo será automatizado para obtener una predicción en tiempo real. La metodología RUP, para elaborar el sistema en el cual será integrado el modelo de predicción automatizado, este sistema se fue elaborado para cubrir las formalidades necesarias con respecto a la evaluación de un estudiante (Autorización).

Metodología CRISP – DM

Fase #1: Comprensión del negocio

Fase #2: Comprensión de los datos

Fase #3: Preparación de los datos

Fase #4: Modelado

Fase #5: Evaluación

Fase #6: Despliegue

Metodología RUP

Iteración #1: Plan de sistemas

Iteración #2: Modelado de negocio

Iteración #3: Modelado casos de uso

Iteración #4: Análisis **Iteración #5:** Diseño

Iteración #6: Implementación y pruebas

Producto acreditable

Interfaces

Se diseñaron las interfaces del sistema web empleando HTML5, PHP7 y CSS3, también se utilizó JavaScript para algunas funcionalidades. Las interfaces más importantes se observan en el *ítem Iteración #5 en el apartado de Resultados – Metodología RUP*.

Arquitectura

Se diseñó la arquitectura del sistema definiendo la estructura y relación de sus componentes para garantizar una correcta operación, tal como se presenta en el *ítem Iteración #5 en el apartado de Resultados – Metodología RUP*.

Infraestructura tecnológica

Se diseñó la arquitectura del sistema definiendo la estructura y relación de sus componentes para garantizar una correcta operación. Las terminales de los usuarios (PC`s) se conectan a internet mediante un router, una vez conectados pueden acceder al sistema de predicción mediante un dominio. Al acceder al sistema establecerán una conexión con el servidor frontend (interfaces, diseño, etc), este se relaciona con el servidor back-end (contiene la lógica de funcionamiento del sistema) el cual también se encuentra relacionado con otros servidores: S. Flask (predicción) y S. Base de Datos (almacenamiento).

Requerimientos de componentes que forman parte de la arquitectura: Pc escritorio (Para los usuarios es necesaria la utilización de una computadora de escritorio con las características mínimas de: procesador Intel Core i3, sistema operativo Windows, memoria de 4GB, disco duro 1TB), Router (Dispositivo que permite conexión a internet ya sea mediante cable Ethernet generalmente 4 puertos en la parte trasera o vía Wi-Fi mediante las bandas de 2.4GHz y 5GHz), Internet (Será de carácter obligatorio el uso de internet para la conexión con el sistema, este internet debe poseer una velocidad mínima de: 5mbps Ethernet/Wifi), Servidor web back-end (Es utilizado para la lógica del funcionamiento del sistema, todo lo que no es visible a simple vista por el usuario, se utilizó: lenguaje de programación PHP), Servidor web front-end (Es utilizado para la interacción del usuario con el sistema “experiencia del usuario”, lo que se ve en pantalla y la navegabilidad son en base a la situación, se utilizó: HTML5 – CSS3 – JS), Servidor flask (Contiene la interfaz que permite la interacción con el modelo de predicción automatizado mediante el lenguaje Python versión 3.7) y Servidor de base de datos (Manejado por MySQL versión 8.0, en este servidor se almacena toda la información que se obtenga de las fuentes conectadas).

Tabla 3: Matriz de consistencia

<u>FORMULACIÓN DEL PROBLEMA</u>	<u>MÉTODO DE INVESTIGACIÓN</u>		
¿De qué manera un sistema de predicción de rendimiento académico en base a diversos factores de influencia brindará información para apoyar la toma de decisiones en la mejora del desempeño académico de los estudiantes del 1° año de secundaria en un colegio de Lambayeque?	<u>TIPO DE INVESTIGACIÓN</u>		
	Tecnológica		
<u>OBJETIVO GENERAL</u>	<u>MÉTODO</u>	<u>DESCRIPCIÓN</u>	
Implementar un sistema de predicción de rendimiento académico en base a los diversos factores de influencia en los estudiantes del 1° año de secundaria en un colegio de Lambayeque.		Estudio y análisis de la problemática con respecto al desempeño académico de los estudiantes que presenta la institución educativa.	
		Obtención de una conclusión en base a premisas identificadas para el correcto planteamiento de la propuesta de solución al problema.	
	Análítico Deductivo	Se pondrá en funcionamiento el sistema de predicción de rendimiento establecido como propuesta de solución, a través de metodologías seleccionadas.	
	Implementación		
<u>TÉCNICAS</u>	<u>INSTRUMENTOS</u>	<u>ELEMENTOS DE LA POBLACIÓN</u>	<u>PROPÓSITO</u>
Entrevista	Guía de entrevista (ver anexo N.º 2)	Subdirector	Obtener información más detallada de la situación y conocer disponibilidad para realizar cambios en sus procesos.
Entrevista	Guía de entrevista (ver anexo N.º 3)	Área de psicología	Conocer las limitaciones que presentan para ayudar a los estudiantes y analizar propuestas que refuercen el modelo.
Revisión Documental	Recurso Electrónico (ver anexo N.º 4)	Área de psicología	Obtener evidencias que demuestren la inestabilidad en la salud mental de los alumnos.

<u>OBJETIVOS ESPECÍFICOS</u>	<u>DESCRIPCIÓN DEL LOGRO DE LOS OBJETVOS ESPECÍFICOS</u>	<u>INDICADORES</u>
Determinar el algoritmo que permita obtener la medida más optima de asertividad en la predicción del rendimiento académico basado en los diversos factores de influencia analizados.	Comparación de la medida de asertividad en la predicción de cada algoritmo, obtenida en base a matrix confusion y uso de comandos.	- Número de predicciones verdaderas para rendimiento 1 - Número de predicciones verdaderas para rendimiento 2 - Numero de predicciones verdaderas para rendimiento 3
Determinar la importancia de cada variable (factores de influencia) dentro del modelo.	Obtención de la medida de importancia de cada variable que interviene en el modelo mediante uso de comandos.	Medida de importancia (cifra numérica) de cada variable que interviene en el modelo.
Evaluar comparativamente la correlación que tienen las dimensiones psicológicas con el rendimiento académico del estudiante.	Comparación de dimensiones psicológicas en base a su coeficiente de correlación con respecto al rendimiento académico, clasificándolas mediante escala de Pearson y herramienta de semaforización.	- Coeficiente de correlación de dimensión Intrapersonal - Coeficiente de correlación de dimensión Interpersonal - Coeficiente de correlación de dimensión Imp. Positiva - Coeficiente de correlación de dimensión Manejo Estrés - Coeficiente de correlación de dimensión Adaptabilidad

Consideraciones éticas

A continuación, se listan los aspectos que se han considerado para la protección y bienestar de los participantes de esta investigación, también con respecto a la seguridad de la información:

La recolección de datos mediante las diferentes técnicas seleccionadas, se realizó luego de la autorización brindada por el subdirector de la institución y bajo el consentimiento de las personas involucradas en cada técnica.

Para la elaboración del modelo de predicción no se tomaron en cuenta los datos con respecto a la identidad de los estudiantes (nombres) con la finalidad de no afectar de ninguna manera su integridad.

Las pruebas que se efectúan a cada estudiante son con previo consentimiento del apoderado mediante el módulo de autorización implementado.

Los resultados que genera el sistema implementado con respecto a cada estudiante, son de libre acceso a los apoderados correspondientes con la finalidad de evidenciar el trabajo que se está realizando.

El acceso a la información generada por el sistema a nivel de reportes es de carácter restringido, este acceso es según los privilegios de los 3 perfiles creados. Existe una validación de contraseña para el ingreso de cada uno al sistema

Resultados y discusión

Resultados de metodologías

Metodología CRISP – DM

Fase #1: Comprensión del negocio

En la primera fase se realizó el reconocimiento del negocio. Primero se recopiló información general de la institución (organigrama, identificación de personal clave, etc.), se entendieron objetivos de la institución y se plantearon criterios de éxito para el proyecto. Posteriormente elaboraron objetivos de minería y se evaluó la situación con respecto a recursos, costes/beneficios. Por último, se diseñó un plan de proyecto en base a fases/riesgos, un cronograma de actividades y una valoración de herramientas/técnicas. Como resultado de esta fase se concluyó que la situación de la institución sí se encuentra dentro de un contexto apto para realizar una solución de minería de datos.

Fase #2: Comprensión de los datos

En la fase de comprensión de datos se pudo estudiar a fondo los datos que se tienen disponibles para realizar la minería, esta fase fue fundamental para evitar problemas imprevistos en la siguiente fase de preparación de datos. Primero se compilaron los datos en formato xlsx, luego se convirtieron a formato csv para poder importarlos al IDE RStudio y de esa manera poder analizarlos en base a su estructura, características de variables y correlación entre ellas. Los resultados que se obtuvieron en esta fase fueron: conocimiento de estructura de data frame

(360 registros con 49 columnas cada uno), características de variables (mínimo, máximo, media, mediana, primer y tercer cuartil), hallazgo variables correlacionadas que pueden perjudicar al modelo en el futuro (Intra-Final x Intra-Cuali / Inter-Final x Inter-Cuali / Adap-Final x Adap-Cuali / Impre-Final x Impre-Cuali) y hallazgo de variables irrelevantes (id_alumno, sexo).

Fase #3: Preparación de los datos

Para esta fase se construyó el data frame final en base a los datos comprendidos anteriormente. Primero fue necesario explorar todos los datos para poder realizar la limpieza de los mismos, en la limpieza se excluyeron las variables con menor relevancia para el modelo: sexo, id_alumno, nombre; se eliminaron las columnas correlacionadas para evitar ruido en el modelo (Intra_Final / Inter_Final / Adap_Final / Impre_Final). Luego se construyeron los nuevos datos, convirtiéndolos a la tipología necesaria: grado = tipo numérico, rendi_academico = tipo numérico (1-2-3). El resultado de esta fase fue la obtención del data frame final con 360 registros y 42 columnas.

Fase #4: Modelado

En esta fase se seleccionaron 2 técnicas de modelado (algoritmos) en función de los objetivos del proyecto, además se requería que las técnicas permitan validar los resultados obtenidos, en base a esto se decidió por los algoritmos: Random Forest y Decision Tree (ambos generan particiones de datos para validación). Primero se generaron 2 particiones del data frame, una para entrenamiento (Train : 288 x 42) y otra para testeo (Test: 72 x 42), la validación se realiza con la columna rendi_academico. Luego se construyó el modelo de Random forest, ejecutando el procedimiento adecuado, los comandos correspondientes para la parametrización y posteriormente la visualización del análisis de los resultados (asertividad, importancia de variables y grado de correlación de dimensiones psicológicas con respecto a rendimiento académico). Por último, se construyó el modelo de Decision Tree ejecutando de la misma manera el procedimiento adecuado, pero con algunas diferencias en la parametrización, la medida de asertividad en su predicción fue inferior a la de Random Forest. El resultado de esta fase fue la determinación del algoritmo más óptimo para la predicción, el cuál es Random forest con 0.95.

```

131
132 Trainrendi_academico<-as.factor(Trainrendi_academico)
133 Testrendi_academico<-as.factor(Testrendi_academico)
134
135 library(randomForest)
136
137 set.seed(12345)
138 predTrain <- predict(RF, Train, type = "class")
139
140
141 mean(predTrain == Trainrendi_academico)
142
143
144 Train_Table<-table(predTrain,Trainrendi_academico)
145 Train_Table
146
147
148 predTest <- predict(RF, Test, type = "class")
149
150
151 mean(predTest == Testrendi_academico)
152
153
154 Test_Table<-table(predTest,Testrendi_academico)
155 Test_Table
156
157

```

```

[1] 1
> Train_Table<-table(predTrain,Trainrendi_academico)
> Train_Table
predTrain  1  2  3
           1 114  0  0
           2  0 106  0
           3  0  0  68
[1] 0.958333
> Test_Table<-table(predTest,Testrendi_academico)
> Test_Table
predTest  1  2  3
          1 29  0  0
          2 2 24  1
          3  0  0 16

```

Fig. 1. Modelo y resultados Random Forest

Fuente: Elaboración propia

diagrama de objetos para cada requerimiento) y Diagrama de objetos general (Relación de entidades/objetos participantes).

Iteración #4: Análisis

Para esta iteración se definieron los requisitos del sistema de manera preliminar, los cuales fueron convertidos en términos de sistema. Primero se elaboró el paquete de análisis en base al sistema establecido y se establecieron funcionalidades principales de acuerdo al paquete, luego se analizaron funcionalidades y relaciones de entidades, por último, se realizó un diagrama de clases general del sistema. Los resultados obtenidos para esta iteración fueron: Diagrama de contexto de análisis (PA PREDICCIÓN), Diagrama de realizaciones de casos de uso de análisis (RCUA: Gestionar Autorización, Evaluar estudiante y Generar reportes) y Diagrama de clases general (Relación de clases participantes).

Iteración #5: Diseño

Lo principal en esta iteración fue el establecimiento de la arquitectura del sistema y el diseño de las funcionalidades, todo lo que se elaboró es consecuencia del análisis ya realizado en la iteración anterior. Los resultados obtenidos fueron: Diagrama de contexto de diseño (SUBSISTEMA DE PREDICCIÓN), Diagrama de realizaciones de casos de uso de diseño del cual se desglosan diagrama de clases y diagrama de secuencia (RCUD: Gestionar autorización, Evaluar estudiante y Generar reporte), Diagrama de clases general (Relación de clases participantes), Diseño de base de datos del sistema, Diagrama de estados (Estudiante Autorizado/NoAutorizado – Activo/NoActivo), Diseño de interfaces del sistema y Diagrama de despliegue (Relacionado a la arquitectura del sistema).

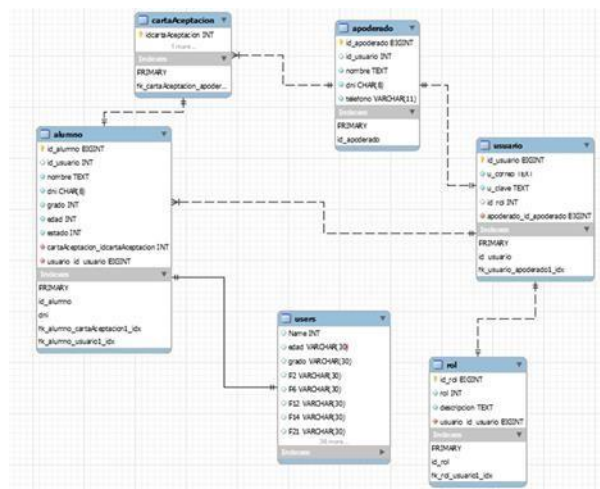


Fig. 4. Diseño de base de datos

Fuente: Elaboración propia

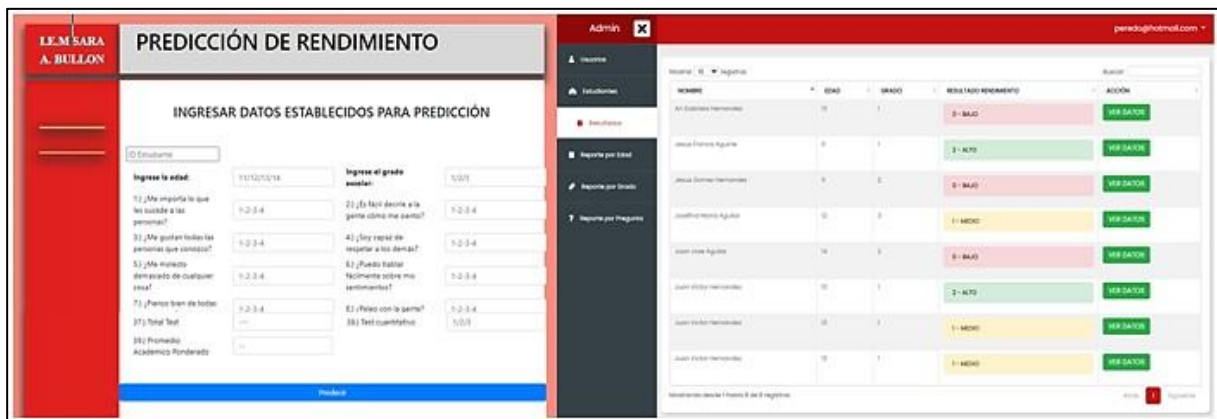


Fig. 5. Interfaces: Predicción y Resultados
Fuente: Elaboración propia

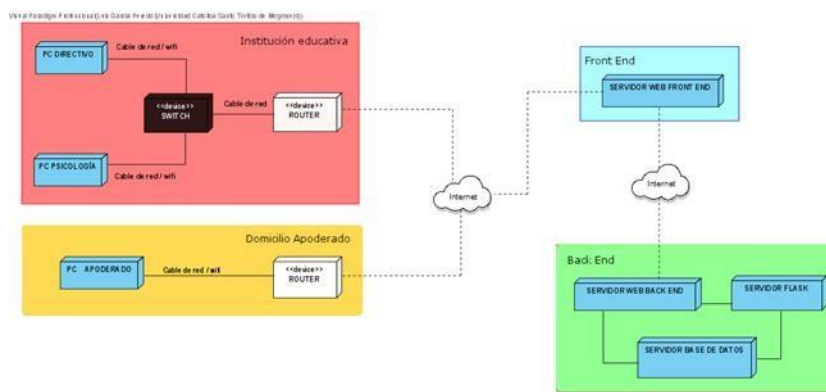


Fig. 6. Diagrama de despliegue
Fuente: Elaboración propia

Iteración #6: Implementación y pruebas

En esta iteración se elaboró el diagrama de componentes del sistema, se registraron capturas de los bloques de código más importantes y se elaboraron pruebas de caja negra/blanca para evaluación del sistema.

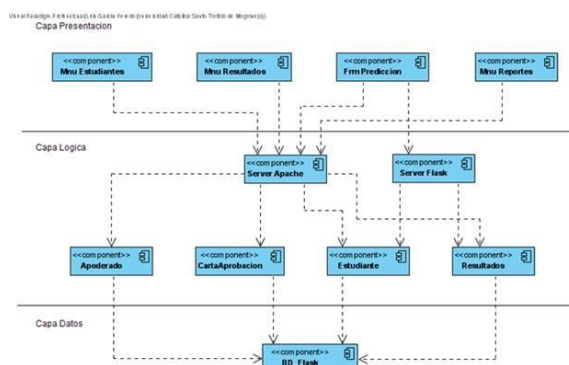


Fig. 7. Diagrama de componentes
Fuente: Elaboración propia

DESCRIPCIÓN DE PRUEBAS DE CAJA NEGRA			DESCRIPCIÓN DE PRUEBAS DE CAJA BLANCA		
Requisito			Requisito		
Sub proceso	Tipo de requisito	Código del requisito	Sub proceso	Tipo de requisito	Código del requisito
Ingreso de datos requeridos	Funcional	RF02	Ingreso	No funcional	RNF01
Descripción del requisito			Descripción del requisito		
Se deben llenar los 39 campos de manera correcta.			El sistema contará con una validación de acceso. El ingreso debe darse mediante un nombre de usuario y contraseña establecida.		
Caso de prueba			Caso de prueba		
Código de prueba	Caso de prueba	Fecha de prueba	Código de prueba	Caso de prueba	Fecha de prueba
PC02	Muestra de resultados al ingresarse los datos correctos	22/07/2020	PCB01	Documentación del código fuente	07/10/2020
Funcionalidad / Característica a evaluar	Acciones de entrada	Resultado esperado	Funcionalidad / Característica a evaluar	Datos de entrada / Acciones de entrada	Resultado esperado
Muestra de resultados	Ingreso correcto de datos en las 39 cajas de texto	Predicción del rendimiento académico según datos ingre	Validación de usuario y contraseña.	Código fuente	Código fuente documentado
Requerimientos de ambiente de pruebas	Equipos: Computadora portátil o de escritorio	Condiciones / Restricciones	Requerimientos de ambiente de pruebas	Equipos: Computadora portátil o de escritorio	Condiciones / Restricciones
		Ninguna	Accesibilidad: Conexión a internet	Código fuente	Ninguno
Seguimiento			Seguimiento		
Resultado obtenido	Estado actual	Observaciones	Resultado obtenido	Estado actual	Observaciones
Predicción de rendimiento académico.	Conforme	Ninguna	Código fuente documentado	Conforme	Ninguna
Correcciones			Correcciones		
Fecha de cambio de estado	Observaciones		Fecha de cambio de estado	Observaciones	
-	-		-	-	

Fig. 8. Pruebas: Caja negra / Caja blanca
Fuente: Elaboración propia

Resultados de objetivos

Objetivo 1: Determinar el algoritmo que permita obtener la medida más óptima de asertividad en la predicción del rendimiento académico basado en los diversos factores de influencia analizados.

Para poder obtener la medida de asertividad de los algoritmos Random Forest y Decision Tree con respecto a la predicción, se utilizó con cada uno la función “mean()” en el IDE (RStudio). Esta función calcula la medida en base a la siguiente fórmula: $A = (V1+V2+V3) / Total$.

Los valores que toma el numerador de la fórmula se obtienen de la “matrix confusion”, un método que permite evaluar los resultados obtenidos de la predicción dentro de una matriz. La cual se genera aplicando el comando “table()” y considerando como parámetros los data sets particionados en entrenamiento y testeo.

```
> mean(predTest == Test$rendi_academico)
[1] 0.9444444
> Test_Table<-table(predTest,Test$rendi_academico)
> Test_Table

predTest  1  2  3
      1 29  1  0
      2  2 23  1
      3  0  0 16
```

Fig. 9. Medida de asertividad Decision Tree
Fuente: Elaboración propia

```
> mean(predTest == Test$rendi_academico)
[1] 0.9583333
> Test_Table<-table(predTest,Test$rendi_academico)
> Test_Table

predTest  1  2  3
      1 29  0  0
      2  2 24  1
      3  0  0 16
```

Fig. 10. Medida de asertividad Random Forest
Fuente: Elaboración propia

Así que como se puede observar en la Fig. 10, la medida de asertividad que se obtiene con el algoritmo Random Forest es la más óptima.

Objetivo 2: Determinar la importancia de cada variable (factores de influencia) dentro del modelo.

El cumplimiento de este objetivo se da en base a la ejecución de las funciones “importance()” y “varImpPlot()” en el IDE (RStudio). Luego de ejecutar ambos comandos se obtuvo la medida de importancia de cada variable (factores de influencia) dentro del modelo mediante una cifra decimal y también una gráfica de puntos que permite visualizar estas cifras de manera más organizada como se muestra en la Fig. 11 y Fig. 12.

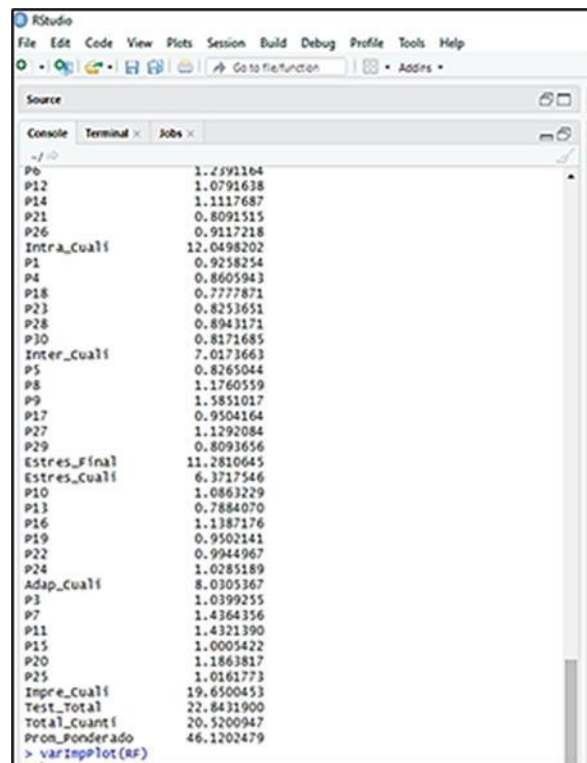


Fig. 11. Medida de importancia de cada variable en el modelo

Fuente: Elaboración propia

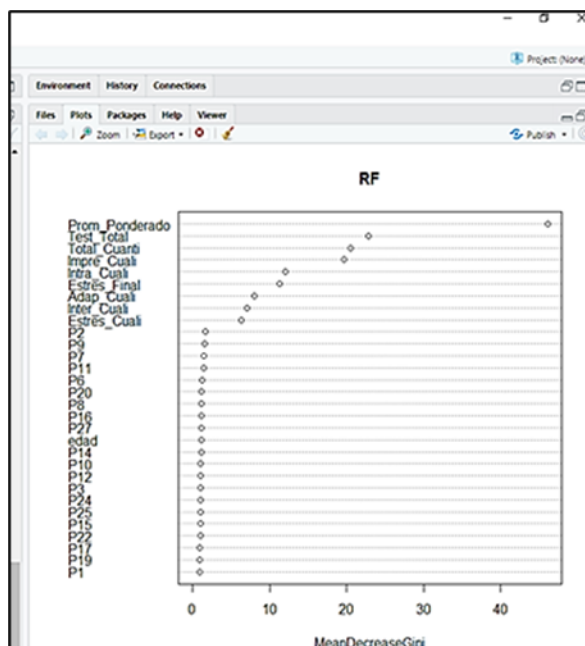


Fig. 12. Gráfica de importancia de variables en el modelo
Fuente: Elaboración propia

El modelo calcula el valor de importancia para cada variable en base a la precisión que estas le aporten, siendo 100 el valor máximo de escalabilidad de importancia para las variables de un modelo, se puede observar que hay un grupo de variables cuyo valor de importancia es por debajo de 2 y otro donde su valor sobrepasa el 11.

Objetivo 3: Evaluar comparativamente la correlación que tienen las dimensiones psicológicas con el rendimiento académico del estudiante.

Para este objetivo se hace uso de la función “cor()” en el IDE (RStudio). El comando facilita la obtención del coeficiente de correlación (r) de cada dimensión psicológica con respecto al rendimiento académico como se observa en la Fig. 13.

```

Console Terminal Jobs
> cor(df$rendi_academico,df$intra_Cuali)
[1] 0.698069
> cor(df$rendi_academico,df$inter_Cuali)
[1] -0.3840557
> cor(df$rendi_academico,df$estres_Cuali)
[1] 0.6264058
> cor(df$rendi_academico,df$adap_Cuali)
[1] 0.5398911
> cor(df$rendi_academico,df$impre_Cuali)
[1] 0.840847
>

```

Fig. 13. Coeficiente de correlación de cada dimensión psicológica con respecto a rendimiento académico
Fuente: Elaboración propia

Además, se utilizó la herramienta de semaforización para poder clasificar los coeficientes obtenidos. Se tomó como parámetro de semaforización el coeficiente de correlación obtenido y los rangos de clasificación fueron establecidos de acuerdo a la escala de Pearson como se muestra en la Tabla 4.

Tabla 4: Escala de Pearson sintetizada Fuente:
Elaboración propia

ESCALA PEARSON SINTETISADA		
RANGO DE VALOR		FUERZA
NEGATIVA	POSITIVA	
-0.3 a -0.1	0.1 a 0.3	DEBIL
-0.5 a -0.3	0.3 a 0.5	MODERADA
-1.0 a -0.5	0.5 a 1.0	FUERTE
-1	1	PERFECTA
0		NULA

Por lo tanto, al analizar la comparativa en la Tabla 5 se puede decir que con respecto al rendimiento académico existe un nivel de correlación positiva fuerte en cuatro dimensiones psicológicas y solo una dimensión presenta un nivel de correlación negativa moderado.

Tabla 5: Comparativa de dimensiones psicológicas en base a correlación Fuente: Elaboración Propia

COMPARATIVA DE DIMENSIONES			
DIMENSIÓN	r	FUERZA	DESCRIPCIÓN
Impresión positiva	0.84	Fuerte	Existe una correlación fuerte entre las variables.
Intrapersonal	0.69	Fuerte	Existe una correlación fuerte entre las variables.
Manejo estrés	0.62	Fuerte	Existe una correlación fuerte entre las variables.

Adaptabilidad	0.53	Fuerte	Existe una correlación fuerte entre las variables.
Interpersonal	-0.38	Moderada	Existe una correlación moderada entre las variables.

Discusiones

En la presente investigación al determinar el algoritmo que permite obtener la medida más óptima de asertividad de la predicción del rendimiento académico basado en los diversos factores de influencia analizados, se pudo encontrar que el algoritmo Random Forest alcanzó una medida de asertividad de 0.95 con respecto a su predicción, mientras que el algoritmo Decision Tree alcanzó una medida de asertividad de 0.94 con respecto a su predicción. Lo que nos permite afirmar que el algoritmo Random Forest realiza una predicción más efectiva, esto quiere decir que obtiene un número mayor de predicciones verdaderas para cada tipo de rendimiento académico (1/2/3), este número de predicciones verdaderas que son analizadas en la diagonal interna de la matrix confusion y posteriormente ingresan a la fórmula para obtener la medida de asertividad. Este resultado es similar al de Mohammed [9] quien en su investigación de minería de datos enfocada a la educación (EMD), obtiene que el algoritmo Random Forest le brinda el nivel más alto de asertividad en la predicción del desenlace del estudiante (pasar / fallar) luego de haberlo comparado con otros algoritmos, siendo Decision Tree uno ellos. Como consecuencia de lo analizado anteriormente, confirmamos que el algoritmo Random Forest genera predicciones más acertadas en comparación al algoritmo Decision Tree debido a su robustez para entrenar el modelo, pero esto puede variar de acuerdo a la situación que se desea abordar.

Luego, al determinar la importancia de cada variable (factores de influencia) dentro del modelo, se halló que existe un grupo de variables que presenta un valor de importancia menor a 2 (P1-P30) y otro que presenta un valor por encima de 11 (Adap_cuali,etc), además hay dos variables cuyo valor de importancia van muy por encima de las demás con 46.1(prom_pnderado) y 22.8 (test_total). El modelo calcula la importancia de cada una según la precisión que estas le generen, en otras palabras, las variables cuyo valor de importancia está por debajo de 2 no las considerada como buenos predictores, por otra parte, las variables cuyo valor está por encima de 11 sí aportan más a su precisión. Esto concuerda en cierta parte con lo que afirma Casas [31], quien obtuvo valores de importancia por debajo de 3 para algunas de las variables de su modelo y considera que estas no cumplen una muy buena función como predictores. Después de este análisis decimos que entre mayor nivel de importancia tenga una variable en el modelo, más precisión aporta y se le puede considerar como buen predictor. Cabe aclarar también que las variables no contribuyen de manera aislada, sino que trabajan en grupo.

Finalmente, al evaluar comparativamente la correlación que tienen las dimensiones psicológicas con el rendimiento académico del estudiante, se obtuvo que, de las cinco dimensiones, cuatro presentan un grado de correlación positiva fuerte (Impresión positiva, Intrapersonal, Manejo estrés, Adaptabilidad) y una presenta un grado de correlación negativa moderado (Interpersonal). Presentar un grado de correlación positiva fuerte significa que a medida que una variable se mueve, la otra variable también lo hace (en una sola dirección), este el caso de las cuatro dimensiones con respecto al rendimiento académico. En cambio, para la única dimensión que obtuvo un grado de correlación negativa moderada significa que se mueve en dirección opuesta a la otra variable, pero de una manera estable. Este análisis refuerza lo establecido por Llibre [14] que en su investigación demostró que existe una relación significativa entre la inteligencia emocional y el logro académico del estudiante. Así también se profundiza en cierto aspecto la investigación de Yamao [11] que entre sus factores de deserción consideró la motivación, quedando este un poco limitado para evidenciar una relación del factor emocional con sus resultados. En tal sentido, bajo lo descrito anteriormente, confirmamos que la inteligencia emocional como tal y las dimensiones que esta abarca, se relacionan y tienen un papel muy importante en los logros de los estudiantes.

Conclusiones

De acuerdo a los resultados que se han podido obtener de la investigación desarrollada, se concluye que:

Se determinó que el algoritmo Random Forest obtuvo la medida de asertividad de 0.95 con respecto a la predicción del rendimiento, siendo esta la más óptima luego de haberla comparado con la del algoritmo Decision Tree, la cual fue de 0.94. Dicha medida se calcula mediante el número de predicciones verdaderas para cada tipo de rendimiento académico (1/2/3). Esta determinación permitió seleccionar el algoritmo con el cual el modelo posteriormente pasó a ser automatizado.

Se logró determinar la importancia para cada variable (factores) que interviene en el modelo, lo que nos permitió identificar que los factores: promedio ponderado (prom_ponderado) y total de test (test_total) obtuvieron los valores de importancia más altos con 46.1 y 22.8, por otro lado, los valores de las dimensiones (Adap_cuali, etc.) oscilan entre 7.0 a 19.0 y los de las respuestas de cada pregunta (P1...Px) entre 0.8 a 1.5.

Se pudo evaluar comparativamente la correlación que se obtuvo de cada una de las dimensiones psicológicas con respecto al rendimiento académico mediante el uso de la herramienta de semaforización y escala de Pearson. Las dimensiones que presentaron una fuerte correlación positiva fueron cuatro: adaptabilidad, manejo de estrés, intrapersonal e impresión positiva, teniendo esta última el valor más alto con 0.84; por otra parte, la dimensión interpersonal fue la única que presentó una correlación negativa moderada con un valor de 0.38.

Recomendaciones

Se recomienda enlazar el sistema de predicción de rendimiento académico con la base de datos de la institución para que pueda ser alimentado de la data correspondiente y a su vez aumente la productividad.

Con respecto a la información generada por el sistema de predicción, se recomienda realizar backups de seguridad. Esto con la finalidad de no perder la información que sirve de base para futuras investigaciones o planteamiento de diferentes modelos.

Se sugiere abarcar más factores que influyan en el rendimiento académico, con el propósito de aumentar la consistencia en los resultados del sistema.

Para el uso correcto del sistema, se aconseja realizar capacitaciones al personal de la institución que tendrá interacción directa con este.

Referencias

- [1] C. Jaspe. (2010, Julio 7). El Rendimiento escolar y estrategias de enseñanza [Online]. Available: <https://bit.ly/2UAANiB>. [Accessed: Oct 15, 2019].
- [2] J. Oliver, “Factores asociados al rendimiento académico”, *Educación*, vol. 31, núm. 1, pp. 43-63, 2007.
- [3] OCDE. Estudiantes de bajo rendimiento. Paris: European Union, 2016.
- [4] “Los países de América Latina con peor rendimiento académico”, en Diario virtual de BBC MUNDO, Febrero de 2016. [Online] Available: <https://bbc.in/36vUehV>. [Accessed: Oct 22, 2019].
- [5] A. Schleicher and D. Salinas, “Estudiantes de bajo rendimiento: Resumen Mexico”, Low Performance Students, 2016. [Online] Available: <https://bit.ly/3cBGnKk>. [Accessed: Oct 16, 2019].
- [6] H. Lamas, “Sobre el rendimiento escolar”, *School Performance*, vol. 3, no. 1, pp. 313386. Junio 2015.
- [7] Ipsos. (2019, Febrero 25). Las preocupaciones del mundo según Ipsos Global [Online]. Available: <https://bit.ly/3kphP8L>. [Accessed: Nov 5, 2019].
- [8] M. Imran, S. Latif and M Saqlain, “Student Academic Performance Prediction using Supervised Learning Techniques”, Article, Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan, 2019. [Online] Available: <https://bit.ly/3HBMQDv>. [Accessed: Oct 24, 2019].
- [9] M. Hikmat and N. Sadiq, “Classifying and Predicting Students Performance using Improved Decision Tree C4.5 in Higher Education Institutes”, Article, University of Zakho and Duhok Polytechnic University, Duhok, Iraq, 2019. [Online] Available: <https://bit.ly/30OeuvR>. [Accessed: Oct 24, 2019]
- [10] A. Holgado, “Detección de patrones de bajo rendimiento académico”, Tesis Maestría, Informática, Universidad Nacional del Altiplano, Puno, Perú, 2018.
- [11] E. Yamao, “Predicción del rendimiento académico mediante minería de datos”, Tesis Maestría, Sistemas y Computación, Universidad San Martín de Porres, Lima, Perú, 2018.
- [12] “¿Cuáles son los factores que afectan el rendimiento escolar en los adolescentes?”, en Diario virtual Perú21, Noviembre de 2018. [Online]. Available: <https://bit.ly/3eWvpQc>. [Accessed: Sep 11, 2019].
- [13] F. Enseñanza. (2011, Enero). Temas para la educación – Inteligencia emocional (12ava ed.). [Online]. Available: <https://bit.ly/2Iz7Xfq>. [Accessed: Sep 25, 2019].
- [14] R. Estrada, “Influencia de la inteligencia emocional en los resultados académicos de estudiantes de las Ciencias Médicas”, *Revista H. Ciencias Médicas*, vol.14, no.2, Abril 2015.
- [15] A. Del Fabro, “¿Por qué el cociente emocional es más importante que el coeficiente intelectual?”, *Psyciencia*, Mayo 2013. [Online]. Available: <https://bit.ly/3cCNPf0>. [Accessed: Sep 26, 2019]
- [16] S. Murcia, “Manejar las emociones, factor importante en el mejoramiento de calidad de vida”, *Logos ciencia y tecnología*, vol. 4, no. 6, pp. 58-67, Diciembre 2012.
- [17] Manual Técnico Test BarOn Ice – Niños y Adolescentes, N. Ugarriza, Lima, Perú, 2010.
- [18] V. San Juan. (2016, Abril 27). Ventajas de los sistemas web (1ra ed.) [Online]. Available: <https://bit.ly/3koGPwZ>. [Accessed: Oct 1, 2019].

- [19] B. Sintés. “¿Qué es PHP?”. Programación en PHP. [Online]. Available: <https://bit.ly/2K4SW60>. [Accessed: Oct 14, 2019].
- [20] Metodoss. “Metodología RUP”. Metodoss C. [Online]. Available: <https://bit.ly/2UmAIPf>. [Accessed: Nov 16, 2019].
- [21] Y. Rodríguez, “Herramientas de Minería de Datos”, *RCCi*, vol. 3, no. 3, pp. 73-80, Diciembre 2009.
- [22] A. Jiménez and H. Álvarez, “Minería de datos en la educación”, Article, Universidad Carlos III de Madrid, Madrid, España, 2010. [Online] Available: <https://bit.ly/3r3MmjL>. [Accessed: Dec 09, 2019].
- [23] J. Molina and J. García, *Técnicas de análisis de datos*, Madrid: HomiSacres, 2006. [Online]. Available: <https://bit.ly/3HKh5Ia>. [Accessed: Dec 11, 2019]
- [24] D. Fernández. (2019, Noviembre 8). Algunos algoritmos para análisis predictivo [Online]. Available: <https://bit.ly/3np6gAB>. [Accessed: Dec 10, 2019].
- [25] Manual CRISP-DM Modeler, IBM, N Castle Dr, United States, 2012.
- [26] P. Rochina. (2016, Noviembre 16). Python vs R para el análisis de datos [Online]. Available: <https://bit.ly/3ncDH9o>. [Accessed: Oct 28, 2019].
- [27] R. APD. (2019, Abril 14). ¿Cuáles son los tipos de algoritmos del machine learning? [Online]. Available: <https://bit.ly/3ktCit6>. [Accessed: Dec 1, 2019].
- [28] E. GRAPH. (2015). Machine Learning en Python [Online]. Available: <https://bit.ly/2UB4akR>. [Accessed: Nov 30, 2019]
- [29] J. Muñoz. (2017, Noviembre 17). ¿Qué es Flask? [Online]. Available: <https://bit.ly/3niAZ26>. [Accessed: Dec 4, 2019].
- [30] C. De La Cruz, “Metodología de la investigación tecnológica en ingeniería”, *Revista Ingenium*, vol. 1, no. 1, pp. 43-46, Junio 2016.
- [31] P. Casas. *Libro vivo de ciencia de datos*. Kindle ed., USA, 2019.

Anexos

Anexo N.º 01. Constancia de aprobación del producto acreditable de la entidad donde se ejecutó la tesis



INSTITUCIÓN EDUCATIVA EMBLEMÁTICA "SARA A. BULLÓN" - 10110

Dios Ciencia y Liderazgo

DECRETO LEY DE CREACIÓN N.º. 14992 de 1964









"Año de la lucha contra la corrupción e impunidad"

Lambayeque, 08 de noviembre de 2019

CARTA N°001-2019

Señor
Ing. Huilder Mera Montenegro
 Director de Escuela de Ingeniería de Sistemas y Computación
 Universidad Católica Santo Toribio de Mogrovejo

Por este medio de la presente hago constatar que el estudiante **GARCÍA PEREDO, LUIS JESÚS** identificado con **DNI 73772317**, de la escuela de **INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**, ha sido aceptado en nuestra Institución Educativa para llevar a cabo la realización de su proyecto de tesis.

Para la continuidad del proyecto del alumno en mención, le brindaremos la información necesaria, las facilidades pertinentes, permisos y apoyo que amerite el caso.

Atentamente,



I.E. N° 10110 "SARA A. BULLÓN"

Mg. Roberto González Dávila

DIRECTOR
DNI: N° 17530175

Fig. 14. Constancia de aprobación del producto acreditable de la entidad donde se ejecutó la tesis

Fuente: Dirección Académica - Institución

Anexo N.º 02. Entrevista – Subdirector

ENTREVISTA N.º 1

Entrevistado: Sub director

Entrevistador: Luis Jesús García Peredo - Tesista

Objetivo: Obtener información más detallada de la situación y conocer disponibilidad para realizar cambios en sus procesos.

Fecha: 10/10/2019

- 1) ¿Cuál es la situación actual con respecto al desempeño académico de los estudiantes?
Es un poco irregular la situación, generalmente siempre están esos grupos de estudiantes que mantienen sus resultados estables, pero también hay otro grupo de estudiantes que no mejoran su desempeño y lo que nos preocupa es que ese grupo es mucho más amplio, son principalmente estudiantes que se encuentran cursando los primeros años de secundaria.

- 2) ¿Por qué cree que se da esa situación en la institución?
Bueno pueden ser diferentes motivos, pero cuando se habla con los estudiantes que no presentan un buen desempeño, casi siempre nos damos cuenta que están pasando por problemas en casa. También hay alumnos que son un poco inseguros al hablar o no se sienten capaces ellos mismos de mejorar su desempeño.

- 3) ¿De qué manera se ha intentado o que acciones se han tomado para tratar de mejorar esta situación?
Tratamos de brindarles reforzamiento académico, pero no ha dado mucho resultado, también se derivan a los estudiantes con el área de psicología para que pueda evaluar un poco más allá, algunos estudiantes si han mostrado algunas mejoras con respecto a su actitud luego de haber sido evaluados por esa área.

- 4) ¿Por qué cree que no han dado mucho resultado los intentos de mejora?

Va más por el tema de la falta de información, se tiene una idea del por qué se da la situación, pero eso no es suficiente. Si se tuviera información un poco más profunda sobre las causas o el contexto, las cosas podrían tomar otro rumbo creo yo.

- 5) ¿Cree que el uso de la tecnológica pueda servir como una herramienta de apoyo para intentar mejorar la situación?

Sí por supuesto, me llama mucho la atención ese tema, la tecnología siempre es una muy buena herramienta si es que uno sabe utilizarla y sacarle provecho. Sería muy importante si se llega a implementar algo aquí.

- 6) ¿De qué manera cree que un sistema de predicción de rendimiento académico apoyaría a la toma de acciones para ayudar a los estudiantes con su desempeño académico? Considero que un sistema así sería algo muy beneficioso para la institución porque nos permitiría trabajar de una manera más eficaz utilizando información más relevante de la situación, además, pienso que nos permitiría adelantarnos a un resultado y así poder tener un panorama más claro con respecto a qué tipo de acciones tomar, ya no esperar que la situación se dé sino tratar de prevenirla.

Anexo N.º 03. Entrevista – Área de psicología

ENTREVISTA N.º 2

Entrevistado: Área de psicología

Entrevistador: Luis Jesús García Peredo - Tesista

Objetivo: Conocer las limitaciones que presentan para ayudar a los estudiantes y analizar propuestas que refuercen el modelo.

Fecha: 15/10/2019

- 1) ¿Cuál es el rol del área dentro del intento de mejora del desempeño académico de los estudiantes?

Nosotros trabajamos con los estudiantes que son derivados por los docentes o por el departamento de TOE debido a que no presentan mejoras en su desempeño. Mayormente son estudiantes de 1º y 2º año, primero dialogamos con ellos y en base a eso procedemos a evaluarlos de acuerdo a lo que nosotros vamos identificando.

- 2) ¿Consideran que existen algunas limitaciones para evaluar a los estudiantes?

Sí, el tiempo es la principal, nosotros evaluamos a todos los niveles de la institución, pero los estudiantes que exigen un poco más de tiempo son los de los años mencionados anteriormente. No se puede evaluar de la misma manera a todos ya que cada uno es una situación distinta, pero dentro de nuestras posibilidades tratamos de abarcar la mayor cantidad de estudiantes.

- 3) ¿Por qué consideran que en 1º y 2º año hay más estudiantes que no presentan un buen desempeño académico? Y ¿Por qué toma más tiempo trabajar con ellos? En esos años los estudiantes están en una edad difícil, se presentan muchos cambios propios del desarrollo, a eso se le suman los problemas que puedan estar pasando en casa debido a que la mayoría vienen de hogares humildes donde existen algunas ausencias, todo eso afecta directamente a lo que ellos puedan realizar. Toma más tiempo trabajar con ellos porque hay que abarcar muchos aspectos para poder llegar a tener un diagnóstico

claro que nos permita ayudarlos de manera eficiente y que esa ayuda luego se vea reflejada en su desempeño académico.

4) ¿En qué tiempos del año escolar se evalúan más a estos estudiantes?

Casi siempre es a mitad y a fines de año, debido a que en esos tiempos recién se va haciendo más evidente que tipo de desempeño están presentando. Esos son los tiempos que nos limitan un poco por la cantidad de estudiantes.

5) ¿Qué tipo de evaluación mayormente les hacen a estos estudiantes?

No es una sola evaluación, son diferentes tipos y enfocadas a diferentes factores. Pero generalmente evaluamos la autoestima, autocontrol, responsabilidad, etc.

6) ¿Consideran que un sistema de predicción de rendimiento académico basado en factores psicológicos y académicos sea un buen complemento para su trabajo con los estudiantes? Y ¿Qué factores psicológicos creen que se deberían abarcar?

Claro que sí, principalmente por el tema del tiempo, nos permitiría comenzar a trabajar con los estudiantes desde inicio de año y de acuerdo con los factores que se consideren ya se iría formando un esquema de trabajo con ellos. Los factores que se deberían considerar deben ser muy variados, hay una evaluación muy completa enfocada a la inteligencia emocional que se divide por escalas y que abarca casi la mayor parte de puntos importantes, "BARON ICE", sería algo muy valioso si se puede elaborar algo considerando esos factores psicológicos.

Anexo N.º 04. Evaluaciones de algunos factores psicológicos por parte del área de psicología

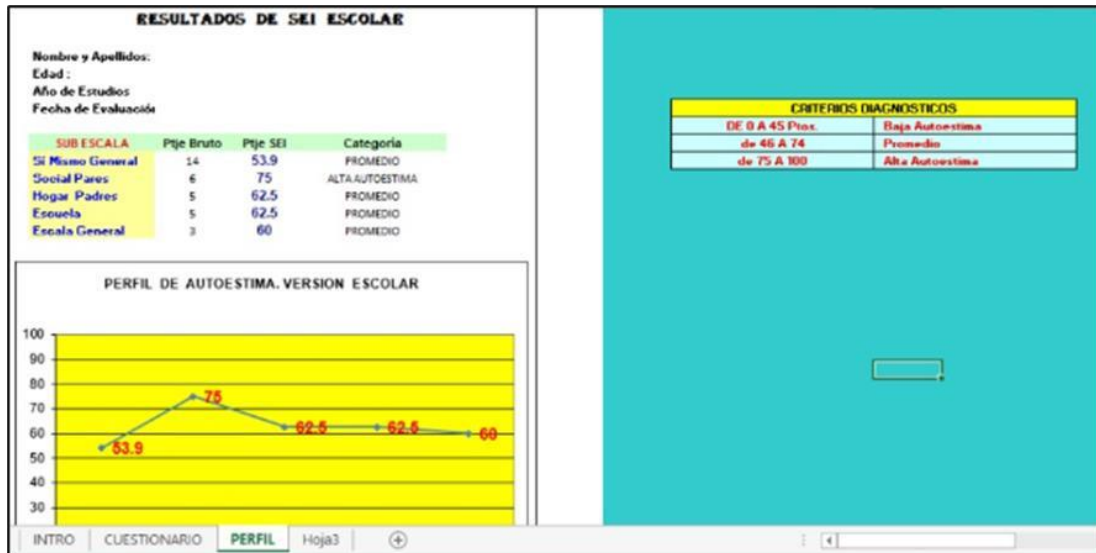


Fig. 15. Evaluación 1 – Autoestima
Fuente: Área de psicología

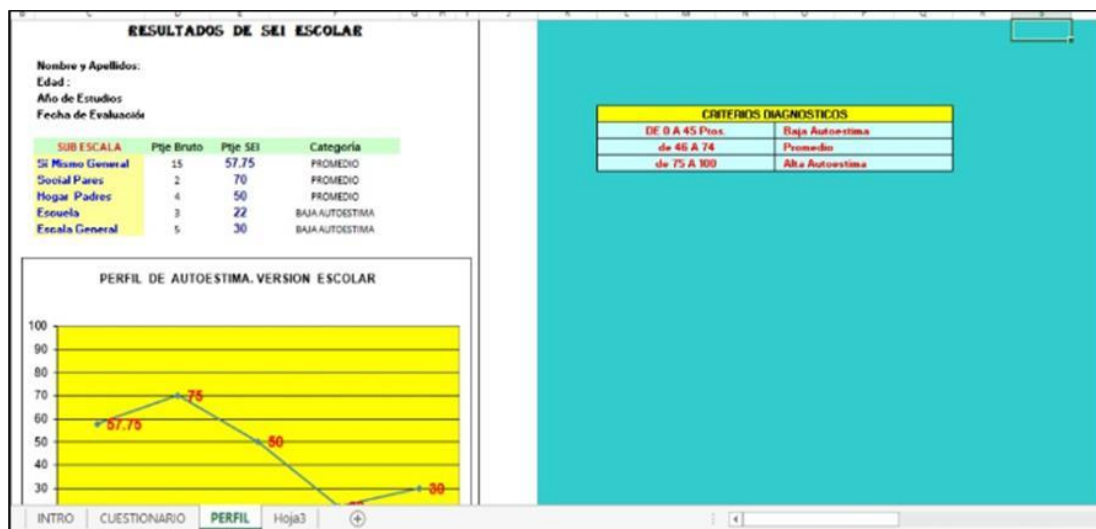


Fig. 16. Evaluación 2 – Autoestima
Fuente: Área de psicología

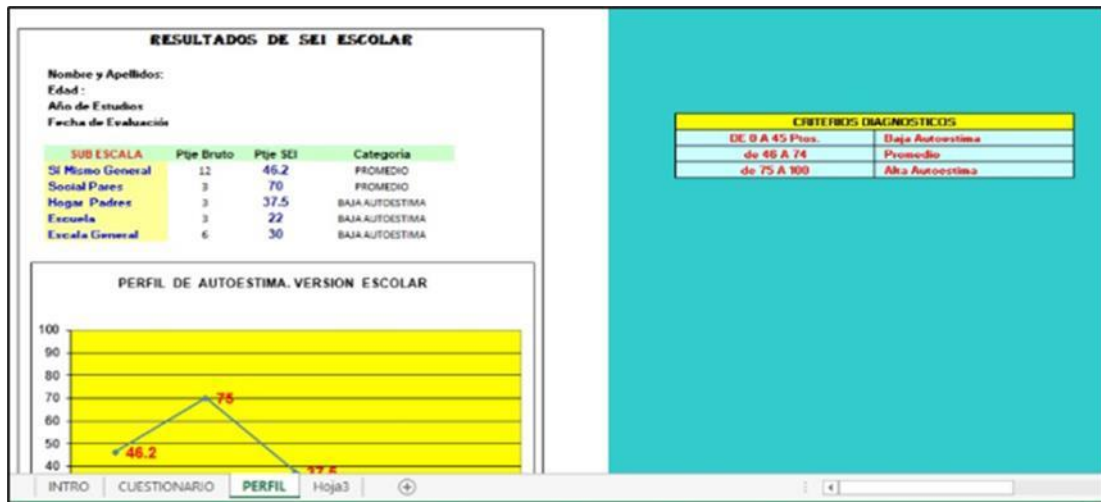


Fig. 17. Evaluación 3 – Autoestima
Fuente: Área de psicología