FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENERÍA DE SISTEMAS E INFORMÁTICA

Tesis

Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba

Para optar el título profesional de Ingeniero de Sistemas e Informática

Autor:
Jules Mao Flores Satalaya
https://orcid.org/0009-0001-6553-0809

Asesor:
Ing. Dr. Carlos Enrique López Rodríguez
https://orcid.org/0000-0001-7847-6859

Tarapoto, Perú

2024

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENERÍA DE SISTEMAS E INFORMÁTICA

Tesis

Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba

Para optar el título profesional de Ingeniero de Sistemas e Informática

Autor

Jules Mao Flores Satalaya

Sustentado y aprobado el 01 de febrero del 2024, por los siguientes jurados:

Presidente de Jurado
Ing. Dr. Jorge Damián Valverde Iparraguirre

Secretario de Jurado
Lic. Jose Luis Ramirez del Aguila

Vocal de Jurado
Ing. Dr. John Antony Ruíz Cueva

Tarapoto, Perú

2024

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

En los ambientes del Aula Magna de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de San Martín, a las 20:00 horas del día jueves 01 de febrero del año 2024, se reunieron los **miembros del Jurado Calificador,** integrado por:

Presidente	:	**ING. DR. JORGE DAMIAN VALVERDE IPARRAGUIRRE**
Secretario	:	**LIC. JOSE LUIS RAMIREZ DEL AGUILA**
VOCAL	:	**ING. DR. JOHN ANTONY RUIZ CUEVA**

Para evaluar la Tesis: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL EN LA INSTITUCIÓN EDUCATIVA 00116 ALTO PERÚ - MOYOBAMBA; presentada por el Bachiller JULES MAO FLORES SATALAYA, participando en calidad de asesor el **Ing. Dr. Carlos Enrique López Rodríguez.**

Los señores miembros del Jurado, después de haber atendido la sustentación y evaluada las respuestas a las preguntas formuladas y terminada la réplica; luego de debatir entre sí, reservada y libremente lo declaran....APROBADO....., por ...UNANIMIDAD...., con el calificativo de .....BUENO....., equivalente a .....QUINCE....., en fe de lo cual firmamos la presente acta, siendo las ....21.30..... horas del mismo día, con lo que se dio por terminado el Acto de Sustentación.

ING. DR. JORGE DAMIAN VALVERDE IPARRAGUIRRE
Presidente

LIC. JOSE LUIS RAMIREZ DEL AGUILA
Secretario

ING. DR. JOHN ANTONY RUIZ CUEVA
Vocal

Constancia de asesoramiento

El que suscribe el presente documento, Ing. Dr. Carlos Enrique López Rodríguez

Hace constar:

Que, he revisado la tesis titulada: **Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba**, en fechas del cronograma a fin de optimizar y agilizar la investigación, elaborada por el Señor:

Bachiller en Ingeniería de Sistemas e Informática: **Jules Mao Flores Satalaya**

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

Tarapoto, 01 de febrero de 2024

Atentamente:

.....

Ing. Dr. Carlos Enrique López Rodríguez

Declaratoria de autenticidad

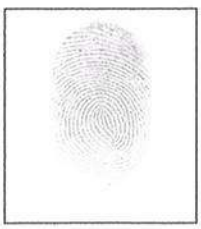
Jules Mao Flores Satalaya, con DNI N° **45218746**, egresado de la Escuela Profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de San Martín, autor de la tesis titulada: **Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba**.

Declaro bajo juramento que:

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

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

Tarapoto, 01 de febrero de 2024.



.....

Jules Mao Flores Satalaya
DNI N° 45218746

Ficha de identificación

Título de la tesis **Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba**	**Área de investigación:** Ingeniería y Tecnología **Línea de investigación:** Estrategia de tecnología de la información y comunicación (TIC) y sistemas constructivos convencionales y no convencionales para el desarrollo sostenible **Sublínea de investigación:** Desarrollo de software y toma de decisiones **Grupo de investigación:** (indicar Resolución) **Tipo de investigación:** Básica ☐, Aplicada ☒, Desarrollo experimental ☐

Autor: **Jules Mao Flores Satalaya**	Facultad de Ingeniería de Sistemas e Informática Escuela Profesional de Ingeniería de Sistemas e Informática https://orcid.org/0009-0001-6553-0809

Asesor: **Ing. Dr. Carlos Enrique López Rodríguez**	**Dependencia local de soporte:** Facultad de Ingeniería de Sistemas e Informática Escuela Profesional de Ingeniería de Sistemas e Informática Unidad o Laboratorio Ingeniería de Sistemas e Informática https://orcid.org/0000-0001-7847-6859

Dedicatoria

¡Que nadie se quede afuera, se los dedico a todos!

Sobre todo, a mis padres y personas que siempre están conmigo en las buenas y en las malas, dándome apoyo altruista.

Mao

Agradecimientos

A mi familia, cimiento de mi existencia y fuente de sabiduría, agradezco profundamente por haberme otorgado el privilegio de forjarme en esta venerable institución académica. Su constante apoyo y aliento han sido mi faro en este viaje hacia el conocimiento.

A mi asesor, quien no solo me ha orientado en la elaboración de esta obra de culminación, sino que ha sido un mentor constante a lo largo de mi travesía universitaria. Su inquebrantable compromiso y estímulo han sido un pilar fundamental en mi crecimiento profesional y en la continua búsqueda de la excelencia.

A la Universidad Nacional de San Martín - Facultad de Ingeniería de Sistemas e Informática, manantial de oportunidades y tesoro de saberes, mi gratitud sincera por haberme brindado un espacio donde nutrirme de conocimiento y crecer como individuo.

Mao

Índice general

Índice de tablas

Índice de figuras

RESUMEN

Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba

El presente estudio tuvo como objetivo determinar la influencia de los algoritmitos aprendizaje supervisado en la deserción estudiantil de la Escuela de Posgrado - Universidad Enrique Guzmán y Valle; el estudio fue de tipo aplicado, nivel descriptivo y diseño metodológico no experimental. La población compuesta por 10659 estudiantes, la muestra fue dada por conveniencia siendo el total de 5758 estudiantes. Las técnicas fueron el análisis documental y la observación, los instrumentos correspondieron a la ficha documental y el cuadro de resumen de predicciones. Se analizaron 110 archivos donde registran la deserción estudiantil, que mediante procedimientos de ETL se pudo consolidar un dataset con información de 387 estudiantes, asimismo se determinó que la situación de matrícula, los cursos aprobados, los cursos desaprobados, el comportamiento, grado culminado y el nacimiento registrado son variables influyentes en gran significancia en la deserción estudiantil. Seguidamente se analizaron 5 técnicas de minería de datos, siendo los algoritmos J48, Random Forest, Vecinos Mas Cercanos, Función Logística y Perceptrón Multicapa, desempeñándose con mayor efectividad el algoritmo J48. Se evaluaron 116 casos, de los cuales el modelo basado en técnicas de minería de datos no acertó 5 casos, siendo equivalente al 4.31%. Por lo tanto se concluyó que la deserción estudiantil se puede predecir a través de los datos académicos y sociales, utilizando los algoritmos de aprendizaje automático hasta en un 95.69% de efectividad.

Palabras clave: Técnicas, minería de datos, predicción, deserción estudiantil.

ABSTRACT

Application of data mining techniques to predict student desertion in the Educational Institution 00116 Alto Perú - Moyobamba.

The objective of this study was to apply data mining techniques to predict student dropout in the Educational Institution 00116 Alto Perú - Moyobamba; the study was applied, with a descriptive level and non-experimental methodological design. The population consisted of 412 students, the sample was given by convenience making a total of 387 students. The techniques used were documentary analysis and observation; the instruments used were the documentary file and the prediction summary table. A total of 110 files recording student desertion were analyzed, and a dataset with information on 387 students was consolidated by means of ETL procedures. It was also determined that enrollment status, courses passed, courses failed, behavior, grade completed and birth registered are highly significant influential variables in student desertion. Subsequently, 5 data mining techniques were analyzed, being the J48, Random Forest, Nearest Neighbors, Logistic Function and Multilayer Perceptron algorithms, with the J48 algorithm performing more effectively. A total of 116 cases were evaluated, of which the model based on data mining techniques failed in 5 cases, equivalent to 4.31%. Therefore, it was concluded that academic performance can be predicted through academic and social data, using machine learning algorithms up to 95.69% of effectiveness.

Keywords: Techniques, data mining, prediction, student dropout.

CAPÍTULO I
INTRODUCCIÓN A LA INVESTIGACIÓN

En un escenario global, la educación se vincula directamente con el acelerado progreso económico de un país (Kim & Kim, 2018). En donde la educación ha desempeñado un papel esencial para lograr un rápido crecimiento económico en diversos países desarrollados (Urquizo, 2014). Siendo una tarea monumental para las instituciones educativas poder identificar la eventualidad de que los estudiantes abandonen sus estudios (Pérez, 2020). Frente a este desafío, la deserción estudiantil representa un problema complejo y diverso a nivel global, impactando de diversas maneras en todas las instituciones educativas (Lázaro et al., 2020). En este sentido, la deserción académica representa un desafío muy grande y fundamental en el ámbito educativo, que afecta a todos los niveles y formas del sistema educativo (Quiñones et al., 2020). Por otro lado, el Gobierno, mediante políticas públicas, debe prevenir la deserción académica, catalogándose no como problema individual, más bien como un fenómeno que podría acarrear graves consecuencias sociales si no se aborda adecuadamente (Moreno, 2013). Entonces la deserción escolar trae consigo varios problemas para la sociedad pues se presenciará menos recursos humanos preparados para afrontar problemas sociales, económicos y políticos (Necochea et al., 2017).

En América Latina, las tasas de abandono escolar temprano son significativamente altas, dado que los alumnos dejan la institución educativa antes de completar el periodo básico (Hopenhayn, 2002). Siendo la deserción escolar un desafío que afecta a naciones en desarrollo, como es el caso de Ecuador, donde los estudiantes se ven compelidos a abandonar sus estudios debido a una variedad de motivos (Sinchi & Gómez, 2018). Asimismo, en Argentina, el 70% de los alumnos que ingresan a una institución educativa no culminan sus estudios (Vera et al., 2020). De la misma manera la deserción estudiantil es notorio en Chile el cual tuvo aumento del 43% (Ministerio de Educación de Chile, 2020). Asimismo, la deserción escolar constituye un desafío complejo en las instituciones educativas de Colombia, impactando de manera uniforme en todo el sistema educativo del país (Varón, 2017). En Bolivia el problema es más preocupante siendo un porcentaje muy significativo de los adolescentes que no concluyen el ciclo de educación escolar (Morales & Vargas, 2018). Ante ello se puede decir que la deserción escolar sigue siendo un desafío importante para América Latina (Carvajal, 2016).

Análogamente en el Perú la educación es una necesidad primordial que todo individuo debe tener, y es fundamental para que como nación pueda lograr un nivel cultural y social elevado (Urquizo, 2014). Por lo tanto, la deserción escolar representa un problema de índole social que se ha convertido en un foco primordial de efectos adversos que afectan a la sociedad (Hernández et al., 2017). En este entorno, en Perú, la partida de los estudiantes antes de finalizar el ciclo educativo conlleva significativas pérdidas tanto a nivel individual como social (Espíndola & León, 2002). La deserción educativa emerge como una preocupación de gran envergadura para el país, como lo señala el MINEDU, pues se estima que hasta julio de 2020 el abandono en la educación primaria aumentó del 1.3% al 3.5%, afectando a unos 128.000 escolares, mientras que en el nivel secundaria supero el 3.5% para llegar al 4%, afectando a unos 102.000 escolares. Además, esta cifra se va incrementado por los diversos traslados de 337.870 escolares de I.E privadas a públicas. Paralelamente, según datos de la ENAHO, hacia fin del 2020, los principales factores que motivan a los escolares a abandonar sus estudios académicos continúan siendo las dificultades económicas (75.2%), problemas en familia (12.3%), y el desinterés (4%) (ComexPeru, 2020). Tras dicho resultado el Gobierno Peruano conjuntamente con sus órganos e instituciones deben reducir la deserción de los estudiantes con la reformulación de estrategias eficaces en la educación peruana (Gallego et al., 2021).

Finalmente, en el escenario regional, el departamento de San Martín la tasa de deserción entre los periodos 2015-2019 en el nivel inicial el promedio regional fue de 2.5% en el 2015 y en el periodo 2019 se arribó a un porcentaje de 1.9 %, lo que significa que ha disminuido el abandono o deserción escolar en el nivel inicial. Asimismo, en el nivel primaria el promedio regional fue del 2.0% en el año 2015 y en el periodo 2019 se arribó a un porcentaje del 1.5%. De la misma forma en el nivel secundaria el promedio regional de deserción escolar fue de 5.7% en el año 2015 y en el año 2019 se arribó a un porcentaje de 4.9% (Dirección Regional de Educación de San Martín, 2021). Entonces se puede decir que la deserción escolar en la región San Martín ha tenido un comportamiento con tendencia de bajada, sin embargo, la deserción educativa aún sigue representando un problema grande el cual las instituciones educativas deben de lidiar.

En el escenario local, se tiene a la Institución Educativa 00116 Alto Perú - Moyobamba el cual es una institución que cuenta con los 3 niveles educativos, donde el nivel inicial cuenta con 57 alumnos, en el nivel primaria con 128 alumnos y secundaria la cantidad de 132 estudiantes, haciendo un total de 317 alumnos (Ministerio de Educación del Perú, 2022), asimismo cabe indicar que dicha I.E se encuentra ubicado en el CCPP Alto Perú

ubicado en el distrito de Soritor, en la provincia de Moyobamba, el cual pertenece a la zona rural de la jurisdicción, y la deserción estudiantil no ha sido la excepción en la institución, siendo producto de muchos factores debido que la gran parte de los alumnos se dedican a otras labores, adheridos a problemas económicos, problemas familiares y por falta de motivación para seguir asistiendo a las aulas de la institución educativa. Dicha problemática viene ser preocupante para el personal estratégico de la institución, ya que no pueden controlar el alto índice de deserción escolar.

Ante ello se resalta la relevancia de la detección temprana de la deserción escolar para formular estrategias claves permitiendo tomar decisiones importantes en la gestión educativa (Quiñones et al., 2020). En ese contexto se puso énfasis en recurrir a las técnicas de minería de datos con el fin de predecir la deserción escolar a futuro en base a datos históricos tanto sociales, económicos y académicos de los estudiantes, con ello permitir apoyar las actividades de toma de decisiones en la institución educativa (Castro et al., 2019). Cabe indicar que la minería de datos educativos ha mejorado los sistemas de educación al predecir y analizar los aspectos conductuales de la enseñanza y los puntajes educativos de los estudiantes (Tan & Lin, 2021). El cual brindan un análisis de exhaustivo de la información y contribuyen a las decisiones educativas (Yağcı, 2022).

Frente a toda la problemática abordada se formuló como hipótesis general H_i: La deserción estudiantil puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú - Moyobamba. Además se contó con la formulación del problema ¿De qué manera la aplicación de técnicas de minería de datos predice la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba?. Y finalmente como objetivo general: Aplicar técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú – Moyobamba. Generándose objetivos específicos como: Analizar los archivos que registran la deserción estudiantil; determinar variables influyentes en la deserción estudiantil; determinar las técnicas de minería de datos a aplicar que mejoren la predicción de la deserción estudiantil y analizar los resultados obtenidos de la aplicación de técnicas de minería de datos en la predicción de la deserción estudiantil.

.

CAPÍTULO II
MARCO TEÓRICO

2.1. Antecedentes de la investigación

En el escenario internacional se tuvo a:

Kim et al. (2021), "*A study in the early prediction of ict literacy ratings using sustainability in data mining techniques*" (articulo cientifico). Tuvo como objetivo examinar la sostenibilidad de las técnicas de minería. Las técnicas empleada se basan en algoritmos OneR, J48, embolsado, bosque aleatorio, perceptrón multicapa y optimización mínima de datos utilizadas para predecir los resultados del aprendizaje. El resultado de predicción temprana más alto de aproximadamente un 69 % de precisión se obtuvo para el algoritmo SMO cuando se usaban 47 atributos. En términos generales, a través de métodos de gestión de datos, estos hallazgos contribuirán a la detección temprana de estudiantes en situación de riesgo durante su proceso educativo, y facilitarán el diseño de estrategias educativas y de aprendizaje adaptadas a las necesidades individuales de cada alumno.

Ferreira & Camões (2019), "*Prediction of restrained shrinkage crack width of slag mortar composites usin data mining techniques*" (Artículo científico). Tuvo como objetivo desarrollar modelos de minería de datos para predecir anchos de grietas por contracción restringida de compuestos cementosos de mortero de escoria. Como herramienta de modelado, se utilizó el entorno R para la ejecución de estas técnicas de DM. Se probaron y analizaron varios algoritmos utilizando todas las combinaciones de los parámetros de entrada. Se concluyó que, utilizando uno o tres parámetros de entrada, los modelos de redes neuronales artificiales (ANN) tienen el mejor rendimiento. Sin embargo, la mejor capacidad de pronóstico se obtuvo con el modelo de máquinas de vectores de soporte (SVM) usando solo dos parámetros de entrada.

Chaurasia et al. (2018), "*Prediction of benign and malignant breast cancer using data mining techniques*" (articulo cientifico). Tuvieron como objetivo elaborar un informe sobre el cáncer de mama, aprovechando las tecnologías para desarrollar modelos predictivos de cáncer de mama. Se utilizaron 3 técnicas populares de DM (Naïve Bayes, Red Neuronal de Base Radial, J48) para construir los modelos de predicción utilizando un extenso conjunto de datos (683 casos de cáncer de mama). Los resultados obtenidos señalaron que Naïve Bayes demostró un mejor rendimiento, con una precisión de

clasificación del 97,36%, seguido por Red Neuronal de Base Radial en segundo lugar, con una precisión del 96,77%, y J48 en tercer lugar, con una precisión del 93,41%.

Kaur et al. (2021), "*An integrated approach for cancer survival prediction using data mining techniques*" (articulo cientifico). Tuvo como objetivo investigar diferentes predictores de supervivencia disponibles para el pronóstico del cáncer utilizando técnicas de DM. Se ha recopilado y utilizado un conjunto de datos de 140 pacientes con cáncer de ovario avanzado que contiene datos de diferentes perfiles de datos (clínica, tratamiento y calidad de vida general) para predecir el cancer en los pacientes. El enfoque propuesto de clasificación y algoritmo de minería secuencial modificado con una precisión del 76,4 % funcionó mejor que el enfoque existente sin minería secuencial, con una precisión de alrededor del 70 %. Los resultados también fueron validados estadísticamente por los médicos. Se recomienda que los investigadores tomen en cuenta tanto la calidad de vida de los sujetos como la secuencia temporal de los tratamientos al desarrollar un modelo predictivo para pacientes con cáncer.

Aluko et al. (2021), "*Prediction of restrained shrinkage crack width of slag mortar composites using data mining techniques*" (articulo cientifico) El objetivo fue evaluar la efectividad del uso de modelos de DM para predecir el rendimiento académico de estudiantes de arquitectura en Nigeria. Los hallazgos indican que el rendimiento escolar previo es un indicador sólido del desempeño futuro en estudiantes de arquitectura. Además, se destaca que el SVM es una herramienta eficaz para prever el rendimiento académico en esta población estudiantil. Se observó que las calificaciones en matemáticas, biología y física influyen notablemente en el rendimiento académico de los estudiantes de arquitectura. Aunque el modelo SVM mostró buenas predicciones, no alcanzó una precisión del 100%. Estos resultados sugieren que existen otros predictores que influyen en el rendimiento escolar de los alumnos de arquitectura que no fueron considerados en los modelos desarrollados.

Meghyasi & Rad (2020), "*Customer churn prediction in telecommunication industry using data mining methods*" (artículo científico). Tuvo como objetivo proporcionar un método híbrido basado en el algoritmo genético y la red neuronal modular para predecir el abandono de clientes en las industrias de telecomunicaciones y utilizar los datos de Irancell como muestra. El resultado de precisión de este estudio, que es del 95,5 %, obtiene el rango de precisión más alto en comparación con el resultado de otros métodos. El uso de redes neuronales modulares con dos módulos de redes neuronales y agregación promedio para la salida conduce a una mayor precisión de clasificación y precisión y recuperación del método propuesto.

Pérez (2020), "*Comparison of data mining techniques to identify signs of student desertion, based on academic performance*" (artículo científico). El propósito fue contrastar distintas técnicas para identificar la deserción estudiantil a partir de los registros académicos de estudiantes inscritos en el programa de Ingeniería de Sistemas de una universidad en Colombia, abarcando un total de siete años. Se compararon técnicas de regresión logística, arboles de decisión y Naive Bayes con el fin de determinar la técnica más efectiva en la detección de desertores. Además, se evaluó la usabilidad y precisión de la herramienta Watson Analytics para un usuario sin experiencia previa. El estudio reflejó que el empleo de técnicas simples es suficiente para obtener precisiones optimas en esta tarea. Asimismo, estos hallazgos se comparten con la comunidad estudiantil con el fin de aportar a la minimización de la deserción estudiantil.

Tan & Lin (2021), "*A new QoE-based prediction model for evaluating virtual education systems with COVID-19 side effects using data mining*" (artículo científico). El objetivo fue introducir un nuevo modelo predictivo para identificar aspectos técnicos relacionados con la enseñanza y el e-learning en plataformas de educación virtual mediante el uso de técnicas de DM. Se emplearon técnicas supervisadas y de minería de reglas de asociación para identificar factores de calidad de experiencia (QoE) eficaces en estos sistemas. Los resultados experimentales mostraron que el modelo predictivo propuesto satisface con los criterios de exactitud, precisión y recall necesarios para predecir los comportamientos vinculados con el aprendizaje asi como la enseñanza en entornos de educación virtual y reuniones virtuales.

Yağcı (2022), "*Educational data mining : prediction of students ' academic performance using machine learning algorithms*" (artículo científico). El propósito fue plantear un nuevo modelo respaldado por algoritmos de machine learning para predecir las calificaciones de las pruebas finales de alumnos universitarios basándose en sus calificaciones previas en exámenes parciales. Se exploraron diversas técnicas de machine learning, como, KNN, RF, SVM, Naive Bayes y Regresión Logística. Se emplearon distintos predictores, algoritmos y enfoques para conocer el desempeño estudiantil de los alumnos. Los resultados validan la viabilidad de utilizar algoritmos de ML para predecir el rendimiento estudiantil de los alumnos.

En el ámbito nacional se contó con:

Bedregal et al. (2020), "*Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria*" (artículo científico). El objetivo fue emplear técnicas de clasificación utilizando IBM SPSS Modeler para anticipar posibles

casos de deserción estudiantil. Se utilizó la matrix confusión para realizar la comparación y la evaluación de la precisión de los diversos modelos, encontrándose que el modelo CHAID 1 alcanzó la precisión del 90.24%. Se determinó que el factor más significativo en la deserción es el índice de rendimiento global, y se resaltó la eficacia de las Técnicas de DM para detectar tendencias y prever el desempeño académico de los alumnos.

Menacho (2020), "*Técnicas de minería de datos aplicadas a la plataforma educativa Moodle*" (artículo científico). El objetivo fue presentar un enfoque metodológico para emplear Técnicas de DM en la plataforma Moodle. Los resultados mostraron que los estudiantes con calificaciones bajas en pruebas y tareas, así como un acceso limitado a Moodle, tienen más probabilidades de no pasar el curso. También se identificó un grupo de estudiantes con deficiente rendimiento que podrían beneficiarse de una retroalimentación adicional. La aplicación de la metodología propuesta, particularmente mediante el uso del algoritmo de árbol C4.5, sugirió que los profesores deberían considerar las tareas y los cuestionarios como actividades que pueden influir en la aprobación de un estudiante.

Bedregal et al. (2020), "*Análisis del rendimiento académico de los estudiantes de Ingeniería de Sistemas , posibilidades de deserción y propuestas para su retención*" (artículo científico). El objetivo fue analizar el rendimiento académico de los grupos de estudiantes de la Escuela Profesional de Ingeniería de Sistemas de una universidad del sector público, abarcando los periodos desde 2011 hasta 2016, mediante datos obtenidos de 976 alumnos. Se utilizaron herramientas de minería de datos, incluyendo un software personalizado y otro comercial, para investigar el desempeño académico de los alumnos y detectar patrones relevantes. Se concluyó que la evaluación del rendimiento académico de un estudiante no debe limitarse únicamente a sus calificaciones, sino que también debe considerar su conducta académica, su desempeño en relación con su cohorte y su progreso en la aprobación de asignaturas.

Huatangari & Carrasco (2020), "*Rendimiento académico empleando minería de datos*" (artículo científico). El propósito consistió en predecir el desempeño estudiantil de los alumnos inscritos en el programa de Ingeniería de Industrias Alimentarias de la Universidad Nacional de Jaén, empleando técnicas de DM. Se recopiló la matriz de datos a partir de registros institucionales y oficinas de la universidad. La metodología aplicada siguió el enfoque CRISP-DM. Tres algoritmos de minería de datos, J48graft, J48 y PART, fueron utilizados para identificar patrones que predicen el rendimiento académico de los estudiantes. Estos algoritmos fueron ejecutados con el software

Weka, logrando un porcentaje de clasificación correcta superior al ochenta y tres por ciento (83%).

Quiñones et al. (2020), "*Modelo para la estimación de la deserción estudiantil Awajún y Wampis empleando minería de datos*" (artículo científico). El objetivo fue emplear técnicas de minería de datos para crear modelos destinados a predecir la deserción de estudiantes de las comunidades Awajún y Wampis en la Universidad Nacional de Jaén. Se adoptó la metodología CRISP-DM, que abarcó la comprensión del problema de la deserción, que afectaba al 45% de la población estudiantil, la identificación de variables pertinentes, la construcción de una matriz de datos con información de cuarenta y nueve estudiantes, la creación de modelos utilizando el software Weka y la evaluación del rendimiento del modelo. Se determinaron cinco variables que tenían un impacto en el abandono, como es: créditos aprobados, cursos aprobados, ciclo de ingreso, comunidad de origen y promedio. Seguidamente, se propusieron 3 modelos con precisiones clasificadas del 87.8%. Se llegó a la conclusión de que si un estudiante tenía menos de diez cursos aprobados o menos de 27 créditos aprobados, era probable que abandonara la universidad.

Finalmente en el ambio regional/local se plasmo a:

Chacaliaza (2021), "*Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA*" (tesis de pregrado). El propósito fue crear un modelo mediante el empleo de técnicas de DM para segmentar clientes en la empresa Suministros del Oriente SA. El estudio se llevó a cabo con un enfoque aplicado de tipo descriptivo y diseño descriptivo-correlacional. Se siguió la metodología KDD para el diseño del modelo. La conclusión obtenida fue que el modelo, fundamentado en técnicas de DM, logró segmentar los clientes de la empresa estudiada.

2.2. Fundamentos teóricos

2.2.1 Minería de datos

El titánico flujo de datos que caracteriza los nuevos entornos comerciales y las dinámicas de mercado exigen que las empresas dispongan de herramientas confiables para recopilar, organizar y analizar estos datos de manera clara y precisa, con el objetivo de dar eficiencia a la toma de decisiones. En consonancia con las ideas expuestas por Gutiérrez & Molina (2016), se entiende la minería de datos como una herramienta tecnológica y un enfoque de modelado matemático diseñado para simplificar la comprensión del contenido completo de una base de datos. Los datos, en su forma

inicial, constituyen la base, pero se convierten en información relevante cuando el usuario les atribuye un significado específico.

Según Perez & Santin (2007), la minería de datos implica el descubrimiento de nuevas relaciones y patrones significativos mediante el análisis de extensas colecciones de datos. Por otro lado, Carrasco (2011) afirma que este proceso consiste en extraer información pertinente de los datos, centrándose en el conocimiento genuino. Weiss & Indurkhya (1998), definen la minería de datos como la búsqueda de información valiosa en grandes conglomerados de datos, destacando su naturaleza colaborativa entre humanos y computadoras. En la actualidad, se ha comprobado la efectividad de una primera generación de algoritmos de minería de datos a través de diversas aplicaciones en el mundo real (Riquelme et al., 2006).

La minería de datos busca analizar datos desde diversas perspectivas estratégicas dentro de una organización, con el propósito de convertirlos en información valiosa y conocimiento, lo cual puede contribuir a aumentar la facturación, expandir el margen operativo, entre otros beneficios (Vallejo et al., 2018).

La minería de datos centra sus bases en la estadística, la administración de bases de datos y los modelos de IA, lo que la relaciona íntimamente con las ciencias de la computación. Su principal propósito radica en producir conocimiento en base a un conjuntos de datos (Martínez & Palencia, 2021).

Considerada como uno de los avances más innovadores, el DM se encarga de extraer patrones ocultos, relevantes y de naturaleza implícita en los datos, convirtiéndose en un método ágil para examinar grandes volúmenes de información (Ruiz & Romero, 2018).

La minería de datos se emplea en diversas disciplinas con el fin de descubrir patrones y modelos ocultos en las bases de datos. Aunque su aplicación es comúnmente observada en los campos de negocios y marketing, su uso y aplicación dependen en última instancia del conocimiento y habilidades de quienes la emplean, quienes deben traducir este conocimiento en información práctica para los niveles superiores de la organización (Salazar & Girón, 2021).

2.2.1.1 Proceso de extracción del conocimiento (KDD)

Según Usama et al. (2001) el Descubrimiento de Conocimiento en Bases de Datos (KDD, por sus siglas en inglés) constituye un conjunto de procesos fundamentales para la identificación de patrones novedosos, válidos y de gran utilidad mediante del análisis de datos.

El proceso KDD, abarca la búsqueda general de conocimiento valioso utilizando datos, mientras que la Minería de Datos (DM) representa una etapa particular dentro de este proceso. El DM implica la utilización de algoritmos especializados para identificar patrones significativos en los datos. La diferenciación entre el proceso de KDD y la fase de minería de datos dentro de este proceso es un punto central en este ámbito (Flores et al., 2019).

Por otro lado Romeu (2010), define el proceso KDD como la extracción y búsqueda de conocimiento basada en los datos, mientras que el DM se considera un componente del KDD, el cual emplea la inteligencia artificial para configurar un modelo.

El proceso KDD implica una serie de etapas bien definidas, cada una de las cuales es crucial para convertir los datos en conocimiento (Quiroz & Vlencia, 2012).

Actualmente, existe cierta confusión entre DM y KDD. Sin embargo, es importante destacar que el DM constituye solo una parte de este proceso, como se ilustra a continuación:

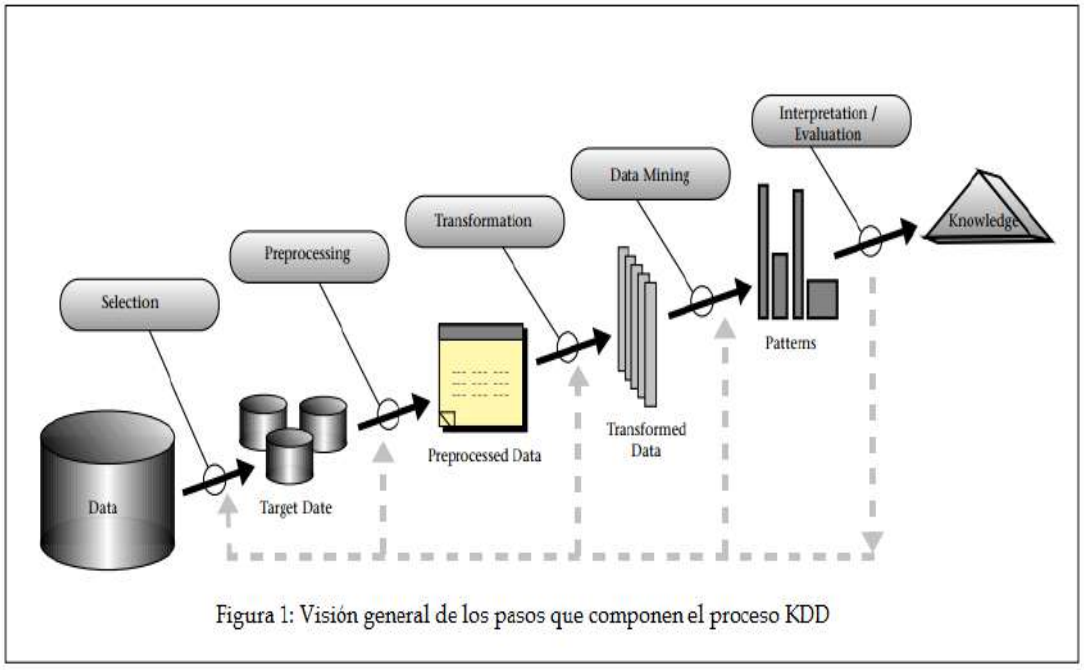


Figura 1
Comparación de los conceptos de minería de datos
Fuente: (Romeu, 2010)

El KDD tiene diversos campos de aplicación entre ello se puede usar para obtener segmentos de clientes morosos, riesgosos, etc., también para establecer relaciones entre variables, para conocer los perfiles de alumnos con buen rendimiento, deserción, bajo rendimiento tomando en consideración datos socioeconómicos y para encontrar patrones de compra de las personas (Timaran et al., 2016).

2.2.1.2 Fases del KDD

Según Timaran et al. (2016), el proceso KDD, representado en la figura adjunta, es interactivo e iterativo, implicando múltiples pasos donde el usuario participa activamente en la toma de numerosas decisiones.

Las etapas del proceso se resumen a continuación:

- ✓ Selección.
- ✓ Preprocesamiento/limpieza.
- ✓ Transformación/reducción.
- ✓ Minería de datos (data mining).
- ✓ Interpretación/evaluación

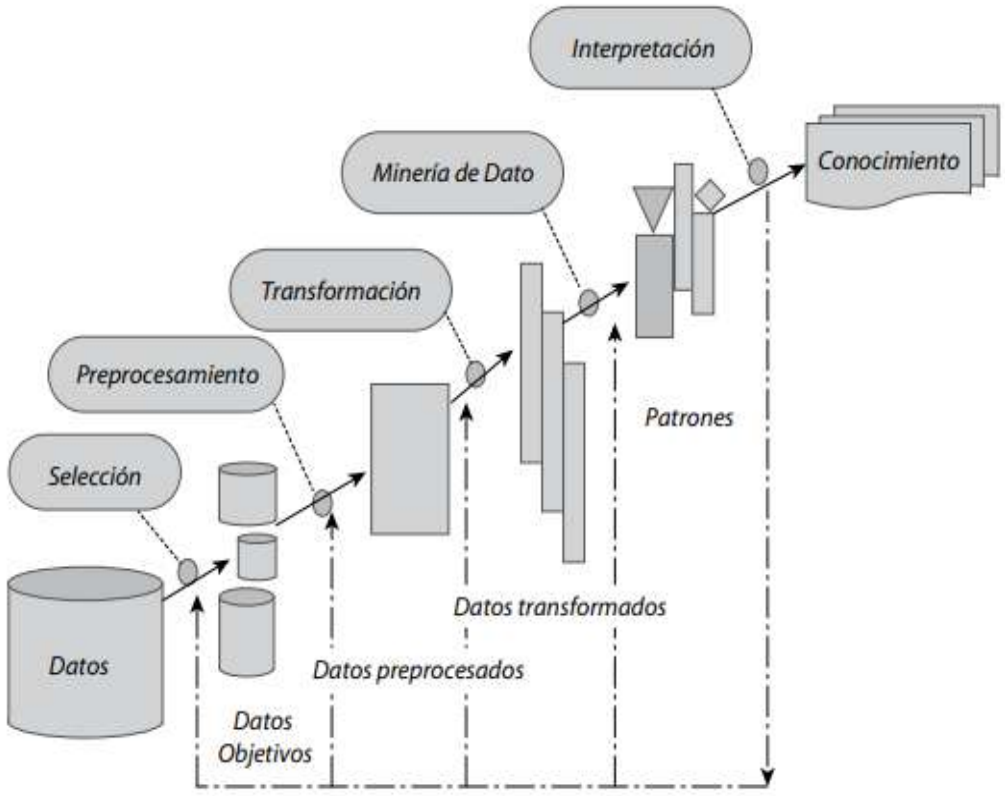


Figura 2
Proceso KDD
Fuente: (Timaran et al., 2016)

En la fase de selección, una vez que se ha determinado el conocimiento importante y se han establecido los objetivos del KDD desde un enfoque del usuario final, se procede a crear un conglomerado de datos objetivo. Este conglomerado puede consistir en el conjunto completo de datos o una muestra representativa de este, según los fines empresariales.

En la etapa de preprocesamiento/limpieza, se realiza una evaluación de la calidad de los datos y se llevan a cabo acciones básicas como la eliminación de datos inconsistentes. Además, se eligen estrategias para manejar datos nulos, desconocidos, duplicados, entre otros. La intervención con el analista o el usuario es esencial en esta fase. Por ejemplo, los datos inconsistentes son aquellos valores que se alejan notablemente del rango de valores esperados, lo cual puede ser resultado de errores humanos o modificaciones en el sistema. Los datos desconocidos son aquellos a los que no se les asigna ningún valor en el contexto del mundo real, mientras que los datos nulos son permitidos por los sistemas de gestión de bases de datos relacionales.

Durante la etapa de transformación/reducción de datos, se exploran atributos relevantes para representar los datos conforme al propósito del proceso. Se utilizan enfoques como la reducción de dimensiones para disminuir la cantidad efectiva de variables o identificar representaciones consistentes de los datos. Por ejemplo, la reducción horizontal implica la eliminación de filas idénticas o la discretización de valores continuos.

La fase de minería de datos tiene como objetivo buscar y descubrir patrones de interés, utilizando técnicas como clasificación, agrupamiento y patrones secuenciales. Estas técnicas generan modelos predictivos o descriptivos que pueden predecir valores futuros o explicar los datos existentes.

Al llegar a la etapa de interpretación/evaluación, se detallan los patrones descubiertos y se pueden llevar a cabo iteraciones adicionales. Se visualizan los patrones extraídos, se eliminan aquellos que no son relevantes o redundantes, y se explican los patrones útiles de manera que sean comprensibles para el usuario. Además, se consolida el conocimiento descubierto para integrarlo en otros sistemas o para su documentación y presentación a las partes interesadas.

2.2.1.3 Clasificación de las técnicas de minería de datos

De acuerdo con Joshi (1997), manifiesta, que la minería de datos se compone por los siguientes:

Clustering: Implica el análisis de datos para formar grupos de reglas que facilitan la clasificación de datos futuros.

Reglas de asociación: Consisten en conjuntos de reglas o condiciones que representan objetos extraídos de una BD.

Análisis de secuencias: Revela patrones que se dan en una secuencia específica.

Reconocimiento de patrones: Se encarga de relacionar una base de datos de entrada con otra fuente de información relevante. Se utiliza para conocer las causas de los problemas e identificar posibles salidas a estos inconvenientes.

Predicción: Se emplea para anticipar el futuro ya sea de una variable o diversas variables, utilizando información histórica y presente. Es importante destacar que las técnicas de predicción suelen ser altamente robustas.

Simulación: Compara el estado presente de una variable con su posible evolución en el futuro.

Optimización: Se ocupa de la reducción o aumento de una función que está influenciada por múltiples variables.

Clasificación: Facilita la atribución de un elemento a una categoría particular mediante la creación de perfiles característicos para cada clase, definidos mediante algoritmos o reglas que consideran múltiples variables. Además, se evalúa la importancia o influencia de estas variables en la clasificación. Este proceso simplifica la clasificación de un nuevo elemento una vez que se tienen los valores de las variables asociadas.

Mientras que para Cabena et al. (1998), la minería de datos implica cuatro operaciones principales respaldadas por técnicas habitualmente empleadas:

Modelización predictiva, que emplea técnicas como:

a) Clasificación

b) Predicción de valores

Segmentación de bases de datos, que se vale de técnicas como:

a) Clustering poblacional

b) Clustering mediante redes neuronales

Análisis de relaciones, que utiliza técnicas como:

a) Descubrimiento de asociaciones

b) Descubrimiento de secuencias de patrones

c) Descubrimiento de secuencias temporales similares

Detección de desviaciones, que incluye:

a) Técnicas estadísticas

b) Técnicas de visualización

Según Romeu (2010), las técnicas de aprendizaje pueden clasificarse de la siguiente manera:

Métodos inductivos:

Son aquellos que, a partir de los datos iniciales y el conocimiento generado, construyen modelos capaces de generar resultados basados en esos datos.

Técnicas predictivas:

Interpolación: Implica la creación de una función continua sobre múltiples dimensiones.

Predicción secuencial: Consiste en predecir el siguiente valor de una secuencia cuando las observaciones están ordenadas de manera secuencial.

Aprendizaje supervisado: En estas técnicas, cada observación está compuesta por varios valores de atributos, a los cuales se les asigna una clase correspondiente. Se genera un clasificador basado en las clases proporcionadas, siendo un caso particular de interpolación donde la función genera un valor discreto en lugar de continuo.

Técnicas descriptivas:

Aprendizaje no supervisado: Se refiere a un conjunto de observaciones que no tienen clases asociadas. Su objetivo es detectar regularidades en los datos, como agrupaciones de datos similares, contornos de grupos, asociaciones o valores anómalos.

Métodos abductivos:

Se centran en obtener los datos de origen a partir de los valores generados y las reglas. El objetivo es explicar la evidencia en relación con los sucesos ocurridos, similar a la labor de un investigador privado que infiere los hechos a partir de las consecuencias y ciertas reglas

2.2.1.4 Técnicas de minería de datos

Arboles de decisiones

La técnica de los árboles de decisión es ampliamente reconocida en el campo de la minería de datos debido a su capacidad para representar de manera precisa los problemas con un número limitado de clases. Además, estos modelos se caracterizan por ser proposicionales y altamente comprensibles (Hernandez et al., 2004).

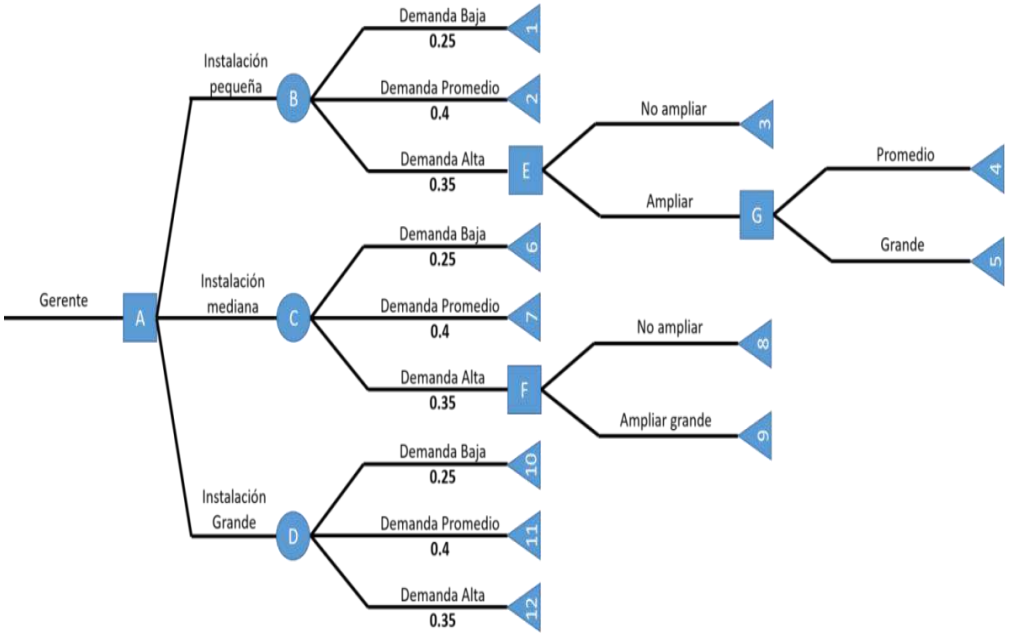


Figura 3
Árbol de decisión
Fuente: (Romeu, 2010)

En contraste, Mazo & Bedoya (2010), describen los árboles de decisión como estructuras donde cada nodo interno realiza una evaluación de uno o más atributos, y cada rama refleja el resultado de esa evaluación, mientras que las hojas indican las clases asociadas.

C4.5

Según Quinlan (1993), este algoritmo posibilita la creación de un árbol de decisión mediante la subdivisión recursiva de los datos. Durante la construcción del árbol, se sigue un método de búsqueda en profundidad que evalúa todas las pruebas posibles para dividir el conjunto de datos, eligiendo aquella que proporcione la mayor ganancia de información.

Para atributos que son discretos, se realiza una prueba que contempla un número n de resultados, donde n refleja la cantidad de valores que puede asumir dicho atributo. En el caso de atributos continuos, se lleva a cabo una prueba binaria para cada uno de los valores que el atributo pueda tener en los datos.

A diferencia de C4.5, C5.0 no es un software de código abierto, pero presenta una mejora sustancial en cuanto a su rendimiento.

Métodos Bayesianos

Hernandez et al. (2004), manifiesta que es un método muy utilizado en resolución de problemas de IA, mediante ello se logra el aprendizaje automático, el cual se considera

como una técnica practica para la inferencia de datos, el cual están basados en cálculos de probabilidad haciendo uso el teorema de bayes.

Los métodos bayesianos su base fundamental está en base a las distribuciones de probabilidad, el cual permite medir la incertidumbre de la información a modelar. Esta técnica brinda una metodología pragmática para la predicción e inferencia, asimismo ayuda en la toma de decisiones que incluyen magnitudes inciertas (Hernandez et al., 2004).

Redes neuronales artificiales

Prosiguiendo con Hernandez et al. (2004) menciona que estos algoritmos incluyen 02 clases de aprendizaje, siendo el primero el aprendizaje supervisado, el cual brinda una serie de datos como entrada y el resultado sirve para problemas de clasificación y regresión. Por otro lado, se tiene al aprendizaje no supervisado que solo se le brinda un conglomerado de información de entrada y el modelo debe tener auto aprendizaje para emitir una solución, esta función es relevante para actividades de agrupamiento

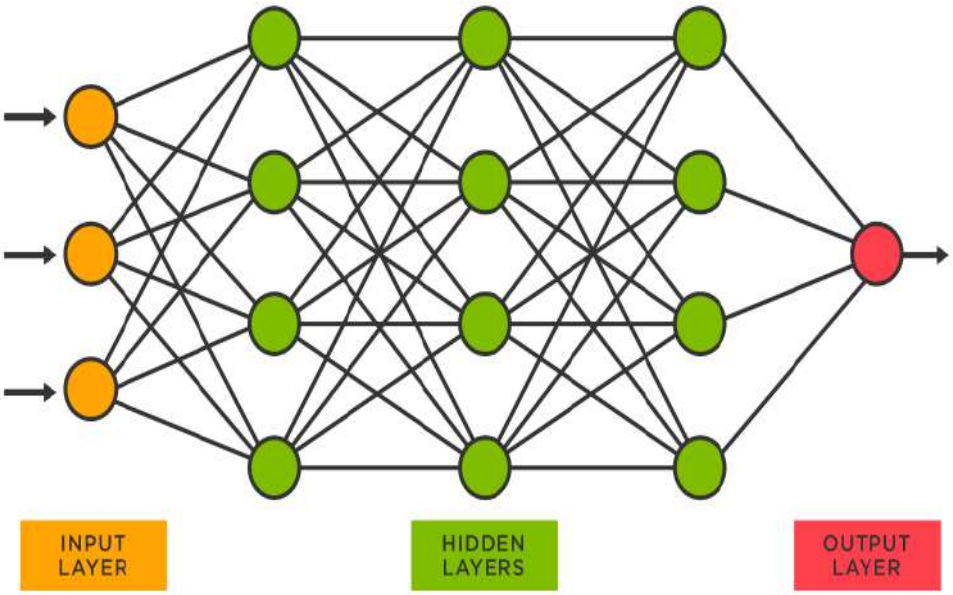


Figura 4
Neurona artificial
Fuente: (Romeu, 2010)

Según lo expresado, las redes neuronales artificiales presentan dos modalidades de aprendizaje:

a) Aprendizaje supervisado: Estos procedimientos algorítmicos demandan que cada conjunto de datos de entrada tenga su respectivo conjunto de datos de salida asociado. Durante el proceso de entrenamiento, se introduce un conjunto de datos de entrada en la red, se calcula su salida y se contrasta con la salida esperada. La disparidad

resultante se emplea para retroalimentar la red y adaptar los pesos según un algoritmo elaborado para reducir al mínimo el error (Basoqain, 2008).

b) Aprendizaje no supervisado: En tales sistemas, la red se enfrenta a un conjunto específico de datos de entrada y debe adquirir conocimiento por sí misma para ofrecer una respuesta. Esta forma de aprendizaje resulta especialmente beneficiosa para actividades de agrupamiento (Basoqain, 2008).

K – means

Este algoritmo, conocido como K-Medias o K-Means, es ampliamente utilizado en la agrupación de datos debido a su rapidez y eficiencia. Opera mediante un enfoque de agrupamiento por proximidad, donde se inicia con un conjunto predefinido de prototipos y un conjunto de ejemplos sin etiquetar que deben ser agrupados (Moody & Darken, 1989).

La meta fundamental del algoritmo K-Means es colocar los centros o prototipos en el espacio de tal manera que los datos que pertenecen al mismo grupo compartan características similares. Cuando se presenta un nuevo ejemplo, este se compara con los prototipos ya posicionados y se asigna al prototipo más cercano, utilizando típicamente la distancia euclidiana como medida. El objetivo es reducir al mínimo la varianza total dentro de cada grupo o la función de error cuadrático para lograr los resultados más óptimos.

Series de tiempo

Las series de tiempo son un tipo de conocimiento obtenido a través de la recopilación de datos o la observación de intervalos regulares de tiempo. Se utilizan para realizar predicciones asumiendo que no habrá cambios. Algunos conceptos clave en este ámbito son:

a) Tendencia: Representa el crecimiento o decrecimiento a largo plazo en un periodo extenso.

b) Estacionalidad: Se refiere a oscilaciones regulares a corto plazo, con frecuencia inferior al año.

c) ETS (Exponential smoothing state): Desde los años 50, los métodos de suavización exponencial han sido ampliamente conocidos y son los más comúnmente empleados en la industria. Recientemente, han experimentado mejoras significativas con la introducción de modelos de espacio de estado, técnicas de cálculo de probabilidades y métodos para la selección del modelo (Shi et al., 2020).

d) Holwinters: Es una variante del suavizado exponencial que elimina el sesgo de predicción de una serie de tendencia al incluir un componente de tendencia en la media móvil (Banda & Garza, 2014).

2.2.1.5 Metodología para la aplicación de minería de datos

a) SEMMA

Según Corrientes & Curuzú (2013), esta metodología se distingue principalmente por su estructura, la cual deriva su nombre de las etapas establecidas para los procesos de explotación de información. Estas etapas incluyen el muestreo (sample), la exploración (explore), la modificación (modify), el modelado (model) y la valoración (assess). Esta metodología fue desarrollada por SAS Institute Inc., una empresa líder en el desarrollo de software para inteligencia empresarial. SEMMA está diseñada para ser implementada con la herramienta de minería de datos "SAS Enterprise Miner".

b) CRISP – DM (Cross Industry Standard Process for Data Mining)

CRISP-DM proporciona un marco para la gestión de proyectos de Minería de Datos, organizando las tareas en una serie de fases definidas para alcanzar los objetivos del proyecto. Estas fases son iterativas y cíclicas, lo que permite retroceder a una fase anterior según sea necesario. La metodología se fundamenta en un modelo jerárquico de procesos que establece un ciclo de vida para los proyectos de explotación de información.

De acuerdo con Espinosa (2020) las fases de CRISP-DM son:

Entendimiento del negocio: En esta fase se busca comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial.

Entendimiento de los datos: Se refiere a la selección y adaptación de los datos para identificar posibles problemas de calidad y obtener conjuntos de datos apropiados para el análisis.

Preparación de los datos: Aquí se lleva a cabo la transformación de los datos seleccionados, incluyendo procesos como la limpieza, estructuración, integración y formateo.

Modelado y evaluación: En esta etapa se eligen y aplican técnicas de Minería de Datos, construyendo y evaluando modelos para su posterior interpretación.

Despliegue del proyecto: En esta última etapa se aprovechan los modelos desarrollados para integrarlos en los procesos de toma de decisiones de la organización, difundir el conocimiento extraído y otros aspectos relacionados.

2.2.2 Deserción estudiantil

Es complicado ofrecer una definición precisa de este concepto debido a la carencia de parámetros teóricos claros que lo delimiten. Se trata del fenómeno en el cual un joven deja de asistir o abandona la institución educativa en la que se ha matriculado para cursar el año escolar. Esta falta de precisión dificulta a los investigadores y a los encargados de formular políticas públicas tomar medidas efectivas para reducir su impacto (Rochin, 2021).

No obstante, algunos académicos como Spady (1970) se basan en la noción general de asociar la deserción universitaria con cualquier individuo que deje una institución educativa antes de obtener su diploma. Según Gómez (1998), este abandono se considera "un fracaso personal temprano cuyas repercusiones persisten a lo largo de la vida" (p. 54). Para Franklin & Kochan (2000) este término se refiere a cuando una persona, inscrita en algún programa el año anterior, decide no continuar sus estudios sin haberse transferido a otra institución.

Se explorarán los aspectos de la estructura familiar que influyen en la deserción escolar, como los problemas económicos y migratorios, ya que la familia desempeña un papel fundamental en el desarrollo de los niños..

Factores familiares

La familia, al igual que las células en el cuerpo humano, constituye la unidad fundamental de la sociedad. Es el vínculo primario entre el individuo y la sociedad, estableciendo conexiones esenciales. Más que la simple suma de sus partes, la familia se percibe como un sistema integral y dinámico, conformado por individuos de diversos sexos y etapas de desarrollo. Tanto a nivel físico como emocional, las familias experimentan cambios a lo largo del tiempo, los cuales influyen tanto en sus miembros de manera individual como en el grupo en su conjunto (Espinoza et al., 2012).

Factor económico

"Cada año, la deserción escolar experimenta un incremento, influenciado por la difícil situación económica de numerosos hogares, así como por la falta de apoyo y comprensión parental hacia sus hijos. Además, la inquietud de los adolescentes por llevar una vida acelerada también contribuye a este fenómeno" (Lozano & Maldonado, 2020). La falta de empleo entre los padres obstaculiza la continuidad educativa de los niños y motiva a muchos adolescentes a abandonar la escuela en un intento, muchas veces infructuoso, de incorporarse al mercado laboral.

Factor migratorio

La migración o la ausencia de los padres también pueden ser causas de repetición de año, abandono escolar, traumas emocionales, inseguridades y ruptura familiar. Los niños que crecen sin el afecto, cuidado y protección necesarios, así como sin un ambiente hogareño cálido y acogedor, pueden tener dificultades para desenvolverse adecuadamente, lo cual afecta tanto a ellos como a sus progenitores (Coronel, 2013).

Factores de Salud

"Un niño que enfrenta desventajas físicas o sufre de enfermedades, inevitablemente se ve impedido de participar de manera activa y espontánea en su proceso educativo junto con otros niños" (Venegas et al., 2017). La salud de los niños es crucial para su experiencia escolar, ya que influye directamente en su bienestar físico y mental, lo que a su vez afecta su participación y desarrollo en el entorno educativo, reduciendo el riesgo de abandono escolar.

2.3. Definición de términos básicos

- **Minería de datos:** Campo de las ciencias informáticas dedicadas a la búsqueda de conocimiento a partir de patrones en conjuntos extensos de datos (Troche, 2014).
- **KDD:** Knowledge Discovery in Databases (KDD) implica un proceso automatizado que fusiona la exploración y el análisis. En este proceso, se extraen patrones en forma de reglas o funciones de los datos, con el propósito de que sean analizados por el usuario (Timaran et al., 2016)
- **Base de datos:** Las bases de datos contienen extensos volúmenes de datos organizados en registros, lo que facilita la eficiencia en las operaciones de inserción, búsqueda, actualización y eliminación de información (Valverde et al., 2019).
- **Deserción:** La deserción estudiantil, que afecta adversamente el avance del país en diversas áreas sociales y científicas, constituye un problema significativo (Barrero, 2015).
- **CRISP-DM:** Es un estándar utilizado globalmente tanto en el ámbito industrial como académico para proyectos de minería de datos (Espinosa-Zúñiga, 2020).
- **Lenguaje R:** En los últimos tiempos, se ha apreciado un significativo avance y mejora en las herramientas de visualización dentro del lenguaje R, el cual fue diseñado específicamente para el análisis de datos (Chan & Galli, 2020).
- **SPSS:** Ampliamente empleado en el análisis cualitativo de datos, este método se extiende a prácticamente todas las disciplinas científicas, sobresaliendo por su eficacia, facilidad de uso y comprensión intuitiva (Rivadeneira et al., 2020).
- **Weka:** Se trata de un software que contiene un conjunto de técnicas de machine learning destinados a actividades de minería de datos. Incluye una variedad de

herramientas para la preparación de datos, así como para la regresión, clasificación, agrupamiento, extracción de reglas de asociación así como la visualización (Espinoza, 2018).

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1. Ámbito y condiciones de la investigación

3.1.1. Ubicación política

El área de investigación está situada en la localidad de Alto Perú, que forma parte del distrito de Soritor, ubicado en la provincia de Moyobamba, dentro del departamento de San Martín.

3.1.2. Ubicación geográfica

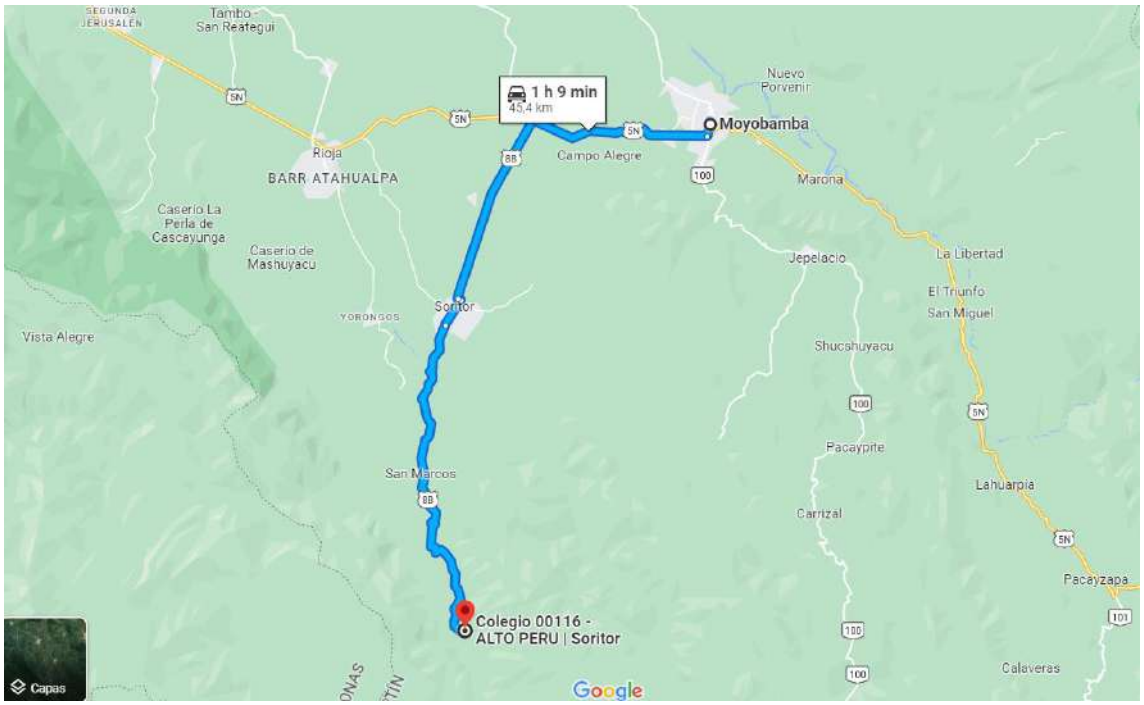


Figura 5
Distancia desde Moyobamba hasta la I.E 00116
Fuente: Google Maps

El lugar de estudio se ubica geográficamente en la latitud y longitud de -6.3005837170426355, -77.09522364533142, además se encuentra con altitud de 994 msnm y zona horaria: UTC-5. Específicamente se trata de la I.E N° 00116 – Alto Perú que es una institución rural perteneciente a la UGEL Moyobamba.

3.1.3. Periodo de ejecución

La investigación se desarrolló a lo largo de un lapso de ocho meses, desde mayo hasta diciembre de 2022.

3.1.4. Autorizaciones y permisos

No aplica

3.1.5. Control ambiental y protocolos de bioseguridad

No aplica

3.1.6. Aplicación de principios éticos internacionales

El estudio se llevó a cabo en la I.E N° 00116 Alto Perú, ubicada en Soritor, Moyobamba, San Martín. Se adhirió estrictamente a las pautas de la séptima edición del estilo APA y siguió las directrices establecidas por la Universidad Nacional de San Martín. Además, se guio por principios éticos de alcance internacional. Se garantizó el respeto a los participantes, quienes contribuyeron de manera voluntaria. Se evaluó cuidadosamente el principio de beneficencia, asegurando que los resultados fueran en beneficio tanto de la institución como de los individuos participantes. Se actuó de acuerdo con el principio de no maleficencia, evitando causar daño alguno a los participantes. Por último, se respetó la autonomía de los sujetos de investigación y se mantuvo la integridad de la información de acuerdo a su contexto.

3.2. Sistema de variables

3.2.1. Variables principales

Variable independiente (VI): Técnicas de minería de datos

Definición conceptual: De acuerdo con Pérez & Santin (2007), inicialmente, la minería de datos se describe como el proceso de descubrir relaciones y patrones significativos y nuevos al analizar grandes volúmenes de datos.

Definición operacional: La minería de datos o exploración de datos representa un campo interdisciplinario que combina la estadística y las ciencias de la computación. Las técnicas relacionadas con la minería de datos comprenden una diversidad de métodos que utilizan algoritmos para descubrir patrones en conjuntos extensos de datos.

Variable dependiente (VD): Deserción estudiantil

Definición conceptual: Para Gómez (1998), describe esta interrupción como un "fracaso personal temprano con repercusiones duraderas" (p. 54). Según Franklin & Kochan (2000) este fenómeno ocurre cuando un individuo, matriculado en un programa el año anterior, opta por no continuar con sus estudios sin haberse transferido a otra institución.

Definición operacional: La deserción escolar se define como la renuncia completa de los estudiantes a participar en el proceso de educación que ofrece una institución educativa.

Tabla 1
Operacionalización de variables

VARIABLES	DIMENSIONES	INDICADORES	ESCALA
Técnicas de minería de datos	Técnicas	N° Técnicas analizadas	Razón
	Factores	Sociales Académicos	Nominal / Ordinal
Deserción estudiantil	Predicción	Confiabilidad de la predicción	Razón
	Estimación	Tiempo para generar estimación.	

Fuente: Elaboración propia

3.2.2. Variables secundarias

No aplica

3.3. Procedimientos de la investigación

a) Tipo y nivel de la investigación

El estudio fue de naturaleza aplicada, ya que el investigador identificó y comprendió claramente el problema en cuestión, utilizando la investigación para abordar preguntas específicas (Hernández et al., 2014). Se centró en abordar el desafío de la deserción estudiantil que enfrenta la Institución Educativa 0016 Alto Perú - Moyobamba, con el objetivo de predecir este fenómeno mediante el uso de técnicas de minería de datos.

Descriptivo: implicó la delineación de un evento o fenómeno para comprender su estructura o modo de funcionamiento (Tamayo y Tamayo, 2003).

Mediante la incorporación de técnicas de DM, se logró analizar y comprender el comportamiento de la deserción estudiantil, además de anticiparla basándose en los patrones identificados por los algoritmos utilizados.

b) Población y muestra

La población se define como un conjunto específico de casos que está claramente definido, accesible y delimitado, que sirve como fundamento para la selección de la muestra. Esta muestra debe cumplir con criterios predefinidos con anterioridad (Arias et al., 2016).

El elemento de estudio determinado como población es el elemento de registro de 412 alumnos que cursaron los 5 completos entre los períodos 2012 – 2022 en el nivel secundario, de la I.E. 0016 Alto Perú – Moyobamba.

La representatividad de una muestra facilita la extrapolación y, por lo tanto, la generalización de los resultados obtenidos en esa muestra a la población accesible y, a su vez, a la población objetivo (Otzen & Manterola, 2017).

En ese sentido la muestra fue compuesta por 387 alumnos del nivel secundaria de la Institución Educativa 0016 Alto Perú – Moyobamba. El cual se empleó el muestreo no probabilistico por conveniencia para determinar dicha muestra, considerandose a criterios de inclusión como estudiantes con información completa y que contaron con 5 años de estudios, y por otro lado excluyendose a alumnos que no contaron con los 5 años completos de estudio, estudiantes con registros incompletos, estudiantes trasladados y fallecidos.

Inherente a la unidad de análisis, correspondió a los datos de cada estudiante de la institución especificada.

c) Diseño de la investigación

Diseño no experimental: Se ejecutó sin manipular alguna variable de estudio para observar algún cambio en la otra. Solo se analizó toda la base de datos de información inherente a los alumnos que la institución educativa facilitó para encontrar patrones mediante técnicas de minerías de datos el cual nos permitió verificar la predicción de la deserción estudiantil.

El diseño trabajado tiene el siguiente esquema:

Donde:

X: Técnicas de minería de datos

Y: Predicción de la deserción estudiantil

d) Técnicas e instrumentos

Respecto a las técnicas:

Análisis documental: Se refiere a la extracción de información de una variedad de fuentes como libros, documentos académicos y artículos, los cuales contienen teorías, técnicas y métodos para abordar diversos problemas.

Observación: Se trata de la captura visual de eventos que ocurren en un entorno real, en el que los acontecimientos se clasifican según un esquema específico, adaptado al problema en cuestión. En esta instancia, la observación se empleó para detectar de forma precisa todos los resultados fiables de las predicciones.

Inherente a los instrumentos:

Ficha documental: Esquema en digital para acumular todo el conglomerado de documentos, registros, reportes con un orden y coherencia.

Cuadro resumen de predicciones: Se elaboró una tabla de síntesis de predicciones que permitió recopilar los resultados predictivos obtenidos mediante las técnicas de DM empleadas en el estudio.

e) Análisis estadístico

Se emplearon las siguientes herramientas para el procesamiento y análisis de los datos recopilados durante la investigación:

- IBM SPSS Statistics 26, un software de análisis estadístico.
- Weka 3.8.6, es una plataforma de software diseñada específicamente para el aprendizaje automático.

El análisis estadístico de los datos se realizó de la siguiente forma:

Se utilizaron tablas para evaluar las técnicas predictivas en la minería de datos.

Se emplearon gráficos estadísticos para evaluar las técnicas predictivas de DM.

3.3.1 Objetivo específico 1: Analizar los archivos que registran la deserción estudiantil

A1: Se accedió al sistema SIAGIE para recopilar las actas de evaluación y las nóminas de matrícula de los estudiantes de los periodos comprendidos del 2012 al 2022.

A2: Se descargaron todos los reportes en pdf con la información requerida.

A3: Se realizó la conversión de archivos a formato Xlsx, para mejor manipulación de los datos.

A4: Se realizó un consolidado de todos los archivos en un solo dataset.

A5: Se trasladó la información a una base de datos construida en MySQL/MariaDB.

3.3.2 Objetivo específico 2: Determinar variables influyentes en la deserción estudiantil

A6: Se realizaron consultas en SQL para contar con el consolidado limpio y consistente

A7: Se empleó el software SPSS para determinar los factores influyentes dentro de la deserción estudiantil, mediante las pruebas estadística de relación (Prueba Chi Cuadrada de Pearson, etc.).

A8: Se realizó la interpretación de los datos obtenidos

3.3.3 Objetivo específico 3: Determinar las técnicas de minería de datos a aplicar que mejoren la predicción de la deserción estudiantil

A9: Se consolidó el dataset en formato arff., formato que soporta el software Weka 3.8.6.

A10: Se procedió a cargar en el software de minado para luego explorar sus datos y validar su consistencia.

A11: Se hizo uso de algoritmos de ML para la clasificación, utilizando diversos algoritmos (bayesianos, lineales, redes neuronales y arboles de decisión).

3.3.4 Objetivo específico 4: Analizar los resultados obtenidos de la aplicación de técnicas de minería de datos en la predicción de la deserción estudiantil.

A12: Se realizó la comparación de los algoritmos y su comportamiento en la clasificación con datos training y test.

A13: Se compararon los resultados haciendo uso una hoja de cálculo verificando los datos predichos con los datos reales.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

El estudio se llevó a cabo en la Institución Educativa 00116, una institución pública que abarca tres niveles educativos: Inicial, Primaria y Secundaria, dentro del marco de la Educación Básica Regular (EBR). Esta institución se localiza en el CCPP Alto Perú, ubicado en el distrito de Soritor, en la provincia de Moyobamba, región San Martín. Es importante destacar que la investigación se centró exclusivamente en los datos correspondientes a los alumnos del nivel secundario durante el período comprendido entre los años 2012 y 2022.

4.1. Metodología CRISP-DM (Cross Industry Standard Process for Data Mining)

Para facilitar la comprensión del estudio y cumplir con los objetivos de la investigación, se utilizó la metodología CRISP-DM para el procesamiento, análisis e interpretación de los datos proporcionados por la Institución Educativa. Esta metodología, descrita en el marco teórico, abarca de las siguientes etapas:

A. FASE 1: COMPRENSIÓN DEL NEGOCIO

Se realizaron todas las tareas establecidas en esta fase del proceso de minería de datos, que tiene como finalidad reconocer y definir los objetivos y requisitos del proyecto desde una perspectiva empresarial. Luego, estos objetivos se convertirán en metas técnicas y en un plan detallado del proyecto.

A.1. Comprensión del contexto y determinación de objetivos

El fin del DM en este proyecto es generar pronósticos confiables utilizando la información disponible sobre los estudiantes de una institución educativa de nivel secundario.

Además en esta actividad tiene como fin enfocar la comprensión del contexto y de los objetivos del proyecto de minería de datos, para posteriormente transformar este conocimiento de datos en un problema, plasmándose un plan como base para la concretización de los objetivos, en ese contexto se describe lo siguiente:

- ✓ **Contexto de la Institución Educativa:** Las instituciones educativas tienen como objetivo primordial proporcionar una educación integral que abarque aspectos de formación moral, física, emocional e intelectual, reflejados comúnmente en las calificaciones obtenidas por los estudiantes en diversas áreas o asignaturas. Además, deben cumplir con los compromisos de gestión escolar (CGE) establecidos por el MINEDU, entre los cuales se encuentra la retención anual e interanual de estudiantes en la institución educativa. Este compromiso se refiere a la capacidad del sistema educativo para garantizar que los estudiantes

permanezcan en las aulas y completen los ciclos y niveles educativos en los tiempos previstos, adquiriendo las competencias y conocimientos correspondientes. Sin embargo, diversos factores, como los académicos, sociales y económicos, pueden influir en la deserción estudiantil. En este sentido, las instituciones educativas carecen de herramientas de apoyo que les permitan detectar tempranamente posibles estudiantes desertores y no existen muchos estudios exhaustivos sobre el comportamiento de los alumnos que proporcionen patrones o conclusiones para predecir el comportamiento de los futuros estudiantes.

- ✓ **Contexto del proyecto:** Los objetivos del proyecto está relacionado a la predicción de datos para que los estudiantes de otro ingreso se puedan estimar de forma fiable a partir de los datos que ya se cuenta de alumnos. Con la información establecida se podrían hacer diversos tipos de predicciones según la necesidad de la institución, pero en este estudio se han definido como objetivo: predecir la deserción estudiantil a partir de datos de un conglomerado de estudiantes de la Institución Educativa 00116, para que el personal directivo pueda tomar acciones y con ello mejorar la retención escolar, mejorando el compromiso de gestión escolar en la institución educativa.

Para desarrollar un modelo predictivo de la deserción estudiantil, se dispuso de una base de datos que abarcaba información académica, económica y social de los alumnos desde 2012 hasta 2022. Esta base de datos fue facilitada por el director de la institución educativa y se extrajo íntegramente de las nóminas de matrícula y de las actas de evaluación.

A.2. Evaluación de la situación

Se dispone de una BD compuesta por archivos obtenidos del Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE). Esta BD contiene información detallada de los estudiantes que han cursado en la institución educativa desde 2012 hasta 2022. Por lo tanto, se dispone de la cantidad suficiente de información para abordar la problemática. La información abarca datos académicos, como las calificaciones de cada asignatura y la cantidad de asignaturas aprobadas y desaprobadas. Además, incluye datos económicos y sociales que serán relevantes para el proceso de minería de datos.

Asimismo, en las bases teóricas se establecieron diversas técnicas de DM, donde se tiene a las técnicas de clasificación y de regresión, que son precisamente técnicas implementadas para la predicción, que en base de algoritmos se podrá determinar la deserción estudiantil, cabe indicar que la estadística descriptiva será de gran apoyo para conocer los resultados en su fase inicial. Además que con la metodología CRISP-DM,

se tendrá mejor comprensibilidad y flexibilidad para entender a mayor detalle los resultados, ya que permite operar de forma ordenada.

- ✓ **Inventario de recursos:** para el proceso de minado se cuenta con los siguientes recursos:

  Se tuvo a disposición el software Weka en su versión 3.8.6, desarrollado en la Universidad de Waikato de Nueva Zelanda, diseñado para el aprendizaje automático y la minería de datos. Weka es un programa de código abierto distribuido bajo la licencia GNU-GPL y proporciona una amplia gama de herramientas de visualización y algoritmos para el análisis de datos y la construcción de modelos predictivos. Además, cuenta con una interfaz gráfica de usuario que facilita el acceso a sus funciones.

  En cuanto a los recursos de hardware, se cuenta con una computadora portátil con las siguientes especificaciones:

  Marca: Asus

  Modelo: X515EA

  Procesador: Intel(R) Core (TM) i5-1135G7 @ 2.40GHz (8 CPUs), ~2.4GHz

  Memoria RAM: 8,00 GB

  Almacenamiento: 480,00 GB

  Tarjeta Gráfica: Intel(R) Iris(R) Xe Graphics

  Sistema Operativo: Windows 11 Home Single Language 64-bit

  La fuente de datos es una base de datos con información de los estudiantes en la institución educativa desde los años 2012 al 2022.

- ✓ **Costes y beneficios:** Los datos del proyecto no cuentan con costes adicionales a la institución educativa, ya que es información propia de la institución, desde que el estudiante se matricula en ella.

  Respecto al beneficio, no puede medirse en términos económicos para la institución educativa de manera directa, pero se puede suponer que es de forma indirecta ya que como objetivo del proyecto que viene ser la predicción de la deserción estudiantil, entonces se va ver reflejado en el mejoramiento de la retención escolar y con ello mejores condiciones para los estudiantes, traduciéndose en prestigio para la institución.

A.3. Determinación de los objetivos de minería de datos aplicados al proyecto

Los objetivos de la minería de datos en la aplicación paren el aprendizaje automático en el proyecto de este estudio, se detallan las siguientes actividades:

- Ejecutar la limpieza de los datos que fueron proporcionados por la institución educativa.
- Conocer los factores relevantes que influyen en la deserción estudiantil en los alumnos.
- Determinar y analizar la mejor técnica de aprendizaje automático que realice la predicción de la deserción estudiantil.
- Realizar la predicción de la deserción estudiantil con la técnica más optima de predicción.

La predicción se fundamenta en dos pasos principales: en primer lugar, se lleva a cabo un análisis estadístico de los datos para identificar los factores que tienen un impacto en la deserción estudiantil. Luego, se emplea el software WEKA para evaluar el rendimiento de diversas técnicas y así seleccionar el predictor más adecuado para realizar la predicción de manera óptima.

A.4. Desarrollo del plan del proyecto

El proyecto se desglosará en las siguientes tareas, con el fin de estructurar su planificación y estimar su duración. Estas actividades se han definido siguiendo los lineamientos de la metodología CRISP-DM para garantizar una gestión eficiente del proyecto.

Tabla 2

Plan de ejecución del proyecto mediante metodología CRISP-DM

Fase	Duración	Herramientas	Riesgo
Fase 1: Comprensión del negocio	20 días	Entrevista con el director y docentes de la I.E.	Ninguno
Fase 2: Comprensión de los datos	20 días	SIAGIE Foxit PhantomPDF Nitro PDF Microsoft Excel MySQL/MariaDB	Inconsistencia de datos
Fase 3: Preparación de los datos	20 días	Microsoft Excel Weka 3.8.6	Datos incompletos
Fase 4: Modelado	30 días	Weka 3.8.6	Modelos deficientes
Fase 5: Evaluación	20 días	Weka 3.8.6	Predicción deficiente
Fase 6: Despliegue	20 días	-	Ninguno

B. FASE 2: COMPRENSIÓN DE LOS DATOS

En la etapa 2 de la metodología CRISP-DM, denominada "comprensión de los datos", se da inicio con la recolección inicial de los datos, marcando el primer acercamiento al problema. Este proceso facilita la familiarización con los datos y la evaluación de su calidad, además de la identificación de subconjuntos preliminares relevantes. Asimismo, permite la identificación de relaciones destacadas que pueden ser fundamentales para la formulación de hipótesis iniciales.

En la labor del estudio, se puede reflejar que la gran parte de autores guardan coincidencia que existen 2 factores que afecta a los alumnos, causando la deserción estudiantil, los cuales son:

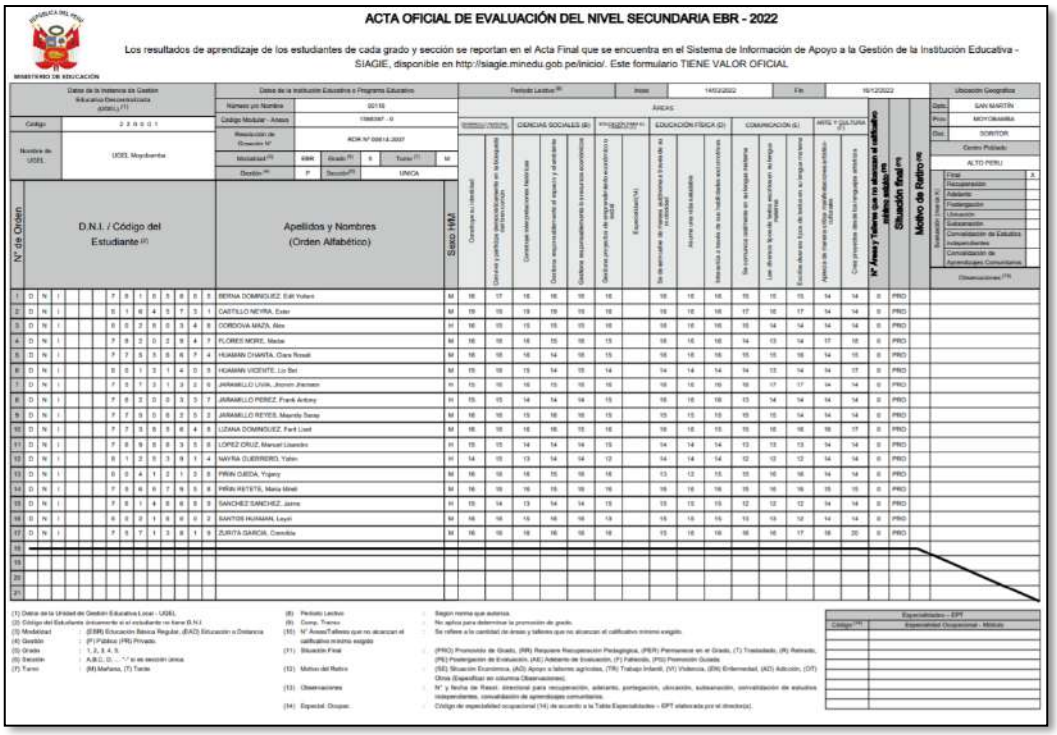
- Factores académicos
- Factores socioeconómicos

B.1. Recolección de datos iniciales

Los datos se encuentran en archivos pdf, el cual fueron extraídos del sistema SIAGIE tomando en consideración los grados, las secciones y los años correspondiente al 2012 al 2022, del nivel secundario.

Para el acta de evaluación → Módulo evaluación / Acta consolidad de evaluación / Generación y envío de acta

Figura 6
Generación de acta de evaluación - SIAGIE



ACTA OFICIAL DE EVALUACIÓN DEL NIVEL SECUNDARIA EBR - 2022

Los resultados de aprendizaje de los estudiantes de cada grado y sección se reportan en el Acta Final que se encuentra en el Sistema de Información de Apoyo a la Gestión de la Institución Educativa - SIAGIE, disponible en http://siagie.minedu.gob.pe/inicio/. Este formulario TIENE VALOR OFICIAL

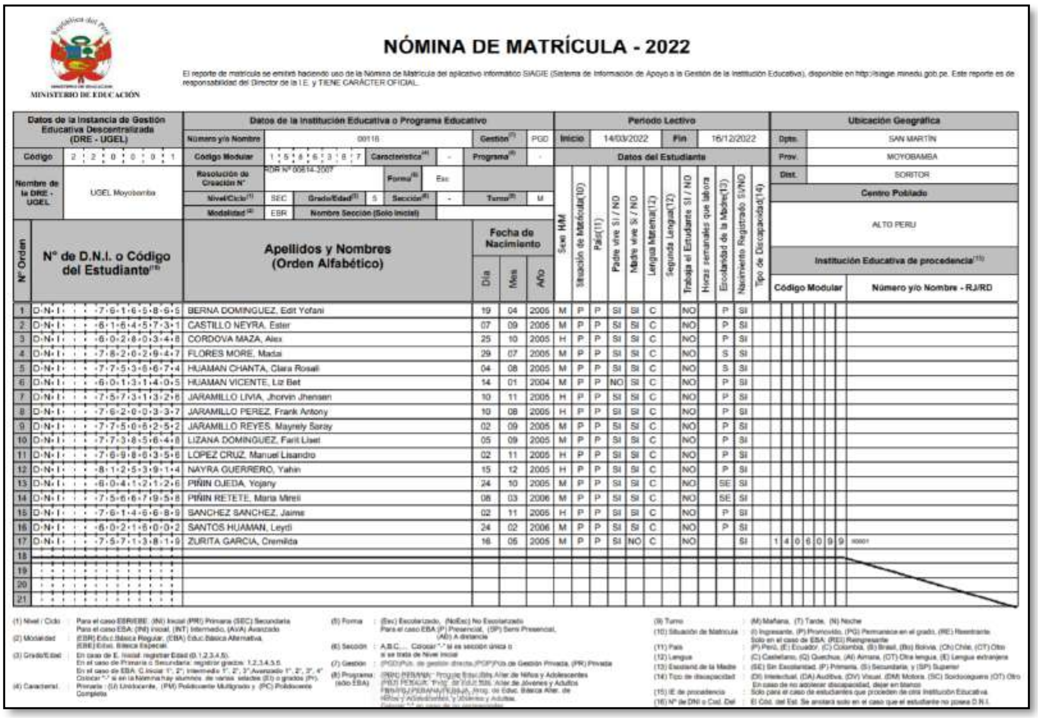
Figura 7

Reporte de acta de evaluación - SIAGIE

Para la nómina de matrícula → Módulo matrícula / Nómina / Nómina de matrícula.

Figura 8

Generación de nómina de matrícula – SIAGIE



NÓMINA DE MATRÍCULA - 2022

N° Orden	N° de D.N.I. o Código del Estudiante	Apellidos y Nombres (Orden Alfabético)	Día	Mes	Año	Sexo H/M	Situación de Matrícula	País	Padre vive SI / NO	Madre vive SI / NO	Lengua Materna	Trabaja el Estudiante SI / NO	Escolaridad de la Madre	Nacimiento Registrado SI/NO
1	DNI 76165865	BERNA DOMINGUEZ, Edit Yofani	19	04	2005	M	P	P	SI	SI	C	NO	P	SI
2	DNI 61645731	CASTILLO NEYRA, Ester	07	09	2005	M	P	P	SI	SI	C	NO	P	SI
3	DNI 60280348	CORDOVA MAZA, Alex	25	10	2005	H	P	P	SI	SI	C	NO	P	SI
4	DNI 78202947	FLORES MORE, Madai	29	07	2005	M	P	P	SI	SI	C	NO	S	SI
5	DNI 77536674	HUAMAN CHANTA, Clara Rosali	04	08	2005	M	P	P	SI	SI	C	NO	S	SI
6	DNI 60131405	HUAMAN VICENTE, Liz Bet	14	01	2004	M	P	P	NO	SI	C	NO	P	SI
7	DNI 75731326	JARAMILLO LIVIA, Jhorvin Jhensen	10	11	2005	H	P	P	SI	SI	C	NO	P	SI
8	DNI 76200337	JARAMILLO PEREZ, Frank Antony	10	08	2005	H	P	P	SI	SI	C	NO	P	SI
9	DNI 77506252	JARAMILLO REYES, Mayrely Saray	02	09	2005	M	P	P	SI	SI	C	NO	P	SI
10	DNI 77385648	LIZANA DOMINGUEZ, Farit Liset	05	09	2005	M	P	P	SI	SI	C	NO	P	SI
11	DNI 76986356	LOPEZ CRUZ, Manuel Lisandro	02	11	2005	H	P	P	SI	SI	C	NO	P	SI
12	DNI 81253914	NAYRA GUERRERO, Yahin	15	12	2005	H	P	P	SI	SI	C	NO	P	SI
13	DNI 60412126	PIÑIN OJEDA, Yojany	24	10	2005	M	P	P	SI	SI	C	NO	SE	SI
14	DNI 75667958	PIÑIN RETETE, Maria Mireli	08	03	2006	M	P	P	SI	SI	C	NO	SE	SI
15	DNI 76146689	SANCHEZ SANCHEZ, Jaime	02	11	2005	H	P	P	SI	SI	C	NO	P	SI
16	DNI 60216002	SANTOS HUAMAN, Leydi	24	02	2006	M	P	P	SI	SI	C	NO	P	SI
17	DNI 75713819	ZURITA GARCIA, Cremilda	16	05	2005	M	P	P	SI	NO	C	NO		SI

Figura 9.

Reporte de nómina de matrícula - SIAGIE

Los datos procesados en bruto respecto a datos iniciales, están compuesto con la siguiente información:

Tabla 3

Recolección inicial de datos

Fuente de datos	Descripción
	Datos UGEL
	Modalidad
	Código alumno
	Apellidos y nombres
	Sexo
	Gestión
	Grado y sección
	Turno
Acta de evaluación	Periodo lectivo
	Notas de las áreas
	N° Áreas/Talleres desaprobados
	Situación final
	Motivo de retiro
	Observaciones
	Especialidad ocupacional – Módulo
	Nombres de los docentes de las áreas
	Fecha aprobación de acta

Nómina de matrícula	Característica
	Forma
	Programa
	Situación de matrícula
	País
	Padre vive
	Madre vive
	Lengua materna
	Segunda lengua
	Trabaja estudiante
	Horas semanales que labora
	Escolaridad de la madre
	Nacimiento registrado
	Tipo de discapacidad
	Institución educativa de procedencia
	Resumen
	Fecha aprobación de la nómina

Fuente: SIAGIE - 2022

La predicción estará en relación a la:

- **Condición** (Abandona y Permanece).

B.2. Descripción de los datos

Tabla 4

Descripción de los datos – Acta de evaluación

Dato	Descripción del dato
Datos UGEL	Datos de la Unidad de Gestión Educativa Local – Moyobamba.
Modalidad	(EBR) Educación Básica Regular, (EAD) Educación a Distancia.
Código alumno	Es el identificador de cada estudiante que puede ser su DNI o un Código (Ejemplo: 47193879, 02056461700070).
Apellidos y nombres	Campo que contiene los apellidos y nombres de los estudiantes (Ejemplo: Torres Villegas Sara Inés).
Sexo	Expresado en Hombre (H) y Mujer (M).
Gestión	Expresado en (P) Público (PR) Privado.
Grado y sección	Son diversos niveles que atraviesan los estudiantes, donde grado se expresa en (1°, 2°, 3°, 4° y 5°) y sección expresada en (A, B, C, D, etc. o Sección Única si solo existe una sección).
Turno	Expresado en (M) Mañana, (T) Tarde.
Periodo lectivo	Va depender según la norma que autoriza, es un rango de fecha (Ejemplo: Inicio: 14/03/2022 y Fin: 16/12/2022).

Notas de las áreas	Son los calificativos medidos bajo la escala vigesimal (0 hasta 20), el cual significa el rendimiento académico de los estudiantes por área (Ejemplo: 08, 12, 20, 19, etc.).
N° Áreas / Talleres desaprobados	Se refiere a la cantidad de áreas y talleres que no alcanzan el calificativo mínimo exigido, y son datos cuantitativos.
Situación final	Categorizados en (PRO) Promovido de Grado, (RR) Requiere Recuperación Pedagógica, (PER) Permanece en el Grado, (T) Trasladado, (R) Retirado, (PE) Postergación de Evaluación, (AE) Adelanto de Evaluación, (F) Fallecido, (PG) Promoción Guiada.
Motivo de retiro	Considerado como la causante de la deserción en los estudiantes, el cual esta expresado en (SE) Situación Económica, (AG) Apoyo a labores agrícolas, (TR) Trabajo Infantil, (VI) Violencia, (EN) Enfermedad, (AD) Adicción, (OT) Otros (Especificar en columna Observaciones).
Observaciones	Se considera información inherente a N° y fecha de Resol. directoral para recuperación, adelanto, postergación, ubicación, subsanación, convalidación de estudios independientes, convalidación de aprendizajes comunitarios.
Especialidad ocupacional – Módulo	Este campo esta referente al código de especialidad ocupacional de acuerdo a la Tabla Especialidades – EPT elaborada por el director(a).
Nombres de los docentes de las áreas	Se consigna los apellidos y nombres de los docentes responsables de cada área de estudio.
Fecha aprobación de acta	Es la fecha en la que se aprueba el acta de evaluación, el cual lo aprueba el director de la I.E.

Fuente: SIAGIE - 2022

Tabla 5

Descripción de los datos – Nómina de matrícula

Dato	Descripción del dato
Característica	Brinda información de la característica del I.E en el nivel Primaria: (U) Unidocente, (PM) Poli docente Multigrado y (PC) Poli docente Completo.
Fecha de nacimiento	Este atributo nos permite obtener la edad de los estudiantes de la I.E (Ejemplo: 01/05/2008).
Forma	Esta en categorizados en (Esc) Escolarizado, (NoEsc) No Escolarizado; para el caso EBA: (P) Presencial, (SP) Semi Presencial, (AD) A distancia
Programa	Solo aplica para EBA.

Situación de matrícula	Representado en categorías de: (I) Ingresante, (P) Promovido, (PG) Permanece en el grado, (RE) Reentrante. Solo en el caso de EBA: (REI) Reingresante
País	Considerado la nacionalidad del estudiante, que puede ser (P) Perú, (E) Ecuador, (C) Colombia, (B) Brasil, (Bo) Bolivia, (Ch) Chile, (Ot) Otro.
Padre vive	Representado en Si / No
Madre vive	Representado en Si / No
Lengua materna	Están expresados en (C) Castellano, (Q) Quechua, (AI) Aimara, (OT) Otra lengua, (E) Lengua extranjera
Segunda lengua	Campo que indica si el alumno es bilingüe y puede estar expresado en (C) Castellano, (Q) Quechua, (AI) Aimara, (OT) Otra lengua, (E) Lengua extranjera.
Trabaja estudiante	Representado en Si / No
Horas semanales que labora	Expresada en datos numéricos y corresponde a la cantidad de horas semanales que los alumnos invierten en labores extra académicos.
Escolaridad de la madre	Es el nivel alcanzado de la madre del estudiante y están expresado en: (SE) Sin Escolaridad, (P) Primaria, (S) Secundaria, y (SP) Superior.
Nacimiento registrado	Representado en Si / No
Tipo de discapacidad	Considerada la deficiencia de los estudiantes, el cual está representado por (DI) Intelectual, (DA) Auditiva, (DV) Visual, (DM) Motora, (SC) Sordoceguera (OT) Otro.
Institución educativa de procedencia	Se consigna el nombre o numero de la I.E de los estudiantes que proceden de otro I.E.
Resumen	Información que contiene la cantidad de alumnos matriculados, distribuidos por sexo.
Fecha aprobación de la nómina	Es la fecha en la que se aprueba la nómina, el cual lo aprueba el director de la I.E.

Fuente: SIAGIE - 2022

B.3. Exploración de los datos

Después de la descripción inicial de los datos, se avanza hacia la etapa de exploración, la cual consiste en la aplicación de pruebas estadísticas fundamentales para revelar las propiedades inherentes de los datos. Se elaboran tablas de frecuencia y gráficos de distribución que permiten evaluar la consistencia y la integridad de los datos. Esta fase es crucial para determinar la fiabilidad y la exhaustividad de la información recopilada.

En una data inicial se contó con 387 estudiantes comprendidos desde los años 2012 y 2022, de los cuales se hizo las siguientes distribuciones:

Tabla 6

Sexo de los estudiantes

Sexo	Frecuencia	%
Hombre	197	50.90%
Mujer	190	49.10%
Total	387	100.00%

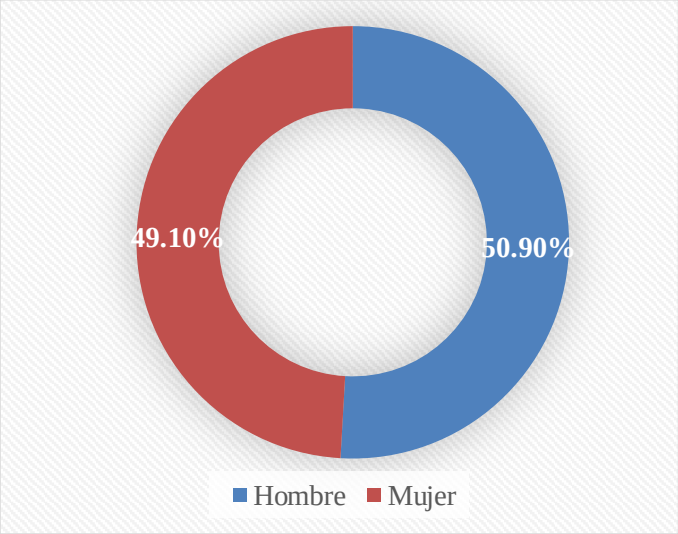


Figura 10

Sexo de los estudiantes

En la tabla 6 y figura 10 se aprecia que del total de estudiantes 197 (50.90%) son del sexo Hombre y 190 (49.10%) corresponden al sexo Mujer.

Tabla 7

Situación de matrícula de los estudiantes

Situación de matrícula	Frecuencia	%
Ingresante	286	73.90%
Repitente	10	2.58%
Promovido	89	23.00%
Reentrante	2	0.52%
Total	387	100.00%

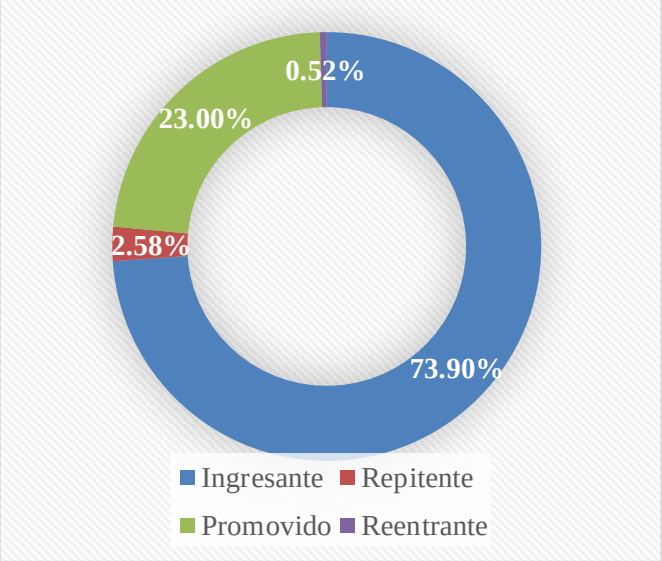


Figura 11
Situación de matrícula de los estudiantes

En la tabla 7 y figura 11, se aprecia que la gran parte pertenecen a Ingresantes con total de 286 (73.90%) estudiantes, 10 (2.58%) corresponden a Repitentes, 89 (23.00%) tienen situación de matrícula Promovido y solamente 2 (0.52%) son Reentrantes; teniendo como datos predominantes a los que tienen situación de Ingresantes.

Tabla 8
Nacionalidad de los estudiantes

País	Frecuencia	%
Perú	387	100.00%
Otros	0	0.00%
Total	387	100.00%

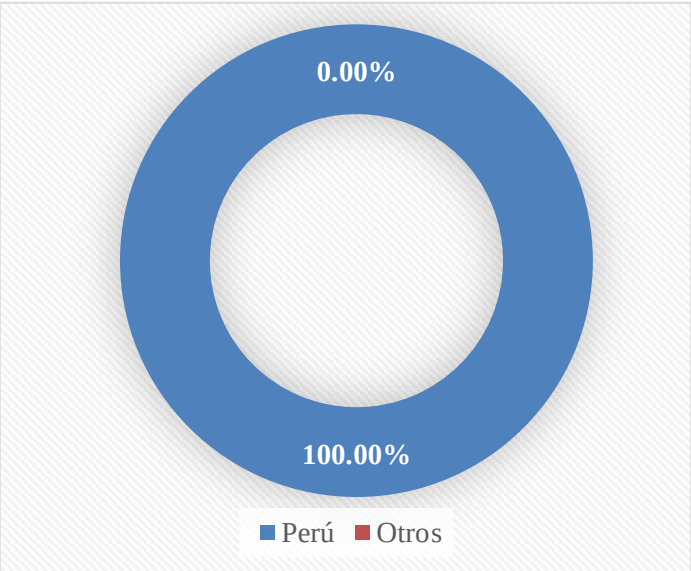


Figura 12
Nacionalidad de los estudiantes

En la tabla 8 y figura 12, se aprecia que el 100% corresponden a estudiantes de nacionalidad peruana.

Tabla 9

Padre vive de los estudiantes

Padre vive	Frecuencia	%
SI	379	97.93%
NO	8	2.07%
Total	387	100.00%

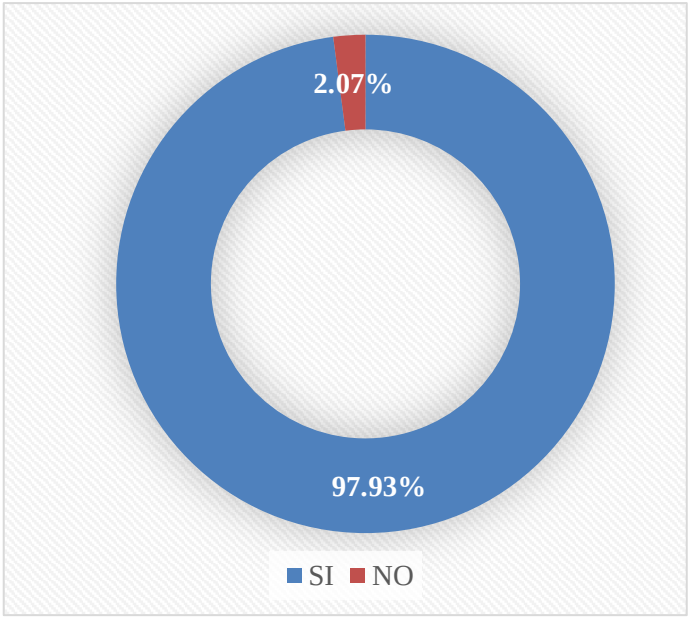


Figura 13

Padre vive de los estudiantes

En la tabla 9 y figura 13, se evidencia que de 379 (97.93%) estudiantes su padre vive, mientras que de 8 (2.07%) alumnos su padre no vive.

Tabla 10

Madre vive de los estudiantes

Madre vive	Frecuencia	%
SI	385	99.48%
NO	2	0.52%
Total	387	100.00%

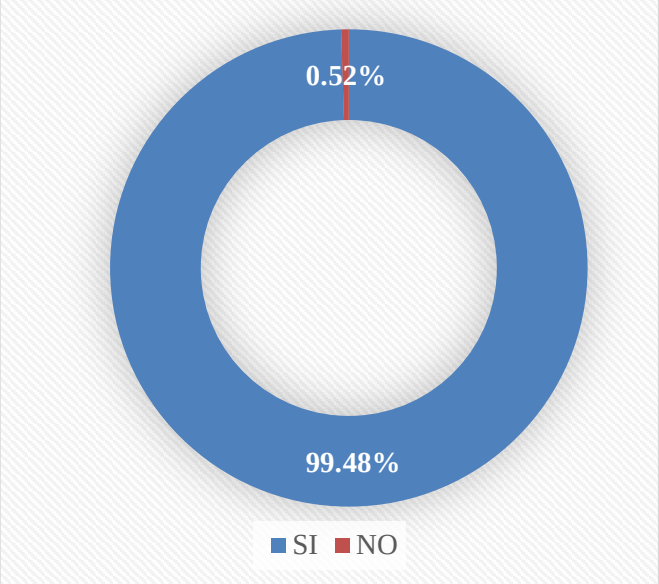


Figura 14
Madre vive de los estudiantes

En la tabla 10 y figura 14, se evidencia que de 385 (99.48%) estudiantes su madre vive, mientras que de 2 (0.52%) alumnos su madre no vive.

Tabla 11
Lengua materna de los estudiantes

Lengua materna	Frecuencia	%
Castellano	387	100.00%
Otro	0	0.00%
Total	387	100.00%

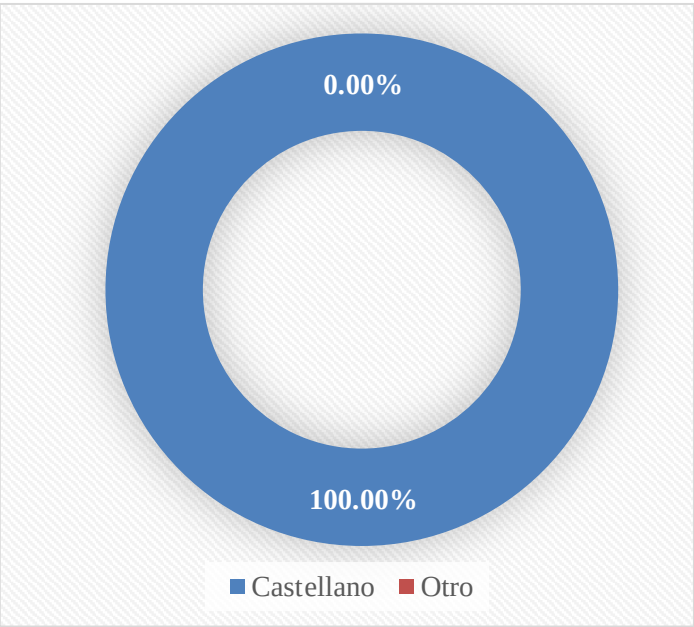


Figura 15
Lengua materna de los estudiantes

En la tabla 11 y figura 15, se visualiza que 387 (100.00%) estudiantes la lengua materna es el castellano.

Tabla 12

Trabaja el estudiante

Trabaja el estudiante	Frecuencia	%
SI	0	0.00%
NO	387	100.00%
Total	387	100.00%

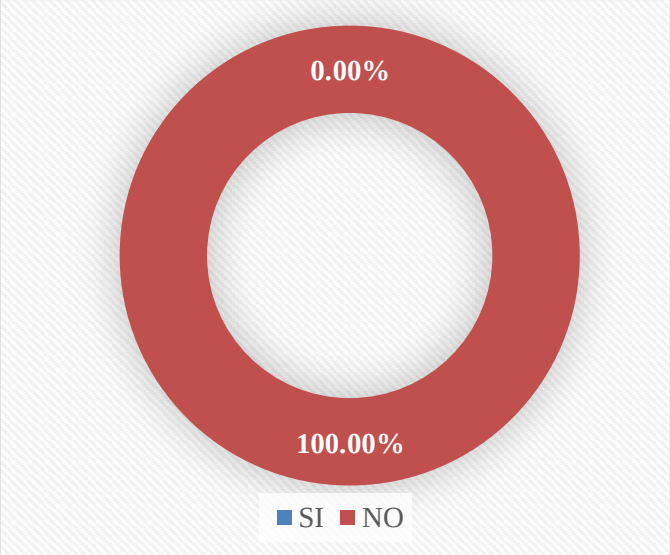


Figura 16

Trabaja el estudiante

En la tabla 12 y figura 16, se evidencia que ningún estudiante labora, siendo la totalidad que se dedica a actividades escolares.

Tabla 13

Escolaridad de la madre de los estudiantes

Escolaridad de la madre	Frecuencia	%
Primaria	316	81.65%
Sin escolaridad	46	11.89%
Secundaria	25	6.46%
Total	387	100.00%

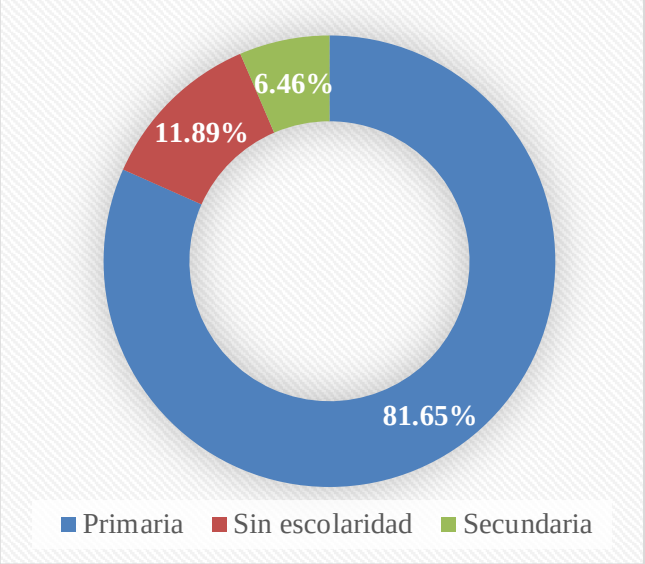


Figura 17
Escolaridad de la madre de los estudiantes

En la tabla 13 y figura 17, los datos reflejan que 316 (81.65%) madres cuentan con primaria, mientras que 46 (11.89%) madres no tienen escolaridad y solo 25 (6.46%) cuenta con nivel secundario.

Tabla 14
Nacimiento registrado de los estudiantes

Nacimiento registrado	Frecuencia	%
SI	253	65.37%
NO	134	34.63%
Total	387	100.00%

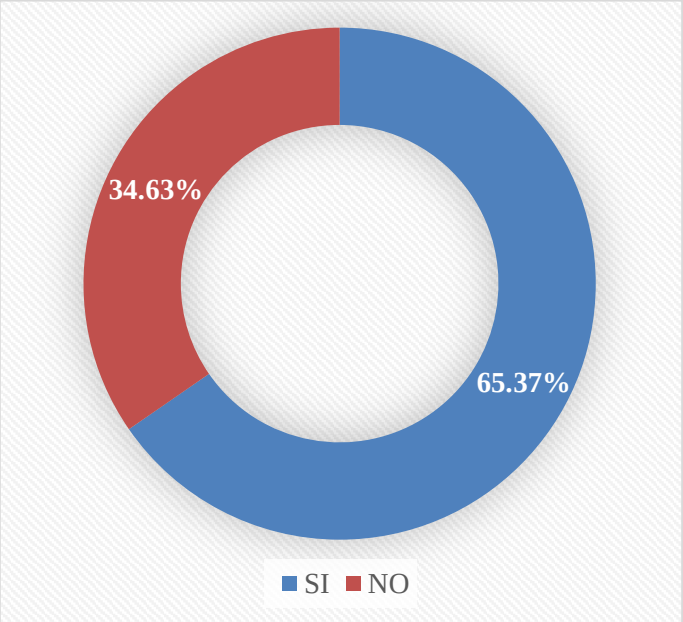


Figura 18
Nacimiento registrado de los estudiantes

En la tabla 14 y figura 18, se aprecia que 253 (65.37%) de los estudiantes si tienen registrado su nacimiento, mientras que 134 (34.63%) no cuenta con registro de nacimiento.

Tabla 15

Tipo de discapacidad de los estudiantes

Tipo de discapacidad	Frecuencia	%
Discapacidad intelectual	1	0.26%
Discapacidad visual	1	0.26%
Sin discapacidad	385	99.48%
Total	387	100.00%

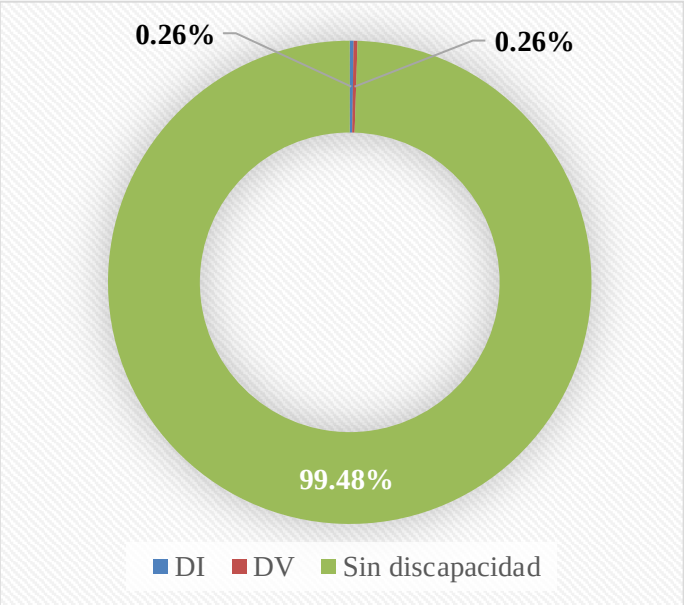


Figura 19

Tipo de discapacidad de los estudiantes

En la tabla 15 y figura 19, en el tipo de discapacidad de los estudiantes, solo un estudiante (0.26%) presenta discapacidad intelectual (DI), también un estudiante (0.26%) presenta discapacidad visual (DV) y 385 (99.48%) no presentan ningún tipo de discapacidad.

Tabla 16

Áreas aprobadas de los estudiantes

Áreas aprobadas	Frecuencia	%
00 - 15	102	26.36%
16 - 30	64	16.54%
31 - 45	79	20.41%
46 - 60	107	27.65%
61 a más	35	9.04%
Total	387	100.00%

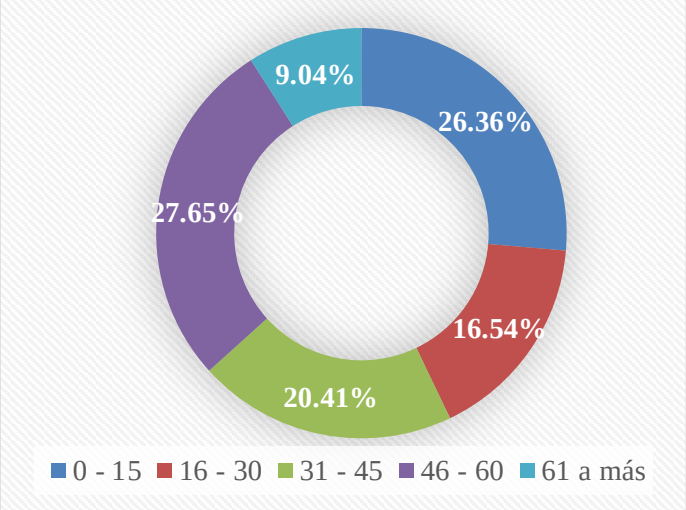


Figura 20
Áreas aprobadas de los estudiantes

En la tabla 16 y figura 20, la cantidad de 102 (26.36%) estudiantes aprobaron de 0–15 áreas, asimismo 64 (16.54%) estudiantes alcanzaron aprobar de 16–30 áreas, también 79 (20.41%) estudiantes aprobaron 31–45 áreas, 107 (27.65%) estudiantes alcanzaron aprobar entre 46–60 áreas, siendo la más predominante debido a su mayor frecuencia, finalmente solo 35 (9.04%) aprobaron más de 61 áreas.

Tabla 17
Áreas desaprobadas de los estudiantes

Áreas desaprobadas	Frecuencia	%
0	292	75.45%
1	32	8.27%
2	21	5.43%
3	16	4.13%
4	11	2.84%
5	10	2.58%
6	4	1.03%
10	1	0.26%
Total	387	100.00%

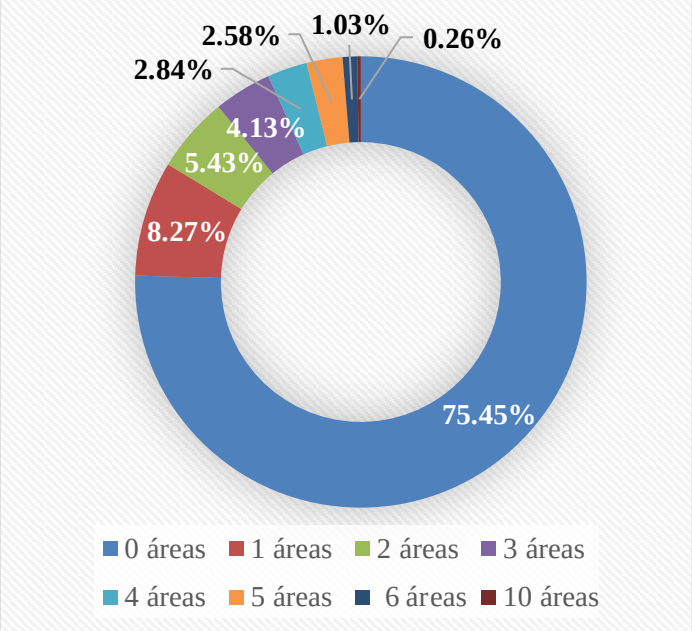


Figura 21
Áreas desaprobadas de los estudiantes

En la tabla 17 y figura 21, se refleja que la mayor parte de los estudiantes no tiene ningún área desaprobada el cual corresponden a 292 (75.45%) alumnos, 32 (8.27%) alumnos cuentan con 1 área desaprobada, 21 (5.43%) cuentan con al menos 2 áreas desaprobadas, 16 (4.13%) estudiantes cuenta con 3 áreas desaprobadas, 11 (2.84%) presentan 4 áreas desaprobadas, 10 (2.58%) estudiantes tienen 5 áreas desaprobadas, 4 (1.03%) tienen 6 áreas desaprobadas y solamente 1 (0.26%) estudiante cuenta con 10 áreas desaprobadas.

Tabla 18
Estado final de los estudiantes

Estado final del alumno	Frecuencia	%
Completado	183	47.29%
Retirado	83	21.45%
Trasladado	15	3.88%
En proceso	106	27.39%
Total	387	100.00%

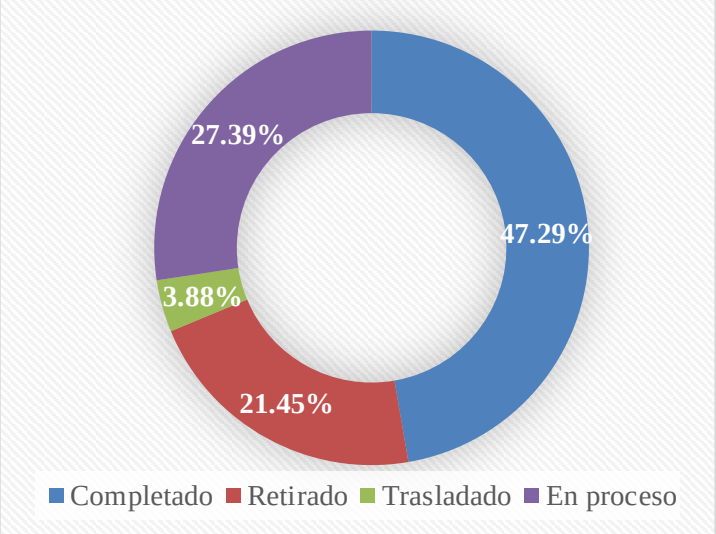


Figura 22
Estado final de los estudiantes

En la tabla 18 y figura 22, la cantidad de 183 (47.29%) completaron sus estudios, 83 (21.45%) se retiraron, 15 (3.88%) se trasladaron y 106 (27.39%) se encuentra en proceso.

Tabla 19
Grado culminado de los estudiantes

Grado culminado	Frecuencia	%
0 grados	7	1.81%
1 grado	56	14.47%
2 grados	55	14.21%
3 grados	38	9.82%
4 grados	47	12.14%
5 grados	184	47.55%
Total	387	100.00%

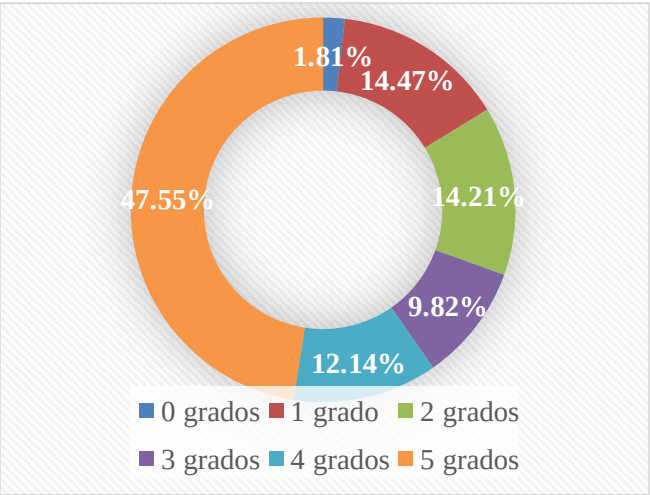


Figura 23
Grado culminado de los estudiantes

En la tabla 19 y figura 23, presentan que 7 (1.81%) estudiantes no culminaron ningún grado, 56 (14.47%) alumnos culminaron un grado, 55 (14.21%) alumnos culminaron dos grados, 38 (9.82%) alumnos culminaron tres grados, 47 (12.14%) alumnos culminaron cuatro grados y 184 (47.55%) estudiantes culminaron los cinco grados.

Tabla 20

Comportamiento promedio de los estudiantes

Comportamiento promedio	Frecuencia	%
AD	7	1.81%
A	344	88.89%
B	30	7.75%
C	6	1.55%
Total	387	100.00%

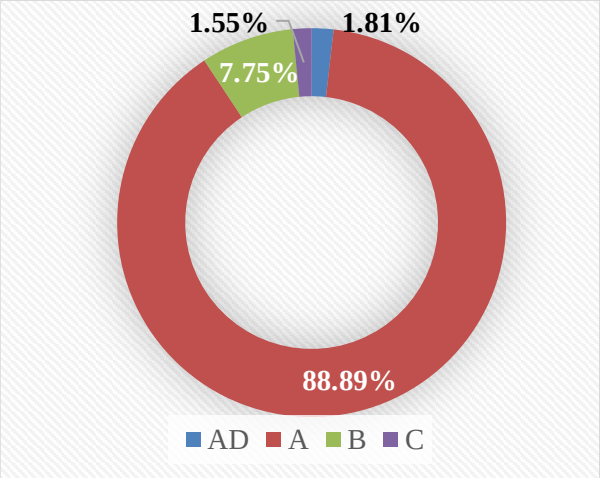


Figura 24

Comportamiento promedio de los estudiantes

En la tabla 20 y figura 24, se aprecia que 7 (1.81%) estudiantes presentaron comportamiento promedio “AD”, 344 (88.89%) alumnos tuvieron comportamiento “A”, 30 (7.75%) estudiantes contaron con promedio “B” y solamente 6 (1.55%) estudiantes con promedio “C”.

B.4. Validación de la calidad de los datos

Durante esta labor, se ejecutó una evaluación exhaustiva de la calidad de los datos, mediante la formulación y respuesta de una serie de interrogantes diseñados para validar los estándares de calidad establecidos:

¿Están completos? Los datos proveídos por la I.E lamentablemente no se encuentran completos, información de los estudiantes se encuentra vacías como fecha de nacimiento, padre vive, madre vive, escolaridad de la madre, etc.; sin embargo significa una pequeña parte de la base de datos el cual se podría exonerar para la generación del modelo de DM.

¿Cubren todos los casos requeridos? Los datos del conglomerado del dataset, si cumplen y cubren las características requeridas, catalogándose como completos, en cada instancia del estudio.

¿Son correctos o contienen errores? La información no presenta errores.

¿Hay valores omitidos? Si existen valores omitidos, pero representan en mínimo porcentaje, el cual se completó para tener un dataset completo.

C. FASE 3: PREPARACIÓN DE LOS DATOS

En esta etapa del proceso CRISP-DM, se aborda la preparación de datos con el objetivo de construir un conjunto de datos adecuado para aplicar las técnicas de minería de datos. Esto implica seleccionar el conjunto de datos a utilizar, realizar limpieza para mejorar la calidad, agregar nuevos atributos y ajustarlos al formato óptimo para evitar cualquier inconveniente con las herramientas de modelado.

C.1. Selección de los datos más relevantes

Se llevaron a cabo pruebas inferenciales de procesamiento estadístico en el conjunto de datos para identificar los factores más relevantes asociados con la deserción estudiantil. Se calcularon medidas como la prueba de Chi cuadrado de Pearson, la V de Cramer y la contingencia. Además, se proporciona una tabla que muestra la importancia de cada factor en relación con la deserción estudiantil.

Tabla 21

Factores relevantes en la deserción estudiantil

Variable	Chi Cuadrado		V de Cramer	Contingencia
	Valor	P – valor		
Sexo	1.692	0.193	0.066	0.066
Situación de matrícula	9.815	0.020	0.159	0.157
País	-	-	-	-
Padre vive	0.061	0.805	0.013	0.013
Madre vive	0.549	0.459	0.038	0.038
Lengua materna	-	-	-	-
Trabaja estudiante	-	-	-	-
Escolaridad de la madre	6.144	0.046	0.126	0.126
Nacimiento registrado	5.811	0.016	0.123	0.122
Tipo de discapacidad	0.549	0.760	0.038	0.038
Áreas aprobadas	195.155	0.000	0.710	0.579
Áreas desaprobadas	33.639	0.000	0.295	0.283
Grado culminado	110.045	0.000	0.533	0.471
Comportamiento	77.432	0.000	0.447	0.408

Fuente: Dataset – SPSS v.25

En la tabla 21, muestra que según el análisis cuando el p-valor es menor a 0.05, significa que las variables están relacionadas con la deserción estudiantil, por lo tanto podemos confirmar que:

Las variables: situación de matrícula, escolaridad de la madre, nacimiento registrado, áreas aprobadas, áreas desaprobadas, grado culminado y el comportamiento tienen relación con la deserción estudiantil.

Además los coeficientes tanto como la V de Cramer y la contingencia permitieron ver el grado de relación de las variables con la deserción estudiantil. Por lo tanto, según el coeficiente de contingencia, se recomienda enfocar la predicción en las variables que mostraron una correlación superior a 0.10, que son las siguientes: situación de matrícula, escolaridad de la madre, nacimiento registrado, áreas aprobadas, áreas desaprobadas, grado culminado y el comportamiento tienen relación con la deserción estudiantil.

C.2. Limpieza de los datos

En esta actividad, se hace la descripción de las variables más relevantes que nos van a permitir predecir la deserción estudiantil y para lograr dicho fin se eliminaron registros de la data inicial que presentaban errores e inconsistencias en las fechas de nacimiento, alumnos fallecidos, alumnos trasladados. Se hizo uso de la herramienta Microsoft Excel para limpiar toda la información con incoherencias.

1. Sexo: (Nominal – dicotómico)

Relevancia: Baja

2. Situación de matrícula: (Nominal – politómico)

Relevancia: Media

3. País: (Nominal - dicotómico)

Relevancia: Baja

4. Padre vive: (Nominal - dicotómico)

Relevancia: Baja

5. Madre vive: (Nominal - dicotómico)

Relevancia: Baja

6. Lengua materna: (Nominal - dicotómico)

Relevancia: Baja

7. Trabaja el estudiante: (Nominal - dicotómico)

Relevancia: Baja

8. Escolaridad de la madre: (Nominal – politómico)

Relevancia: Media

9. Nacimiento registrado: (Nominal - dicotómico)

Relevancia: Media

10. Tipo de discapacidad: (Nominal - politómico)

Relevancia: Baja

11. Áreas aprobadas: (Razón)

Relevancia: Alta

12. Áreas desaprobadas: (Razón)

Relevancia: Alta

13. Grado culminado: (Ordinal)

Relevancia: Alta

14. Comportamiento: (Ordinal)

Relevancia: Alta

C.3. Construcción de nuevos datos

En esta actividad solo se destacó la transformación del campo Código Alumno/DNI del alumno por un ID. Dicha transformación consistió en dar un código único a cada alumno con valores numéricos, que permitieron tener mayor comprensión del dataset.

C.4. Integración de los datos

Se integraron los datos de las fuentes de información, ya que se trabajaron de dos documentos diferentes que contienen información relevante para la predicción de la deserción estudiantil.

C.5. Formateo de los datos

Los datos de los campos del dataset, han sido codificados con valores comprensibles ya que el Danta mining exige que los datos deben ser claros y precisos. En inició los datos estaban compuestos por caracteres alfabéticos y numéricos, posteriormente se normalizó usando una estructura de códigos que permitan mejor entendimiento de la data y mejor eficiencia en el procesamiento del DM. Quedando de la siguiente forma:

Tabla 22

Formateo de datos

Variable	Valores
Sexo	Hombre
	Mujer
Situación de matrícula	Ingresante
	Repitente
	Promovido
	Reentrante
País	Perú

	Otro
Padre vive	Si
	No
Madre vive	Si
	No
Lengua materna	Castellano
	Otro
Trabaja estudiante	Si
	No
	Primaria
Escolaridad de la madre	Secundaria
	SE
Nacimiento registrado	Si
	No
	DI
Tipo de discapacidad	DV
	SD
Áreas aprobadas	Valores numéricos entre 0 a 64
Áreas desaprobadas	Valores numéricos entre 0 a 10
	0
	1
Grado culminado	2
	3
	4
	5
	C
Comportamiento	B
	A
	AD

D. FASE 4: MODELADO

En esta etapa del CRISP-DM, se elige la(s) técnica(s) más idóneas para los propósitos planteados del DM. Posteriormente realizado las pruebas para los modelos elegidos, se procede a la aplicación de dichas técnicas sobre el conglomerado de información para la generación del modelo y con ello se procede a evaluar el cumplimiento de los criterios de éxito o fracaso.

D.1. Selección de las técnicas del modelado

Para esta actividad se usará el software Weka 3.8.6, el cual se tomará en cuenta los algoritmos más relevantes para la predicción; donde los datos obtenidos por el software, tendrán el formato mostrado a continuación:

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         363               93.7984 %
Incorrectly Classified Instances        24                6.2016 %
Kappa statistic                          0.8175
Mean absolute error                      0.0936
Root mean squared error                  0.2221
Relative absolute error                 27.6924 %
Root relative squared error             54.1017 %
Total Number of Instances              387

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.867    0.043    0.847      0.867    0.857      0.818    0.952     0.847     Si
                 0.957    0.133    0.964      0.957    0.960      0.818    0.952     0.980     No
Weighted Avg.    0.938    0.113    0.939      0.938    0.938      0.818    0.952     0.951

=== Confusion Matrix ===

   a   b   <-- classified as
  72  11 |   a = Si
  13 291 |   b = No
```

Figura 25
Resultados generados por Weka 3.8.6

Los algoritmos más importantes para la predicción son 5, lo cuales corresponden a:

1. Algoritmo arboles de decisión J48
2. Algoritmo arboles de decisión RANDOM FOREST
3. Algoritmo VECINOS MAS CERCANOS
4. Algoritmo FUNCIÓN LOGÍSTICA
5. Algoritmo PERCEPTRÓN MULTICAPA

```
@relation DE

@attribute codigo real
@attribute sexo {Hombre,Mujer}
@attribute situación_matricula {Ingresante,Repitente,Promovido,Reentrante}
@attribute pais {Perú,Otro}
@attribute padre_vive {Si,No}
@attribute madre_vive {Si,No}
@attribute lengua_materna {Castellano,Otro}
@attribute trabaja_estudiante {Si,No}
@attribute escolaridad_madre {Primaria,Secundaria,SE}
@attribute nacimiento_registrado {Si,No}
@attribute tipo_discapacidad {DI,DV,SD}
@attribute areas_aprobadas real
@attribute areas_desaprobadas real
@attribute grado_culminado real
@attribute comportamiento {AD,A,B,C}
@attribute abandono {Si,No}

@data
1,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,9,2,1,B,Si
2,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Secundaria,No,SD,55,0,5,A,No
3,Hombre,Ingresante,Perú,Si,Si,Castellano,No,SE,No,SD,55,0,5,A,No
4,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,50,5,5,A,No
5,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,55,0,5,A,No
6,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,55,0,5,A,No
7,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,11,0,1,A,Si
8,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,55,0,5,A,No
9,Mujer,Ingresante,Perú,Si,Si,Castellano,No,SE,No,SD,22,0,2,B,Si
10,Mujer,Ingresante,Perú,Si,Si,Castellano,No,SE,No,SD,51,4,5,A,No
11,Hombre,Repitente,Perú,Si,Si,Castellano,No,SE,No,SD,6,5,1,C,Si
12,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,55,0,5,A,No
13,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,11,0,1,B,Si
14,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,10,1,1,B,Si
15,Mujer,Repitente,Perú,Si,Si,Castellano,No,Primaria,No,SD,44,0,5,A,No
16,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,9,2,1,B,Si
17,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Secundaria,No,SD,55,0,5,A,No
18,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,54,1,5,A,No
19,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,53,2,5,A,No
20,Mujer,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,11,0,1,A,Si
21,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,55,0,5,A,No
22,Hombre,Ingresante,Perú,Si,Si,Castellano,No,Primaria,No,SD,50,5,5,A,No
23,Mujer,Promovido,Perú,Si,Si,Castellano,No,Secundaria,No,SD,22,0,2,A,Si
24,Mujer,Repitente,Perú,Si,Si,Castellano,No,Primaria,Si,SD,33,0,4,A,No
25,Hombre,Reentrante,Perú,Si,Si,Castellano,No,SE,No,SD,8,3,2,B,Si
```

Figura 26

Fragmento del formato arff para la carga en Weka 3.8.6

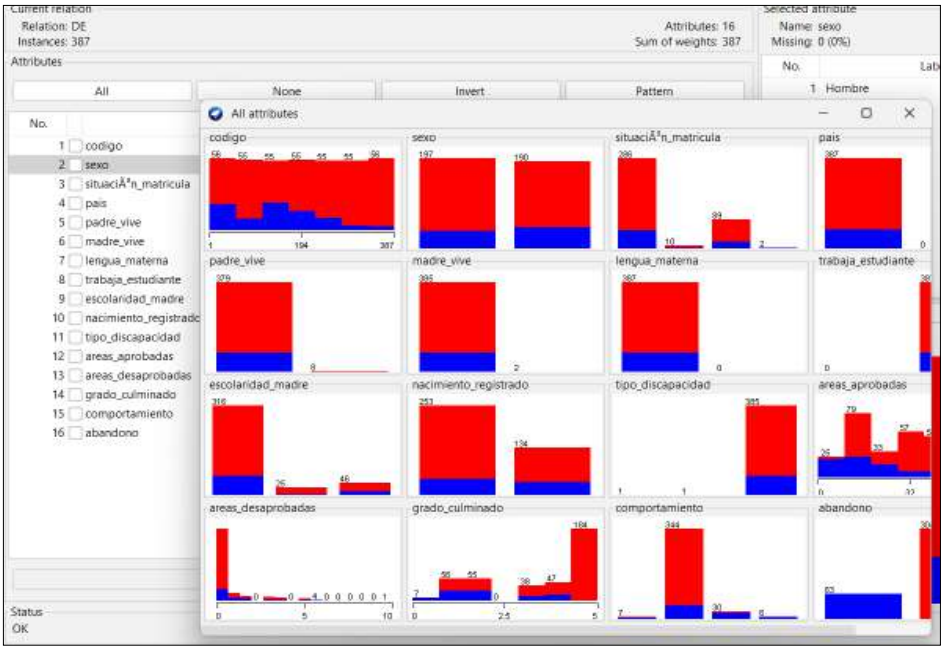


Figura 27

Carga de los datos en el software Weka 3.8.6

En la figura 27, muestra la carga de los datos al software Weka 3.8.6, estableciéndose en base a 16 atributos y 387 registros. Para después poder clasificar los 5 algoritmos enumerados con anterioridad.

✓ ALGORITMO ARBOLES DE DECISIÓN J48

Los resultados del algoritmo J48 generado por el software Weka 3.8.6, se resumen a continuación:

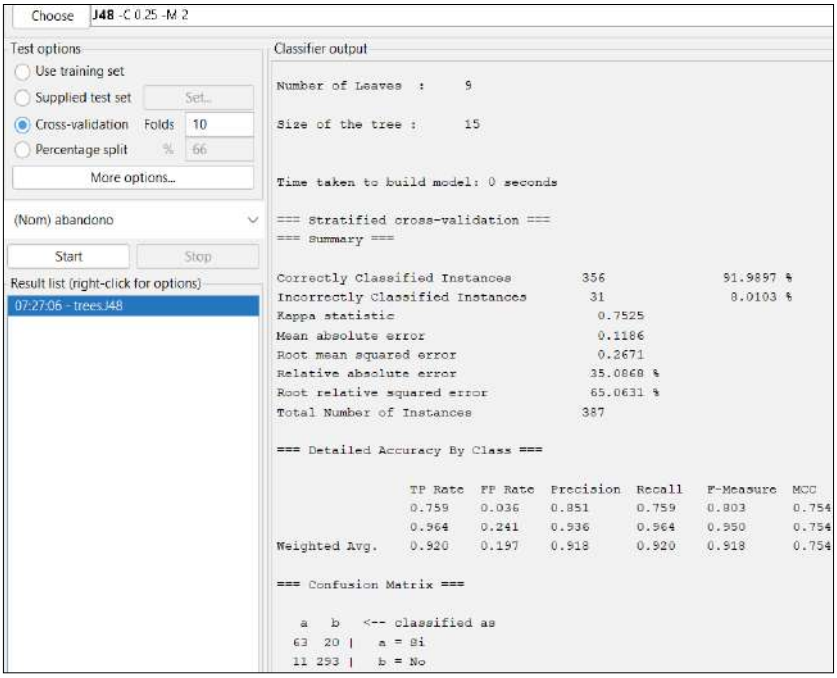


Figura 28
Resultado de la ejecución del algoritmo J48 - Weka 3.8.6

Tabla 23
Resumen del algoritmo arboles de decisión J48

Resumen J48	Casos	Clasificación
Correctly Classified Instances	356	91.99 %
Incorrectly Classified Instances	31	8.01 %
Total	387	100.00%

Fuente: Weka 3.8.6

Tabla 24
Matriz de confusión del algoritmo arboles de decisión J48

confusion Matrix	Si abandona	No abandona	Total
Si abandona	63	20	83
No abandona	11	293	304
Total	74	313	387

Fuente: Weka 3.8.6

En la figura 28 y tabla 23, el algoritmo J48 clasifico de manera correcta hasta un 91.99%, y en la matriz de confusión (Tabla 24) se aprecia que:

El total de 74 registros fueron clasificados correctamente, para la condición “Si abandona”.

El total de 313 registros fueron clasificados correctamente, para la condición “No abandona”.

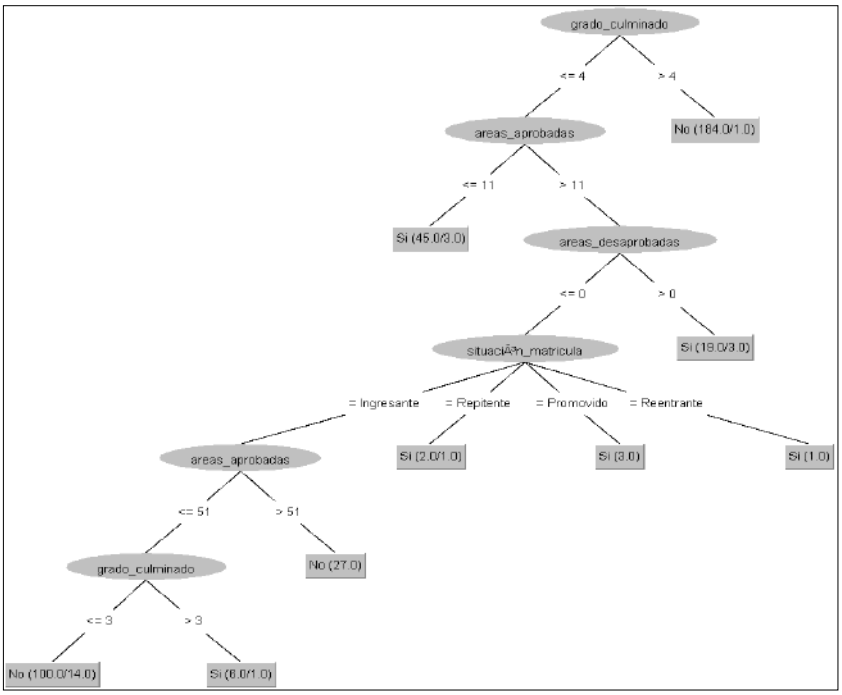


Figura 29
Árbol creado por el algoritmo J48 - Weka 3.8.6

✓ ALGORITMO DE DECISIÓN RANDOM FOREST

Los resultados del algoritmo Random Forest generado por el software Weka 3.8.6, se resumen a continuación:

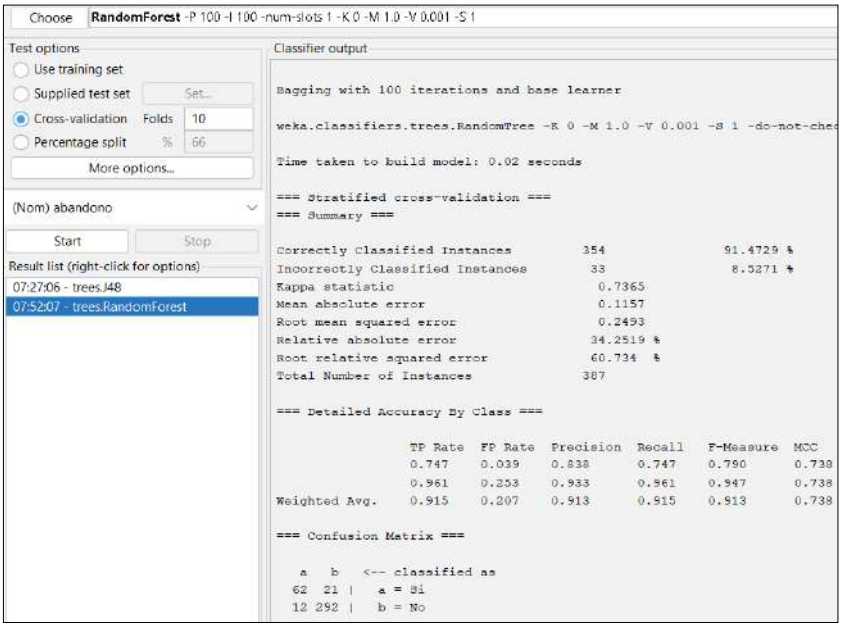


Figura 30
Resultado de la ejecución del algoritmo Random Forest - Weka 3.8.6

Tabla 25

Resumen del algoritmo arboles de decisión Random Forest

Resumen Random Forest	Casos	Clasificación
Correctly Classified Instances	354	91.47 %
Incorrectly Classified Instances	33	8.53 %
Total	387	100.00%

Fuente: Weka 3.8.6

Tabla 26

Matriz de confusión del algoritmo arboles de decisión Random Forest

Confusion Matrix	Si abandona	No abandona	Total
Si abandona	62	21	83
No abandona	12	292	304
Total	74	313	387

Fuente: Weka 3.8.6

En la figura 30 y tabla 25, el algoritmo Random Forest clasifico de manera correcta hasta un 91.47%, y en la matriz de confusión (Tabla 26) se aprecia que:

El total de 74 registros fueron clasificados correctamente, para la condición “Si abandona”.

El total de 313 registros fueron clasificados correctamente, para la condición “No abandona”.

✓ **ALGORITMO VECINOS MAS CERCANOS**

Los resultados del algoritmo Vecinos Mas Cercanos generado por el software Weka 3.8.6, se resumen a continuación:

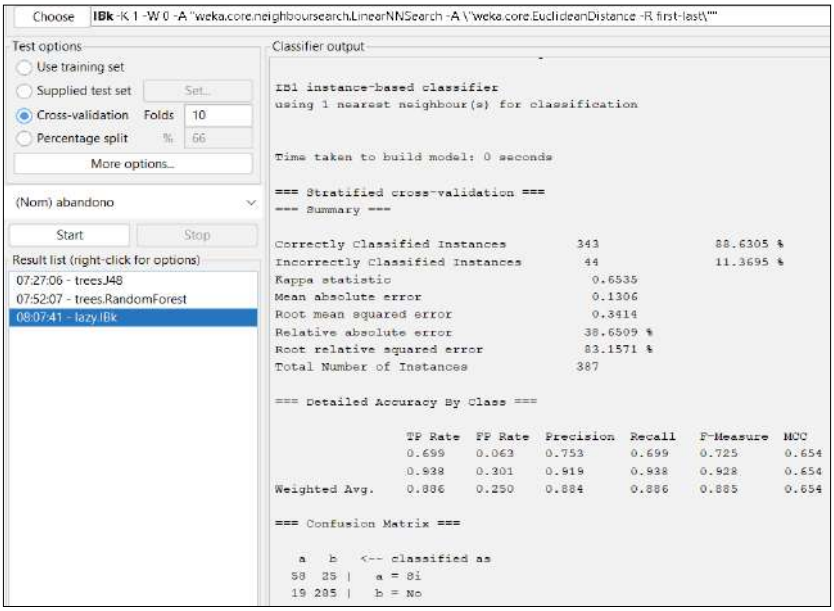


Figura 31

Resultado de la ejecución del algoritmo Vecinos Mas Cercanos - Weka 3.8.6

Tabla 27

Resumen del algoritmo Vecinos Mas Cercanos

Resumen Vecinos Mas Cercanos	Casos	Clasificación
Correctly Classified Instances	343	88.63 %
Incorrectly Classified Instances	44	11.37 %
Total	387	100.00%

Fuente: Weka 3.8.6

Tabla 28

Matriz de confusión del algoritmo Vecinos Mas Cercanos

Confusion Matrix	Si abandona	No abandona	Total
Si abandona	58	25	83
No abandona	19	285	304
Total	77	310	387

Fuente: Weka 3.8.6

En la figura 31 y tabla 27, el algoritmo Vecinos más Cercanos clasifico de manera correcta hasta un 88.63%, y en la matriz de confusión (Tabla 28) se aprecia que:

El total de 77 registros fueron clasificados correctamente, para la condición “Si abandona”.

El total de 310 registros fueron clasificados correctamente, para la condición “No abandona”.

✓ ALGORITMO FUNCIÓN LOGÍSTICA

Los resultados del algoritmo Función Logística generado por el software Weka 3.8.6, se resumen a continuación:

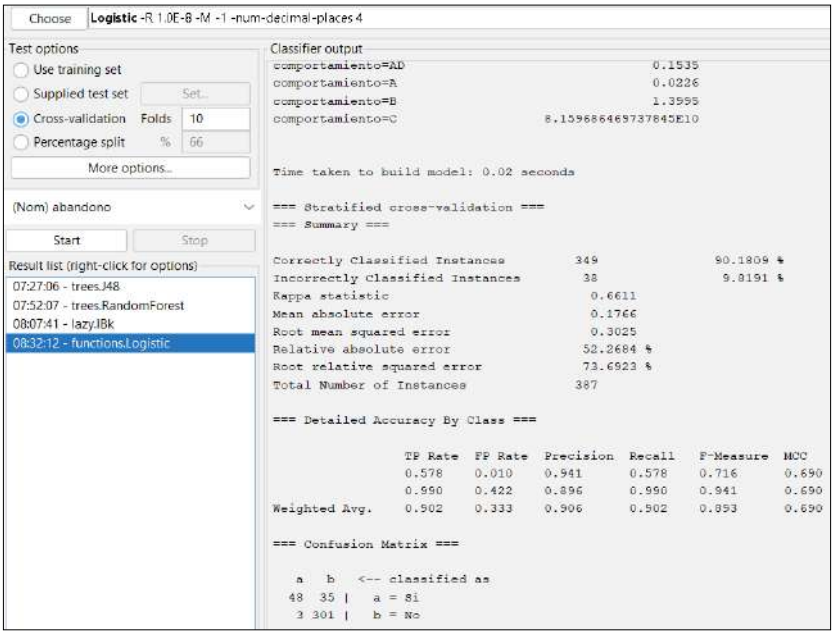


Figura 32

Resultado de la ejecución del algoritmo Función Logística - Weka 3.8.6

Tabla 29

Resumen del algoritmo Función Logística

Resumen Función Logística	Casos	Clasificación
Correctly Classified Instances	349	90.18 %
Incorrectly Classified Instances	38	9.82 %
Total	387	100.00%

Fuente: Weka 3.8.6

Tabla 30

Matriz de confusión del algoritmo Función Logística

Confusion Matrix	Si abandona	No abandona	Total
Si abandona	48	35	83
No abandona	3	301	304
Total	51	336	387

Fuente: Weka 3.8.6

En la figura 32 y tabla 29, el algoritmo Función Logística clasifico de manera correcta hasta un 90.18%, y en la matriz de confusión (Tabla 30) se aprecia que:

El total de 51 registros fueron clasificados correctamente, para la condición “Si abandona”.

El total de 336 registros fueron clasificados correctamente, para la condición “No abandona”.

✓ **ALGORITMO PERCEPTRÓN MULTICAPA**

Los resultados del algoritmo Perceptrón Multicapa generado por el software Weka 3.8.6, se resumen a continuación:

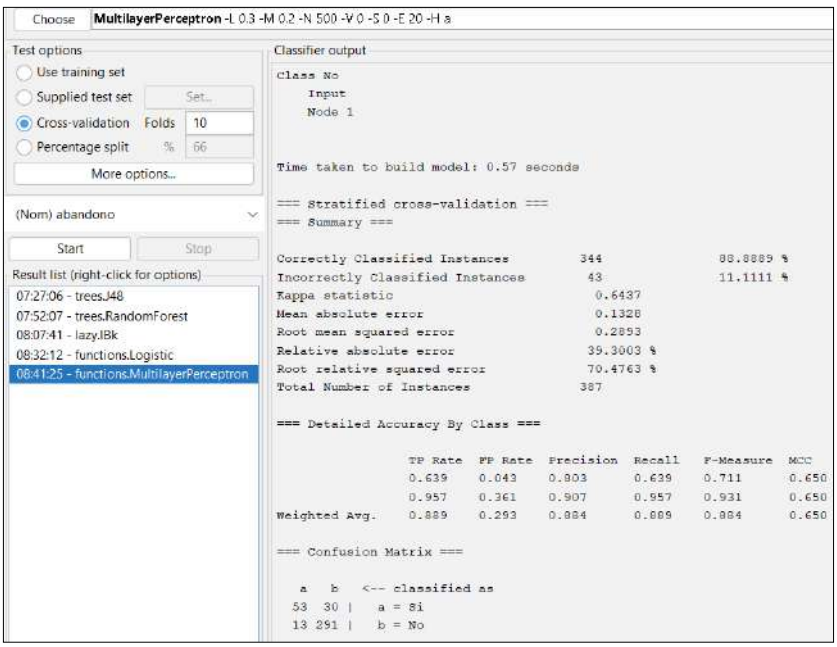


Figura 33

Resultado de la ejecución del algoritmo Perceptrón Multicapa - Weka 3.8.6

Tabla 31

Resumen del algoritmo Perceptrón Multicapa

Resumen Perceptrón Multicapa	Casos	Clasificación
Correctly Classified Instances	344	88.89 %
Incorrectly Classified Instances	43	11.11 %
Total	387	100.00%

Fuente: Weka 3.8.6

Tabla 32

Matriz de confusión del algoritmo Perceptrón Multicapa

Confusion Matrix	Si abandona	No abandona	Total
Si abandona	53	30	83
No abandona	13	291	304
Total	66	321	387

Fuente: Weka 3.8.6

En la figura 33 y tabla 31, el algoritmo Perceptron Multicapa clasifico de manera correcta hasta un 88.89%, y en la matriz de confusión (Tabla 32) se aprecia que:

El total de 66 registros fueron clasificados correctamente, para la condición “Si abandona”.

El total de 321 registros fueron clasificados correctamente, para la condición “No abandona”.

D.2. Diseño de comprobación del modelado

Se generó en base a los resultados hallados con anterioridad, con el procesamiento de las diversas técnicas, en resumen se tuvo:

Tabla 33

Resumen de las técnicas de minería de datos

N°	Técnica	Predicción acertada	Tiempo	Acertados “Abandona” (83 casos)	Acertados “No abandona” (304 casos)
1	J48	91.99 %	0.00 Seg.	74	313
2	Random Forest	91.47 %	0.02 Seg.	74	313
3	Vecinos más Cercanos	88.63 %	0.00 Seg.	77	310
4	Función Logística	90.18 %	0.02 Seg.	51	336
5	Perceptrón Multicapa	88.89 %	0.57 Seg.	66	321

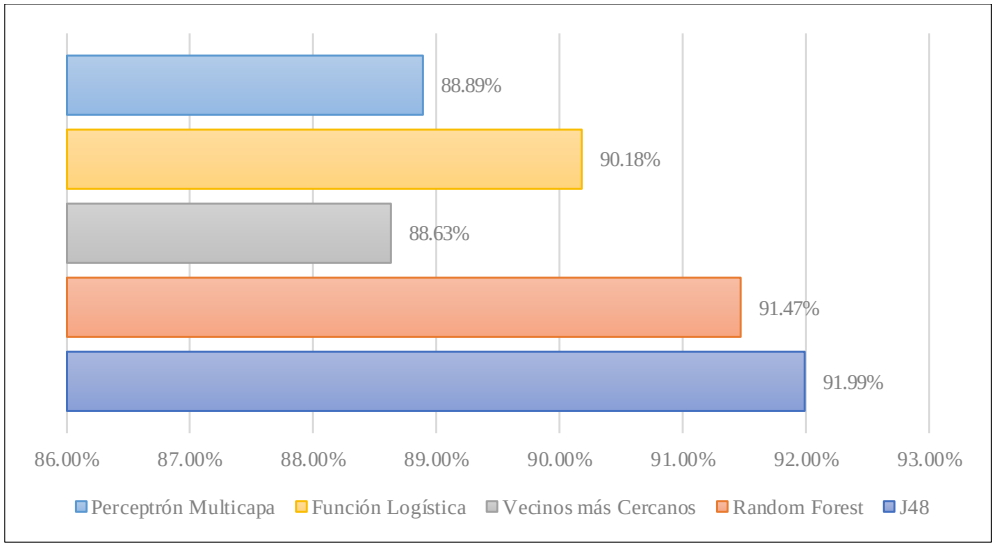


Figura 34
Resumen de las técnicas de minería de datos

En la tabla 33 y figura 34 se evidencia que la técnica más eficiente corresponde al algoritmo J48, el cual tiene un desempeño de predicción del 91.99 % y el algoritmo menor eficiente pertenece al algoritmo Vecinos Mas Cercanos con valor de 88.63 % de performance. Por lo tanto, para este estudio se utilizará el algoritmo de Árbol de Decisión J48 para predecir la deserción estudiantil en la I.E 00116.

D.3. Generación del modelo

Para generar el modelo, se efectuó la técnica elegida con anterioridad, y para ello se particionará la información en 2 grupos:

Primer grupo denominado como grupo de entrenamiento: Compuesto por el 70% del dataset.

Segundo grupo denominado como grupo test: Conformado por el 30% del dataset.

Cabe indicar que para el segundo grupo o grupo del test, el atributo **abandono (Si, No)**, estará definido por el signo de interrogación (?) el cual representa como desconocido y será la técnica quien determine la deserción estudiantil como “Si” o “No”.

D.4. Evaluación y comprobación del modelo

En las acciones de evaluación y comprobación del modelo se implementaron para la técnica del algoritmo árbol de decisión J48 en el software Weka 3.8.6, obteniendo los resultados siguientes:

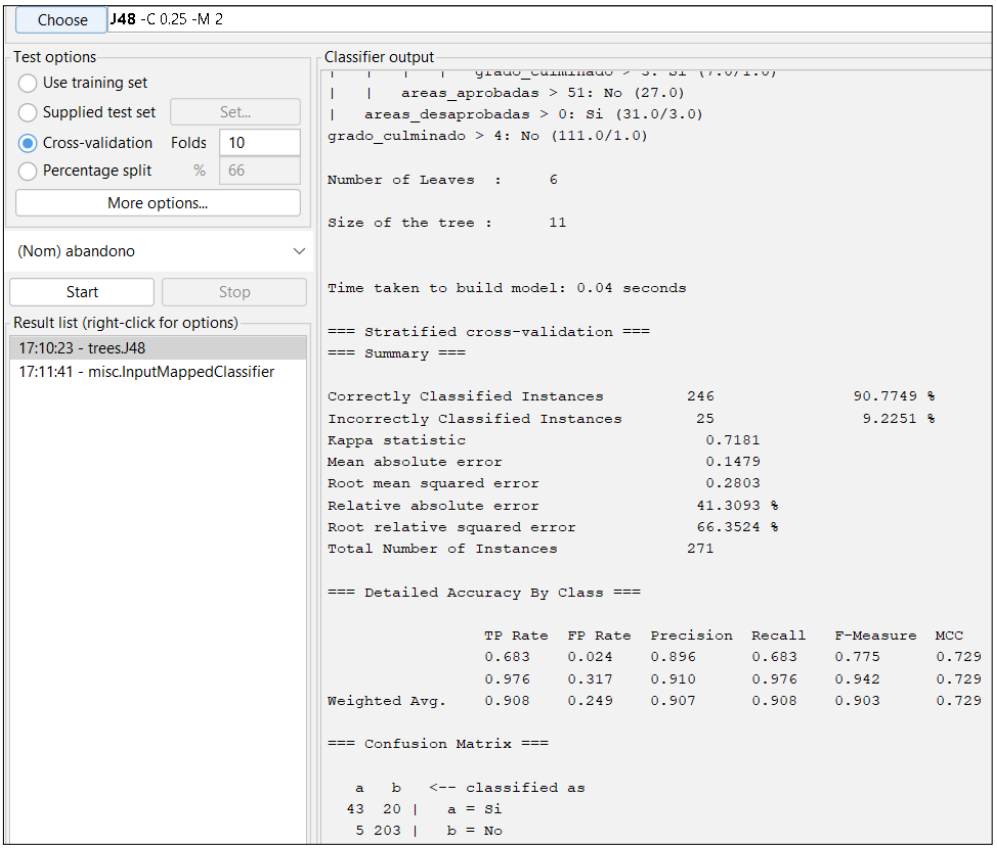


Figura 35
Resultado de la ejecución del algoritmo J48 con datos de entrenamiento - Weka 3.8.6

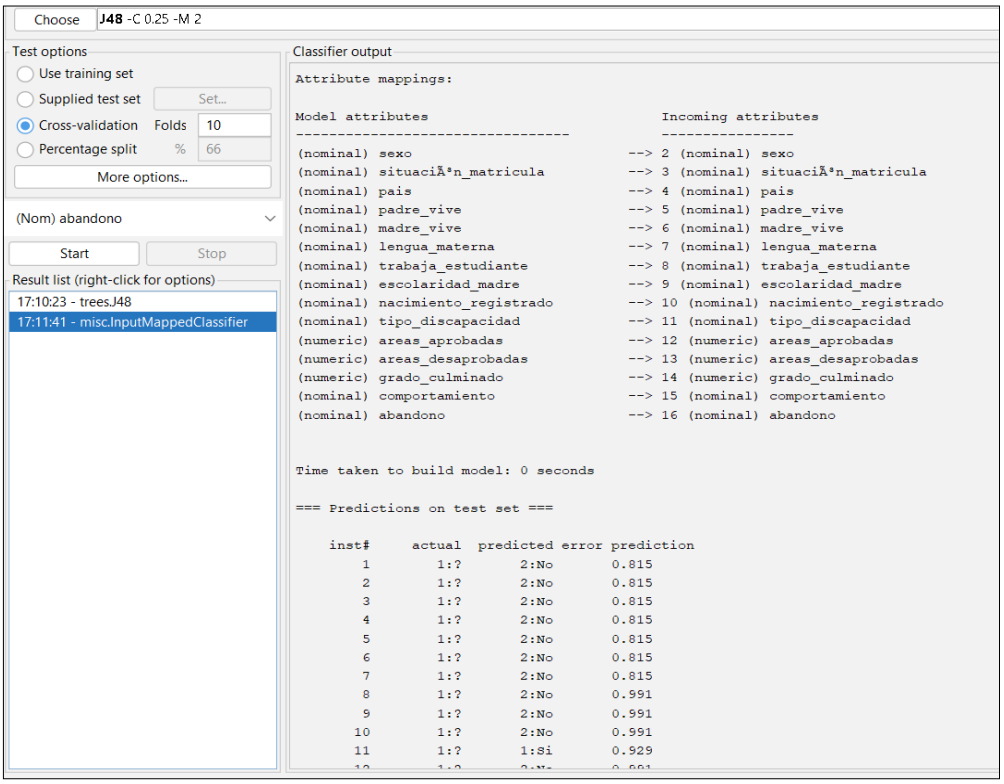


Figura 36
Resultado de la ejecución del algoritmo J48 con datos de test - Weka 3.8.6

E. FASE 5: EVALUACIÓN

En esta fase del proyecto, se analizan uno o varios modelos que brindan una calidad suficiente a la expectativa de la predicción y de la analítica de datos.

E.1. Evaluación de resultados

Los resultados se obtuvieron del reporte generado por Weka 3.8.6, en base a los datos procesados del test, el cual se importa para ser analizados con apoyo del Microsoft Excel.

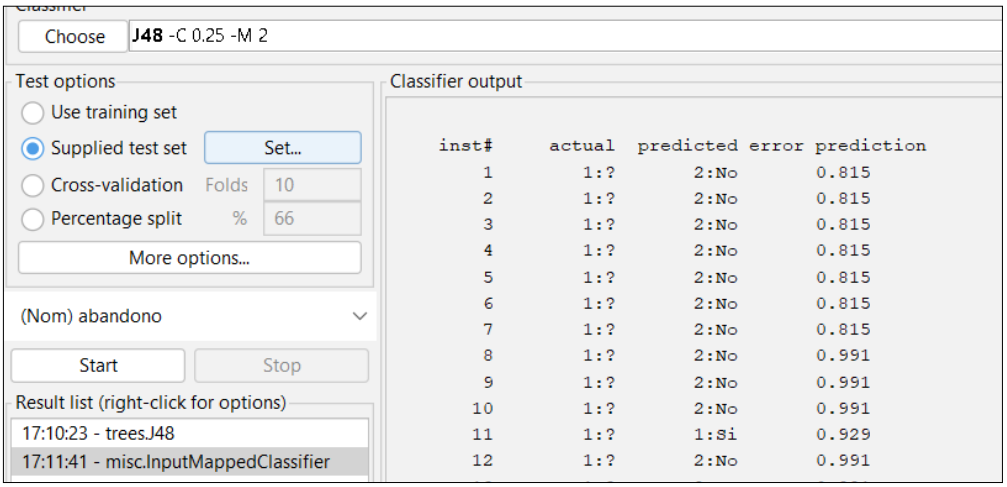


Figura 37

Resultado de la predicción mediante algoritmo J48 - Weka 3.8.6

Con el apoyo de Excel se creó una plantilla electrónica haciendo uso de la función SI, para analizar los aciertos y desaciertos. Se muestra un fragmento de la comparación de resultados:

Tabla 34

Comparación de resultados aciertos y desaciertos con J48

Valor Real	Actual	Precisión	Predicción J48	Observación
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.815	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
Si	1:?	0.929	Si	Acertó
No	1:?	0.991	No	Acertó

No	1:?	0.991	No	Acertó
Si	1:?	0.815	No	No acertó
Si	1:?	0.903	Si	Acertó
Si	1:?	0.903	Si	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
Si	1:?	0.903	Si	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
No	1:?	0.991	No	Acertó
Si	1:?	0.929	Si	Acertó
Si	1:?	0.929	Si	Acertó
Si	1:?	0.929	Si	Acertó
No	1:?	0.991	No	Acertó

E.2. Proceso de revisión

Los procedimientos hasta este punto se han realizado tal y como se había planeado, cabe indicar que se presentó algunos inconvenientes a la hora de realizar el modelo pero de mínima significancia que se logró solucionar y obtener el modelo idóneo.

E.3. Proceso de revisión

Los procedimientos posteriores a estos resultados, será contar con mayor información de los estudiantes ingresantes en los siguientes años, contando con mayor información para la fuente de entrenamiento y se obtenga predicciones con mayor precisión y apoye a la toma de decisiones educativas.

F. FASE 6: DESPLIEGUE

La generación y configuración del modelo no viene ser la parte final del proyecto, lo que se pretende en esta fase es brindar un producto final al cliente y este pueda darle uso en beneficio de la institución. Sin embargo en el presente estudio solo se limitó a diseñar el modelo y corroborar la hipótesis del estudio. Por lo que su despliegue e implementación queda a criterio de la I.E y del investigador continuar con el proyecto en un futuro posterior.

4.2. Análisis inferencial

Contrastación de hipótesis de la investigación

(H_1): La deserción estudiantil puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú - Moyobamba.

(H_o) La deserción estudiantil no puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú - Moyobamba.

Tabla 35

Tabla cruzada con valores reales – valores TMD

			Valores_TMD		Total
			No	Si	
Valores_Reales	No	Recuento	93	3	96
		% dentro de Valores_Reales	96,9%	3,1%	100,0%
	Si	Recuento	2	18	20
		% dentro de Valores_Reales	10,0%	90,0%	100,0%
Total		Recuento	95	21	116
		% dentro de Valores_Reales	81,9%	18,1%	100,0%

En la tabla 35, se evaluó en base a 116 casos de los cuales en 96 casos se evidencia que 93 (96,9%) casos fueron acertados como “No abandona”, asimismo, 3 (3,1%) casos en los valores reales correspondieron a “No abandona” y en los valores por TDM fue de “Si abandona” existiendo un desacierto. Por otra parte de 20 casos restantes, en los valores reales 2 (10,0%) casos fue de “Si abandona” mientras que por TDM fue de “No abandona” habiendo desaciertos, finalmente 21 (90,0%) casos fueron acertados como condición de “Si abandona”.

Tabla 36

Estadísticos de contrastación de hipótesis

Estadísticos	Valor	df	Sig. (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	84,257[a]	1	,000		
Corrección de continuidad[b]	78,499	1	,000		
Prueba exacta de Fisher				,000	,000

Asociación lineal por lineal	83,531	1	,000
N de casos válidos	116		

a. 1 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,62.

b. Sólo se ha calculado para una tabla 2x2

Fuente: Valores de la predicción – SPSS v.25

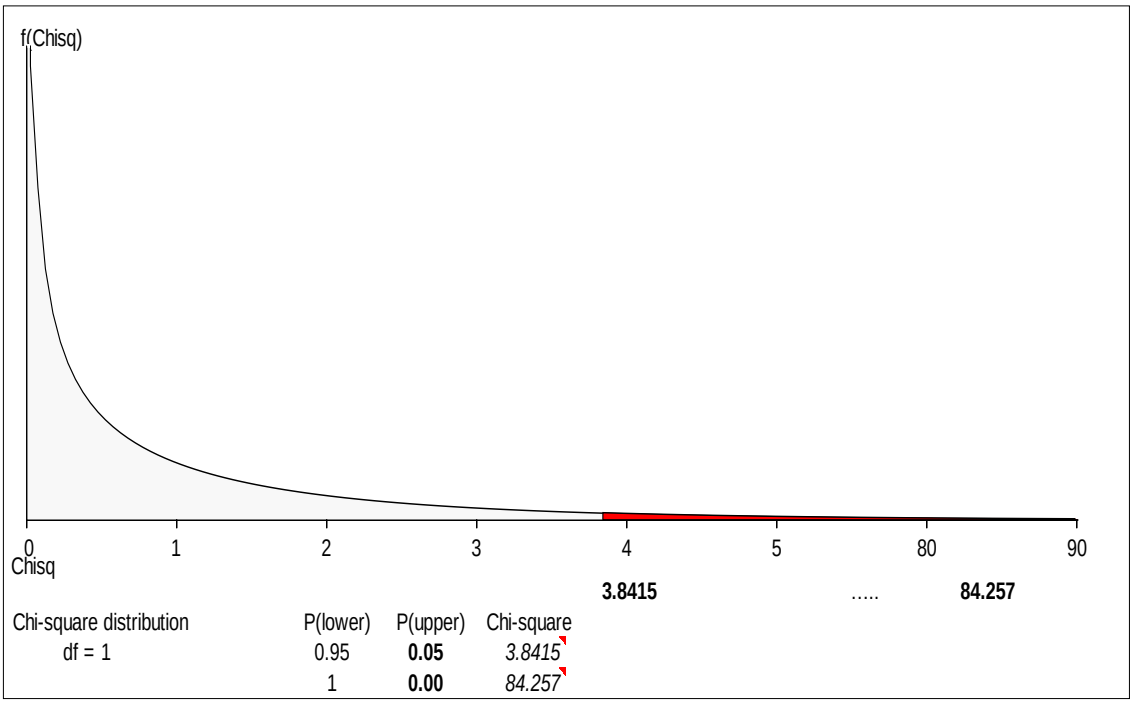


Figura 38

Distribución Chi Cuadrado de Pearson – MegaStat

En la tabla 36 y figura 38, muestra que el valor de Chi-cuadrado de Pearson fue de 84,257, con Sig. bilateral igual a 0.000 mayor al 0.05, además el Chi cuadrado calculado es mayor al Chi cuadrado crítico (84.257 > 3.8415), el cual nos permite rechazar la hipótesis nula y aceptar la hipótesis alterna de la investigación, misma que menciona que la deserción estudiantil puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú – Moyobamba.

4.3. Discusiones

La investigación se efectuó en la Institución Educativa 00116 Alto Perú – Soritor – Moyobamba, donde se analizaron fuentes de datos de archivos de nóminas de matrícula y actas de evaluación de los estudiantes comprendidos de los periodos de 2012 a 2022, que fueron extraídos del sistema SIAGIE. En total se analizaron 110 archivos de un total de 412 estudiantes del nivel secundario. Asimismo, mediante procedimientos de ETL (Extract, Transform and Load), se pudo consolidar un dataset con información de 387 estudiantes el cual permitió concretizar con el objetivo del minado. Se determinaron

diversas variables influyentes de la deserción estudiantil, entre ellos las más relevantes se tuvo la situación de matrícula, los cursos aprobados, los cursos desaprobados, el comportamiento, grado culminado y el nacimiento registrado; en ese contexto el estudio de Bedregal et al. (2020), demuestra que la evaluación del rendimiento académico de un estudiante va más allá de simplemente observar las calificaciones obtenidas; es necesario tener en cuenta su comportamiento académico, su desempeño en comparación con sus compañeros de clase y su progreso en la aprobación de asignaturas a lo largo del tiempo. Asimismo, Aluko et al. (2021) menciona que las variables rendimiento académico previo, las calificaciones obtenidas por los estudiantes en las materias tienen el efecto más significativo en el rendimiento académico. Entendiendose que las mismas variables son importantes para determinar la deserción estudiantil, ya que el deficiente rendimiento academico según el estudio esta directamente relacionada con la deserción estudiantil.

La minería de datos se elaboró en base a la metodología CRISP-DM, respetando todos las fases y procedimientos, permitiendo contar con el orden y mejor entendimiento en el análisis de datos; dicha metodología también se utilizaron en las investigaciones de (Quiñones & Carrasco, 2020), (Quiñones et al., 2020), (Pérez, 2020) y (Bedregal-Alpaca et al., 2020). Asimismo se usó el software Weka como plataforma para el aprendizaje automático y la minería de datos, permitiendo concretizar el objetivo de manera satisfactoria; dicha herramienta también fue usada en los estudios de (Quiñones & Carrasco, 2020), (Quiñones et al., 2020), (Chaurasia et al., 2018), (Tan & Lin, 2021) y (Menacho, 2020).

Se evaluaron 5 técnicas de minería de datos, los cuales fueron: el Algoritmo arboles de decisión J48, Algoritmo arboles de decisión RANDOM FOREST, Algoritmo VECINOS MAS CERCANOS, Algoritmo FUNCIÓN LOGÍSTICA y el Algoritmo PERCEPTRÓN MULTICAPA, de los cuales el mejor desempeño lo tuvo el algoritmo J48, concordando con la investigación de Menacho (2020), quien utilizó el algoritmo C4.5 de manera similar, se pudo sugerir al profesorado que considere los cuestionarios y las tareas como actividades que influyen en la posibilidad de que un estudiante repruebe. También guardó relación el estudio de Huatangari & Carrasco (2020) que usó los algoritmos J48graft, J48 y PART, determinados con el software Weka con un porcentaje de clasificación correcta mayor a 83%. Además, Quiñones et al. (2020), tuvo excelentes porcentajes de aciertos haciendo uso de los algoritmos J48, Ridor y PART que han permitido obtener un porcentaje de instancias bien clasificadas de 87.8%. Por otro lado, en los estudios de (Kim et al., 2021), (Ferreira & Camões, 2019), (Chaurasia et al., 2018), (Pérez, 2020), (Yağcı, 2022), (Bedregal-Alpaca et al., 2020) y (Tan & Lin, 2021), también usaron algunas de estas técnicas para predecir, los cuales tuvieron resultados

satisfactorios en sus predicciones. Pérez (2020) sugiere que la experimentación con algoritmos simples puede ser adecuada para lograr altos niveles de precisión. Además, Tan & Lin (2021) demostraron mediante el uso de estos algoritmos que se puede obtener un modelo de predicción que cumple con los requisitos de precisión necesarios para predecir aspectos conductuales del aprendizaje electrónico.

El modelo desarrollado posibilitó la predicción de la deserción estudiantil, ofreciendo así un respaldo para la toma de decisiones en el ámbito educativo. Este planteamiento coincide con la perspectiva de Kim et al. (2021), quienes sostienen que mediante técnicas de minería de datos es factible identificar a los estudiantes en riesgo desde el inicio del proceso de aprendizaje, lo que facilita la elaboración de estrategias educativas y de aprendizaje adaptadas a las necesidades individuales de cada estudiante. Asimismo, Pérez (2020), destaca que los resultados obtenidos mediante técnicas de minería de datos son valiosos para la comunidad académica al contribuir a reducir la deserción estudiantil. Por otra parte, Yağcı (2022) demuestra la viabilidad de utilizar algoritmos de aprendizaje automático para predecir el rendimiento académico de los estudiantes. En una línea similar, Bedregal et al. (2020) afirman que las técnicas de minería de datos son eficaces para identificar patrones y anticipar el comportamiento académico de los estudiantes, lo que permite identificar tempranamente a aquellos en riesgo y adoptar medidas pertinentes en las instituciones educativas.

Por lo tanto, las técnicas de minería de datos han adquirido una relevancia significativa en el sector educativo. Actualmente, numerosas instituciones educativas, tanto escuelas como universidades en todo el mundo, emplean modelos de minería de datos como herramientas de apoyo para su personal estratégico. Este uso no se limita únicamente al ámbito educativo, sino que se extiende a otras áreas, como se evidencia en el estudio de Kaur et al. (2021), en el campo de la salud y en el trabajo de Chacaliaza (2021), en el ámbito comercial. En consecuencia, la minería de datos abarca un espectro amplio de aplicaciones que muchas organizaciones aún no han explorado por completo, a pesar de representar un recurso valioso que, a corto, mediano o largo plazo, podría brindarles ventajas competitivas significativas.

CONCLUSIONES

1. Se aplicaron técnicas de minería de datos para predecir la deserción estudiantil, el cual se tuvo 95.69% de efectividad, a pesar que la deserción es un tema bastante complejo y que depende de muchas variables como factores de aspecto emocional de los estudiantes, factores familiares, factores de salud, migratorio entre otros, tal como se verificó en la revisión bibliográfica, sin embargo, la deserción estudiantil se puede predecirse a través de los datos académicos y sociales.

2. Se analizaron 110 archivos (Nóminas de matrícula y Actas de evaluación) de 412 estudiantes del nivel secundaria, comprendidos de los años 2012 a 2022, que fueron extraídos del sistema SIAGIE en formato pdf, que mediante procedimientos de ETL (Extract, Transform and Load), se pudo consolidar un dataset con información de 387 estudiantes el cual permitió concretizar con el objetivo del minado.

3. Se determinaron diversas variables influyentes de la deserción estudiantil, entre ellos las más relevantes se tuvo la situación de matrícula, los cursos aprobados, los cursos desaprobados, el comportamiento, grado culminado y el nacimiento registrado, donde las prueba estadística Chi Cuadrado de Pearson, arrojó como valor Sig.= 0.000 demostrando la relación con la deserción estudiantil.

4. Se determinaron 5 técnicas de minería de datos, teniendo al Algoritmo arboles de decisión J48, Algoritmo arboles de decisión Random Forest, Algoritmo Vecinos Mas Cercanos, Algoritmo Función Logística y el Algoritmo Perceptrón Multicapa, con clasificaciones de 91.99 %, 91.47 %, 88.63 %, 90.18 % y 88.89 % respectivamente, donde el mejor desempeño lo tuvo el algoritmo J48.

5. Se evaluaron 116 casos, de 96 casos el 3,1% (3) de los casos en los valores reales correspondieron a “No abandona” y en los valores por TDM fue de “Si abandona” existiendo un desacierto. De 20 casos restantes, en los valores reales el 10,0% (2) de los casos fue de “Si abandona” mientras que por TDM fue de “No abandona” habiendo desaciertos. Entonces el modelo basado en técnicas de minería de datos no acertó 5 casos, siendo equivalente al 4.31% del total de los casos.

RECOMENDACIONES

1. A las instancias encargadas de velar por la educación en el país llevar a cabo proyectos de analítica de datos e inteligencia de negocios sobre asuntos educativos, con el objetivo de apoyar significativamente en la toma de decisiones, el cual se verá reflejado en la calidad educativa.

2. A la Dirección Regional de Educación y Unidad de Gestión Educativa Local generar políticas de obtención de recojo de información de los estudiantes, el cual servirá para consolidar una información más precisa, relevante y con ello poder obtener modelos más finos para las predicciones.

3. A las instituciones educativas implantar medidas para recolección de otras variables que repercuten en la deserción estudiantil, siendo importante conocer la situación económica del alumno, las distancias de su casa a la institución educativa, enfermedades que padecen, tiempo dedicado el estudio, etc.

4. A futuros investigadores aplicar otras técnicas de clasificación, agregando más variables con el objetivo de construir nuevos modelos con mayor efectividad en la predicción de la deserción estudiantil.

5. A la Institución Educativa debe de apoyarse con herramientas finales que pueden ser aplicativos o sistemas para mejorar la comprensión de la información y de los datos.

REFERENCIAS BIBLIOGRÁFICAS

Aluko, R. O., Daniel, E. I., Shamisdeen, O., Aigbavboa, C. O., & Akinsola, A. O. (2021). Towards reliable prediction of academic perfomance of architecture students : using data mining techniques. *Journal of Engineering, Design and Technology*, *16*(3), 385–397. https://doi.org/10.1108/JEDT-08-2017-0081

Arias, J., Villasís, M. Á., & Miranda, M. G. (2016). El protocolo de investigación III: la población de estudio. *Revista Alergia Mexico*, *63*(2), 201–206. https://doi.org/10.29262/ram.v63i2.181

Banda, H., & Garza, R. (2014). Aplicación teórica del método Holt-Winters al problema de Credit Scoring de las instituciones de microfinanzas. *Mercados y Negocios*, *15*(2), 5–21. https://dialnet.unirioja.es/descarga/articulo/5811252.pdf

Barrero, F. (2015). Investigación en deserción estudiantil universitaria. *Educación y Desarrollo Social*, *9*(2), 86–101. https://dialnet.unirioja.es/descarga/articulo/5386219.pdf

Basoqain, X. (2008). *Redes neuronales artificiales y sus aplicaciones*. Escuela Superior de Ingeniería de Bilbao, EHU. https://ocw.ehu.eus/pluginfile.php/40137/mod_resource/content/1/redes_neuro/contenidos/pdf/libro-del-curso.pdf

Bedregal-Alpaca, N., Aruquipa-Velazco, D., & Cornejo-Aparicio, V. (2020). Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *27*, 592–604. https://www.proquest.com/docview/2385757429/fulltextPDF/C2C2E769893D4668PQ/1?accountid=37408

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining. from concept to implementation*. Upper Saddle River, NJ: Prentice Hall.

Carrasco, R. (2011). *Data Mining: aplicaciones económico-financieras* (Editorial Academica Espanola (ed.)).

Carvajal, P. (2016). Revisión de estudios sobre deserción estudiantil Ee educación superior en Latinoamerica bajo la perspectiva de Pierre Bourdieu. *Congresos CLABES*, 10. https://revistas.utp.ac.pa/index.php/clabes/article/view/1324

Castro, L. F., Espitia, E., & Cardona, S. A. (2019). Analysis of student desertion in a

systems and Computing Engineering Undergraduate Program. *Revista Colombiana de Computación*, *20*(1), 72–82. https://doi.org/10.29375/25392115.3608

Chacaliaza, J. H. (2021). *Modelo basado en técnicas de minería de datos para la segmentación de clientes en la empresa distribuidora Suministros del oriente SA* [Universidad Nacional de San Martín]. https://repositorio.unsm.edu.pe/handle/11458/4200

Chan, D., & Galli, M. (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. *Raes*, *12*(20), 123–136. https://dialnet.unirioja.es/servlet/articulo?codigo=7592065

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, *12*(2), 119–126. https://doi.org/10.1177/1748301818756225

ComexPeru. (2020). *230,000 estudiantes dejaron de ir al colegio en 2020*. https://www.comexperu.org.pe/articulo/230000-estudiantes-dejaron-de-ir-al-colegio-en-2020

Coronel, F. (2013). Efectos de la migración en el proceso de aprendizaje-enseñanza y su tratamiento desde la escuela. *Revista Integra Educativa*, *6*(1), 57–77. http://www.scielo.org.bo/pdf/rieiii/v6n1/v6n1a04.pdf

Corrientes, D. L. L. R., & Curuzú, C. E. P. (2013). Metodología de estudio del rendimiento académico mediante la minería de datos. *Revista Científica de Tecnología Educativa*, *3*(1). http://www.uajournals.com/campusvirtuales/journal/4/5.pdf

Dirección Regional de Educación de San Martín. (2021). *Plan Regional de Convivencia Escolar de la Dirección Regional de Educación San Martín*. https://dresanmartin.gob.pe/descargar/archivo/documentos/210331114020_resolucion-directoral-regional-n0-0330-2021-grsm-dre.pdf

Espíndola, E., & León, A. (2002). La deserción escolar en América Latina: un tema prioritario para la agenda regional. *Revista Ibero Americana de Educación*, *30*, 39–62. https://doi.org/10.35362/rie300941

Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería Investigación y Tecnología*, *21*(1), 1–17. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Espinoza, M. A. (2018). WEKA, áreas de aplicación y sus algoritmos: una revisión

sistemática de literatura. *Revista Científica ECOCIENCIA*, *5*, 1–26. https://doi.org/10.21855/ecociencia.50.153

Espinoza, O., Castillo, D., González, L. E., & Loyola, J. (2012). Factores familiares asociados a la deserción escolar en Chile. *Revista de Ciencias Sociales*, *18*(1). https://doi.org/10.31876/rcs.v18i1.24967

Ferreira, F., & Camões, A. (2019). Prediction of restrained shrinkage crack width of slag mortar composites using data mining techniques. *Revista Materia*, *24*(4). https://doi.org/10.1590/s1517-707620190004.0852

Flores, G. A., Cadena, J. A., Quinatoa, E. E., & Villa, M. W. (2019). Minería de datos como herramienta estratégica. *Recimundo Revista Científica Mundo de La Investigación y El Conocimiento*, *3*(1), 955–970. https://doi.org/10.26820/recimundo/3.(1).enero.2019.955-970

Franklin, B. J., & Kochan, S. (2000). Collecting and reporting dropout data in Louisiana. *American Education Research*, 18. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.496.365&rep=rep1&type=pdf

Gallego, M. G., Perez de los Cobos, A. P., & Gómez, J. C. (2021). Identifying students at risk to academic dropout in higher education. *Education Sciences*, *11*, 427. https://doi.org/10.3390/educsci11080427

Gómez, H. (1998). *Educación : la agenda del siglo XXI : hacia un desarrollo humano*.

Gutiérrez, J. A., & Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare*, *3*(2), 33–51. https://doi.org/10.21158/23823399.v3.n2.2015.1440

Hernandez, J., Ramirez, J., & Ferri, C. (2004). *Introducción a la mineria de datos*. Pearson.

Hernandez, M., Alvarez, J., & Aranda, A. (2017). El problema de la deserción escolar en la producción científica educativa. *Revista Internacional de Ciencias Sociales y Humanidades, SOCIOTAM*, *27*(1), 89–112. https://www.redalyc.org/pdf/654/65456040007.pdf

Hernández, R., Fernández, C., & Baptista, M. del P. (2014). *Metodología de la investigación* (Mc Graw Hill Education (ed.); 6ta Edició). http://observatorio.epacartagena.gov.co/wp-

content/uploads/2017/08/metodologia-de-la-investigacion-sexta-edicion.compressed.pdf

Hopenhayn, M. (2002). Educar para la sociedad de la información y de la comunicación: una perspectiva latinoamericana. *Revista Iberoamericana de Educación*, *30*, 1–20. https://doi.org/10.35362/rie300946

Joshi, K. P. (1997). Analysis of data mining algorithms. In *Sequential Analysis*. https://mdsoar.org/bitstream/handle/11603/12772/462.pdf?sequence=1&isAllowed=y

Kaur, I., Doja, M. N., Ahmad, T., Ahmad, M., Hussain, A., Nadeem, A., & Abd El-Latif, A. A. (2021). An integrated approach for cancer survival prediction using data mining techniques. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/6342226

Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability (Switzerland)*, *10*(4), 1–18. https://doi.org/10.3390/su10040954

Kim, K., Kim, H. S., Shim, J., & Park, J. S. (2021). A study in the early prediction of ict literacy ratings using sustainability in data mining techniques. *Sustainability (Switzerland)*, *13*(4), 1–11. https://doi.org/10.3390/su13042141

Lázaro, N., Callejas, Z., & Griol, D. (2020). Factores que inciden en la deserción estudiantil en carreras de perfil Ingeniería Informática. *Revista Fuentes*, *1*(22), 105–126. https://doi.org/10.12795/revistafuentes.2020.v22.i1.09

Lozano, D. F., & Maldonado, L. (2020). Asociación entre factores económicos y sociales con la propensión de deserción escolar en colegios militarizados. *Revista de Estudios y Experiencias En Educación*, *19*(40), 35–52. https://doi.org/10.21703/rexe.20201940lozano2

Martínez, C. J., & Palencia, O. (2021). Modelo de minería de datos para el análisis de la productividad y crecimiento personal en las mujeres emprendedoras: el caso de la Asociación las Rosas. *Suma de Negocios*, *12*(26), 23–30. https://doi.org/10.14349/sumneg/2021.v12.n26.a3

Mazo, C. X., & Bedoya, O. F. (2010). PESPAD: una nueva herramienta para la predicción de la estructura secundaria de la proteína basada en árboles de decisión. *Ingeniería Y Competitividad*, *12*(2), 9–22. https://doi.org/10.25100/iyc.v12i2.2690

Meghyasi, H., & Rad, A. (2020). Customer churn prediction in telecommunication industry using data mining methods. *Innovaciencia Facultad De Ciencias Exactas Físicas Y Naturales*, *8*(1), 1–8. https://doi.org/10.15649/2346075X.999

Menacho, C. H. (2020). Técnicas de minería de datos aplicadas a la plataforma educativa Moodle. *Revista Tierra Nuestra*, *14*(1), 137–146. https://doi.org/10.21704/rtn.v14i1.1509

Ministerio de Educación de Chile. (2020). Deserción escolar: diagnóstico y proyección en tiempos de pandemia. In *Documento de trabajo, Centro de Estudios Mineduc* (Vol. 22). https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2020/10/DOCUMENTO-DE-TRABAJO-22_2020_f01.pdf

Ministerio de Educación del Perú. (2022). *Escale*. https://escale.minedu.gob.pe/web/inicio/padron-de-iiee

Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, *1*(2), 281–294. https://doi.org/10.1162/neco.1989.1.2.281

Morales, J., & Vargas, Y. (2018). Determinantes de la deserción escolar y el trabajo adolescente en Bolivia. *Investigacion & Desarrollo*, *18*(2), 93–110. https://doi.org/10.23881/idupbo.018.2-6e

Moreno, D. M. (2013). La deserción escolar: un problema de carácter social. *Revista In Vestigium Ire*, *6*(1), 115–124. http://revistas.ustatunja.edu.co/index.php/ivestigium/article/view/795

Necochea, Y., Nervi, C., Tuesta, V., Olazabal, L., Rodríguez, J., Gastelo, A., & León, F. (2017). Frecuencia y características del abandono estudiantil en una Escuela de Medicina de Lambayeque, 2006-2014. *Revista Medica Herediana*, *28*(3), 171–177. https://doi.org/10.20453/rmh.v28i3.3184

Otzen, T., & Manterola, C. (2017). Técnicas de muestreo sobre una población a estudio. *International Journal of Morphology*, *35*(1), 227–232. https://doi.org/10.4067/S0717-95022017000100037

Pérez, B. R. (2020). Comparison of data mining techniques to identify signs of student desertion, based on academic performance. *Revista UIS Ingenierías*, *19*(1), 193–204. https://doi.org/10.18273/revuin.v19n1-2020018

Perez, C., & Santin, D. (2007). *Minería de datos. Técnicas y herramientas* (S. A. Ediciones Paraninfo (ed.)).

Quinlan, J. R. (1993). Programs for machine learning. Part II. *Machine Learning*, *16*, 235–240. http://server3.eca.ir/isi/forum/Programs for Machine Learning.pdf

Quiñones, L., & Carrasco, Y. (2020). Rendimiento académico empleando minería de datos. *Revista ESPACIOS*, *41*(44), 277–285. https://doi.org/10.48082/espacios-a20v41n44p17

Quiñones, L., Jara, D. M., Alvarado, N., Milla, M. E., & Gamarra, O. A. (2020). Modelo para la estimación de la deserción estudiantil Awajún y Wampis empleando minería de datos. *Revista de Ciencia y Tecnología*, *34*, 45–50. https://doi.org/10.36995/j.recyt.2020.34.006

Quiroz, N. L., & Vlencia, C. A. (2012). Aplicación del proceso de KDD en el contexto de bibliomining el caso Elogim. *Rev. Interam. Bibliot. Medellín (Colombia) Interamericana de Bibliotecología*, *35*(1), 97–108. http://www.scielo.org.co/pdf/rib/v35n1/v35n1a9.pdf

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de datos: conceptos y tendencias. *Revista Iberoamericana de Inteligencia Artificial*, *10*(29), 11–18. https://www.redalyc.org/pdf/925/92502902.pdf

Rivadeneira, J., De La Hoz, A., & Barrera, M. (2020). Análisis general del SPSS y su utilidad en la estadística. *E-IDEA-Journal of Business Sciences*, *2*(4), 17–25. https://revista.estudioidea.org/ojs/index.php/eidea/article/view/19/19

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

Romeu, P. D. (2010). *Minería de datos aplicada al análisis del tratamiento informativo de la drogadicción* [Universidad CEU Cardenal Herrera]. http://dspace.ceu.es/bitstream/10637/5020/1/Minería de datos aplicada al análisis del tratamiento informativo de la drogadicción_Romeu Guallart,Pablo María.pdf

Ruiz, E. M., & Romero, C. P. (2018). Results obtained in a data mining process applied to a database containing bibliographic information concerning four segments of science. *Journal of Information Systems and Technology Management – Jistem USP*, *15*(1), 1–11. https://doi.org/10.4301/S1807-1775201815003

Salazar, J. I., & Girón, E. (2021). Análisis y aplicación de algoritmos de minería de datos. *Perspectivas*, *1*(21), 71–88. https://revistas.uniminuto.edu/index.php/Pers/issue/view/195

Shi, Y., Tang, S., & Li, J. (2020). A two-population extension of the exponential smoothing state space model with a smoothing penalisation scheme. *Risks*, *8*(3), 1–18. https://doi.org/10.3390/risks8030067

Sinchi, E. R., & Gómez, G. P. (2018). Acceso y deserción en las universidades. Alternativas de financiamiento. *ALTERIDAD Revista de Educación*, *13*(2), 274–287. https://doi.org/10.17163/alt.v13n2.2018.10

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85. https://doi.org/10.1007/BF02214313

Tamayo y Tamayo, M. (2003). *El Proceso de la investigacion cientifica* (Limusa Noriega Editores (ed.); 4ta Edicio). https://www.academia.edu/17470765/EL_PROCESO_DE_INVESTIGACION_CIENTIFICA_MARIO_TAMAYO_Y_TAMAYO_1

Tan, C., & Lin, J. (2021). A new QoE-based prediction model for evaluating virtual education systems with COVID-19 side effects using data mining. *Soft Computing*. https://doi.org/10.1007/s00500-021-05932-w

Timaran, S. R., Hernandez, I., Caicedo, S. J., Hidalgo, A., & Alvarado, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, *8*(26), 63–86. https://doi.org/10.16925/9789587600490

Troche, A. (2014). Aplicación de la minería de datos sobre bases de datos transaccionales. *Fides et Ratio - Revista de Difusión Cultural y Científica*, *7*(7), 58–66. http://www.scielo.org.bo/pdf/rfer/v7n7/v7n7_a05.pdf

Urquizo, C. (2014). La educación como estrategia de desarrollo en el Perú. *Revista Psicológica Herediana*, *9*(1–2), 51. https://doi.org/10.20453/rph.v9i1-2.3006

Usama, F., Georges, G., & Andreas, W. (2001). *Information Visualization in Data Mining and Knowledge Discovery* (1ra Ed.). Morgan Kaufmann.

Vallejo, H. F., Guevara, E., & Medina, S. R. (2018). Minería de datos. *Revista Científica Mundo de La Investigación y El Conocimiento*, *2*(Esp), 339–349. https://doi.org/10.26820/recimundo/2.esp.2018.339-349

Valverde, V., Portalanza, N., & Mora, P. (2019). AnÃ¡lisis descriptivo de base de datos

relacional y no relacional. *Cuadernos de Educación y Desarrollo*, *108*, 1–16. https://www.eumed.net/rev/atlante/2019/06/base-datos-relacional.html

Varón, F. (2017). El fenómeno de la deserción escolar en un contexto local: estudio de la política municipal. *Derecho y Politicas Publicas*, *19*(26), 85–97. https://doi.org/10.16925/di.v19i26.1953

Venegas, G., Chiluisa Chiluisa, M., Castro, S., & Casillas, I. (2017). La deserción en la educación. *Boletín Virtual*, *6*(4), 235–239. https://dialnet.unirioja.es/descarga/articulo/6145622.pdf

Vera, L. M., Niño, J. A., Porras, A. M., Durán, J. N., Delgado, P. A., Caballero, M. C., & Navarro, J. P. (2020). Salud mental y deserción en una población universitaria. *Revista Virtual Universidad Católica Del Norte*, *5821*(60), 137–158. https://doi.org/10.35575/rvucn.n60a8

Weiss, S. M., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide (The Morgan Kaufmann Series in Data Management Systems)* (1ra Ed.). California : Morgan Kaufmann Publishers.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(11). https://doi.org/10.1186/s40561-022-00192-z

ANEXOS

Anexo 01: Árbol de problemas

<table>
<tr><th>PROBLEMA</th><th>OBJETIVO</th><th>HIPÓTESIS</th><th>DISEÑO</th><th>POBLACIÓN Y MUESTRA</th><th>VARIABLES</th><th>DIMENSIONES</th><th>INDICADORES</th></tr>
<tr><td rowspan="4">Problema general
¿De qué manera la aplicación de técnicas de minería de datos predice la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba ?</td><td rowspan="4">Objetivo general
Aplicar técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba.
Objetivos específicos
OE₁: Analizar los archivos que registran la deserción estudiantil.
OE₂: Determinar variables influyentes en la deserción estudiantil.
OE₃: Determinar las técnicas de minería de datos a aplicar que mejoren la predicción de la deserción estudiantil.
OE₄: Analizar los resultados obtenidos de la aplicación de técnicas de minería de datos en la predicción de la deserción estudiantil.</td><td rowspan="4">H_i: La deserción estudiantil puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú - Moyobamba.
H₀: La deserción estudiantil no puede predecirse mediante técnicas de minería de datos en la Institución Educativa 0016 Alto Perú - Moyobamba.</td><td rowspan="4">Diseño no experimental: se realizará sin manipular alguna variable de estudio para observar algún cambio en la otra.

Donde:
X: Técnicas de minería de datos
Y: Predicción de la deserción estudiantil</td><td rowspan="4">Población
Conformado por 412 alumnos entre los períodos 2012 – 2022, en la Institución Educativa 0016 Alto Perú – Moyobamba.
Muestra
La muestra se dio en base al muestreo no probabilístico por conveniencia, el cual fue compuesto por 387 estudiantes del 2012 a 2022.</td><td rowspan="2">Técnicas de minería de datos</td><td>Técnicas</td><td>N° técnicas de minería de datos</td></tr>
<tr><td>Factores</td><td>Sociales
Académicos</td></tr>
<tr><td rowspan="2">Deserción estudiantil</td><td>Predicción</td><td>Confiabilidad de la predicción</td></tr>
<tr><td>Estimación</td><td>Tiempo para generar estimación</td></tr>
</table>

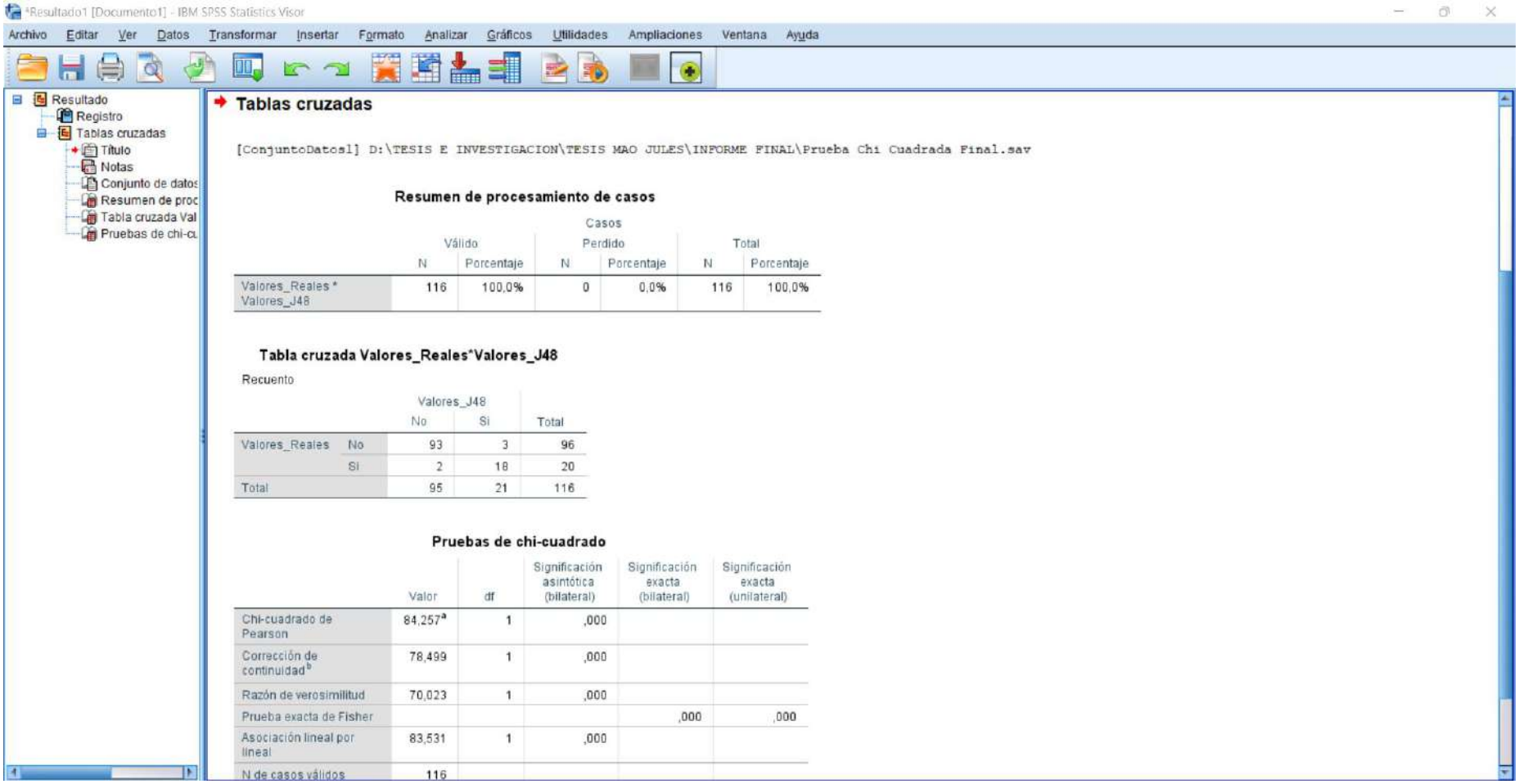
Anexo 02: Tabla de distribución Chi Cuadrado de Pearson

TABLA 3-Distribución Chi Cuadrado χ^2

P = Probabilidad de encontrar un valor mayor o igual que el chi cuadrado tabulado, ν = Grados de Libertad

ν/p	0,001	0,0025	0,005	0,01	0,025	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
1	10,8274	9,1404	7,8794	6,6349	5,0239	3,8415	2,7055	2,0722	1,6424	1,3233	1,0742	0,8735	0,7083	0,5707	0,4549
2	13,8150	11,9827	10,5965	9,2104	7,3778	5,9915	4,6052	3,7942	3,2189	2,7726	2,4079	2,0996	1,8326	1,5970	1,3863
3	16,2660	14,3202	12,8381	11,3449	9,3484	7,8147	6,2514	5,3170	4,6416	4,1083	3,6649	3,2831	2,9462	2,6430	2,3660
4	18,4662	16,4238	14,8602	13,2767	11,1433	9,4877	7,7794	6,7449	5,9886	5,3853	4,8784	4,4377	4,0446	3,6871	3,3567
5	20,5147	18,3854	16,7496	15,0863	12,8325	11,0705	9,2363	8,1152	7,2893	6,6257	6,0644	5,5731	5,1319	4,7278	4,3515
6	22,4575	20,2491	18,5475	16,8119	14,4494	12,5916	10,6446	9,4461	8,5581	7,8408	7,2311	6,6948	6,2108	5,7652	5,3481
7	24,3213	22,0402	20,2777	18,4753	16,0128	14,0671	12,0170	10,7479	9,8032	9,0371	8,3834	7,8061	7,2832	6,8000	6,3458
8	26,1239	23,7742	21,9549	20,0902	17,5345	15,5073	13,3616	12,0271	11,0301	10,2189	9,5245	8,9094	8,3505	7,8325	7,3441
9	27,8767	25,4625	23,5893	21,6660	19,0228	16,9190	14,6837	13,2880	12,2421	11,3887	10,6564	10,0060	9,4136	8,8632	8,3428
10	29,5879	27,1119	25,1881	23,2093	20,4832	18,3070	15,9872	14,5339	13,4420	12,5489	11,7807	11,0971	10,4732	9,8922	9,3418
11	31,2635	28,7291	26,7569	24,7250	21,9200	19,6752	17,2750	15,7671	14,6314	13,7007	12,8987	12,1836	11,5298	10,9199	10,3410
12	32,9092	30,3182	28,2997	26,2170	23,3367	21,0261	18,5493	16,9893	15,8120	14,8454	14,0111	13,2661	12,5838	11,9463	11,3403
13	34,5274	31,8830	29,8193	27,6882	24,7356	22,3620	19,8119	18,2020	16,9848	15,9839	15,1187	14,3451	13,6356	12,9717	12,3398
14	36,1239	33,4262	31,3194	29,1412	26,1189	23,6848	21,0641	19,4062	18,1508	17,1169	16,2221	15,4209	14,6853	13,9961	13,3393
15	37,6978	34,9494	32,8015	30,5780	27,4884	24,9958	22,3071	20,6030	19,3107	18,2451	17,3217	16,4940	15,7332	15,0197	14,3389
16	39,2518	36,4555	34,2671	31,9999	28,8453	26,2962	23,5418	21,7931	20,4651	19,3689	18,4179	17,5646	16,7795	16,0425	15,3385
17	40,7911	37,9462	35,7184	33,4087	30,1910	27,5871	24,7690	22,9770	21,6146	20,4887	19,5110	18,6330	17,8244	17,0646	16,3382
18	42,3119	39,4220	37,1564	34,8052	31,5264	28,8693	25,9894	24,1555	22,7595	21,6049	20,6014	19,6993	18,8679	18,0860	17,3379
19	43,8194	40,8847	38,5821	36,1908	32,8523	30,1435	27,2036	25,3289	23,9004	22,7178	21,6891	20,7638	19,9102	19,1069	18,3376
20	45,3142	42,3358	39,9969	37,5663	34,1696	31,4104	28,4120	26,4976	25,0375	23,8277	22,7745	21,8265	20,9514	20,1272	19,3374
21	46,7963	43,7749	41,4009	38,9322	35,4789	32,6706	29,6151	27,6620	26,1711	24,9348	23,8578	22,8876	21,9915	21,1470	20,3372
22	48,2676	45,2041	42,7957	40,2894	36,7807	33,9245	30,8133	28,8224	27,3015	26,0393	24,9390	23,9473	23,0307	22,1663	21,3370
23	49,7276	46,6231	44,1814	41,6383	38,0756	35,1725	32,0069	29,9792	28,4288	27,1413	26,0184	25,0055	24,0689	23,1852	22,3369
24	51,1790	48,0336	45,5584	42,9798	39,3641	36,4150	33,1962	31,1325	29,5533	28,2412	27,0960	26,0625	25,1064	24,2037	23,3367
25	52,6187	49,4351	46,9280	44,3140	40,6465	37,6525	34,3816	32,2825	30,6752	29,3388	28,1719	27,1183	26,1430	25,2218	24,3366
26	54,0511	50,8291	48,2898	45,6416	41,9231	38,8851	35,5632	33,4295	31,7946	30,4346	29,2463	28,1730	27,1789	26,2395	25,3365
27	55,4751	52,2152	49,6450	46,9628	43,1945	40,1133	36,7412	34,5736	32,9117	31,5284	30,3193	29,2266	28,2141	27,2569	26,3363
28	56,8918	53,5939	50,9936	48,2782	44,4608	41,3372	37,9159	35,7150	34,0266	32,6205	31,3909	30,2791	29,2486	28,2740	27,3362
29	58,3006	54,9662	52,3355	49,5878	45,7223	42,5569	39,0875	36,8538	35,1394	33,7109	32,4612	31,3308	30,2825	29,2908	28,3361

Anexo 03: Prueba Chi Cuadrado de Pearson



Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba

por Jules Mao Flores Satalaya

Fecha de entrega: 14-may-2024 12:22p.m. (UTC-0500)
Identificador de la entrega: 2373422413
Nombre del archivo: tesis_mao_unsm_obs_lev_turnitin_ult.docx (8.02M)
Total de palabras: 20566
Total de caracteres: 116041

Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la Institución Educativa 00116 Alto Perú - Moyobamba

INFORME DE ORIGINALIDAD

22%	22%	4%	10%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.unsaac.edu.pe Fuente de Internet	3%
2	repositorio.uss.edu.pe Fuente de Internet	2%
3	repositorio.unsm.edu.pe Fuente de Internet	2%
4	tesis.unsm.edu.pe Fuente de Internet	2%
5	Submitted to Universidad Nacional de San Martín Trabajo del estudiante	1%
6	uifisi.unsm.edu.pe Fuente de Internet	1%
7	repositorio.undac.edu.pe Fuente de Internet	1%
8	dokumen.tips Fuente de Internet	1%