UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TÍTULO DEL TRABAJO DE TITULACIÓN

"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA ESCUELA DE EDUCACIÓN BÁSICA "26 DE SEPTIEMBRE"

AUTOR

Aguirre Chamba Kelvin Roosbelth

PROYECTO UIC

Previo a la obtención del grado académico en
INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

TUTOR

Ing. Rosero Vásquez Shendry Balmore. Ms.CC

Santa Elena, Ecuador

Año 2024

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TRIBUNAL DE SUSTENTACIÓN

Firmado electrónicamente por:
JOSE MIGUEL SANCHEZ AQUINO

Ing. Jose Sanchez A. Msc.
DIRECTOR DE LA CARRERA

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente porSHENDRY BALMORE ROSERO VASQUEZ
DN: cn=SHENDRY BALMORE ROSERO VASQUEZ gn=SHENDRY BALMORE c=EC
Motivo:Soy el autor de este documento
Ubicación:
Fecha:2024-01-22 11:47-05:00

Ing. Shendry Rosero Vásquez. Ms.CC
TUTOR

Firmado electrónicamente por:
ALICIA GERMANIA ANDRADE VERA

Ing. Alicia Andrade Vera. Msc.
DOCENTE ESPECIALISTA

Firmado electrónicamente por:
MARJORIE ALEXANDRA CORONEL SUAREZ

Ing. Marjorie Coronel S. Mgti.
DOCENTE GUÍA UIC

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por Aguirre Chamba Kelvin Roosbelth, como requerimiento para la obtención del título de Ingeniero en Tecnologías de la Información.

La Libertad, a los 11 días del mes de diciembre del año 2023

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente por SHENDRY BALMORE ROSERO VASQUEZ
DN: cn=SHENDRY BALMORE ROSERO VASQUEZ gn=SHENDRY BALMORE c=EC
Motivo: Soy el autor de este documento
Ubicación:
Fecha:2023-12-11 12:25-05:00

Ing. Shendry Rosero Vásquez. Ms.CC

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Aguirre Chamba Kelvin Roosbelth**

DECLARO QUE:

El trabajo de Titulación, Aplicación de técnicas de minería de datos para la predicción del rendimiento académico de los estudiantes de la Escuela de Educación Básica "26 de septiembre" previo a la obtención del título en Ingeniero en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

La Libertad, a los 11 días del mes de diciembre del año 2023

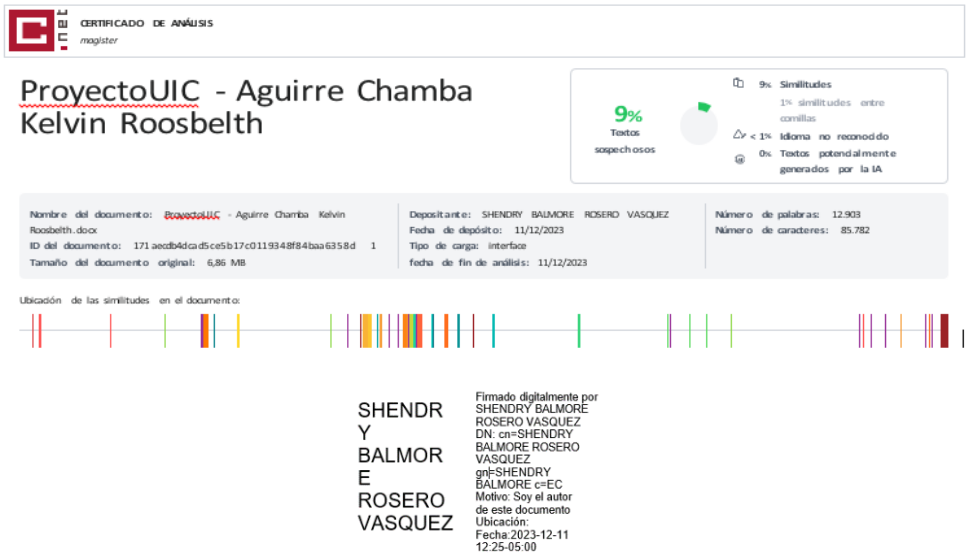
Aguirre Chamba Kelvin Roosbelth

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA

FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN DE ANTIPLAGIO

Certifico que después de revisar el documento final del trabajo de titulación denominado "Aplicación de técnicas de minería de datos para la predicción del rendimiento académico de los estudiantes de la Escuela de Educación Básica 26 de septiembre", presentado por el estudiante, Aguirre Chamba Kelvin Roosbelth fue enviado al Sistema Antiplagio, presentando un porcentaje de similitud correspondiente al 9%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.


CERTIFICADO DE ANÁLISIS
magister

ProyectoUIC - Aguirre Chamba Kelvin Roosbelth

9% Textos sospechosos

9% Similitudes
1% similitudes entre comillas
< 1% Idioma no reconocido
0% Textos potencialmente generados por la IA

Nombre del documento: ProyectoUIC - Aguirre Chamba Kelvin Roosbelth.docx
ID del documento: 171aecdb4dcad5ce5b17c0119348f84baa6358d1
Tamaño del documento original: 6,86 MB

Depositante: SHENDRY BALMORE ROSERO VASQUEZ
Fecha de depósito: 11/12/2023
Tipo de carga: interface
fecha de fin de análisis: 11/12/2023

Número de palabras: 12.903
Número de caracteres: 85.782

Ubicación de las similitudes en el documento:

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente por SHENDRY BALMORE ROSERO VASQUEZ
DN: cn=SHENDRY BALMORE ROSERO VASQUEZ, gn=SHENDRY BALMORE c=EC
Motivo: Soy el autor de este documento
Ubicación:
Fecha:2023-12-11 12:25-05:00

Ing. Shendry Rosero Vásquez. Ms.CC

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

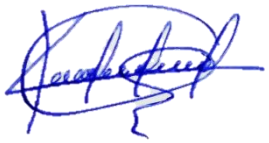
AUTORIZACIÓN

Yo, **Aguirre Chamba Kelvin Roosbelth**

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales del presente trabajo de titulación con fines de difusión pública, además apruebo la reproducción de este trabajo de titulación dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor.

Santa Elena, a los 11 días del mes de diciembre del año 2023



Aguirre Chamba Kelvin Roosbelth

AGRADECIMIENTO

A mis padres y hermanos, por ser mi pilar fundamental a lo largo de mi carrera, por su apoyo incondicional en cada etapa de mi vida y por cada valor de bien que han sembrado en mí para seguir creciendo personalmente.

A mis docentes, ingeniera Marjorie Coronel e ingeniero Shendry Rosero, por brindarme su guía con sus conocimientos para desarrollar este trabajo, por su paciencia y enseñanzas.

Kelvin Roosbelth Aguirre Chamba

DEDICATORIA

Dedico este trabajo a mis padres, Roosbelth y Rocío, y a mis hermanos, Jefferson y Jeremy, por cada esfuerzo que realizan por mí y por la familia. Que sepan que lo valoro mucho y esta es mi forma de retribuirles y agradecerles por todo.

También, dedico este trabajo a mi novia, Alennis Borbor, por cada palabra de aliento en momentos de dificultad durante la ejecución del proyecto.

Kelvin Roosbelth Aguirre Chamba

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La escuela '26 de septiembre.' ha establecido metas orientadas al desarrollo estratégico, que incluyen la expansión de su infraestructura y la obtención de reconocimientos por su destacada calidad académica. Para alcanzar estos objetivos, se ha dado prioridad a la evaluación de la población estudiantil a través de indicadores como el rendimiento académico. Sin embargo, se enfrentan a limitaciones en la evaluación de este indicador debido a la falta de conocimiento óptimo para la toma de decisiones y a la carencia de una gestión y organización adecuada de la información. A pesar de ser una institución privada, la falta de información estructurada ha llevado a resultados poco convincentes debido a la parcialidad y subjetividad.

Con el propósito de abordar este problema, se propuso la aplicación de técnicas de minería de datos utilizando la metodología de Descubrimiento de Conocimiento en Base de Datos (KDD). Este enfoque consta de cinco fases, desde la recolección e integración de datos hasta la minería de datos, con el objetivo de obtener conocimientos valiosos para los administradores de la institución. Se ha empleado la herramienta Microsoft Visual Studio, reconocida por su capacidad para generar procesos de extracción, transformación y carga de datos, facilitando así la integración adecuada de la información. En cuanto a la creación de modelos de minería de datos, se ha destacado el uso de Jupyter Notebook y Orange Data Mining.

Entre los resultados alcanzados, se destaca la implementación de herramientas de inteligencia de negocios para crear un almacén de datos que resuelva problemas de gestión y análisis empresarial. Además, se aplicaron técnicas de minería de datos, como árboles de decisión de regresión, redes neuronales y vectores de soporte mediante la creación de modelos de regresión para predecir el rendimiento académico estudiantil.

Palabras claves: minería de datos, rendimiento académico, almacén de datos.

ABSTRACT

The '26 de septiembre' school has set goals focused on strategic development, including expanding its infrastructure and earning recognition for its outstanding academic quality. To achieve these objectives, priority has been given to evaluating the student population through indicators such as academic performance. However, they face limitations in assessing this indicator due to a lack of optimal knowledge for decision-making and a deficiency in proper information management and organization. Despite being a private institution, the lack of structured information has led to less convincing results due to partiality and subjectivity.

To address this issue, the application of data mining techniques using the Knowledge Discovery in Databases (KDD) methodology has been proposed. This approach consists of five phases, from data collection and integration to data mining, with the aim of obtaining valuable insights for the institution's administrators. Microsoft Visual Studio has been employed as the tool, known for its capability to generate processes for data extraction, transformation, and loading, facilitating the proper integration of information. Regarding the creation of data mining models, the use of Jupyter Notebook and Orange Data Mining has been highlighted.

Among the achieved results, the implementation of business intelligence tools stands out, creating a data warehouse to address management and business analysis issues. Additionally, data mining techniques such as regression decision trees, neural networks, and support vector machines were applied by creating regression models to predict student academic performance.

Keywords: data mining, academic performance, data warehouse

INTRODUCCIÓN

En la constante búsqueda de la excelencia y el avance, las instituciones se enfocan en la mejora continua, un motor esencial para su progreso y adaptabilidad. Este compromiso arraigado refleja la aspiración de las organizaciones por superarse a sí mismas, solventar sus necesidades y elevar constantemente sus estándares de eficiencia y efectividad. De tal manera que, la relevancia del uso de información histórica en las instituciones se plantea como un pilar fundamental en la toma de decisiones y la planificación estratégica. Así, resalta la escuela "26 de septiembre", ubicada en el cantón La Libertad de la provincia de Santa Elena.

La evaluación interna como lo es el aspecto del rendimiento académico demanda que la institución cuente con información detallada y precisa, estableciendo una base de apoyo para la toma de decisiones posteriores. De esta manera, surge como propuesta la implementación de técnicas de minería de datos para facilitar el análisis de gran cantidad de información, beneficiando la toma de decisiones y la gestión y comunicación efectiva de los datos. Por lo tanto, se propone la aplicación de estos procesos analíticos de información con el propósito de lograr la predicción del rendimiento académico de los estudiantes de la institución.

Durante la revisión de trabajos enfocados a esta área, se encuentra "Predicción del desempeño de los estudiantes utilizando técnicas de clasificación", en el cual se puede destacar el uso de técnicas de clasificación. No obstante, los datos utilizados se limitan a una sola asignatura, lo que puede no ser representativo del rendimiento general del estudiante. Otro estudio desarrollado por Eiriku Yamao de la Universidad de San Martín de Porres "Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas", se basa en fuentes de datos con variables demográficas, académicas y socioeconómicas.

El trabajo de titulación se encuentra estructurado por dos capítulos, detallados a continuación:

En el primer capítulo, se aborda la descripción de los antecedentes englobando las principales problemáticas que afectan a la institución. De esta manera, se destaca principalmente la carencia de información estructurada para el desarrollo óptimo de los procesos administrativos educativos, y a su vez, la evaluación del rendimiento académico de los estudiantes. En este contexto, se establecieron los objetivos y la metodología, optando por la aplicación de la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD).

En el capítulo 2, se detallan las herramientas empleadas en el desarrollo de cada etapa del proyecto, basándose en la metodología planteada anteriormente. Dicha metodología está compuesta por cinco fases, las cuales son: la primera, el proceso de recopilación e integración de información. La segunda, que corresponde a la creación de un almacén de datos o data warehouse. La tercera, que hace mención a las técnicas de minería de datos empleadas para el análisis del conjunto de datos. La cuarta que abarca la evaluación de los modelos mediante las métricas de medición de desempeño, tales como error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE), error absoluto medio (MAE), media del error absoluto en porcentaje (MAPE) y coeficiente de determinación (R2). La quinta, que corresponde a la difusión del conocimiento obtenido a las autoridades y administrativos de la institución. Finalmente, se presentan los resultados obtenidos tras la ejecución de cada etapa del proyecto.

CAPÍTULO 1. FUNDAMENTACIÓN

1.1.Antecedentes

La falta de información estructurada dificulta la identificación de problemas y la toma de decisiones en las organizaciones [1]. Esta problemática se ve reflejada en la dificultad para identificar los problemas que existen dentro de las organizaciones, lo que afecta la capacidad de los líderes para tomar decisiones informadas. La falta de información estructurada y accesible en torno al rendimiento académico puede dificultar la identificación de fortalezas y debilidades en el sistema educativo, así como limitar la capacidad de los maestros y administradores escolares para tomar decisiones informadas basadas en datos [2].

La Escuela de Educación Básica "26 de septiembre" se encuentra ubicada en el cantón La Libertad de la provincia de Santa Elena. Fue fundada el 4 de abril de 2005 y actualmente cuenta con un total de 11 trabajadores entre personal docente y administrativo; y en la cual asisten un total de 263 estudiantes. En este contexto, debido a la falta de una adecuada gestión y organización de la información, se da origen a problemas en la colaboración y afectaciones en el aspecto del tiempo y la economía para las actividades de la escuela (Ver Anexo 1).

Los métodos de recolección de información para iniciar el proceso de investigación se basan en entrevistas con el personal administrativo (Ver Anexo 2), de tal manera, que para llevar a cabo los análisis pertinentes y poder realizar predicciones coherentes, se requerirá también del manejo de registros de datos de los estudiantes que proporcionen los detalles y conocimientos necesarios acerca de los mismos.

La falta de información estructurada y organizada sobre el desempeño de los estudiantes es una problemática significativa en el ámbito educativo. A menudo, los datos relacionados con el rendimiento de los alumnos se encuentran dispersos en diferentes sistemas, formatos y fuentes, lo que dificulta su recopilación y análisis eficiente. Esta falta de estructura y coherencia en los datos dificulta a los docentes y

administradores escolares obtener una visión completa y precisa del progreso académico de los estudiantes.

Por esta razón, la inadecuada evaluación del desempeño académico por la falta de datos confiables puede tener graves consecuencias en el sistema educativo. Sin la evaluación adecuada, no se pueden identificar las fortalezas y debilidades de los estudiantes, lo que puede afectar su capacidad para aprender y desarrollarse en su educación futura. Además, sin una evaluación apropiada, los docentes pueden no tener la información necesaria para adaptar su enseñanza y brindar el apoyo adecuado a los estudiantes que lo necesitan. Esto puede resultar en un aumento de la brecha de aprendizaje entre los estudiantes y una disminución en la calidad de la educación en general.

Las limitaciones tecnológicas en el entorno educativo también dificultan la entrega de retroalimentación efectiva y oportuna. Los sistemas de gestión del aprendizaje carecen de herramientas adecuadas para proporcionar comentarios individualizados, lo que deja a los docentes con opciones limitadas para ofrecer una retroalimentación significativa y personalizada a sus alumnos. Esta situación plantea un desafío en el esfuerzo por mejorar el rendimiento académico y garantizar una educación de calidad para cada estudiante.

De tal modo, se puede determinar que la Escuela de Educación Básica "26 de septiembre" enfrenta una serie de problemas debido a la falta de una adecuada gestión y organización de la información del desempeño estudiantil. La dispersión de datos, la falta de herramientas tecnológicas y la limitada capacidad para proporcionar retroalimentación efectiva y personalizada son obstáculos que dificultan el proceso educativo.

Con respecto al ámbito mundial, sobresale el trabajo “Predicción del desempeño de los estudiantes utilizando técnicas de clasificación” desarrollado por Ahmed Alsanad y Toqa Aiman Mukheimer, pertenecientes a King Saud University de Arabia Saudita [3]. La cual tuvo como objetivo predecir el rendimiento académico de los estudiantes utilizando técnicas de clasificación. Se recolectaron datos de una universidad en Asia,

incluyendo su historial académico y otras características personales. Se aplicaron varios algoritmos de clasificación, para construir modelos predictivos del rendimiento académico y se evaluaron los resultados utilizando dos medidas de evaluación de modelos. No obstante, los datos utilizados se limitaron a una sola asignatura, lo que puede no ser representativo del rendimiento general del estudiante.

A nivel de Latinoamérica, según el estudio realizado por Eiriku Yamao de la Universidad de San Martín de Porres, titulado "Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas"[4]. En el cual se utilizó una base de datos que incluía variables demográficas, académicas y socioeconómicas de los estudiantes, y se aplicaron técnicas de análisis exploratorio de datos y modelos de regresión lineal y redes neuronales.

Los resultados mostraron que las variables académicas y socioeconómicas influyen en el rendimiento académico estudiantil, y que la red neuronal fue el modelo que mejor predijo el rendimiento académico. Sin embargo, la limitante de este trabajo es que solo se utilizó datos recopilados de un semestre en particular, por lo que no se podrá verificar la efectividad del modelo en un rango de estudio más amplio.

Desde la perspectiva local, es relevante el trabajo “Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca” de Cesar Gabriel Loja Rodas [5]. La limitación de este estudio está en que utilizó técnicas de minería de datos convencionales, lo que podría limitar la capacidad del modelo para detectar patrones complejos y no lineales en los datos.

Por tanto, la falta o inadecuada evaluación del desempeño académico puede tener graves consecuencias en el sistema educativo y puede afectar negativamente tanto a los estudiantes como a los docentes. Es importante que se realicen evaluaciones regulares y apropiadas para asegurar que los estudiantes estén recibiendo la educación que necesitan para tener éxito en su futuro académico y profesional.

1.2.Descripción del proyecto

En las instalaciones de la Escuela de Educación Básica “26 de septiembre” se ha identificado que existen inconvenientes en la organización y gestión de la información debido a la falta de una herramienta de apoyo para el análisis de datos. En vista del requerimiento de optimizar el proceso de toma de decisiones y fortalecer el rendimiento académico estudiantil, se sugiere aplicar minería de datos con técnicas de aprendizaje supervisado. Esto permitirá predecir su desempeño y tomar medidas más efectivas. Al analizar datos relevantes, se identificarán patrones que influyen en el éxito estudiantil, promoviendo un entorno educativo más eficiente.

La propuesta tecnológica está conformada por cinco fases, las cuales están basadas en la metodología KDD (Knowledge Discovery in Databases) o Descubrimiento de Conocimiento en Bases de Datos, presentada por U. Fayyad, G. Piatestsky-Shapiro y P. Smyth [6], se detalla las actividades de cada fase a continuación:

Fase 1: Construcción del conjunto de datos.

- Analizar la información mediante entrevistas con el personal administrativo de la institución, con el objetivo de recopilar datos relevantes para el estudio.
- Implementar la extracción de datos desde fuentes como hojas de cálculo de Excel, las cuales poseen información personal de los estudiantes, representantes y docentes, como: cédula, apellidos, nombres, dirección, fecha de nacimiento, entre otros. Así mismo, datos sobre calificaciones, promedios y comportamiento.
- Implementar la extracción de información de los Registros Administrativos del Ministerio de Educación – base de datos AMIE (Archivo Maestro de Instituciones Educativas) respecto a la unidad educativa.
- Implementar una nueva base de datos en el ambiente de Microsoft SQL Server, donde se propondrá un diagrama de la misma para organizar la información de manera más eficaz.

- Implementar la importación de registros desde las hojas de cálculo de Excel a la nueva base de datos, utilizando SQL Server Management Studio para garantizar un manejo eficiente de los datos y facilitar su gestión.

Fase 2: Creación de un almacén de datos (Data warehouse).

- Implementar técnicas para corregir errores en el conjunto de datos extraído, con el fin de analizar la información y eliminar los datos sobrantes.
- Diseñar un data warehouse, el cual permitirá consolidar y organizar la información relevante para el estudio.
- Implementar procesos ETL (Extracción, transformación y carga) con el objetivo de obtener un conjunto integral y coherente de datos. Estos procesos permitirán la preparación adecuada de la información para su posterior análisis y ejecución de las técnicas de minería de datos.

Fase 3: Implementación de técnicas de minería de datos.

- Implementar técnicas de minería de datos para realizar análisis predictivo del rendimiento académico de los estudiantes.
- Analizar patrones y correlaciones significativas para determinar las características de interés que estén relacionados de forma directa con el problema a resolver.
- Implementar la transformación de los datos para prepararlos de manera óptima y aplicar el proceso de minería correspondiente.
- Implementar técnicas de aprendizaje automático supervisado para el análisis de los datos, tales como: redes neuronales, árboles de decisión, máquina de vector de soporte.

Fase 4: Evaluación de modelos y análisis de resultados.

- Implementar técnicas de medición del desempeño, tales como: error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio

(RMSE), media del error absoluto en porcentaje (MAPE) y coeficiente de determinación (R^2)

- Implementar dichas técnicas con el fin de realizar comparaciones y determinar que algoritmo posee el menor error para generar predicciones y lograr resultados más efectivos.
- Si se detectan inconsistencias durante el proceso, será esencial identificar las posibles causas y realizar la corrección de las mismas para garantizar la integridad de los resultados y asegurar la precisión y confiabilidad de los análisis posteriores.

Fase 5: Difusión de conocimiento.

- Implementar capacitaciones dirigidas a los administradores de la institución, compartiendo los resultados y hallazgos derivados de los análisis de minería de datos realizados.
- Analizar recursos y proporcionar orientación a los educadores con las herramientas necesarias para comprender y utilizar eficazmente la información analizada
- Implementar estrategias para fomentar la comprensión de los patrones identificados, las tendencias y las posibles implicaciones para la mejora del rendimiento académico de los estudiantes.

Para la realización del proyecto se han considerado las siguientes herramientas:

- ✓ **Python:** Es uno de los lenguajes de codificación líderes en la actualidad para la analítica de datos.
- ✓ **Excel:** Herramienta altamente eficiente que permite procesamiento de gran volumen de datos para extraer información de valor.
- ✓ **Jupyter Notebook:** Aplicación web original para crear y compartir documentos computacionales.

- ✓ **MatplotLib**: Biblioteca de Python que permite la creación de gráficos estáticos e interactivos.
- ✓ **Keras:** Es una API de aprendizaje profundo escrita en Python para redes neuronales.
- ✓ **SQL Server Developer:** Es una edición gratuita con todas las funciones, con licencia para su uso como base de datos de prueba y desarrollo en un entorno que no sea de producción.
- ✓ **SQL Server Management Studio:** Es una plataforma que permite la gestión de infraestructura relacionada con SQL.
- ✓ **Microsoft Visual Studio:** Es una plataforma de lanzamiento creativa que se puede utilizar para editar, depurar y compilar código.
- ✓ **Scikit-learn:** Herramienta de código abierto para el análisis predictivo de datos.
- ✓ **Orange Data Mining:** Es un software que cuenta con múltiples funciones para procesos de minería de datos.

Con base a la resolución RCF-FST-SO-09 No. 03-2021, el proyecto contribuye a la línea de investigación de Tecnología y Sistemas de la Información (TSI), con sub-línea de investigación Inteligencia Computacional [7].

1.3.Objetivos del Proyecto

Objetivo General

Implementar técnicas de minería de datos para el establecimiento de patrones y tendencias en el desempeño académico estudiantil a través de la exploración y verificación de modelos predictivos.

Objetivos Específicos

- Analizar técnicas que permitan la recolección de información de fuentes confiables y de calidad para la identificación de factores que influyen en el rendimiento académico.
- Diseñar un almacén de datos mediante una herramienta que integre los procesos de extracción, transformación y carga (ETL) de manera eficiente.
- Implementar técnicas de análisis de datos y descubrimiento de patrones utilizando herramientas de aprendizaje automático como árboles de decisión, redes neuronales y máquinas de vector de soporte.

1.4.Justificación del Proyecto

La importancia de la minería de datos como una herramienta estratégica en las empresas para mejorar la disponibilidad de la información es cada vez más importante [8]. Los datos cada vez cobran más vida y se han convertido en información vital y estratégica para la toma de decisiones [9]. En tal sentido, las empresas han venido evolucionando y han querido agregarle valor a la gran cantidad de información que tienen almacenada en sus bases de datos. Para ello, se han interesado en automatizar los procesos y poder así descubrir información valiosa, que de otra manera seguiría siendo subutilizada o simplemente desperdiciada [9].

Por otro lado, la importancia de la minería de datos para descubrir patrones y tendencias desde la gestión del conocimiento (GC) en las empresas de hoy, es fundamental dadas las situaciones de competitividad y desarrollo que deben enfrentar éstas [10]. Gestionar la información en las empresas es, hoy en día, una herramienta clave para poder sobrevivir en un mercado en constante cambio. [11]. Riquelme, Ruíz y Gilbert mencionan que “todo apunta a que más temprano que tarde la minería de datos será usada por la sociedad, al menos con el mismo peso que actualmente tiene la Estadística” [12].

La Escuela de Educación Básica “26 de septiembre” posee la necesidad de contar con información estructurada para obtener ayuda en la toma de decisiones mediante el análisis de la misma. En virtud de esto, se propone aplicar técnicas de minería de datos como una herramienta para generar predicciones sobre el rendimiento académico de los estudiantes, basándose en la recopilación de información de diversas fuentes para determinar una evaluación más profunda y rigurosa, maximizando la efectividad de los resultados obtenidos.

Contar con información ordenada en una institución educativa es fundamental para el análisis y toma de decisiones efectivas. El acceso a datos organizados y actualizados proporcionará una serie de beneficios y puntos positivos que contribuirán al mejor funcionamiento de la institución y al éxito de sus estudiantes. Así mismo, permitirá a los responsables educativos tener una visión clara y precisa del desempeño de los estudiantes, tanto a nivel individual como colectivo. Con datos bien estructurados sobre el rendimiento académico, la asistencia y otros aspectos relevantes, será posible identificar patrones y tendencias que ayudarán a comprender las fortalezas y debilidades de los estudiantes, así como las áreas en las que se requerirá intervención o apoyo adicional.

De tal modo, que conllevará a una comunicación más efectiva y transparente entre los diferentes actores educativos, como docentes, directivos, padres y estudiantes. Al tener acceso a datos actualizados, cada parte interesada podrá tomar decisiones basadas en hechos concretos y tener una comprensión clara de la situación. Siendo así que, los reportes generados mediante la minería de datos serán una herramienta poderosa para brindar una comprensión profunda del rendimiento estudiantil y respaldar la mejora continua de la calidad educativa de la institución.

El tema propuesto se encuentra en concordancia con los objetivos del Plan de Creación de Oportunidades 2021-2025 de Ecuador, particularmente en relación al siguiente eje:

Eje 2.- Eje Social.

Objetivo 7.- "Potenciar las capacidades de la ciudadanía y promover una educación innovadora, inclusiva y de calidad en todos los niveles" [13].

Política 7.2.- "Promover la modernización y eficiencia del modelo educativo por medio de la innovación y el uso de herramientas tecnológicas" [13].

1.5.Alcance del Proyecto

El presente proyecto permitirá recopilar, analizar y obtener información basada en la aplicación de distintas técnicas de minería de datos acerca del rendimiento académico de los estudiantes de la Escuela de Educación Básica "26 de septiembre" del Cantón La Libertad con el fin de mejorar la gestión de la información y la obtención de reportes de los estudiantes sobre el rendimiento académico de forma eficaz.

El presente proyecto abarcará las siguientes fases:

- Fase de construcción del conjunto de datos.

En esta fase, se analizará información relevante a través de entrevistas al personal administrativo, se realizará la extracción de datos de hojas de Excel (conteniendo información personal de estudiantes, representantes y docentes) y la base de datos AMIE (conteniendo datos estadísticos sobre la institución) del Ministerio de Educación. Dicha información recolectada será almacenada en una nueva base de datos generada en SQL Server.

- Fase de creación de un almacén de datos (Data warehouse).

En esta fase, se implementarán técnicas para corregir errores en el conjunto de datos extraído, eliminando datos innecesarios. Posteriormente, se creará un data warehouse el cual estará organizado con información de importancia para el estudio, importando los datos mediante procesos ETL.

- Fase de aplicación de técnicas de minería de datos.

En esta fase, se implementará la búsqueda de patrones y tendencias con el fin de determinar las características de interés para el estudio, aplicando tres técnicas de

minería de datos (aprendizaje automático): redes neuronales, árboles de decisión, máquina de vector de soporte.

- Fase de evaluación de modelos y análisis de resultados.

En esta fase, se implementarán cinco técnicas de medición de desempeño a los resultados obtenidos tras aplicar minería de datos: MAE, MSE, RMSE, MAPE Y R2, determinando el algoritmo con menor error para la generación de predicciones.

- Fase de difusión de conocimiento.

En esta fase, se analizarán los resultados obtenidos de la información analizada a través de reuniones para la capacitación del personal administrativo de la unidad educativa, los cuales ayudarán a la toma de decisiones informadas.

Es importante indicar que este análisis no está direccionado a crear nuevos métodos de aprendizaje para mejorar el rendimiento académico de los estudiantes, el trabajo fundamental del presente proyecto es generar predicciones del desempeño académico para los posteriores periodos lectivos, y a través de este análisis generar información de valor, lo que permitirá a las autoridades y docentes de la institución evaluar, mejorar o implementar nuevas técnicas de aprendizaje y facilitar la toma de decisiones.

Además, es necesario resaltar que la implementación de minería de datos no asegura mejorar el rendimiento académico de los estudiantes, sino demostrar las predicciones que se pueden obtener desde el análisis de datos históricos para el beneficio de la institución.

1.6.Metodología del Proyecto

1.6.1. Metodología de la Investigación

Debido a la falta de fuentes sobre procesos para la gestión minuciosa de información direccionadas al tema de minería de datos, se utilizará la metodología de investigación de tipo exploratorio, dado que “su objetivo es ayudar al investigador a definir el problema, establecer hipótesis y definir la metodología para formular un estudio de

investigación definitivo" [14]. De tal forma que se indagarán y analizarán trabajos con el fin de recopilar información asociada a este campo de estudio, permitiendo encontrar características que sirvan de comparación frente al trabajo propuesto.

Con el fin de recolectar datos y conseguir un amplio conocimiento sobre la unidad educativa, se empleará una metodología de tipo diagnóstica [14], obteniendo un panorama de la situación actual de la institución para emplear mejoras en el proyecto que permitan que la toma de decisiones se realice de manera eficaz.

1.6.2. Beneficiarios del proyecto

La población escogida para aplicar esta propuesta tecnológica está conformada por beneficiarios directos (administrativos) y beneficiarios indirectos (docentes, estudiantes del año lectivo en curso, período 2023) de la institución.

Conforme al Archivo Maestro de Instituciones Educativas (AMIE) [15], en donde se recopila información sobre instituciones públicas del país en términos de apertura y cierre de períodos escolares, se registraron los siguientes datos al comienzo del periodo 2022-2023:

Beneficiarios	**Cantidad**
DIRECTOS	
Administrativos	2
INDIRECTOS	
Docentes	9
Estudiantes	228
TOTAL	**239**

Tabla 1: Grupo de beneficiarios del proyecto

1.6.3. Variable

En la propuesta presentada, se busca disminuir el tiempo de evaluación del rendimiento académico de los estudiantes de la Escuela de Educación Básica "26 de septiembre". Por tanto, la variable fundamental y de interés se centra en el tiempo de obtención de reportes (por parte del personal administrativo) acerca del desempeño estudiantil.

1.6.4. Análisis de recolección de datos

A partir de la técnica de observación se obtuvo información relacionada a los estudiantes, como datos personales y notas obtenidas en los distintos años lectivos cursados. Dichos datos se recopilaron mediante hojas de cálculo de Excel y la base de datos AMIE (Archivo Maestro de Instituciones Educativas) del Ministerio de Educación, de la cual se ha obtenido los siguientes registros entre el periodo 2019 al 2022:

Período Lectivo	Estudiantes femenino	Estudiantes masculino	Total estudiantes
2019 – 2020	150	128	246
2020 – 2021	116	85	201
2021 - 2022	120	97	217
2022 - 2023	136	120	239

Tabla 2: Cantidad de estudiantes por sexo y suma total: AMIE

En relación a la técnica de entrevista, esta fue realizada al Lcdo. Carlos Rubén Rivera Ramírez (ver Anexo 2), quien es uno de los principales responsables de la gestión de la unidad educativa al ser el secretario de la institución, obteniendo datos de importancia acerca del modo de evaluación del rendimiento académico de los estudiantes. Del mismo modo, se formularon preguntas a la Lcda. Sandra Ramírez

Quimí (rectora de la institución) para determinar cómo está compuesto el funcionamiento administrativo dentro de la institución (ver Anexo 4).

1.7.Metodología de desarrollo

Para el presente proyecto se establecerá como base la metodología Descubrimiento del Conocimiento en Bases de Datos (KDD) propuesta por Fayyad, Piatetsky y Smyth [16] **,** debido a que el contexto del mismo es el análisis predictivo del desempeño académico estudiantil mediante la aplicación de minería de datos.

Basado en los lineamientos de dicha metodología, el proyecto estará enfocado en cinco fases que se detallan a continuación:

- **Primera fase:** Construcción del conjunto de datos.

Se implementará la recopilación de datos mediante el uso de fuentes como hojas de cálculo de Excel y también de la base de datos AMIE (Archivo Maestro de Instituciones Educativas), oficial del Ministerio de Educación, las cuales contienen información acerca de los estudiantes. Posterior a esto, se creará una nueva base de datos en SQL Server con los datos elegidos para el análisis.

- **Segunda fase:** Creación de un almacén de datos (data warehouse).

Se implementará un data warehouse para consolidar y organizar la información relevante de las fuentes de datos para el estudio del rendimiento académico. Por tanto, se generará dos datamarts (estudiantes y profesores), que serán el conjunto de datos objetivos del que se partirá para el empleo de minería de datos.

- **Tercera fase:** Aplicación de técnicas de minería de datos.

Se implementará una la limpieza y transformación de los datos que serán usados para el estudio, de tal forma, que se eliminarán los datos que no posean relevancia para el enfoque de la propuesta, y de tal manera, no afecten en los resultados que se obtengan posterior al análisis.

Una vez obtenidas las variables pertinentes, se pretenderá generar predicciones de tendencias mediante el empleo de técnicas de aprendizaje supervisado para el análisis

de los datos, las cuales serán: redes neuronales, árboles de decisión, máquinas de Vector Soporte.

- **Cuarta fase:** Evaluación de modelos y análisis de resultados.

Se implementarán técnicas de medición del desempeño como el error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE), error de porcentaje medio absoluto (MAPE) y el coeficiente de determinación (R^2) para analizar los resultados obtenidos y evitar inconsistencias en los mismos al momento de evaluar nuevos conjuntos de datos.

- **Quinta fase:** Difusión de conocimiento.

Se analizarán los resultados y hallazgos obtenidos de los análisis de minería de datos realizados a través de reuniones y capacitaciones, con el fin de brindar información que ayude en la toma de decisiones efectivas para el mejoramiento de planes, estrategias y enfoques que permitan al estudiante elevar su rendimiento académico.

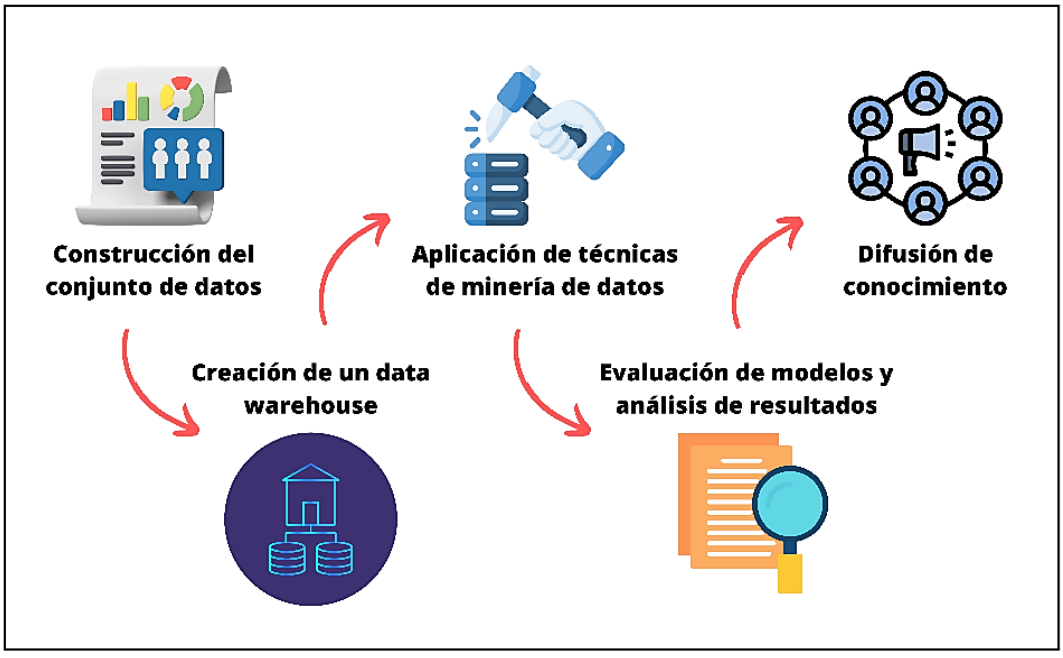


Figura 1: Proceso de la Metodología KDD: Elaboración Propia

CAPÍTULO 2. PROPUESTA

2.1. Marco Contextual

2.1.1. Institución “26 de septiembre”

La Escuela de Educación Básica “26 de septiembre” se fundó el 4 de abril del año 2007 (Ver Anexo 1). Es un establecimiento particular cuyo nivel de educación es la Educación General Básica (EGB). Se encuentra ubicada en el cantón La Libertad de la provincia de Santa Elena.

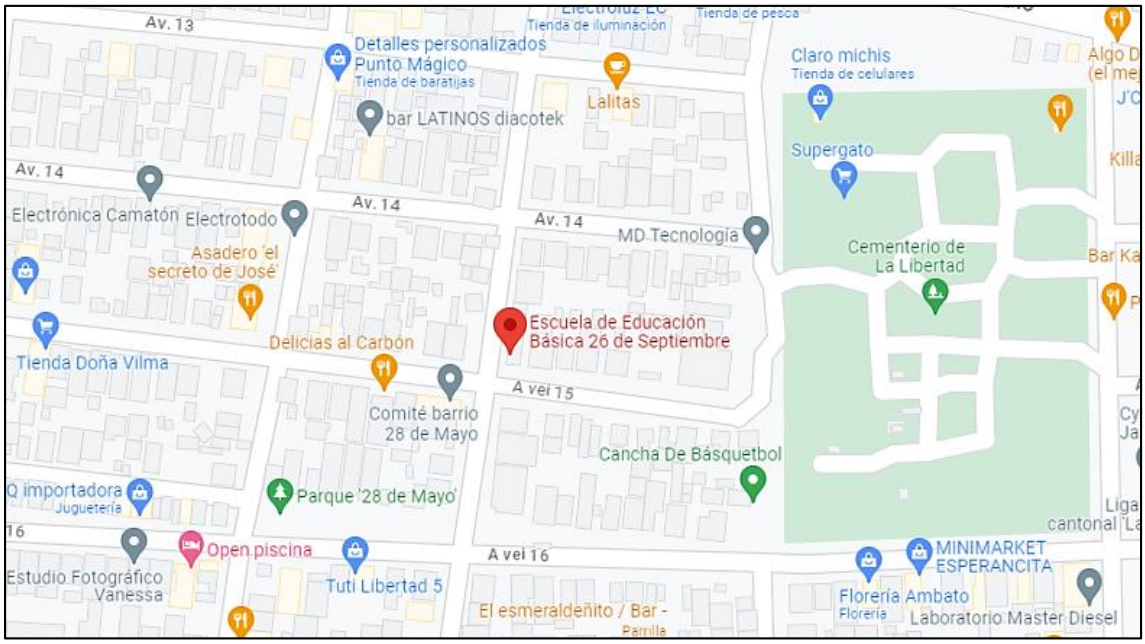


Figura 2: Ubicación de la institución: Google Maps

2.1.2. Misión de la institución

Nuestra misión es garantizar que los estudiantes adquieran los conocimientos básicos que puedan desarrollar capacidades, habilidades, actitudes y valores para su formación integral e incorporación a la sociedad y el aprendizaje, logrando los propósitos educativos (Ver Anexo 5).

2.1.3. Visión de la institución

Ser una escuela donde se imparta una educación que cumpla y sirva de base para el interés de los estudiantes, logrando una formación integral como seres humanos en un desarrollo pleno y armónico con valores que puedan enfrentar los retos de la vida con un equipo de docentes capacitados y comprometidos con el devenir de la educación de los estudiantes, padres de familia, responsables motivados y colaborativos contando con la infraestructura necesaria impartiendo una educación de calidad (Ver Anexo 5).

2.1.4. Valoración del desempeño académico en la entidad educativa y contexto normativo

En la entidad educativa, con el fin de evaluar el desempeño académico, se aplican actividades como cuestionarios que les permita evaluar de manera formativa y sumativa, permitiendo observar la capacidad de aplicación del entendimiento adquirido por los estudiantes, plasmados en los resultados que se obtienen en las pruebas de conocimiento (Ver Anexo 1).

De acuerdo con el artículo 68 de la Ley Orgánica de Educación Intercultural, se considera el desempeño académico como una parte fundamental de la evaluación continua en el ámbito educativo. La ley establece que el rendimiento estudiantil es un elemento esencial que se tiene en cuenta de manera constante a lo largo del proceso de evaluación [17].

2.2. Marco Conceptual

2.2.1. Bases de datos

Una base de datos es un conjunto estructurado de datos que representa entidades y sus interrelaciones de forma única e integrada, a pesar de que debe permitir utilizaciones varias y simultáneas [18].

2.2.2. Base de datos SQL Server

Una base de datos de SQL Server consta de una colección de tablas en las que se almacena un conjunto específico de datos estructurados, conteniendo una colección de filas, también denominadas tuplas o registros, y columnas, también denominadas atributos e, donde cada columna de la tabla se ha diseñado para almacenar un determinado tipo de información; por ejemplo, fechas, nombres, importes en moneda o números [19].

2.2.3. SQL Server Management Studio

SQL Server Management Studio (SSMS) es un entorno integrado para administrar cualquier infraestructura SQL, desde SQL Server hasta Azure SQL Database. SSMS proporciona herramientas para configurar, monitorear y administrar instancias de SQL Server y bases de datos. Se utiliza para implementar, monitorear y actualizar los componentes de la capa de datos utilizados por sus aplicaciones y crear consultas y scripts [20].

2.2.4. Visual Studio

Visual Studio es el IDE más rápido para la productividad. Teniendo como destino cualquier plataforma o dispositivo. Permite compilar cualquier tipo de aplicación, trabajo en equipo y en tiempo real. Diagnosticar y detener problemas antes de que ocurran, haciendo que procesos diarios sean más flexibles y adaptables [21].

2.2.5. Data Warehouse

El almacén de datos es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para ofrecer apoyo a procesos de ayuda a la decisión [22].

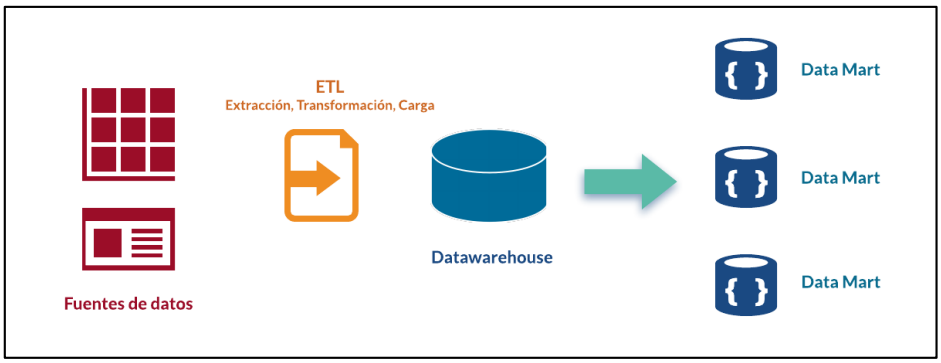


Figura 3: Estructura de un Data Warehouse: Evaluando Software

2.2.6. Data Mart

Un datamart es una parte de un datawarehouse y que le permite construir en menos tiempo una solución de Soporte de Decisiones [23]. Si el Data warehouse integra los datos de toda la organización, el datamart se restringe a un determinado proceso de negocios o departamento [23].

2.2.7. Python

Es un lenguaje de secuencias de comandos poderoso, de código abierto, interpretado y gratuito, ampliamente utilizado en aplicaciones web. Se trata de un lenguaje de programación que, aunque sencillo, ofrece una gran capacidad y estructura, especialmente beneficioso para el desarrollo de aplicaciones de gran envergadura [24].

2.2.8. Jupyter Notebook

Jupyter Notebook es la aplicación web original para crear y compartir documentos computacionales. Ofrece una experiencia sencilla, optimizada y centrada en documentos [25].

2.2.9. Excel

Excel es un programa del tipo Hoja de Cálculo que permite realizar operaciones con números organizados en una cuadrícula [26], además permite:

- Ahorrar tiempo con herramientas mejoradas por inteligencia para expertos y principiantes

- Presentar los datos con claridad mediante gráficos y grafos

2.2.10. Keras

Keras es una plataforma de alto nivel para redes neurales escrita en Python, esta plataforma está enfocada en permitir una experimentación rápida de los datos de entrada [27].

2.2.11. Scikit-learn

Es una biblioteca de aprendizaje automático de software gratuito que permite desarrollar este tipo de sistemas [28]. Además, está compuesto por:

- Herramientas simples y eficientes para el análisis predictivo de datos
- Construido sobre NumPy, SciPy y Matplotlib
- Código abierto, utilizable comercialmente: licencia BSD

2.2.12. MatplotLib

"Matplotlib es una biblioteca completa para crear visualizaciones estáticas, animadas e interactivas en Python" [29], permitiendo funciones como:

- Crear gráficos con calidad de publicación
- Crear figurar interactivas que permitan hacer zoom, desplazarse y actualizarse
- Personalizar el estilo visual y el diseño
- Incrustar en JupyterLab

2.2.13. Orange Data Mininig

Es una herramienta de aprendizaje automático de código abierto y visualización de datos [30].

2.2.14. Inteligencia de Negocios

Es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un datawarehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones [31]. El proceso de Inteligencia de Negocios incluye la comunicación de los descubrimientos y efectuar

los cambios. Las áreas incluyen clientes, proveedores, productos, servicios y competidores [31].

2.2.15. Redes Neuronales

Las redes neuronales son modelos simples del funcionamiento del sistema nervioso. Las unidades básicas son las **neuronas**, que generalmente se organizan en **capas**, Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas [32].

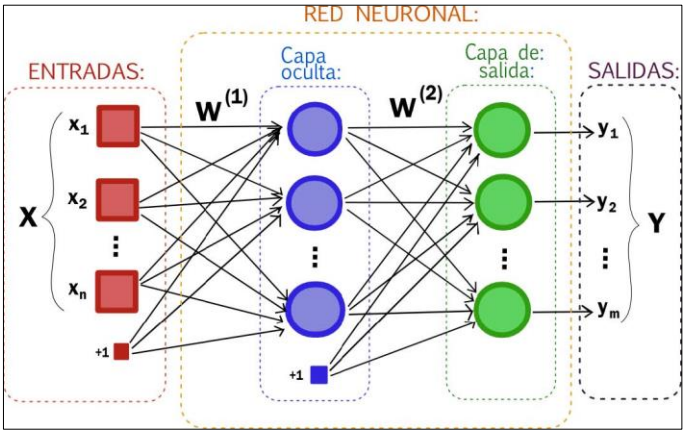


Figura 4: Procesos de una red neuronal: Art From Code

2.2.16. Árboles de decisión

El aprendizaje mediante árboles de decisión es un método de aproximación de una función objetivo de valores discretos en el cual la función objetivo es representada mediante un árbol de decisión [33]. De esta manera, permite tomar la decisión "más acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones.

2.2.17. Máquinas de vectores de soporte

SVM es un sistema de aprendizaje basado en el uso de un espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el

cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización el cual utiliza elementos de la teoría de generalización [34]. Se han encontrado muchas aplicaciones como clasificación de imágenes, reconocimiento de caracteres, clasificación de patrones, identificación de funciones, entre otras [34].

2.2.18. Métricas de rendimiento

Las métricas de rendimiento juegan un papel muy importante en problemas de clasificación donde se busca discriminar diferentes algoritmos Machine y Deep Learning, con la finalidad de facilitar la elección del mejor algoritmo dependiendo del objetivo de investigación [35].

2.3. Marco Teórico

2.3.1. Minería de datos educativos: Una herramienta para analizar patrones de aprendizaje en el ámbito educativo

La minería de datos educativos es una disciplina emergente, que se ocupa del desarrollo de métodos para explorar datos únicos y cada vez a mayor escala obtenidos en entornos educativos, y utiliza esos métodos para comprender mejor a los estudiantes y los entornos en los que aprenden [36]. De tal modo que, el análisis del aprendizaje es la medición, recopilación, análisis y presentación de informes de datos sobre los alumnos y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce [37].

Alejandro Ballesteros Román, Daniel Sánchez Guzmán y Ricardo García mencionan que, con el transcurso de los años, las actividades y el desarrollo de nuevas tecnologías han generado de forma considerable el almacenamiento de información, donde todo ese flujo de información que sea recolectado ha permitido satisfacer las necesidades diarias de las organizaciones, pero ha presentado un problema inherente en las

capacidades humanas para analizar y transformar la información en conocimiento útil y relevante que apoye a la toma de decisiones [38].

2.3.2. Influencia de la Minería de datos Educacional en el rendimiento académico

Cuando se aborda la evaluación y mejora del rendimiento académico, se examinan diversos factores que pueden influir en él. Estos incluyen, entre otros, aspectos socioeconómicos, el alcance de los programas de estudio, las metodologías de enseñanza empleadas, la viabilidad de implementar una enseñanza personalizada, los conocimientos previos de los estudiantes y sus capacidades de pensamiento formal [39].

Por otra parte, el rendimiento académico puede ser categorizado en dos sentidos: uno estricto y otro amplio. En lo que refiere al estricto, entendido como parámetro social y legal, las calificaciones obtenidas conforman un indicador sobre los conocimientos que se han adquiridos; en cuanto al sentido amplio se lo va a relacionar con éxito, el retraso o abandono de la educación formal [40].

A través de técnicas avanzadas de análisis y modelos predictivos, la minería de datos educativos (MDE) identifica patrones y tendencias en el rendimiento de los estudiantes; traduciéndose en la personalización del aprendizaje, la detección temprana de problemas académicos y la implementación de intervenciones específicas para cada estudiante [41].

2.3.3. Modelos de Predicción del Rendimiento Escolar

Johan Van Der Molen en su estudio [42] detalla de manera clara los algoritmos que se toman en cuenta como modelos para la predicción del desempeño escolar de estudiantes. Éstos se detallan a continuación:

Árboles de decisión, se llevan a cabo a través de un algoritmo que selecciona la partición de las observaciones de entrenamiento, buscando maximizar alguna métrica que evalúe la calidad de clasificación

Vector de máquina de soporte, hace referencia a la clasificación de las observaciones de entrenamiento mediante un hiperplano separador en el espacio generado por un grupo de variables explicativas.

Redes neuronales, Actúan de manera similar a nuestro cerebro al aprender de la experiencia y el pasado, aplicando este conocimiento para abordar situaciones novedosas. Este aprendizaje se obtiene como resultado del adiestramiento ("training") y éste permite la sencillez y la potencia de adaptación y evolución ante una realidad cambiante y muy dinámica; una vez adiestradas las redes de neuronas pueden hacer previsiones, clasificaciones y segmentación; presentando además una eficiencia y fiabilidad similar a los métodos estadísticos y sistemas expertos, si no mejor, en la mayoría de los casos [43].

2.4. Requerimientos

2.4.1. Requerimientos Funcionales

Según las funciones y fases del proyecto, los requisitos que deben ser destacados son los siguientes

Código	Requerimiento
RF – 01	El proyecto contará con 2 roles: el Ingeniero de datos y el Analista de datos.
RF – 02	El Ingeniero de datos recolectará la información.
RF – 03	El Analista de datos desarrollará modelos de predicción del rendimiento académico.
RF – 04	El proyecto deberá extraer datos de fuentes de información como hojas de Excel.
RF – 05	El proyecto permitirá almacenar la información recolectada en una nueva base de datos generada en SQL Server.
RF – 06	El sistema debe ser compatible con el sistema operativo Windows 10
RF – 07	Se deberá contar con el programa Visual Studio para los procesos ETL.

RF – 08	EL proyecto implementará técnicas de limpieza de datos para corregir errores y eliminar información innecesaria del conjunto de datos.
RF – 09	El proyecto organizará la información en el Data Warehouse de manera estructurada y coherente.
RF – 10	Se implementará un datamart para la importación de los datos de estudio del proyecto
RF – 11	El datamart contará con tabla de hechos y tablas de dimensiones en su estructura.
RF – 12	El proyecto realizará procesos de extracción, transformación y carga (ETL) para importar datos en el Data Warehouse de manera eficiente.
RF – 13	Se deberá extraer un archivo de tipo CSV que contenga los datos a analizar.
RF – 14	El archivo CSV debe ser compatible con la plataforma Jupyter Notebook y Orange Data Mining.
RF – 15	Se requerirá la extensión de Python para la codificación de los modelos de predicción.
RF – 17	Se requerirá el software Orange Data Mining para la elaboración de las técnicas de minería de datos.
RF – 18	El proyecto aplicará técnicas de aprendizaje automático basadas en árboles de decisión para el análisis de datos.
RF – 19	El proyecto aplicará técnicas de aprendizaje automático basadas en redes neuronales para el análisis de datos.
RF – 20	El proyecto aplicará técnicas de aprendizaje automático basadas en máquinas de vector de soporte para el análisis de datos.
RF – 21	El proyecto calculará y presentará el Error Absoluto Medio (MAE) como medida de desempeño de los modelos.
RF – 22	El proyecto calculará y presentará el Error Cuadrático Medio (MSE) como medida de desempeño de los modelos.

RF – 23	El proyecto calculará y presentará la Raíz del Error Cuadrático Medio (RMSE) como medida de desempeño de los modelos.
RF – 24	El proyecto calculará y presentará la Raíz del Error Cuadrático Medio (RMSE) como medida de desempeño de los modelos.
RF – 25	El proyecto calculará y presentará la Media del Error Absoluto en porcentaje (MAPE) como medida de desempeño de los modelos.
RF - 26	El proyecto calculará y presentará el Coeficiente de Determinación (R^2) como medida de desempeño de los modelos.
RF – 27	El proyecto realizará una comparación de los resultados de los algoritmos de aprendizaje automático y determinará el que tiene el menor error para la generación de predicciones.
RF – 28	El proyecto permitirá la presentación de los resultados obtenidos a través de matrices de correlación, diagramas de Pareto, diagramas de caja y gráficos de densidad.
RF – 29	El proyecto facilitará la organización de reuniones para la capacitación del personal administrativo de la escuela.
RF – 30	El proyecto apoyará a las autoridades y docentes en la toma de decisiones informadas relacionadas con el rendimiento académico.

Tabla 3: Requerimientos Funcionales

2.4.2. Requerimientos no Funcionales

Tipo	Código	Requerimientos
Requisitos de interfaces	RNF – 01	El proyecto debe utilizar recursos mínimos en el equipo, tales como: • Procesador: Intel Core i3 11th • Memoria RAM: 8 GB • Espacio en disco: 50 GB

Requerimiento de disponibilidad	RNF – 02	Estará disponible para su uso durante las horas clave de trabajo de la institución educativa.
Requerimiento de escalabilidad	RNF - 03	El proyecto será capaz de adaptarse a nuevas fuentes de datos y a un aumento en la cantidad de usuarios.
Requerimiento de almacenamiento	RNF – 04	El proyecto contará con una base de datos en SQL Server donde se guardará toda la información
Requerimiento de Rendimiento	RNF - 05	El sistema será capaz de procesar y analizar grandes volúmenes de datos

Tabla 4: Requerimientos no Funcionales

2.5. Componentes de la propuesta

2.5.1. Etapa 1: Recolección de información

Recopilación de información

En el desarrollo del proyecto de investigación, se utilizaron dos fuentes de datos distintas:

Archivos de Microsoft Excel, en los cuales se almacena información correspondiente entre los años 2019-2022 acerca de los estudiantes, representantes y docentes de la institución. Conteniendo datos específicos que han sido registrados por la misma institución a través de los periodos lectivos, mediante el registro digital en hojas de cálculo con un formato específico.

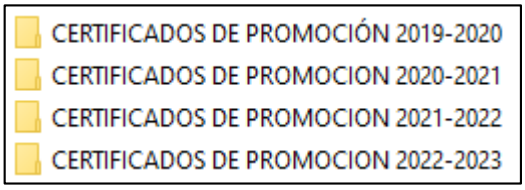


Figura 5: Certificados de promoción entre los años 2019-2022: Elaboración Propia

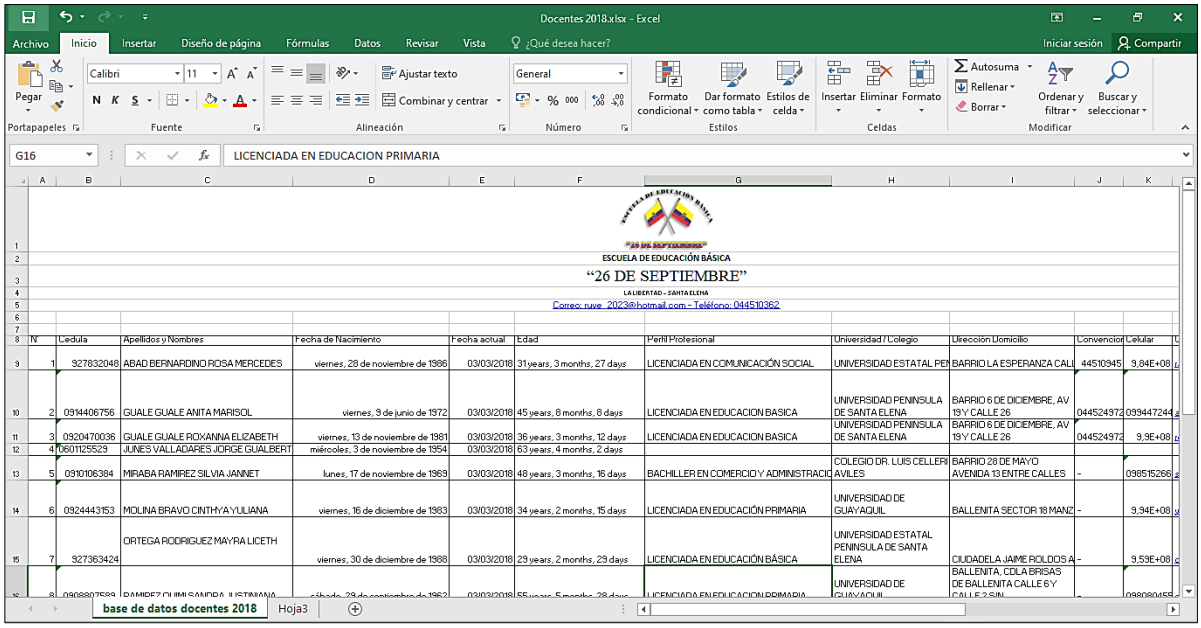


Figura 6: Datos personales de docentes de la institución: Elaboración Propia

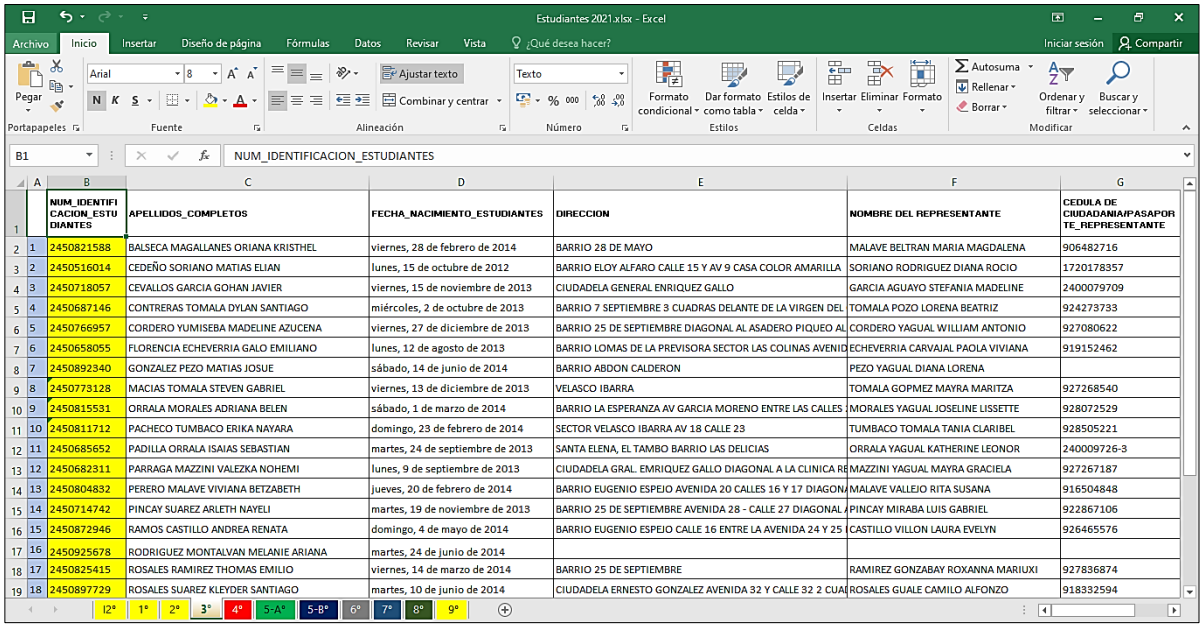
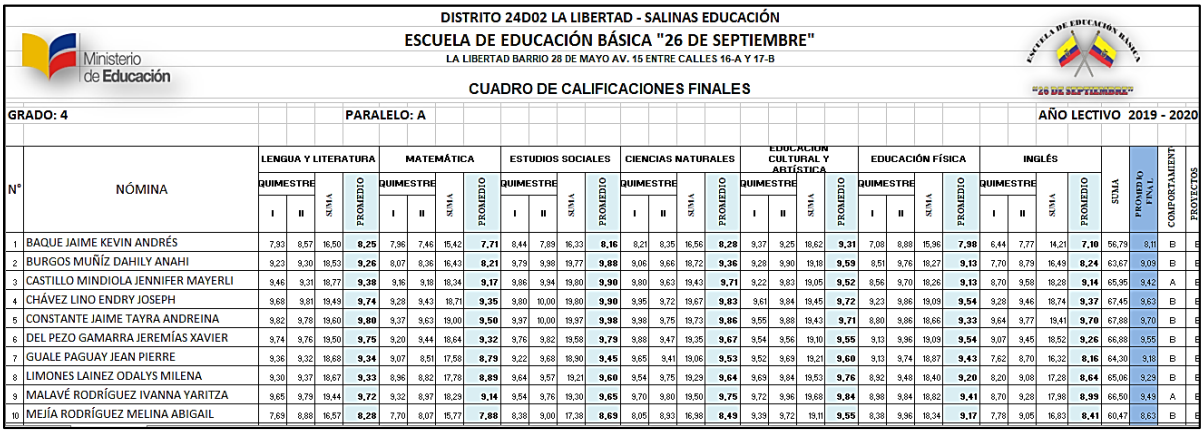


Figura 7: Datos personales de los estudiantes de la institución: Elaboración Propia

De la misma manera, dentro de estos archivos se almacena las notas obtenidas por los estudiantes en las diferentes materias, quimestres y períodos lectivos.



DISTRITO 24D02 LA LIBERTAD - SALINAS EDUCACIÓN

ESCUELA DE EDUCACIÓN BÁSICA "26 DE SEPTIEMBRE"

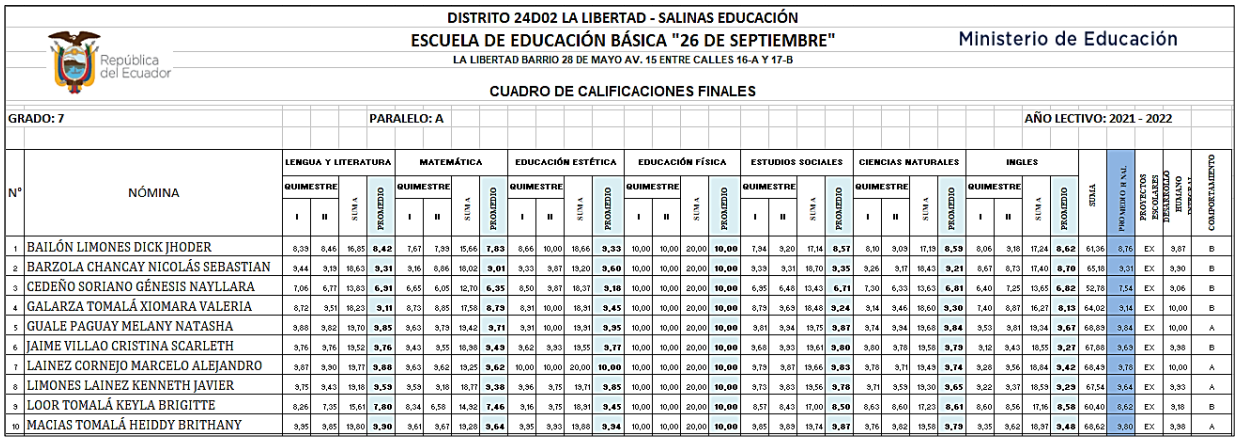
LA LIBERTAD BARRIO 28 DE MAYO AV. 15 ENTRE CALLES 16-A Y 17-B

CUADRO DE CALIFICACIONES FINALES

GRADO: 4 PARALELO: A AÑO LECTIVO 2019 - 2020

N°	NÓMINA	LENGUA Y LITERATURA				MATEMÁTICA				ESTUDIOS SOCIALES				CIENCIAS NATURALES				EDUCACIÓN CULTURAL Y ARTÍSTICA				EDUCACIÓN FÍSICA				INGLÉS				SUMA	PROMEDIO FINAL	COMPORTAMIENTO	PROYECTOS
		QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO				
		I	II			I	II			I	II			I	II			I	II			I	II			I	II						
1	BAQUE JAIME KEVIN ANDRÉS	7,93	8,57	16,50	**8,25**	7,96	7,46	15,42	**7,71**	8,44	7,89	16,33	**8,16**	8,21	8,35	16,56	**8,28**	9,37	9,25	18,62	**9,31**	7,08	8,88	15,96	**7,98**	6,44	7,77	14,21	**7,10**	56,79	8,11	B	B
2	BURGOS MUÑÍZ DAHILY ANAHI	9,23	9,30	18,53	**9,26**	8,07	8,36	16,43	**8,21**	9,79	9,98	19,77	**9,88**	9,06	9,66	18,72	**9,36**	9,28	9,90	19,18	**9,59**	8,51	9,76	18,27	**9,13**	7,70	8,79	16,49	**8,24**	63,67	9,09	B	B
3	CASTILLO MINDIOLA JENNIFER MAYERLI	9,46	9,31	18,77	**9,38**	9,16	9,18	18,34	**9,17**	9,86	9,94	19,80	**9,90**	9,80	9,63	19,43	**9,71**	9,22	9,83	19,05	**9,52**	8,56	9,70	18,26	**9,13**	8,70	9,58	18,28	**9,14**	65,95	9,42	A	B
4	CHÁVEZ LINO ENDRY JOSEPH	9,68	9,81	19,49	**9,74**	9,28	9,43	18,71	**9,35**	9,80	10,00	19,80	**9,90**	9,95	9,72	19,67	**9,83**	9,61	9,84	19,45	**9,72**	9,23	9,86	19,09	**9,54**	9,28	9,46	18,74	**9,37**	67,45	9,63	B	B
5	CONSTANTE JAIME TAYRA ANDREINA	9,82	9,78	19,60	**9,80**	9,37	9,63	19,00	**9,50**	9,97	10,00	19,97	**9,98**	9,98	9,75	19,73	**9,86**	9,55	9,88	19,43	**9,71**	8,80	9,86	18,66	**9,33**	9,64	9,77	19,41	**9,70**	67,88	9,70	B	B
6	DEL PEZO GAMARRA JEREMÍAS XAVIER	9,74	9,76	19,50	**9,75**	9,20	9,44	18,64	**9,32**	9,76	9,82	19,58	**9,79**	9,88	9,47	19,35	**9,67**	9,54	9,56	19,10	**9,55**	9,13	9,96	19,09	**9,54**	9,07	9,45	18,52	**9,26**	66,88	9,55	B	B
7	GUALE PAGUAY JEAN PIERRE	9,36	9,32	18,68	**9,34**	9,07	8,51	17,58	**8,79**	9,22	9,68	18,90	**9,45**	9,65	9,41	19,06	**9,53**	9,52	9,69	19,21	**9,60**	9,13	9,74	18,87	**9,43**	7,62	8,70	16,32	**8,16**	64,30	9,18	B	B
8	LIMONES LAINEZ ODALYS MILENA	9,30	9,37	18,67	**9,33**	8,96	8,82	17,78	**8,89**	9,64	9,57	19,21	**9,60**	9,54	9,75	19,29	**9,64**	9,69	9,84	19,53	**9,76**	8,92	9,48	18,40	**9,20**	8,20	9,08	17,28	**8,64**	65,06	9,29	B	B
9	MALAVÉ RODRÍGUEZ IVANNA YARITZA	9,65	9,79	19,44	**9,72**	9,32	8,97	18,29	**9,14**	9,54	9,76	19,30	**9,65**	9,70	9,80	19,50	**9,75**	9,72	9,96	19,68	**9,84**	8,98	9,84	18,82	**9,41**	8,70	9,28	17,98	**8,99**	66,50	9,49	A	B
10	MEJÍA RODRÍGUEZ MELINA ABIGAIL	7,89	8,88	16,57	**8,28**	7,70	8,07	15,77	**7,88**	8,38	9,00	17,38	**8,69**	8,05	8,93	16,98	**8,49**	9,39	9,72	19,11	**9,55**	8,38	9,96	18,34	**9,17**	7,78	9,05	16,83	**8,41**	60,47	8,63	B	B

Figura 8: Notas de estudiantes, periodo lectivo 2019 - 2020: Elaboración Propia



DISTRITO 24D02 LA LIBERTAD - SALINAS EDUCACIÓN

ESCUELA DE EDUCACIÓN BÁSICA "26 DE SEPTIEMBRE"

LA LIBERTAD BARRIO 28 DE MAYO AV. 15 ENTRE CALLES 16-A Y 17-B

CUADRO DE CALIFICACIONES FINALES

GRADO: 7 PARALELO: A AÑO LECTIVO: 2021 - 2022

N°	NÓMINA	LENGUA Y LITERATURA				MATEMÁTICA				EDUCACIÓN ESTÉTICA				EDUCACIÓN FÍSICA				ESTUDIOS SOCIALES				CIENCIAS NATURALES				INGLES				SUMA	PROMEDIO FINAL	PROYECTOS ESCOLARES	PROYECTOS DESARROLLO HUMANO	COMPORTAMIENTO
		QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO	QUIMESTRE		SUMA	PROMEDIO					
		I	II			I	II			I	II			I	II			I	II			I	II			I	II							
1	BAILÓN LIMONES DICK JHODER	8,39	8,46	16,85	**8,42**	7,67	7,99	15,66	**7,83**	8,66	10,00	18,66	**9,33**	10,00	10,00	20,00	**10,00**	7,84	9,20	17,14	**8,57**	8,10	9,09	17,19	**8,59**	8,06	9,18	17,24	**8,62**	61,36	8,76	EX	9,87	B
2	BARZOLA CHANCAY NICOLÁS SEBASTIAN	9,44	9,19	18,63	**9,31**	9,16	8,86	18,02	**9,01**	9,33	9,87	19,20	**9,60**	10,00	10,00	20,00	**10,00**	9,39	9,31	18,70	**9,35**	9,26	9,17	18,43	**9,21**	8,67	8,73	17,40	**8,70**	65,18	9,31	EX	9,90	B
3	CEDEÑO SORIANO GÉNESIS NAYLLARA	7,06	6,77	13,83	**6,91**	6,65	6,05	12,70	**6,35**	8,50	9,87	18,37	**9,18**	10,00	10,00	20,00	**10,00**	6,35	6,48	13,43	**6,71**	7,30	6,33	13,63	**6,81**	6,40	7,25	13,65	**6,82**	52,78	7,54	EX	9,06	B
4	GALARZA TOMALÁ XIOMARA VALERIA	8,72	9,51	18,23	**9,11**	8,73	8,85	17,58	**8,79**	8,91	10,00	18,91	**9,45**	10,00	10,00	20,00	**10,00**	8,79	9,69	18,48	**9,24**	9,14	9,46	18,60	**9,30**	7,40	8,87	16,27	**8,13**	64,02	9,14	EX	10,00	B
5	GUALE PAGUAY MELANY NATASHA	9,88	9,82	19,70	**9,85**	9,63	9,79	19,42	**9,71**	9,91	10,00	19,91	**9,95**	10,00	10,00	20,00	**10,00**	9,81	9,94	19,75	**9,87**	9,74	9,94	19,68	**9,84**	9,53	9,81	19,34	**9,67**	68,89	9,84	EX	10,00	A
6	JAIME VILLAO CRISTINA SCARLETH	9,76	9,76	19,52	**9,76**	9,43	9,55	18,98	**9,49**	9,62	9,93	19,55	**9,77**	10,00	10,00	20,00	**10,00**	9,68	9,93	19,61	**9,80**	9,80	9,78	19,58	**9,79**	9,12	9,43	18,55	**9,27**	67,88	9,69	EX	9,98	B
7	LAINEZ CORNEJO MARCELO ALEJANDRO	9,87	9,90	19,77	**9,88**	9,63	9,62	19,25	**9,62**	10,00	10,00	20,00	**10,00**	10,00	10,00	20,00	**10,00**	9,79	9,87	19,66	**9,83**	9,76	9,71	19,49	**9,74**	9,28	9,56	18,84	**9,42**	68,49	9,78	EX	10,00	A
8	LIMONES LAINEZ KENNETH JAVIER	9,75	9,43	19,18	**9,59**	9,59	9,18	18,77	**9,38**	9,96	9,75	19,71	**9,85**	10,00	10,00	20,00	**10,00**	9,73	9,83	19,56	**9,78**	9,71	9,59	19,30	**9,65**	9,22	9,37	18,59	**9,29**	67,54	9,64	EX	9,93	A
9	LOOR TOMALÁ KEYLA BRIGITTE	8,26	7,35	15,61	**7,80**	6,34	6,58	14,92	**7,46**	9,16	9,75	18,91	**9,45**	10,00	10,00	20,00	**10,00**	8,57	8,43	17,00	**8,50**	8,63	8,60	17,23	**8,61**	8,60	8,56	17,16	**8,58**	60,40	8,62	EX	9,98	B
10	MACIAS TOMALÁ HEIDDY BRITHANY	9,95	9,65	19,60	**9,90**	9,61	9,67	19,28	**9,64**	9,95	9,93	19,88	**9,94**	10,00	10,00	20,00	**10,00**	9,85	9,89	19,74	**9,87**	9,76	9,82	19,58	**9,79**	9,35	9,62	18,97	**9,48**	68,62	9,80	EX	9,98	A

Figura 9: Notas de estudiantes, periodo lectivo 2021-2022: Elaboración Propia

Esquema de base de datos

A partir de estos datos, se propone una nueva base de datos en el ambiente de SQL Server, la cual contiene información específica de los estudiantes; y de la misma forma, datos que llevan relación con los representantes y docentes de los mismos. El esquema se muestra a continuación:

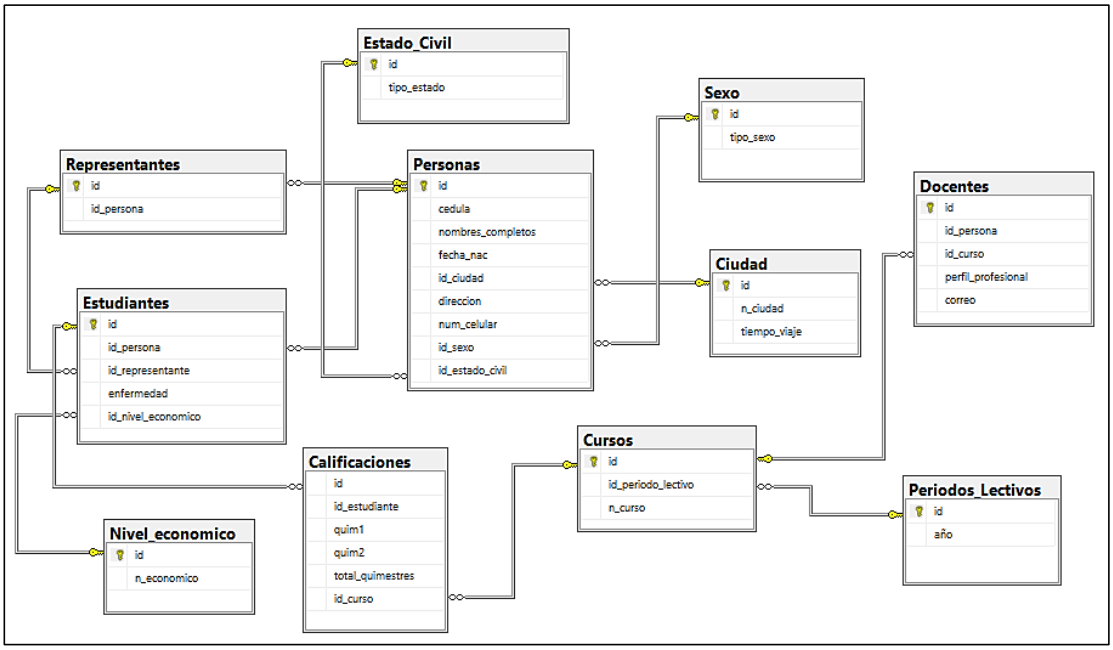


Figura 10: Base de datos propuesta para la unidad educativa: Elaboración Propia

Este nuevo modelo está compuesto por once tablas:

1) Estudiantes
2) Representantes
3) Nivel_economico
4) Estado_civil
5) Personas
6) Calificaciones
7) Cursos
8) Sexo
9) Ciudad
10) Docentes
11) Periodos_Lectivos

A continuación, se detallan las columnas que integran las tablas fundamentales creadas:

Columna	Tipo	Descripción
id	int	Identificador de persona
cedula	varchar (10)	Cedula de la persona
apellidos	varchar (50)	Apellidos de la persona
nombres	varchar (50)	Nombres de la persona
fecha_nac	date	Fecha de nacimiento de la persona
id_ciudad	int	Identificador de ciudad de la persona
direccion	varchar (100)	Dirección de la persona
num_celular	varchar(50)	Número de celular de persona
id_genero	int	Identificador de género de la persona
id_estado_civil	int	Identificador del estado civil de la persona

Tabla 5: Tabla "Persona"

Columna	Tipo	Descripción
id	int	Identificador de estudiante
id_persona	int	Identificador de la persona estudiante
id_representante	int	Identificador de representante
enfermedad	varchar (50)	Enfermedad del estudiante
id_nivel_economico	int	Identificador del nivel económico del estudiante

Tabla 6: Tabla "Estudiante"

Columna	Tipo	Descripción
id	int	Identificador de la persona representante
id_persona	int	Identificador de persona

Tabla 7: Tabla "Representante"

Columna	Tipo	Descripción
id	int	Identificador de nivel económico del estudiante

n_economico	int	Nivel económico del estudiante

Tabla 8: Tabla "Nivel_economico"

Columna	Tipo	Descripción
id	int	Identificador de calificaciones
id_estudiante	int	Nivel económico del estudiante
quim1	float	Nota del quimestre 1 del estudiante
quim2	float	Nota del quimestre 2 del estudiante
total_quimestres	float	Nota total de los quimestres del estudiante
id_curso	int	Identificador del curso del estudiante

Tabla 9: Tabla "Calificaciones"

Columna	Tipo	Descripción
id	int	Identificador del estado civil de la persona
tipo_estado	int	Tipo de estado civil de la persona

Tabla 10: Tabla "Estado_Civil"

Columna	Tipo	Descripción
id	int	Identificador del curso del estudiante
id_periodolectivo	int	Identificador del periodo lectivo del estudiante
n_curso	varchar (20)	Nombre del curso del estudiante

Tabla 11: Tabla "Cursos"

Columna	Tipo	Descripción
id	int	Identificador del género de la persona
tipo_sexo	varchar (10)	Tipo de género de la persona

Tabla 12: Tabla "Sexo"

Columna	Tipo	Descripción
id	int	Identificador de la ciudad de la persona
n_ciudad	varchar (20)	Nombre de la ciudad de la persona
tiempo_viaje	int	Tiempo de viaje definido para cada ciudad

Tabla 13: Tabla "Ciudad"

Columna	Tipo	Descripción
id	int	Identificador del periodo lectivo
año	varchar (50)	Descripción del año lectivo

Tabla 14: Tabla "Periodos_Lectivos"

Columna	Tipo	Descripción
id	int	Identificador del docente
id_persona	int	Identificador de la persona docente
id_curso	int	Identificador del curso del docente
perfil_profesional	varchar (50)	Nombre del perfil profesional del docente
correo	varchar (50)	Correo del docente

Tabla 15: Tabla "Docentes"

Creación de base de datos

Posterior a la creación del esquema, se procedió con la creación de la base de datos propuesta mediante el motor de base de datos SQL Server con el nombre "ESCUELABD".

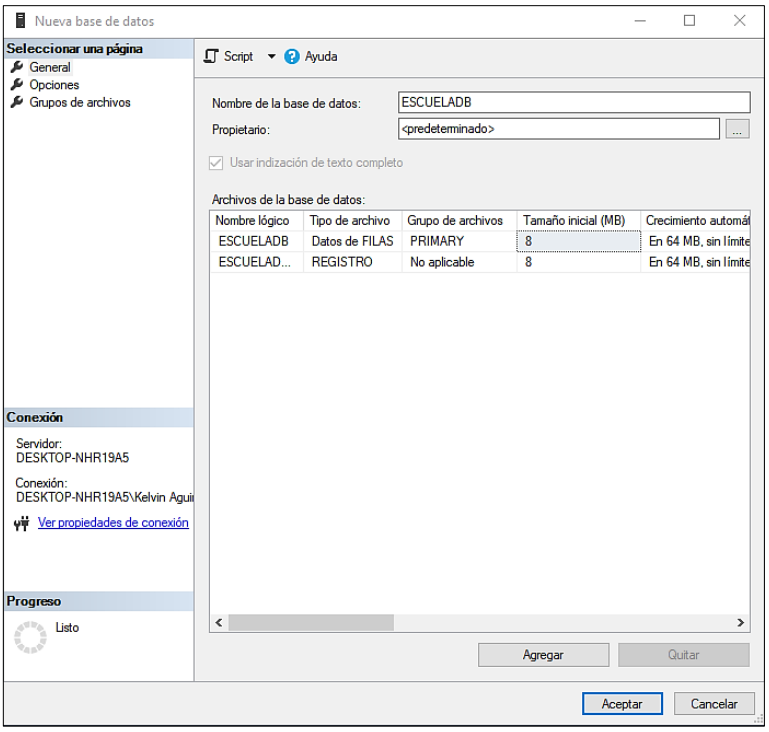


Figura 11: Creación de la base de datos "ESCUELADB" : Elaboración Propia

Se realizó la respectiva creación de las tablas ya mencionadas por las cuales estaría compuesta la base de datos.

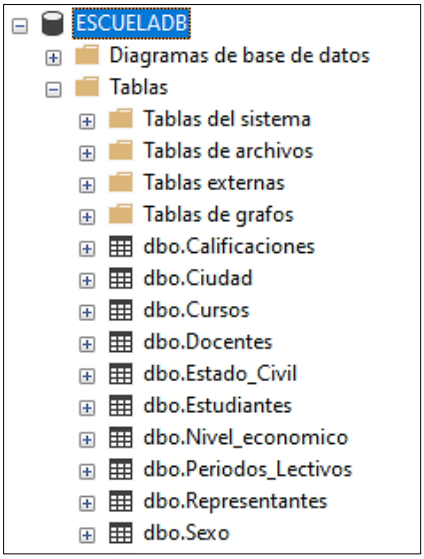
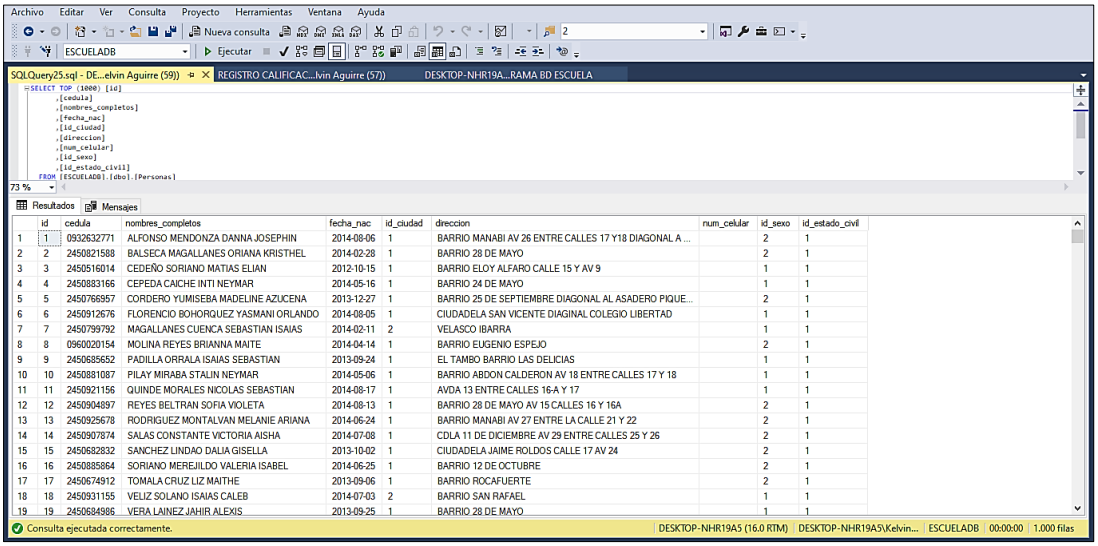


Figura 12: Creación de tablas en la base de datos "ESCUELADB" : Elaboración Propia

Se procedió con la importación de los datos que almacenan los archivos de Excel, alojándolos en la nueva base de datos “ESCUELADB”.



	id	cedula	nombres_completos	fecha_nac	id_ciudad	direccion	num_celular	id_sexo	id_estado_civil
1	1	0932632771	ALFONSO MENDONZA DANNA JOSEPHIN	2014-08-06	1	BARRIO MANABI AV 26 ENTRE CALLES 17 Y18 DIAGONAL A ...		2	1
2	2	2450821588	BALSECA MAGALLANES ORIANA KRISTHEL	2014-02-28	1	BARRIO 28 DE MAYO		2	1
3	3	2450516014	CEDEÑO SORIANO MATIAS ELIAN	2012-10-15	1	BARRIO ELOY ALFARO CALLE 15 Y AV 9		1	1
4	4	2450883166	CEPEDA CAICHE INTI NEYMAR	2014-05-16	1	BARRIO 24 DE MAYO		1	1
5	5	2450766957	CORDERO YUMISEBA MADELINE AZUCENA	2013-12-27	1	BARRIO 25 DE SEPTIEMBRE DIAGONAL AL ASADERO PIQUE...		2	1
6	6	2450912676	FLORENCIO BOHORQUEZ YASMANI ORLANDO	2014-08-05	1	CIUDADELA SAN VICENTE DIAGINAL COLEGIO LIBERTAD		1	1
7	7	2450799792	MAGALLANES CUENCA SEBASTIAN ISAIAS	2014-02-11	2	VELASCO IBARRA		1	1
8	8	0960020154	MOLINA REYES BRIANNA MAITE	2014-04-14	1	BARRIO EUGENIO ESPEJO		2	1
9	9	2450685652	PADILLA ORRALA ISAIAS SEBASTIAN	2013-09-24	1	EL TAMBO BARRIO LAS DELICIAS		1	1
10	10	2450881087	PILAY MIRABA STALIN NEYMAR	2014-05-06	1	BARRIO ABDON CALDERON AV 18 ENTRE CALLES 17 Y 18		1	1
11	11	2450921156	QUINDE MORALES NICOLAS SEBASTIAN	2014-08-17	1	AVDA 13 ENTRE CALLES 16-A Y 17		1	1
12	12	2450904897	REYES BELTRAN SOFIA VIOLETA	2014-08-13	1	BARRIO 28 DE MAYO AV 15 CALLES 16 Y 16A		2	1
13	13	2450925678	RODRIGUEZ MONTALVAN MELANIE ARIANA	2014-06-24	1	BARRIO MANABI AV 27 ENTRE LA CALLE 21 Y 22		2	1
14	14	2450907874	SALAS CONSTANTE VICTORIA AISHA	2014-07-08	1	CDLA 11 DE DICIEMBRE AV 29 ENTRE CALLES 25 Y 26		2	1
15	15	2450682832	SANCHEZ LINDAO DALIA GISELLA	2013-10-02	1	CIUDADELA JAIME ROLDOS CALLE 17 AV 24		2	1
16	16	2450885864	SORIANO MEREJILDO VALERIA ISABEL	2014-06-25	1	BARRIO 12 DE OCTUBRE		2	1
17	17	2450674912	TOMALA CRUZ LIZ MAITHE	2013-09-06	1	BARRIO ROCAFUERTE		2	1
18	18	2450931155	VELIZ SOLANO ISAIAS CALEB	2014-07-03	2	BARRIO SAN RAFAEL		1	1
19	19	2450684986	VERA LAINEZ JAHIR ALEXIS	2013-09-25	1	BARRIO 28 DE MAYO		1	1

Figura 13: Base de datos “ESCUELADB”: Elaboración Propia

2.5.2. Etapa 2: Creación del almacén de datos (data warehouse)

Esquema del data warehouse

Una vez establecida la base de datos en SQL Server, el siguiente paso implica la creación del almacén de datos, el cual será la base para la obtención del conjunto de datos esencial en el proceso de minería de datos. Es fundamental que este almacén posea una estructura organizada para facilitar y garantizar que la información extraída sea relevante para el respectivo análisis.

El enfoque de Kimball para el ciclo de vida del almacén de datos, también conocido como el enfoque de estilo de vida dimensional empresarial, posibilita que las herramientas de inteligencia empresarial exploren diversos esquemas en estrella, generando así información confiable [44]. Es decir, en este modelo se parte de la creación de datamarts para luego generar un almacén de datos.

Por tanto, para este proyecto se eligió el enfoque de Kimball, creándose un datamart en el cual se almacenan las dimensiones de importancia para el análisis.

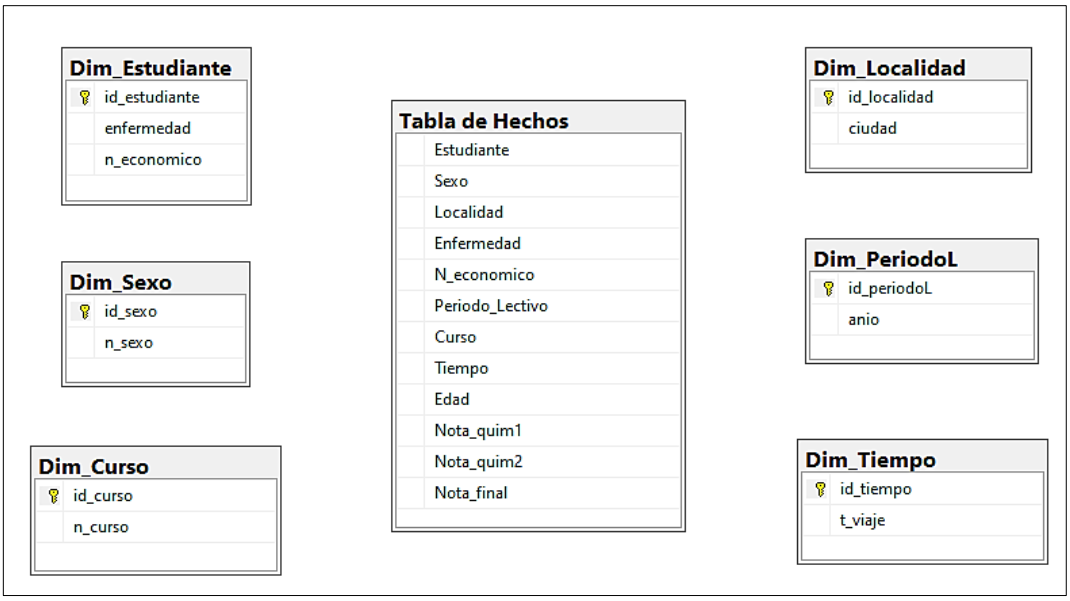


Figura 14: Estructura de Datamart: Elaboración Propia

Creación del data warehouse

Para el respectivo llenado del data warehouse primero se debe crear la base de datos en donde se alojará el mismo. Definiéndose como “DW_ESCUELADB”.

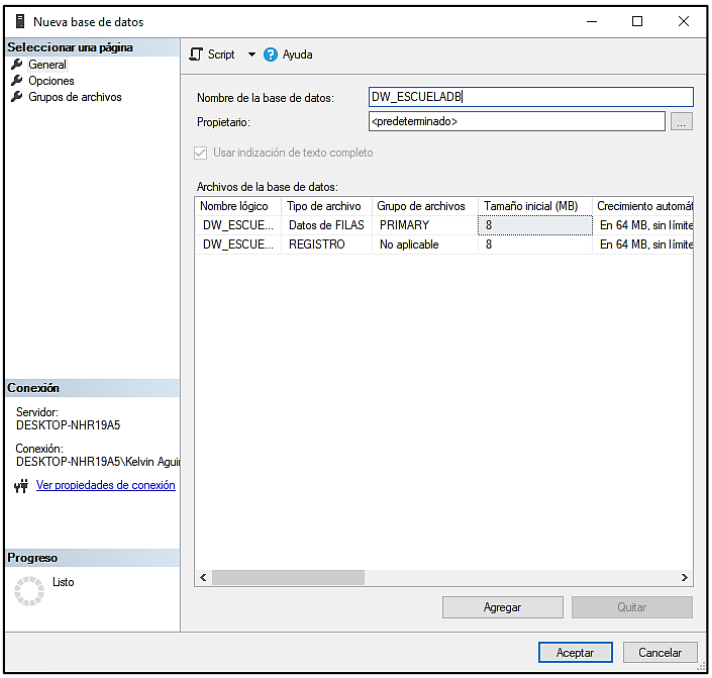


Figura 15: Creación de base de datos para el datawarehouse: Elaboración Propia

Procesos ETL en Microsoft Visual Studio 2019 (Integration Services)

La ejecución de Extracción, Transformación y Carga de datos se realizaron en el ambiente de Visual Studio 2019, instalando la característica de integración de servicios para este software.

Se creó un nuevo proyecto con el nombre “DW”, en el cual se realizarán los procesos ETL.

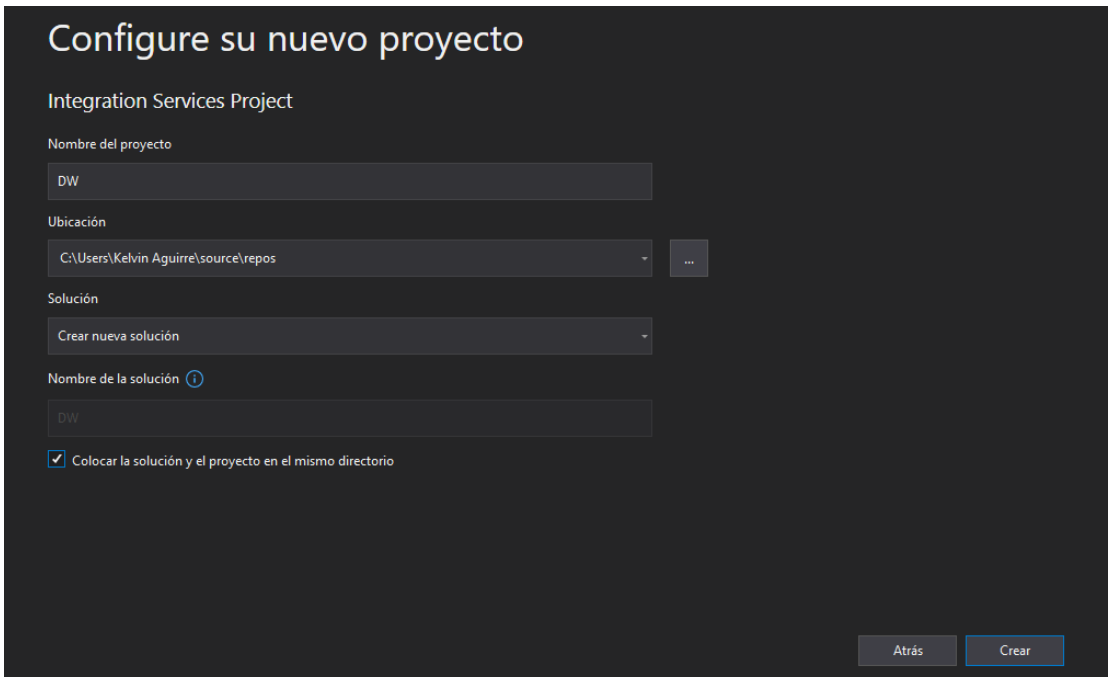


Figura 16: Creación de proyecto en Microsoft Visual Studio 2019: Elaboración Propia

Dimensiones establecidas

A continuación, se pueden ver los procesos ETL en las dimensiones creadas para el análisis.

Se agregaron tareas de flujo para realizar los procesos ETL, destinadas para las respectivas dimensiones y tabla de hechos que se crearon en el datamart, entre las cuales están:

- Dim_Estudiante
- Dim_Sexo
- Dim_Curso
- Dim_Localidad
- Dim_PeriodoLectivo
- Dim_Tiempo
- Tabla_Hechos

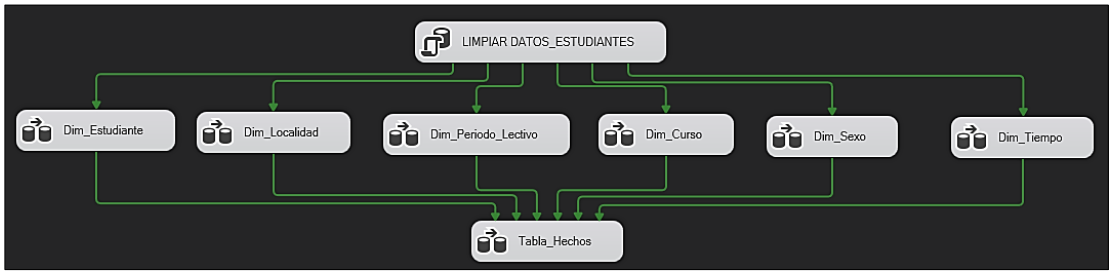


Figura 17: Dimensiones y tabla de hechos: Elaboración Propia

En el administrador de conexiones, se añade la base de datos "ESCUELADB", en la que se almacena toda la información con respecto a la institución educativa y la cual será el origen de datos.

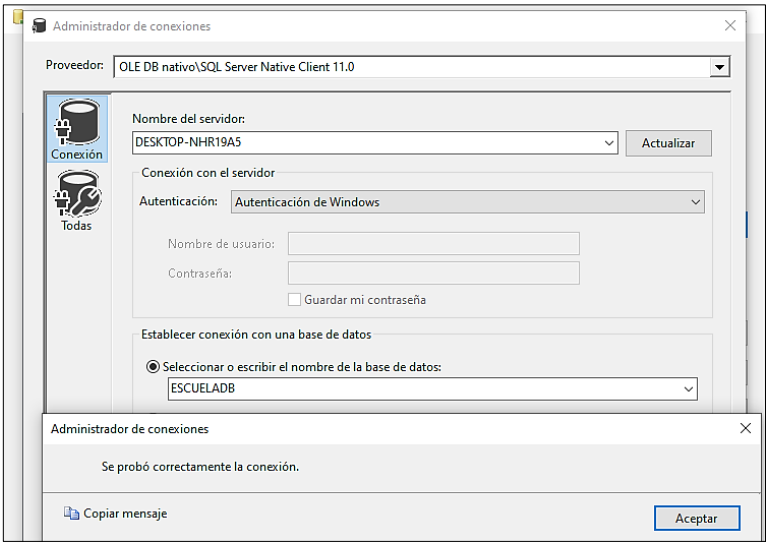


Figura 18: Configuración y conexión con base de datos de origen "ESCUELADB: Elaboración Propia

Cada una de las dimensiones y tablas de hechos creadas están compuestas por un origen y un destino de tipo OLE DB, y de un conversor de datos.

Dim_Estudiante

Dentro de la primera Dimensión que corresponde a Dim_Estudiante seleccionamos el origen y destino de tipo OLE DB.

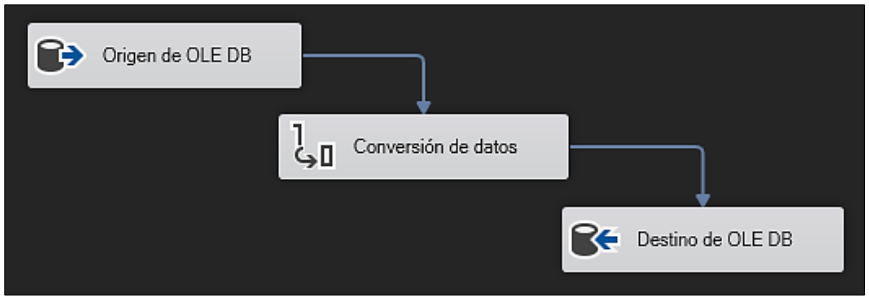


Figura 19: Componentes para el desarrollo de los procesos ETL: Elaboración Propia

El origen de datos será la base de datos denominada "**ESCUELADB**", en la que se almacena toda la información con respecto a la institución educativa.

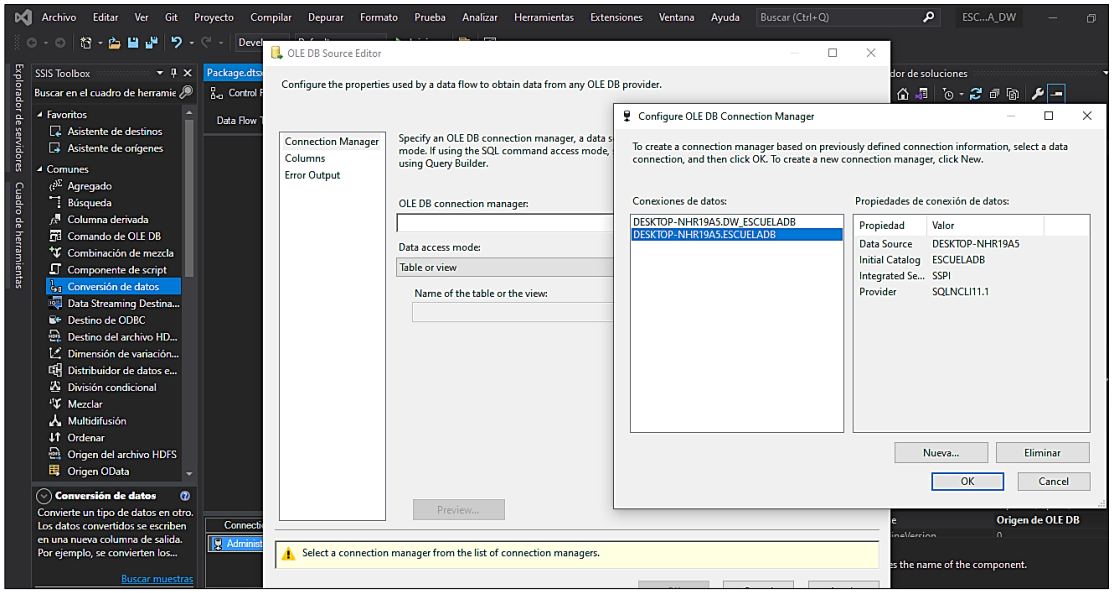


Figura 20: Origen de datos "ESCUELADB" : Elaboración Propia

El destino de datos será la base de datos que fue creada para el data warehouse, denominada "**DW_ESCUELADB**".

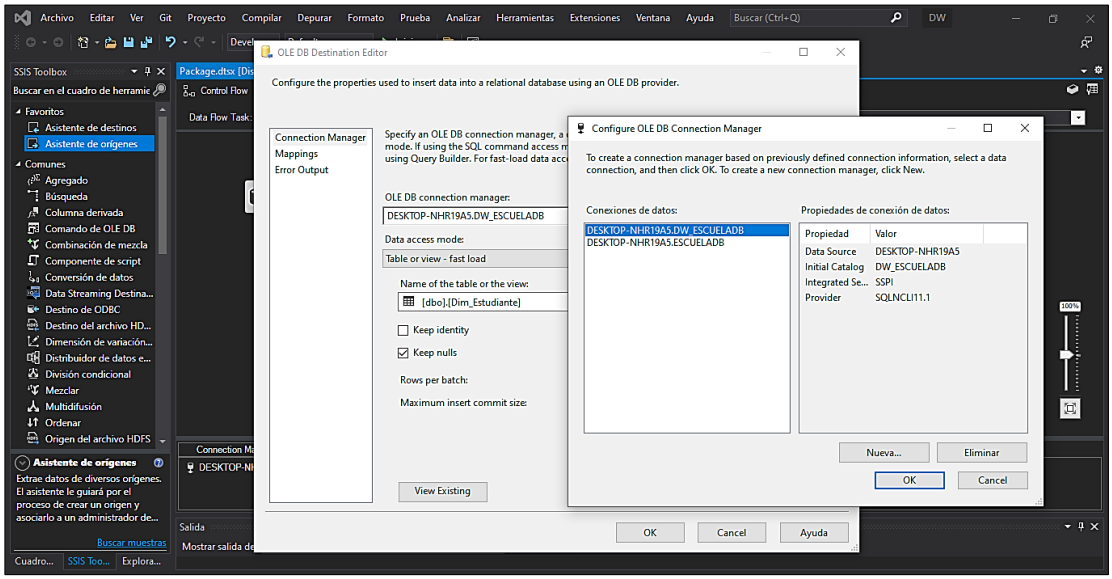


Figura 21: Destino de datos "DW_ESCUELADB: Elaboración Propia

Para realizar la extracción de los datos de interés se hace uso de sentencias SQL, describiendo que parámetros queremos recopilar, para este caso de **Dim_Estudiante**, se extrajo el id_estudiante, fecha_nac, enfermedad, id_nivel_economico.

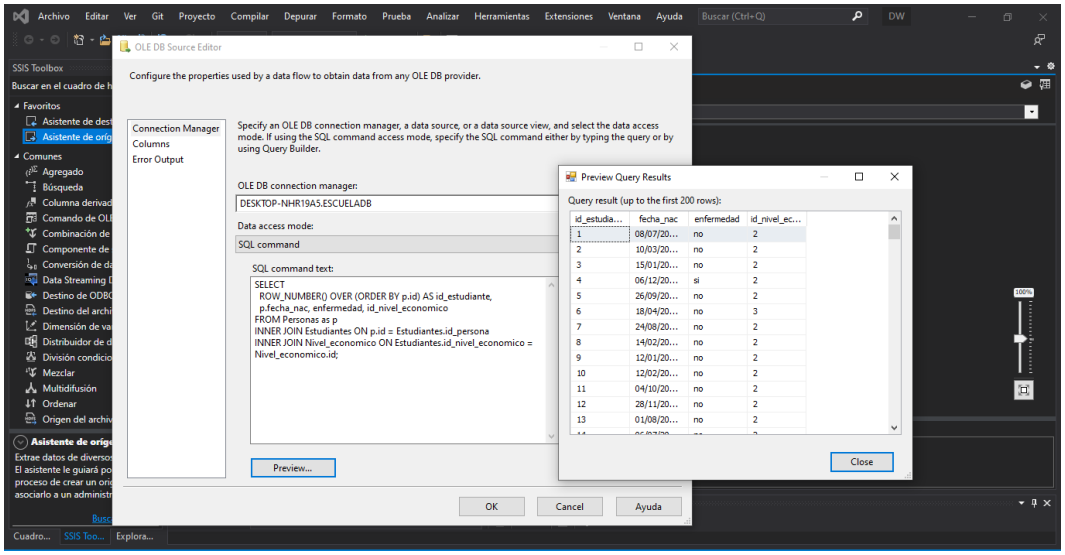


Figura 22: Extracción de datos "Dim_Estudiante": Elaboración Propia

Para evitar la incompatibilidad, se agrega un conversor de datos, el cual nos ayudará a controlar los errores de compatibilidad que se puedan producir por el tipo de dato al realizar los procesos ETL.

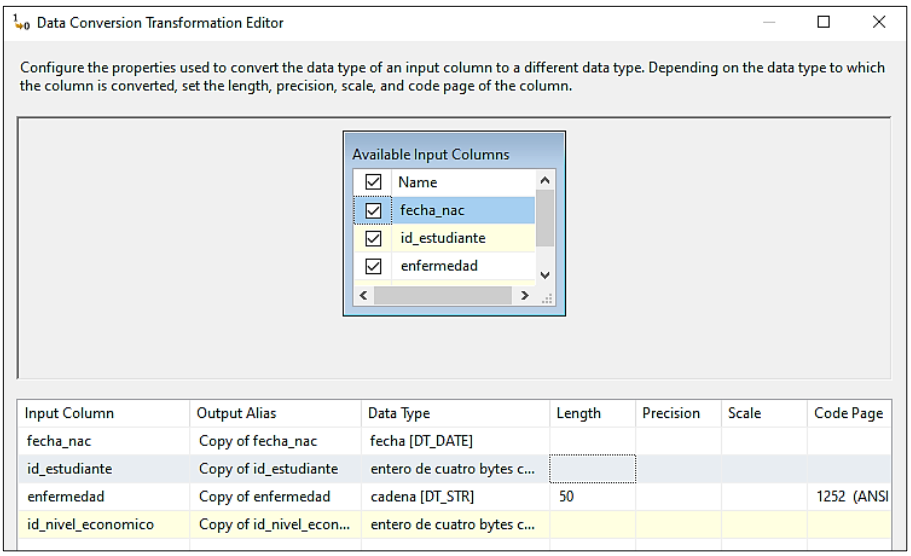


Figura 23: Conversión de datos "Dim_Estudiante": Elaboración Propia

En el destino OLE DB, se realiza la relación en el apartado de Mappings, seleccionando las copias generadas las cuales ya cuentan con el conversor de datos implementado.

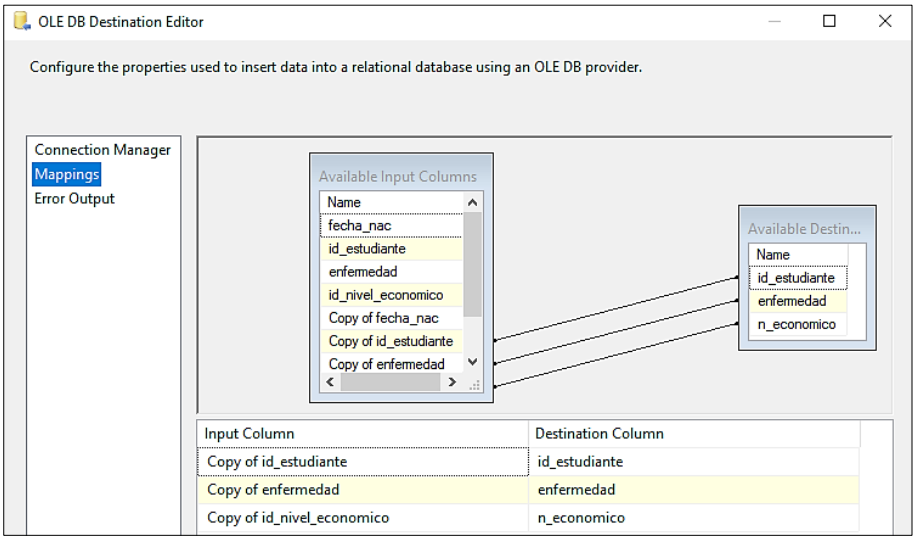


Figura 24: Relaciones entre columnas de entrada y destino "Dim_Estudiante": Elaboración Propia

Dim_Sexo

En esta dimension se extrajo los datos de interés: id_sexo, tipo_sexo.

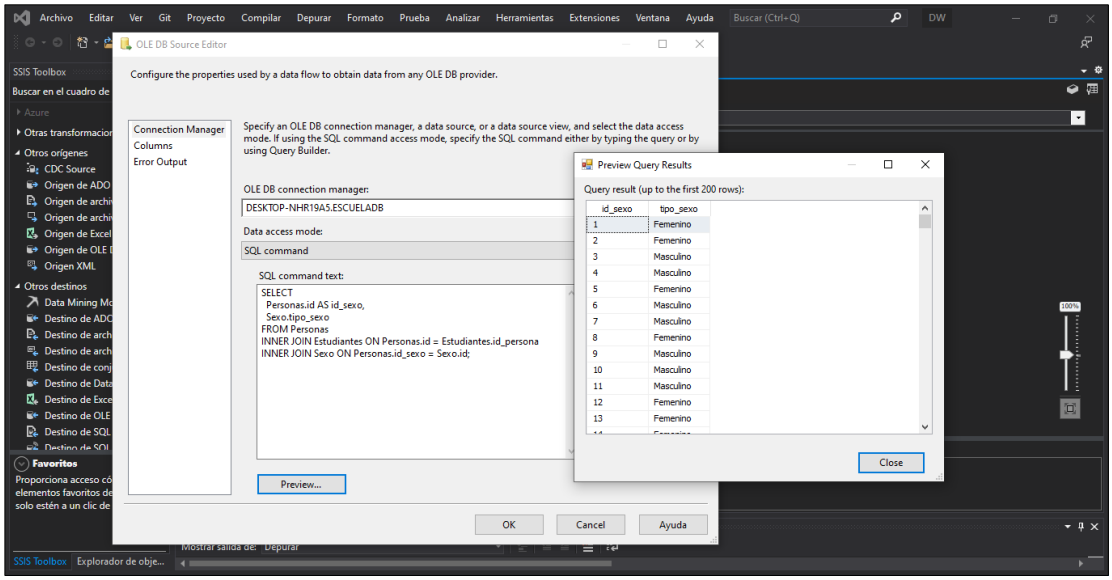


Figura 25: Extracción de datos "Dim_Sexo": Elaboración Propia

Dim_Curso

En esta dimension se extrajo los datos de interés: id_curso, n_curso.

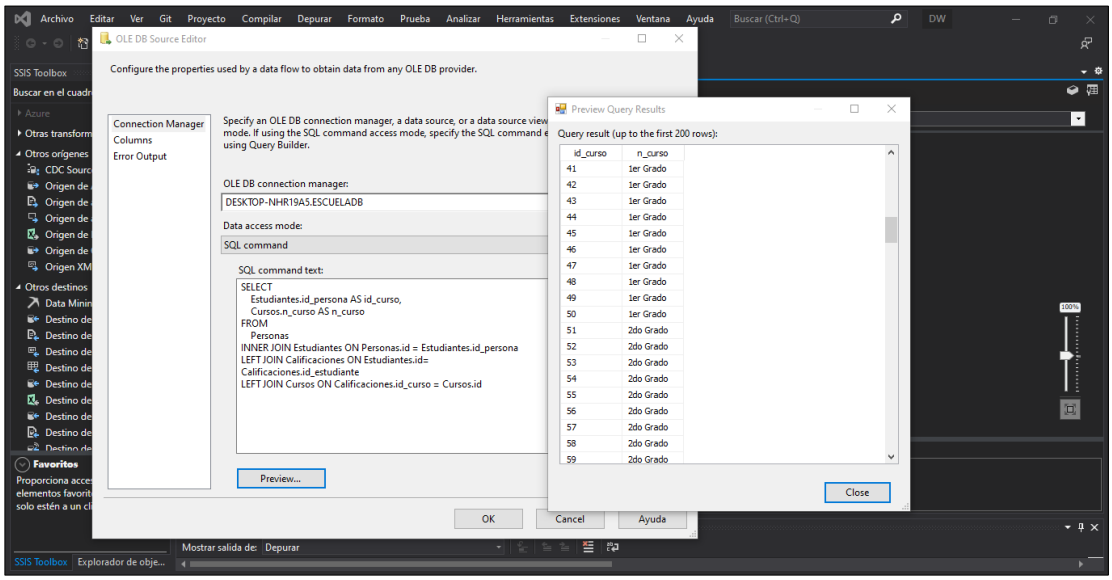


Figura 26: Extracción de datos "Dim_Curso": Elaboración Propia

Dim_Localidad

En esta dimension se extrajo los datos de interés: id_localidad, id_ciudad.

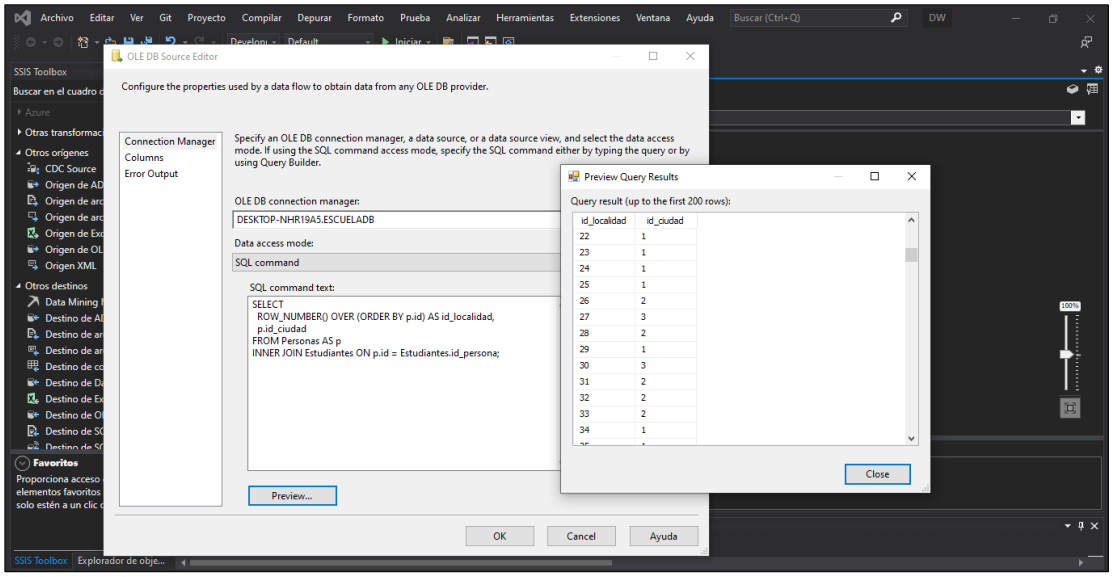


Figura 27: Extracción de datos "Dim_Localidad": Elaboración Propia

Dim_PeriodoLectivo

En esta dimension se extrajo los datos de interés: id_periodo_lectivo, año.

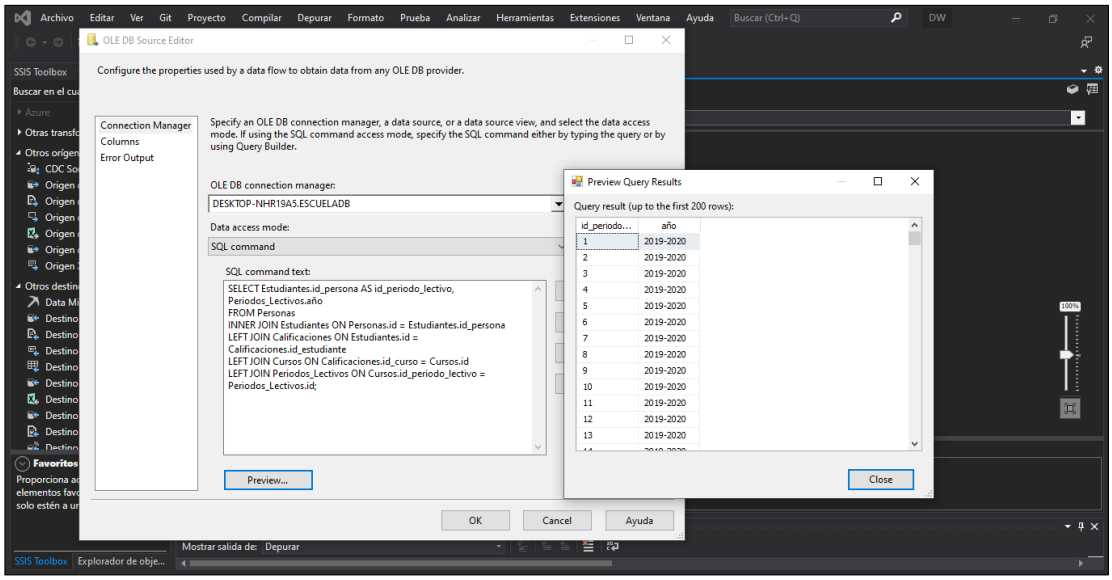


Figura 28: Extracción de datos "Dim_PeriodoLectivo": Elaboración Propia

Dim_Tiempo

En esta dimension se extrajo los datos de interés: id_tiempo, tiempo_viaje.

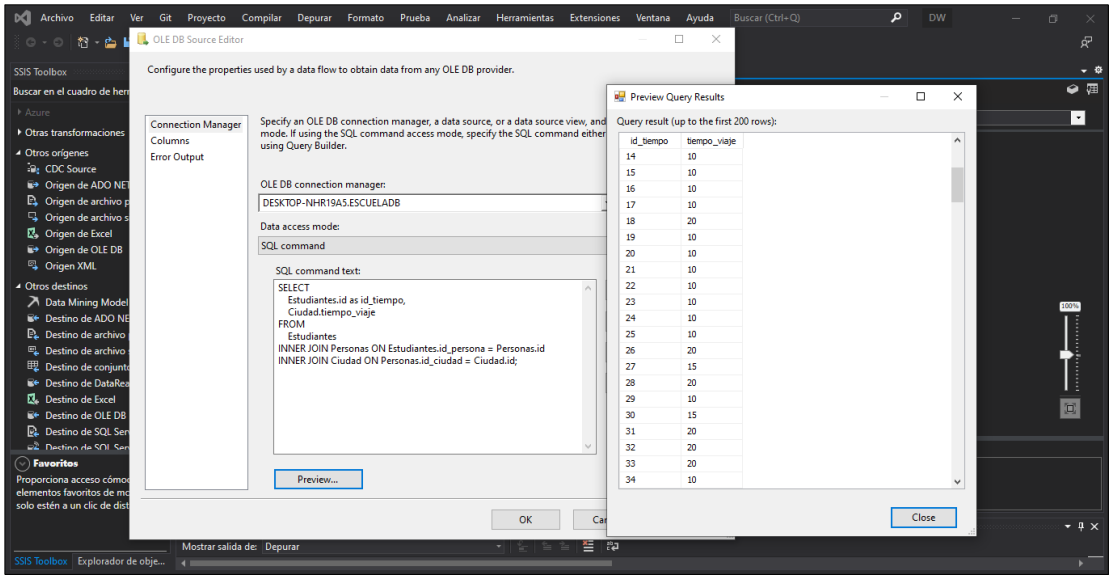


Figura 29: Extracción de datos "Dim_Tiempo": Elaboración Propia

Los procesos detallados anteriormente, se realizan para cada una de las dimensiones por las que está compuesto el Datamart; elegir un origen OLE DB (base de datos "ESCUELABD"), extraer los datos de interés mediante sentencias SQL e implementar el conversor de datos. Por último, elegir el destino OLE DB (base de datos "DW_ESCUELABD"), y seleccionar la tabla a la que se quiere enviar la información extraída.

Tabla de Hechos

Luego de haber realizado el llenado de las tablas correspondientes a las dimensiones del Datamart, se realiza el proceso de llenado de la tabla de hechos.

Mediante sentencias SQL, se extrae toda la información recopilada anteriormente en cada una de las dimensiones, para elaborar un único conjunto de datos que almacene todas las variables a analizar.

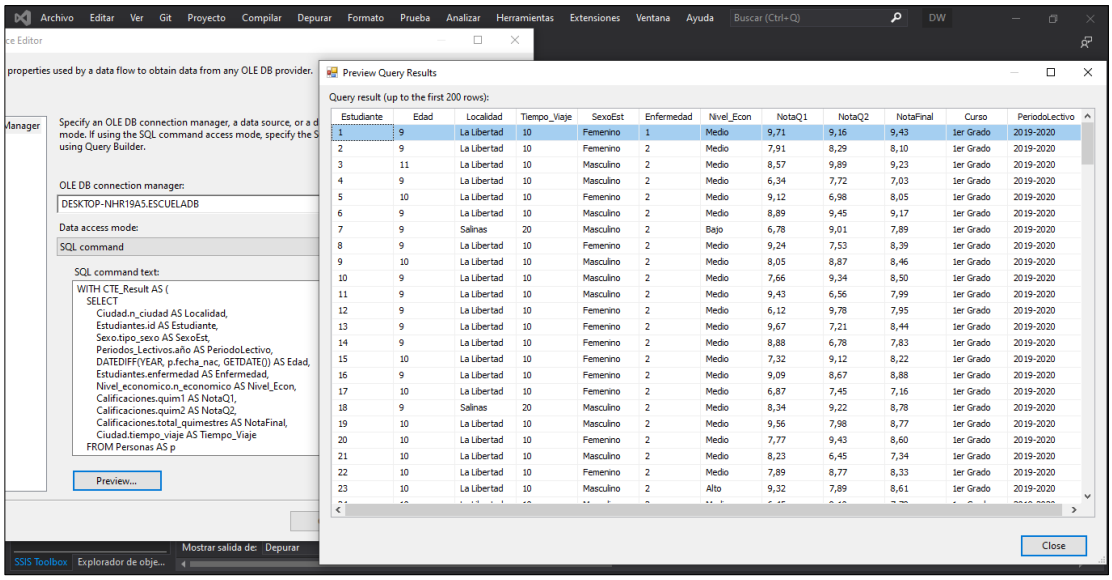


Figura 30: Entrada de origen "Tabla_Hechos": Elaboración Propia

De la misma forma que en las dimensiones, se implementa el convertidor de datos para evitar la incompatibilidad en los mismos al realizar los procesos ETL.

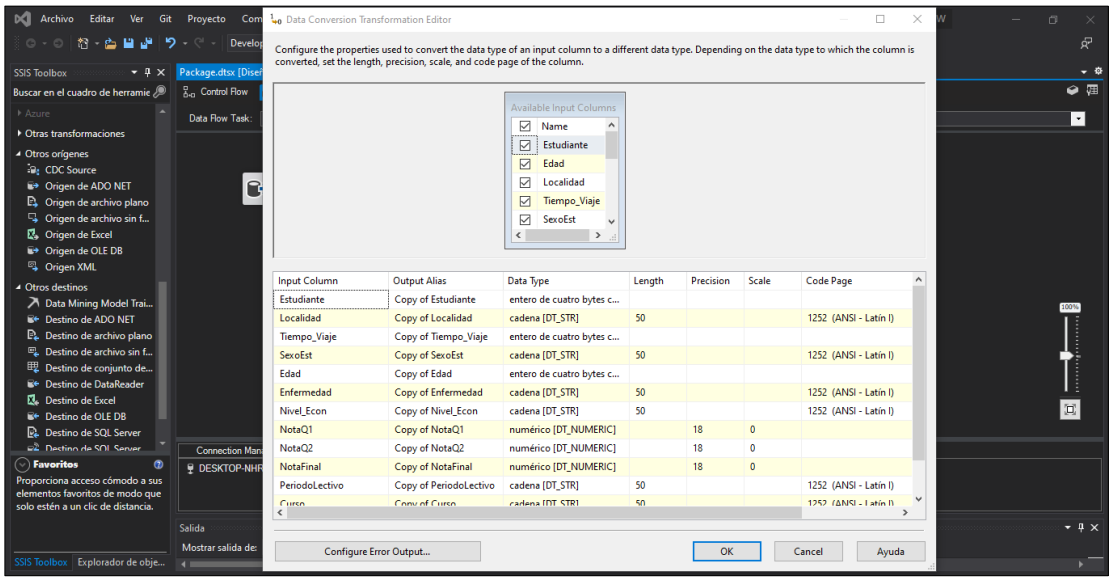


Figura 31: Conversión de datos "Tabla_Hechos": Elaboración Propia

Se agregó también una tarea de ejecución SQL con el nombre “LIMPIAR DATOS_ESTUDIANTES”, la cual almacena la función TRUNCATE que permitirá realizar una limpieza en las tablas del datawarehouse cada vez que el proceso ETL sea ejecutada, evitando la repetición de datos.

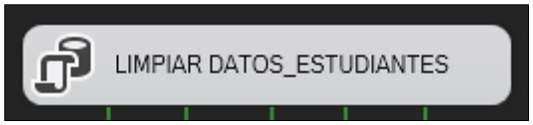


Figura 32: Tarea de ejecución SQL: Elaboración Propia

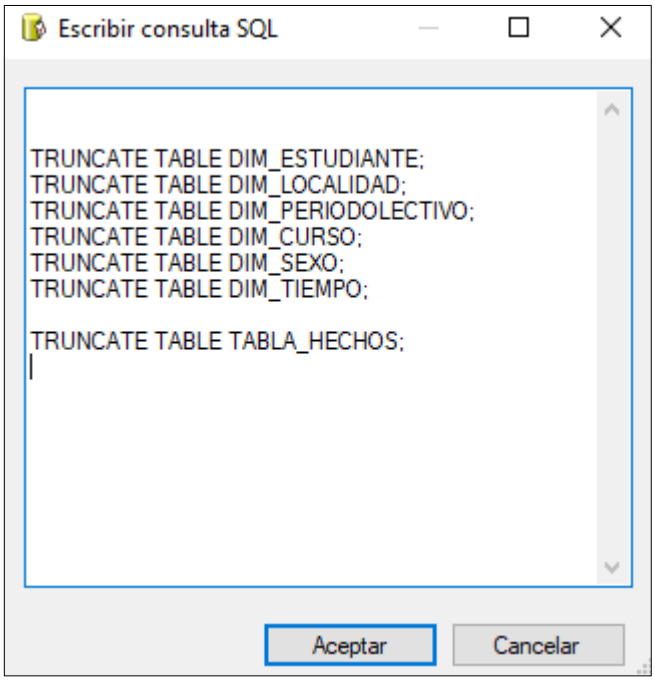


Figura 33: Función de limpieza Truncate: Elaboración Propia

Una vez realizado los procesos ETL en cada una de las dimensiones y tabla de hechos, se procede con la ejecución del esquema para lograr el envío del conjunto de datos, y que este sea almacenado en la base de datos que se creó para el datawarehouse.

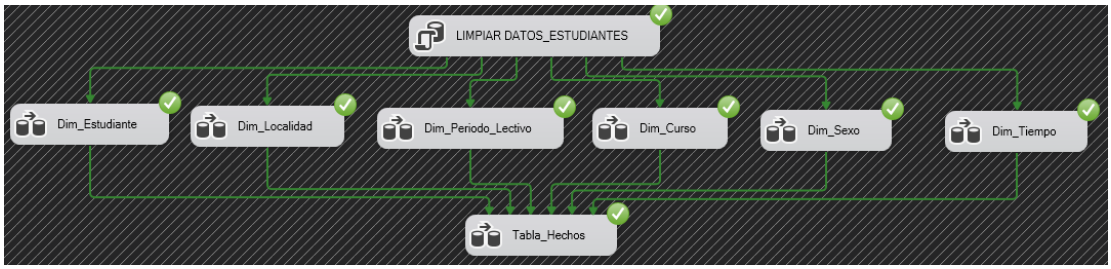


Figura 34: Ejecución de procesos ETL: Elaboración Propia

De esta manera, se puede apreciar que los datos extraídos en cada una de las dimensiones y en la tabla de hechos se encuentran almacenados en el data warehouse.

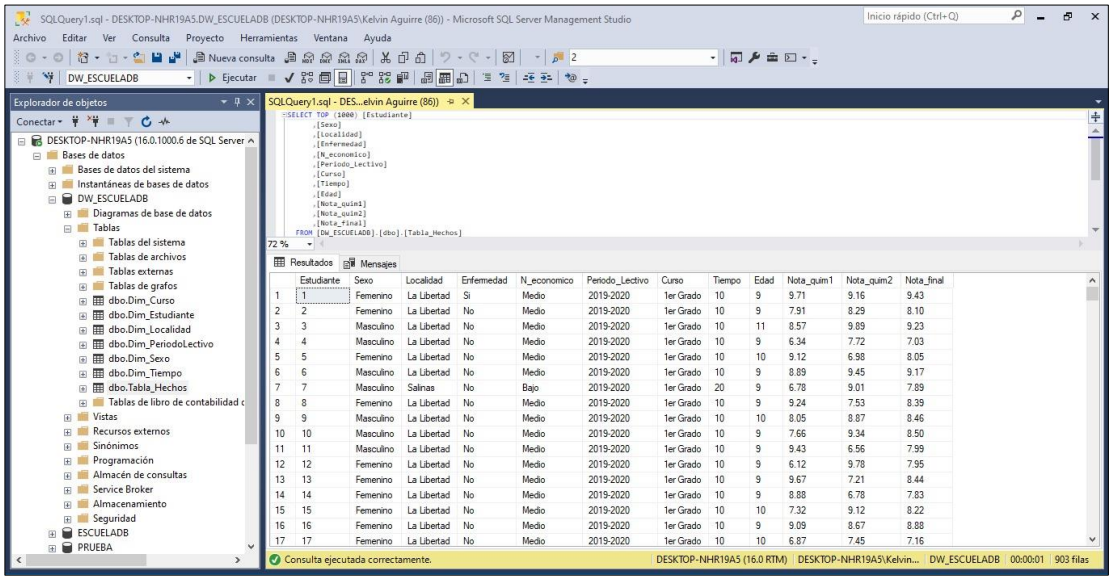


Figura 35: Base de Datos del data warehouse: Elaboración Propia

Luego de haber realizado el proceso de integración de los datos, tanto en la base de datos como en el data warehouse, se llevó a cabo creación del conjunto de datos objetivo o dataset, mediante la generación de un archivo en formato .csv

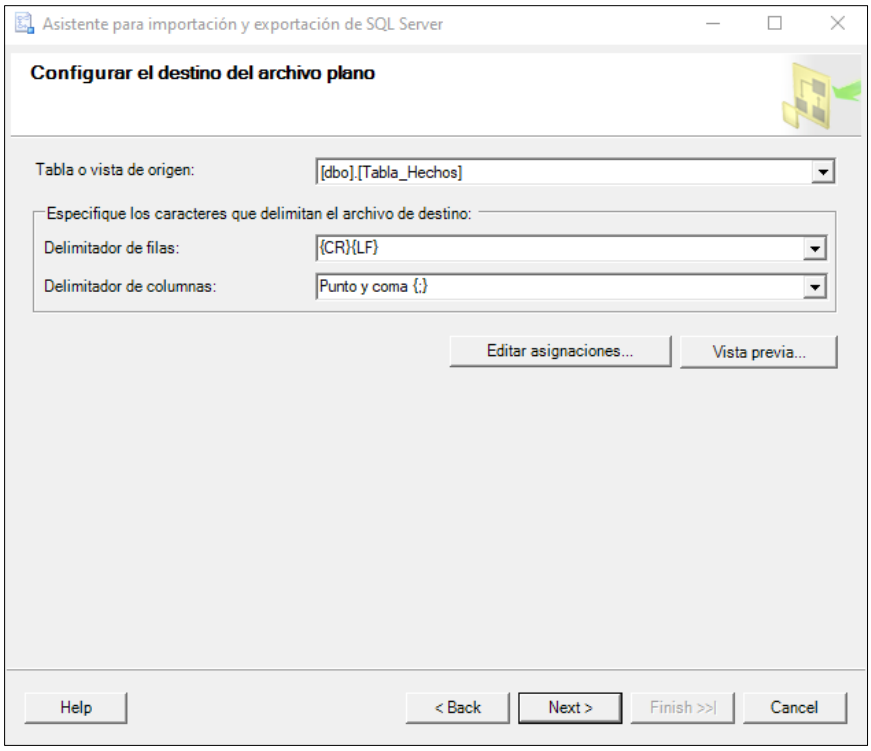


Figura 36: Proceso de creación del archivo .csv: Elaboración Propia

Figura 37: Archivo .csv del data warehouse: Elaboración Propia

2.5.3. Etapa 3: Aplicación de minería de datos

La etapa de aplicación de minería de datos está compuesta de tres subetapas, las cuales se definen a continuación:

1	Exploración de los datos
2	Conversión de datos
3	Minería de datos

Tabla 16: Subetapas de minería de datos.

Exploración de datos

La meta de esta subetapa consiste en, a partir del conjunto de datos creado durante la construcción del archivo CSV, determinar qué variables serán adecuadas para el proceso de minería de datos. Al mismo tiempo, se busca identificar las relaciones entre las variables independientes y aquellas que se desea predecir.

Las bibliotecas fundamentales requeridas para llevar a cabo estos procedimientos incluyen las siguientes:

N°	LIBRERÍAS
1	Pandas
2	NumPy
3	MatplotLib
4	Seaborn

Tabla 17: Librerías de python para minería de datos.

Una vez realizada la importación de las librerías de Python mencionadas, se hace uso del método "**read.csv**" junto al parámetro "**delimiter**" declarado con el valor ";". El objetivo del uso de este método es proceder con la lectura del archivo de tipo csv que se extrajo anteriormente y en donde están almacenados los datos del Data warehouse, acompañado del parámetro delimiter, el cual realiza una separación entre los valores para una mejor comprensión y procesamiento de los mismos.

```
datos = pd.read_csv('FICHERO DW.csv',delimiter=';')
```

Figura 38: Método read.csv y parámetro delimiter: Elaboración Propia

Tras la importación del fichero, se pudo comprobar que la cantidad total de variables a analizar es de 11, de las cuales, 6 son variables categóricas y 6 son variables numéricas.

```
dtypes: int64(6), object(6)
```

Figura 39: Tipos de variables: Elaboración Propia

Una vez identificada la cantidad y naturaleza de las variables con las que se trabajará, se procede a evaluar la relación entre las variables independientes y la variable dependiente. Para las variables numéricas, se utiliza la matriz de coeficientes de correlación de Pearson. Este análisis se visualiza mediante un gráfico de calor, utilizando la biblioteca seaborn de Python, que facilita la observación de las relaciones existentes.

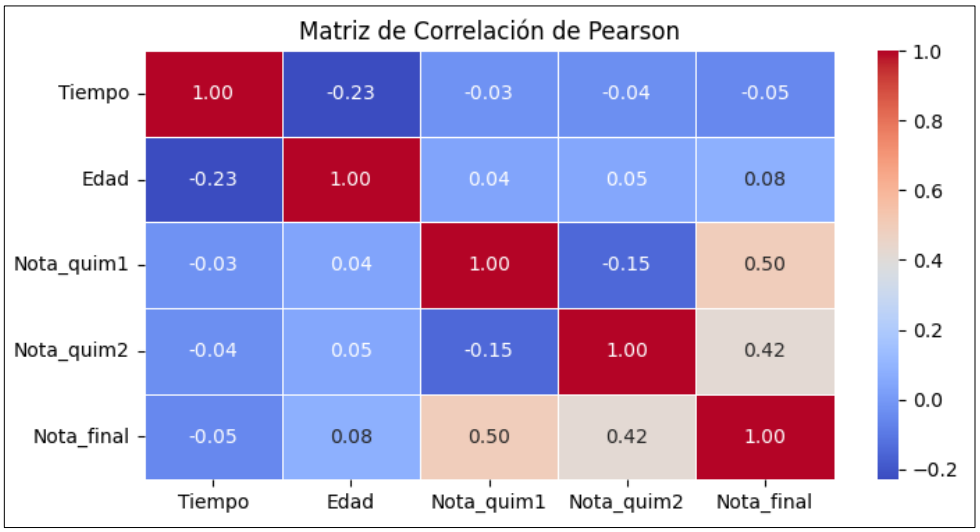


Figura 40: Correlación entre variables numéricas: Elaboración Propia

Los resultados que arrojó la gráfica de la Matriz de Correlación de Pearson, indica que las variables que tienen mayor asociación lineal a la variable dependiente **Nota_final** son **Nota_quim1** y **Nota_quim2**.

Una vez realizado el análisis de las variables numéricas, se procede también a analizar las variables categóricas del conjunto de datos extraído. De tal manera, se elaboraron diagramas de Caja para una mejor comprensión de la distribución de Nota_final en relación con las variables categóricas.

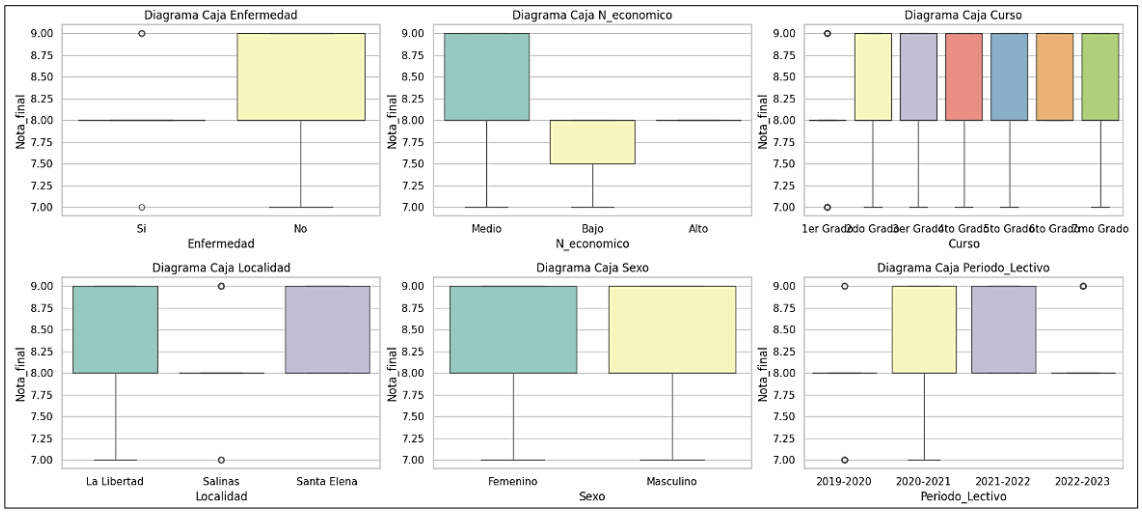


Figura 41: Diagramas de Caja de variables categóricas: Elaboración Propia

También se analiza cómo están distribuidas las variables independientes que han sido seleccionadas durante la subetapa de exploración de datos, mediante gráficos de densidad.

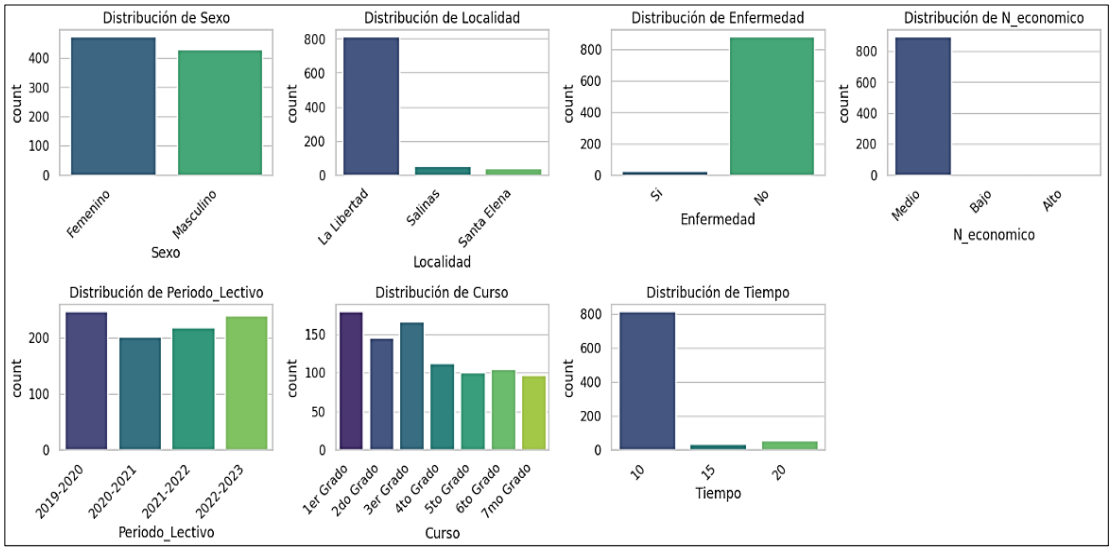


Figura 42: Distribución de Nota_final para cada variable independiente: Elaboración Propia

Luego de los procesos realizados en esta subetapa, el resultado es un nuevo conjunto de datos con 9 variables.

Conversión de datos

Del conjunto de datos se verificó el tipo de variables por las que está compuesta, visualizándose la cantidad de variables existentes y su respectivo tipo de datos.

```
dtypes: int64(6), object(5)
```

Figura 43: Cantidad y tipos de variable del conjunto de datos: Elaboración Propia

Para la aplicación correcta de minería de datos, todas las variables del conjunto de datos deberán convertirse a tipo numéricas, ya que para este caso de predicción se utilizará regresiones, y posteriormente, se aplicarán los algoritmos como Árboles de Decisión de regresión y Vectores de Soporte de Regresión.

Por lo tanto, para la conversión de variables categóricas (o de texto) a variables numéricas, se realizó el proceso One Hot Encoding, y específicamente, se aplicó el método "get.dummies()". Este método toma una o más columnas categóricas de un DataFrame y crea nuevas columnas binarias (dummy variables) para cada categoría única en esas columnas. Estas nuevas columnas binarias indicarán la presencia o ausencia de cada categoría en la fila correspondiente. Siendo así que, el conjunto de datos quedó de la siguiente manera:

	Tiempo	Nota_quim1	Nota_quim2	Nota_final	Femenino	Masculino	La Libertad	Salinas	Santa Elena	No	Si	Alto	Bajo	Medio	2019-2020	2020-2021	2021-2022	2022-2023	1er Grado	2do Grado	3er Grado
0	10	9	9	9	1	0	1	0	0	0	1	0	0	1	1	0	0	0	1	0	0
1	10	7	8	8	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	0
2	10	8	9	9	0	1	1	0	0	1	0	0	0	1	1	0	0	0	1	0	0
3	10	6	7	7	0	1	1	0	0	1	0	0	0	1	1	0	0	0	1	0	0
4	10	9	6	8	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	0

Figura 44: Proceso One Hot Encoding aplicado al conjunto de datos: Elaboración Propia

Minería de datos en Orange Data Mining

Para empezar con la aplicación de las técnicas de minería de datos, primero se debe dividir las variables en dos conjuntos, siendo el primer conjunto el que contiene la variable objetivo (target) y el segundo conjunto, las variables restantes que se analizarán.

Definidos dichos conjuntos, en el software de Orange, el punto de partida será importar el archivo que almacena el conjunto de datos extraído. Definiendo Nota_final como el target.

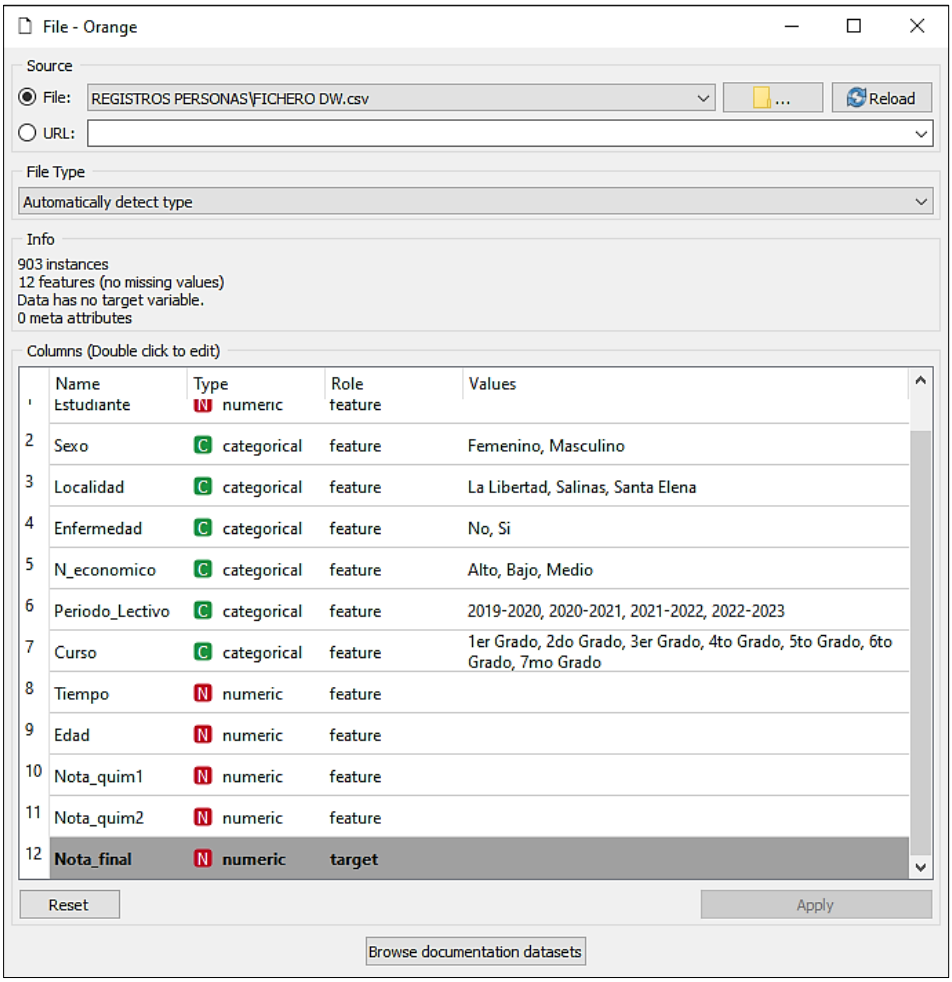


Figura 45: Importación de conjunto de datos en Orange Data Mining: Elaboración Propia

Realizando la comprobación de las estadísticas de los datos, se puede observar que el porcentaje de datos perdidos (Missing) en cada variable es del 0%, lo cual es de importancia para el correcto análisis.

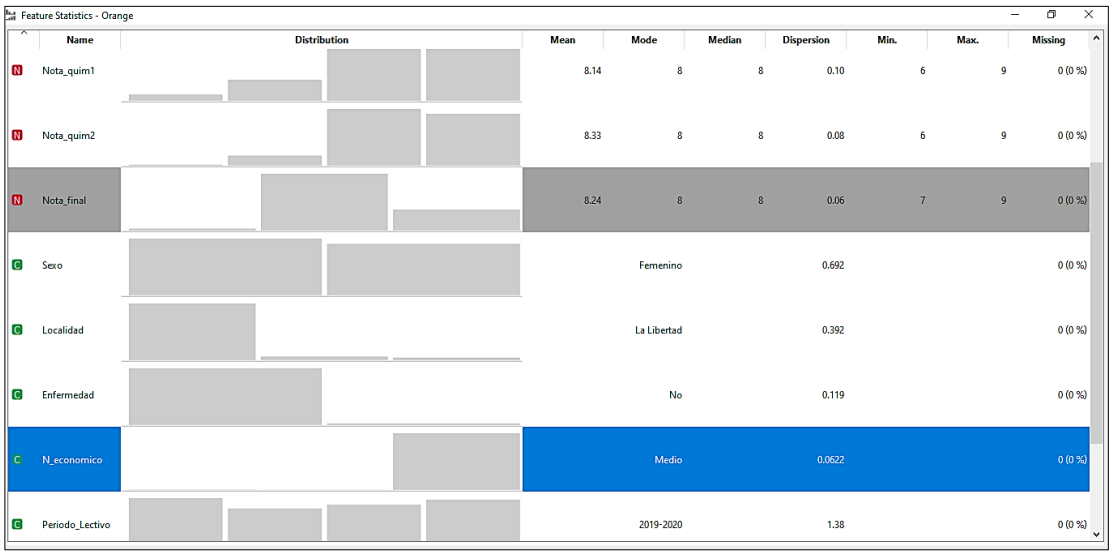


Figura 46: Estadísticas de los datos: Elaboración Propia

Implementación de árboles de decisión

Luego del establecimiento de la variable objetivo, el siguiente paso que se realizó fue enlazar el archivo de datos a un widget de selección de columnas, en el que se definió los siguientes puntos:

- Variables que se tomarían en cuenta para el análisis
- Variables que serían ignoradas
- Variable objetivo que se aplicaría para el árbol de decisión.

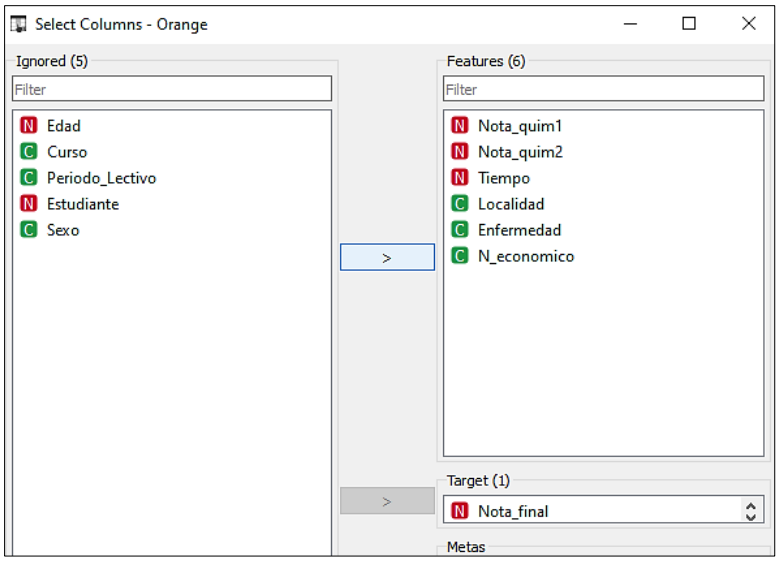


Figura 47: Selección de variables – Árbol de decisión: Elaboración Propia

Posteriormente, se realizó la importación del modelo de árbol de decisiones, en el cual se definieron los parámetros para la configuración del modelo de regresión, siendo esenciales para determinar si el rendimiento del modelo es satisfactorio o si existe el riesgo de sobreajuste. Por lo tanto, se detallan a continuación:

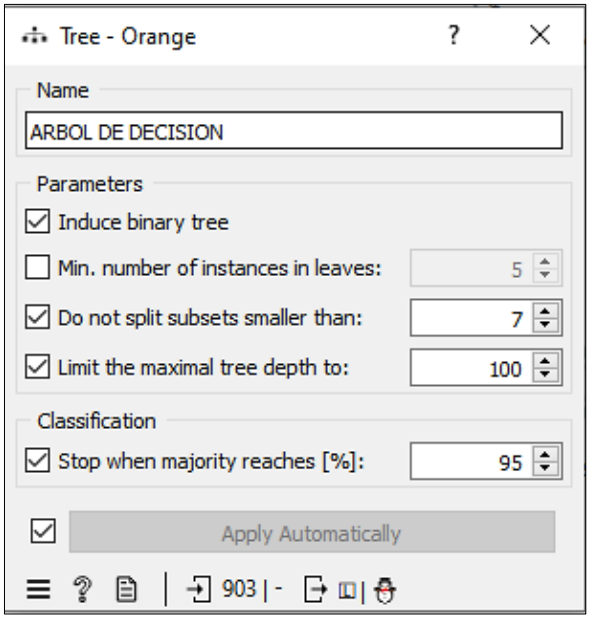


Figura 48: Parámetros del árbol de decisión: Elaboración Propia

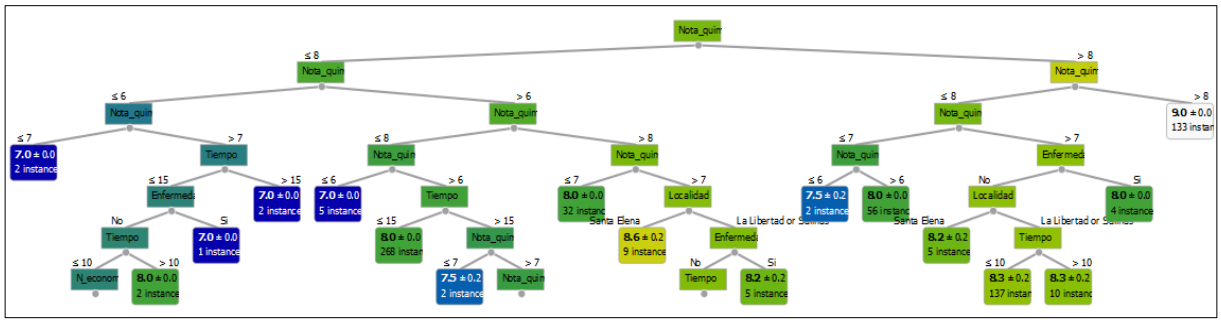


Figura 49: Visualización del árbol de decisión de regresión: Elaboración Propia

Implementación de redes neuronales

Para este modelo se establecieron los mismos datos de variables a evaluar y variable objetivo o target. Además de la configuración con los siguientes parámetros por defecto:

- Neuronas en capas ocultas = 100
- Activación: Unidad Lineal Rectificada “ReLU”
- Solucionador: Adam
- Regularización: a = 0.0001
- Número máximo de iteraciones = 200

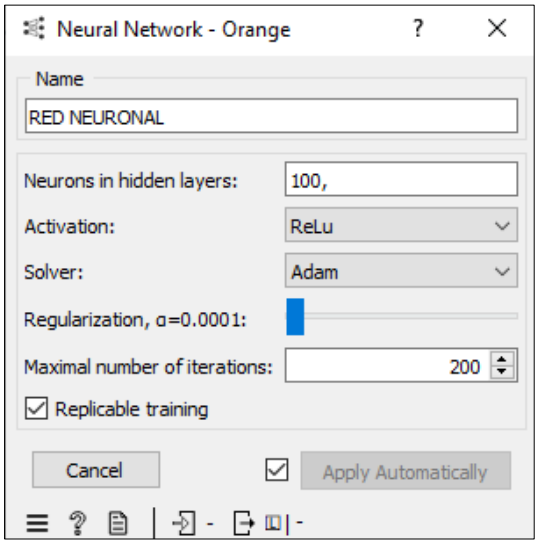


Figura 50: Parámetros de la red neuronal: Elaboración Propia

Implementación de Máquinas de Vectores de Soporte (SVM)

En este modelo se empleó los conjuntos definidos para los anteriores modelos, siendo estos:

- Variables escogidas para el análisis
- Variables ignoradas
- Variable objetivo

Posterior a esto, se agregó el modelo SVM bajo los siguientes parámetros de configuración:

- Tipo: SVM
- Constante (C) : 1,00
- Épsilon de pérdida de regresión: 0,10
- Kernel: RBF (Función de base radial)
- Tolerancia numérica: 0,0010
- Límite de iteraciones: 100

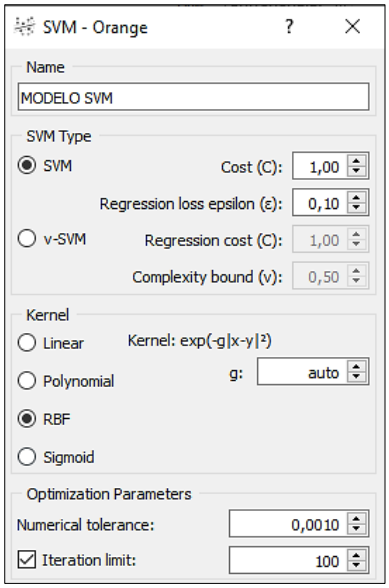


Figura 51: Parámetros de Máquina de Vectores de Soporte: Elaboración Propia

Una vez configurado el modelo con los parámetros detallados anteriormente, se ajustó el modelo y se generó la variable predictiva a partir del conjunto de datos para el análisis.

2.5.4. Etapa 4: Evaluación de los modelos

Para la evaluación de los modelos respecto a su funcionamiento, se usaron las siguientes métricas:

- Error cuadrático medio (MSE)
- Raíz del error cuadrático medio (RMSE)
- Error absoluto medio (MAE)
- Media del error absoluto en porcentaje (MAPE)
- Coeficiente de determinación (R^2)

Estas métricas de evaluación de modelos se centran en medir la precisión y la calidad de las predicciones realizadas por un modelo de regresión. MSE Y RMSE calculan la magnitud de las diferencias entre los valores predichos y los reales, penalizando de manera significativa los errores más grandes. MAE proporciona una medida promedio de las desviaciones absolutas entre las predicciones y las observaciones reales, sin tener en cuenta la dirección de los errores. Por último, MAPE mide el porcentaje promedio de desviación entre las predicciones y los valores reales.

Para recopilar los resultados de las métricas, se enlazó cada modelo a la función de Prueba y Puntuación.

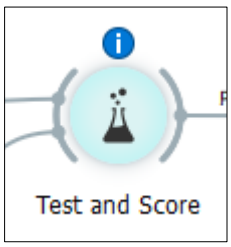


Figura 52: Función Prueba y Puntuación: Elaboración Propia

Métricas de medición de desempeño en árbol de decisión

Los resultados para el árbol de decisión fueron los siguientes:

Métrica	Resultado
MAE	0.175
MSE	0.088
RMSE	0.296
MAPE	0.021
R^2	0.616

Tabla 18: Resultados de métricas en árboles de decisión

Métricas de medición de desempeño en redes neuronales

Los resultados para la red neuronal fueron los siguientes:

Métrica	Resultado
MAE	0.312
MSE	0.176
RMSE	0.420
MAPE	0.038
R^2	0.229

Tabla 19: Resultados de métricas en red neuronal

Métricas de medición de desempeño en máquina de vectores de soporte

Los resultados para la máquina de vectores de soporte fueron los siguientes:

Métrica	Resultado
MAE	0.306
MSE	0.131
RMSE	0.363
MAPE	0.306
R^2	0.425

Tabla 20: Resultados de métricas en SVM

2.5.5. Etapa 5: Difusión de conocimiento

Se realizó una reunión virtual a través de la plataforma de videoconferencia Zoom, para impartir la capacitación a los directivos y encargados administrativos de la institución (Ver Anexos 6 y 7), en la cual se trataron los siguientes puntos clave:

1. **Introducción a la problemática**

Se partió desde la introducción a la problemática, donde se explicó el contexto actual de la institución y la manera en que las herramientas tecnológicas pueden contribuir al desarrollo y optimización de los procesos educativos administrativos, brindando un apoyo en la toma de decisiones. Además, se dio a conocer los problemas que se encontraron y cuáles fueron los métodos de recolección de información.

2. **Objetivos de la propuesta**

Se dio a conocer cuáles fueron los objetivos planteados para la propuesta tecnológica, definiendo la razón de cada uno y su influencia en el desarrollo de las etapas del proyecto.

3. **Metodología y técnicas aplicadas**

Se detalló cuáles fueron los datos que se utilizó para aplicar cada una de las metodologías y técnicas aplicadas en las respectivas etapas del proyecto, realizando la presentación de los resultados obtenidos por los mismos a través de las predicciones.

4. **Resultados**

Se detalló las métricas que fueron empleadas para evaluar cada uno de los modelos, definiendo cual era el mejor y de mayor rendimiento según los valores obtenidos, además de realizar la comparación de los valores reales con los valores generados mediante predicción.

2.6. Resultados

2.6.1. Resultados de la evaluación de los modelos

Una vez realizado el proceso que se definió en la cuarta etapa, en la que se describía la evaluación de cada modelo para verificar cual es el que poseía mayor efectividad en la generación de predicciones, se obtuvieron los siguientes resultados:

Técnicas de minería de datos	**Métricas de medición de desempeño**				
	MAE	**MSE**	**RMSE**	**MAPE**	**R^2**
Árboles de decisión	0.175	0.088	0.296	0.021	0.616
Redes neuronales	0.312	0.176	0.420	0.038	0.229
SVM	0.306	0.131	0.363	0.306	0.425

Tabla 21: Resultados de las métricas de medición del desempeño

Según los resultados obtenidos, la técnica de minería de datos que posee mayor efectividad y un menor valor en cada una de las métricas, es el árbol de decisión, y cuenta con los siguientes valores:

- MAE = 0.175
- MSE = 0.088
- RMSE = 0.296
- MAPE = 0.021
- R^2 = 0.616

Por lo tanto, se pudo verificar que este modelo era el más óptimo al ser implementado. Adicionalmente, se verificó su rendimiento al comparar los valores reales del conjunto de datos con los valores predichos generados por el modelo.

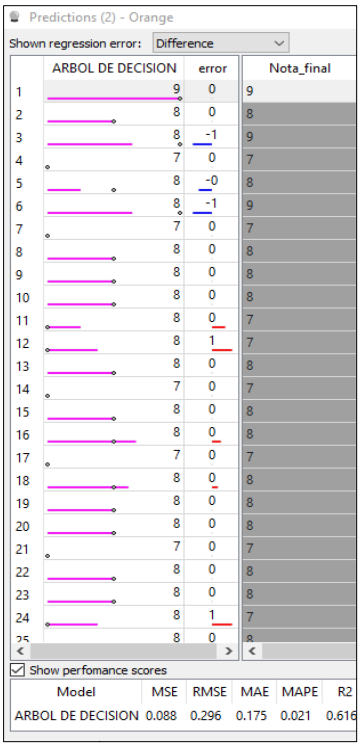


Figura 53: Valores reales y valores generados por predicción del modelo: Elaboración Propia

La columna "error" proporciona una medida cuantitativa de la diferencia entre los valores generados por la predicción del modelo de árbol de decisión y los valores reales. En la que se puede observar su precisión al generar valores iguales en su mayoría.

Para calcular el porcentaje de error absoluto entre una nota real y una predicción, se hace uso de la siguiente fórmula:

$$Error\ porcentual\ absoluto = \frac{Nota\ real - Predicción}{Nota\ Real}\ x\ 100$$

$$Ejemplo\ 1: Error\ porcentual\ absoluto = \frac{9-8}{9}\ x\ 100 = \mathbf{11.11}\ \%$$

$$Ejemplo\ 2: Error\ porcentual\ absoluto = \frac{7-8}{7}\ x\ 100 = \mathbf{14.29}\ \%$$

2.6.2. Patrones obtenidos

Luego de haber obtenido la representación gráfica del árbol de decisión (Figura 46), se pudo determinar cuáles son los patrones encontrados al momento de analizar el rendimiento de los estudiantes, detallándose a continuación:

- Si la calificación del estudiante durante el primer quimestre (Nota_quim1) es mayor a 8, y la nota del segundo quimestre (Nota_quim2) es mayor a 8, el promedio será de 9.0.
- Si la calificación del estudiante durante el primer quimestre (Nota_quim1) es mayor a 8, pero la nota del segundo quimestre (Nota_quim2) es menor a 8. Se relacionará a las siguientes condiciones:
  - ✓ Si Nota_quim2 es menor o igual a 8, 7 y 6, el promedio será 7.5, caso contrario, si es mayor a 6 el promedio será 8.0
  - ✓ Si Nota_quim2 es menor o igual a 8, pero mayor a 7. Se evaluará la enfermedad:
    - Si el estudiante posee enfermedad el promedio será de 8.0.
    - Si el estudiante no posee enfermedad, se analizará la Localidad:
      - Si el estudiante es de Santa Elena, el promedio será de 8.2.
      - Si el estudiante es de La Libertad o Salinas, se analiza el tiempo de viaje:
        - Si el tiempo de viaje es menor, igual o mayor a 10, el promedio será de 8.3.
- Si la calificación del estudiante durante el primer quimestre (Nota_quim1) es menor o igual a 8, y la nota del segundo quimestre (Nota_quim2) es menor o igual a 6. El promedio será de 7.0

- Si la calificación del estudiante durante el primer quimestre (Nota_quim1) es menor o igual a 8, Nota_quim2 menor o igual a 6, pero mayor a 7. Se evaluará lo siguiente:
    - Si el tiempo de viaje es mayor a 15, el promedio será de 7.0.
    - Si el tiempo de viaje es menor o igual a 15, se analizará la variable de Enfermedad:
        - Si el estudiante posee enfermedad, el promedio será de 7.0, caso contrario se evaluará nuevamente el tiempo de viaje:
            - Si el tiempo de viaje es mayor a 10, el promedio será de 8.0
            - Si el tiempo de viaje es menor o igual a 10, se evaluará el nivel económico:
                - Si el nivel económico del estudiante es Bajo o Medio, el promedio será de 7.8
                - Si el nivel económico del estudiante es Alto, el promedio será de 8.0.

2.6.3 Resultados de la variable

La variable que se planteó para el proyecto se basa en el tiempo de obtención de reportes relacionado al rendimiento académico de los estudiantes. En la entrevista realizada al secretario de la institución (Ver Anexo 2), se determinó que, el rendimiento de los estudiantes se evalúa de manera formativa y sumativa, pero estos procedimientos conllevan de aproximadamente seis días, ya que las evaluaciones diagnósticas se las realizan durante toda la semana, además de que solo se toman en cuenta variables numéricas como las notas. Es por ello, que la solución planteada a través de esta propuesta, permitió entender y conocer que con la ayuda de las técnicas de minería de

datos implementada en esta área del rendimiento académico, se lograría mejorar el tiempo de obtención de los reportes, reduciéndolo a dos días.

Tiempo de obtención de reportes	
Cantidad de días (Antes)	**Cantidad de días (Después)**
6	2

Tabla 22: Comparación de tiempo de obtención de reportes

CONCLUSIONES

- Mediante el uso de técnicas de recopilación de datos, como observación y entrevista, se logró obtener información básica sobre la institución, identificando los métodos de evaluación y el tiempo de obtención de reportes acerca del rendimiento académico estudiantil.
- Al ejecutarse la subetapa de exploración de datos, y realizándose una matriz de correlación de Pearson, se verificó que las variables que tienen mayor asociación lineal a la nota final del estudiante son las notas del quimestre 1 y 2, obteniendo una correlación positiva de 0.50 y 0.42 respectivamente, lo que indica que dichas variables estudiadas guardan relación directa con la variable dependiente.
- Luego de la implementación del modelo de vectores de soporte y al ser evaluado por las métricas de medición de desempeño, se obtuvo los siguientes valores de error: MAE (0.306), MSE (0.131), RMSE (0.363), MAPE (0.306), R^2 (0.425), dando como resultado ser el segundo modelo con menor error en las predicciones.
- Luego de la implementación del modelo de redes neuronales y al ser evaluado por las métricas de medición de desempeño, se obtuvo los siguientes valores de error: MAE (0.312), MSE (0.176), RMSE (0.420), MAPE (0.038), R^2 (0.229), dando como resultado ser el tercer modelo con menor error en las predicciones.
- Se comprobó que el modelo que obtuvo un mejor rendimiento luego de ser evaluado por las métricas de medición del desempeño, fueron los árboles de decisión. Dando los siguientes valores de error: MAE (0.175), MSE (0.088), RMSE (0.296), MAPE (0.021), R^2 (0.616)**.** La verificación de que el modelo creado alcanzó su nivel óptimo se evidencia al comparar los valores generados por predicción con los valores reales, estableciendo así su eficacia.

- Se verificó que el modelo de árboles de decisiones presentó un porcentaje de error con un rango entre el 11.11% y 14.29% al comparar los valores reales con los valores generados mediante predicción, y aplicando la fórmula del error porcentual absoluto.
- El desarrollo de la propuesta permitió corroborar que, el tiempo empleado para la obtención de los reportes acerca del rendimiento académico de los estudiantes se disminuyó de 6 a 2 días.

RECOMENDACIONES

- Se recomienda ampliar el alcance de la recopilación de datos para obtener una comprensión más detallada del rendimiento académico estudiantil en la institución. Se sugiere la implementación de métodos de evaluación más estructurados y cuantificables, como el seguimiento continuo de indicadores clave de rendimiento (KPIs).
- Debido a los resultados obtenidos con la matriz de correlación de Pearson y que al observarse que, si bien se pueden establecer relaciones fuertes entre variables, se recomienda que se implemente un modelo de regresión lineal multivariante para determinar si los índices correlaciones se mantienen.
- Aplicar una metodología distinta, orientada al área de inteligencia de negocios y minería de datos, verificando si el desarrollo de esta clase de procesos se puede optimizar en una menor cantidad de fases.
- Se recomienda aplicar otras técnicas de minería de datos, con el fin de evaluar su rendimiento al ser implementados y verificar si hay un modelo que tenga un menor valor de error en la generación de análisis predictivo.

BIBLIOGRAFÍA

[1] R. A. R., Inteligencia de negocios: Un enfoque práctico, Bogotá: Ecoe Ediciones, 2004.

[2] S. Vincent-Lancrin, J. Urgel, S. Kar y G. Jacotin, Measuring Innovation in Education 2019: What Has Changed in the Classroom? Educational Research and Innovation, Paris: OECD, 2019.

[3] T. A. M. Ahmed Alsanad, «Predicting Students Performance Using Classification Techniques,» *International Journal of Computer Science and Network Security,* vol. 22, nº 9, 2022.

[4] E. Yamao, Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de las Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú: USMP, 2018.

[5] C. G. L. Rodas, Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca, Cuenca: Repositorio Institucional UCuenca, 2019.

[6] G. P.-S. P. S. Usama Fayyad, «From Data Mining to Knowledge Discovery in Databases,» *AI Magazine,* vol. 17, nº 3, p. 37, 1996.

[7] Consejo de la Facultad de Sistemas y Telecomunicaciones, «Resolución RCF-FST-SO-09 No. 03-2021,» 2021.

[8] A. A. Viviana Vanessa Ruiz Díaz de Salvioni, «La importancia de la Minería de Datos como una herramienta estratégica en las Empresas,» *Ciencia Latina Internacional,* vol. 7, nº 2, 2023.

[9] B. B. Martínez, «Minería de Datos,» [En línea]. Available: http://bbeltran.cs.buap.mx/NotasMD.pdf.

[10] M. L. T. M. H. M. S. Carlos Eduardo Marulanda Echeverry, «Minería de datos en gestión del conocimiento de pymes de Colombia,» *Revista Virtual Universidad Católica del Norte,* pp. 224-237, 27 03 2017.

[11] J. L. Cano, Business Inteligence: Competir con Información, Madrid: Fundación Cultural Banesto, 2007.

[12] R. R. K. G. José C. Riquelme, «Minería de datos: Conceptos y Tendencias,» *Revista Iberoamericana de Inteligencia Artificial*, vol. 10, nº 29, pp. 11-18, 2006.

[13] S. N. d. Planificación, «Plan de Creación de Oportunidades 2021 - 2025,» 2023. [En línea]. Available: https://observatorioplanificacion.cepal.org/sites/default/files/plan/files/Plan-de-Creaci%C3%B3n-de-Oportunidades-2021-2025-Aprobado_compressed.pdf.

[14] C. R. R. K. M. S. Hugo Sánchez Carlessi, Manual de términos en investigación científica, tecnológica y humanística, Lima: Universidad Ricardo Palma, 2018.

[15] M. d. Educación, «Base de datos AMIE,» 2023. [En línea]. Available: https://educacion.gob.ec/base-de-datos/.

[16] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «De la minería de datos al descubrimiento de conocimiento en bases de datos,» *KDD*, vol. 17, nº 3, p. 18, 1996.

[17] M. d. Educación, «Ley Orgánica de Educación Intercultural,» 05 2021. [En línea]. Available: https://educacion.gob.ec/wp-content/uploads/downloads/2017/05/Ley-Organica-Educacion-Intercultural-Codificado.pdf. [Último acceso: 25 09 2023].

[18] R. Camps Paré, L. A. Casillas Santillán, D. Costal Costa, M. Gibert Ginestá, C. Martín Escofet y O. Pérez Mora, Bases de datos, Cataluña: Eureca Media, SL, 2005.

[19] Microsoft, «Microsoft,» [En línea]. Available: https://learn.microsoft.com/es-es/sql/relational-databases/databases/databases?view=sql-server-ver16. [Último acceso: 25 09 2023].

[20] Microsoft, «Microsoft,» [En línea]. Available: https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16. [Último acceso: 25 09 2023].

[21] Microsoft, «Microsoft: Visual Studio,» [En línea]. Available: https://visualstudio.microsoft.com/es/vs/. [Último acceso: 25 09 2023].

[22] A. Abelló Gamazo, J. Curto Díaz, Á. Rius Gavídia, M. Serra Vizern, J. Samos Jiménez y J. Vidal Gil, «Introducción al Data Warehouse,» [En línea]. Available: https://openaccess.uoc.edu/bitstream/10609/136246/5/Disen%C2%BFo%20y

%20construccio%C2%BFn%20de%20un%20almace%C2%BFn%20de%20datos_Mo%C2%BFdulo1_Introduccio%C2%BFn%20al%20Data%20Warehouse.pdf.

[23] E. Y. Del Castillo Gabriel y J. P. Sandoval Ordoñez, IMPLEMENTACIÓN DE UN DATAMART PARA LA TOMA DE DECISIONES PARA LAS VENTAS DE CONTENEDORES EN EL ÁREA COMERCIAL EN LA EMPRESA SPACEWISE PERÚ, Lima: Repositorio Académico USMP, 2016.

[24] R. Gupta, Making Use of Python, New York: Wiley Publishing Inc, 2002.

[25] Jupyter, «Jupyter,» [En línea]. Available: https://jupyter.org/. [Último acceso: 25 09 2023].

[26] Microsoft, «Microsoft Excel,» [En línea]. Available: https://www.microsoft.com/es/microsoft-365/excel. [Último acceso: 25 09 2023].

[27] Gidahatari, «Keras,» [En línea]. Available: https://gidahatari.com/ih-es/tutorial-completacion-datos-hidrologicos-inteligencia-artificial-python-keras#:~:text=Keras%20es%20una%20plataforma%20de,de%20los%20datos%20de%20entrada.. [Último acceso: 25 09 2023].

[28] S. learn, «Scikit learn,» [En línea]. Available: https://scikit-learn.org/stable/. [Último acceso: 25 09 2023].

[29] Matplotlib, «Matplotlib,» [En línea]. Available: https://matplotlib.org/. [Último acceso: 25 09 2023].

[30] O. D. Mining, «Orange Data Mining,» [En línea]. Available: https://orangedatamining.com/. [Último acceso: 2023].

[31] S. R. Gardner, «Building the Data Warehouse,» *Association for Computing Machinery. Communications of the ACM,* vol. 41, pp. 52-61, 1998.

[32] IBM, «IBM: El modelo de redes neuronales,» 17 08 2021. [En línea]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model.

[33] J. A. Alonso Jiménez y M. A. Gutiérrez Naranjo, «Arboles de decisión,» 2001. [En línea]. Available: https://www.cs.us.es/~jalonso/cursos/ra-00/temas/tema-12.pdf.

[34] J. A. R. Trejo, «Las maquinas de vectores de soporte para identificación en línea,» 09 2006. [En línea]. Available: https://www.ctrl.cinvestav.mx/~yuw/pdf/MaTesJAR.pdf.

[35] K. J. Danjuma, «Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients,» 17 04 2015. [En línea]. Available: https://arxiv.org/abs/1504.04646.

[36] E. D. M. Society, «Educational Data Mining,» 2011. [En línea]. Available: https://educationaldatamining.org/. [Último acceso: 2023].

[37] S. f. L. A. Research, «Solaresearch,» 2016. [En línea]. Available: https://www.solaresearch.org/about/what-is-learning-analytics/. [Último acceso: 2023].

[38] A. Ballesteros Román, D. Sánchez Guzmán y R. García Salcedo, «Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo,» *Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada,* nº 694, 2013.

[39] G. O. Benítez, «Las Asignaturas Pendientes y el Rendimiento Académico:¿Existe Alguna Relación?,» 2000.

[40] P. G. Imig, «Rendimiento académico: un recorrido conceptual que aproxima a una definición unificada para el ámbito superior,» *Universidad Nacional de Mar del Plata,* pp. 87-102, 2020.

[41] M. E. E. H. L. R. J. C. B. Osvaldo M. Spositto, «Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil,» 02 07 2010. [En línea]. Available: https://repositoriocyt.unlam.edu.ar/handle/123456789/1267.

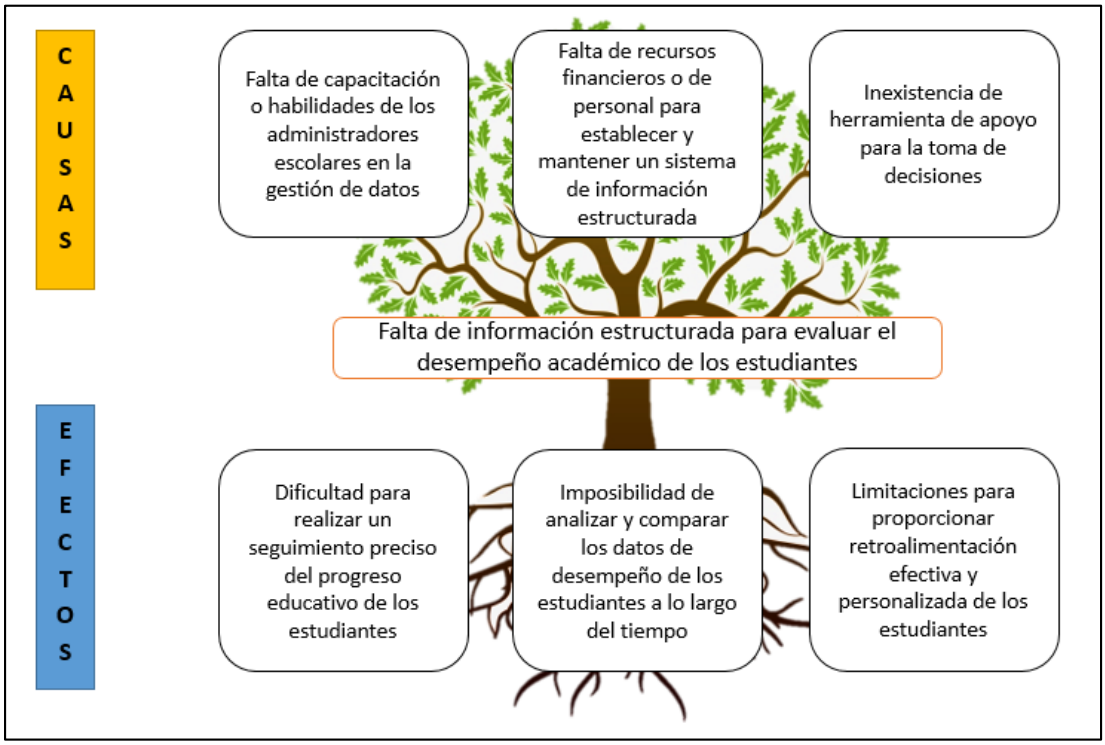
[42] J. V. D. M. Moris, «MINERÍA DE DATOS EDUCACIOMODELOS DE PREDICCIÓN DEL DESEMPENO ESCOLAR EN ALUMNOS DE ENSEÑANZA BÁSICA,» 04 01 2013. [En línea]. Available: https://repositorio.uchile.cl/bitstream/handle/2250/113034/cf-vandermolen_jm.pdf?sequence=1&isAllowed=y.

[43] G. H. J. Molina López José Manuel, «Técnicas de Análisis de Datos: Aplicaciones Prácticas Utilizando Microsoft Excel y Weka,» *Universidad de Jaén,* pp. 153-154, 2006.

[44] T. Naeem, «Astera,» 15 11 2023. [En línea]. Available: https://www.astera.com/es/type/blog/data-warehouse-concepts/#What-Are-the-Two-Data-Warehouse-Concepts-Kimball-vs-Inmon-Explained.

ANEXOS

Anexo 1: Árbol de problemas



Anexo 2: Formato de entrevista

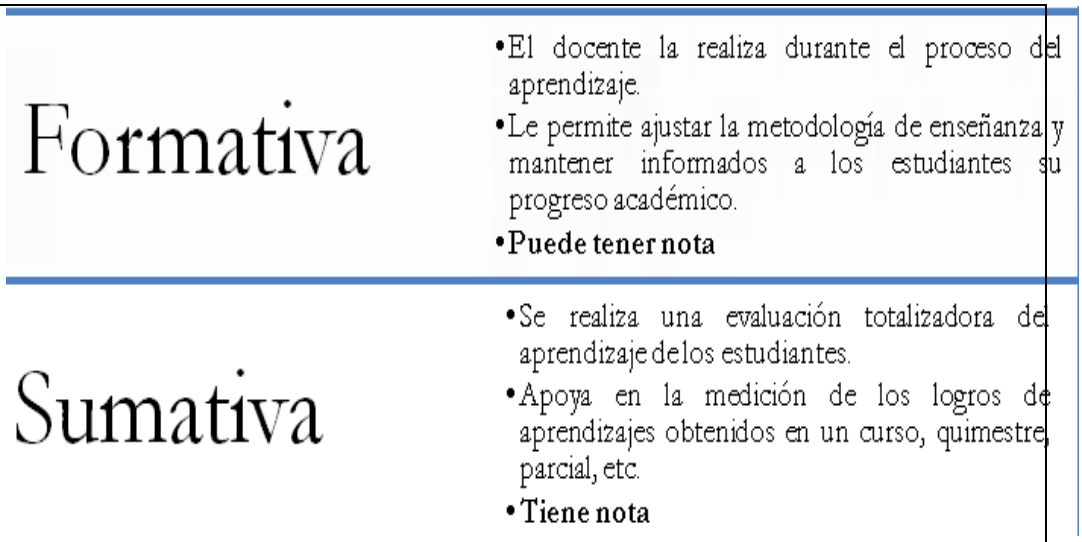
ENTREVISTA	
Entrevistador:	Kelvin Roosbelth Aguirre Chamba
Entrevistado:	Lcdo. Carlos Rubén Rivera Ramírez
Rol del entrevistado:	Secretario de la Escuela de Educación Básica 26 de septiembre
Objetivo:	Obtener información básica sobre la unidad educativa, así como de su organización.

Fecha:	07/06/2023
Ubicación de la escuela:	La Libertad – Santa Elena

1. **¿Cuál es la fecha de fundación de la escuela?**
   4 de abril de 2007
2. ¿**Cuántos trabajadores hay en la escuela, incluyendo a los docentes, personal administrativo y de apoyo?**
   Hay 11 trabajadores en total en la institución.
3. **¿Cuántos de estos trabajadores son docentes y cuántos son personal administrativo o de apoyo?**
   10 docentes y 1 administrativo
4. **¿Cuál es la cantidad de alumnos que asisten actualmente a la escuela?**
   Hay 236 alumnos actualmente
5. **¿Existen desafíos o dificultades en la toma de decisiones relacionadas con el desempeño académico de los estudiantes?**
   Las dificultades que se presentan en la parte académica de los estudiantes es cuando los padres los descuidan, no presentan tareas, las evaluaciones son bajas por que los estudiantes manifiestan que los padres no los hacen estudiar. Aunque es muy bajo esa cantidad, dando ejemplos de 3 a 5 en el total de estudiantes.
6. **¿Cuál es el nivel de rendimiento académico actual de los estudiantes?**
   No tenemos un promedio general actualmente ya que el primer parcial culmina la segunda semana de julio, pero en un pequeño sondeo, el promedio va desde 8,50 a 9 por año básico.

7. **¿Cómo se evalúan los logros o el rendimiento académico de los estudiantes en la institución?**

Se evalúan de manera Formativa y Sumativa



8. **¿El proceso de análisis de resultados respecto al rendimiento académico es ágil?**

   Utilizamos un formato de Excel en donde ya está configurada sus celdas con fórmulas, en donde los docentes solo van ubicando las notas y el promedio sale automáticamente de la misma manera está vinculado con el reporte final.

9. **¿Qué expectativas a futuro tiene usted sobre la institución?**

Desde nuestro inicios hemos aspirado mucho en nuestra institución, cabe recalcar que solo éramos un jardín de infantes y a futuro queríamos tener escuela, cuando logramos eso aspiramos a tener colegio, esto se hizo realidad hace 2 periodos lectivos, luego queríamos ampliar el patio y también se logró, luego aspiramos a tener proyectores en todas aulas y se logró, cámaras de seguridad ya tenemos en toda la institución, queremos tener a futuro un departamento de psicología y psicopedagogía es lo que aspiramos a futuro.

10. **¿Considera que la aplicación de técnicas de minería de datos para predecir el rendimiento académico de los estudiantes permitirá mejorar**

el análisis y proporcionar una comprensión más completa de los mismos?

Por supuesto, toda aportación para mejorar el desarrollo de la comunidad educativa siempre es beneficioso.

Anexo 3: Entrevista al Licenciado Carlos Rivera Ramírez (Secretario de la institución)

Anexo 4: Entrevista a la Licenciada Sandra Ramírez Quimí (Rectora de la institución)

Anexo 5: Misión y visión de la institución.

Anexo 6: Etapa de difusión de conocimientos

Anexo 7: Muestra de resultados obtenidos por los modelos de minería de datos

Técnicas de minería de datos	Métricas de medición de desempeño				
	MAE	MSE	RMSE	MAPE	R^2
Árboles de decisión	0.175	0.088	0.296	0.021	0.616
Redes neuronales	0.312	0.176	0.420	0.038	0.229
SVM	0.306	0.131	0.363	0.306	0.425