UNIVERSIDAD ESTATAL
PENÍNSULA DE SANTA ELENA

FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN

TRABAJO DE TITULACIÓN

Propuesta Tecnológica, previo a la obtención del Título de:

INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

"Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.'"

AUTOR:

VILLAO BALÓN ALEX JOAO

PROFESOR TUTOR:

ING. WALTER OROZCO IGUASNIA

LA LIBERTAD - ECUADOR

2021

AGRADECIMIENTO

A mis queridos padres y mi hermana por su apoyo incondicional durante mi carrera, por haber forjado parte de mi carácter para con los demás, por mostrarme que en esta vida hay obstáculos que se deben superar para un bien mayor.

A la universidad por ser mi centro de acopio de conocimientos.

A mis docentes, Marjorie Coronel y Walter Orozco, por guiarme a lo largo del trabajo final, por su tiempo y apoyo. También, a los docentes a lo largo de mi carrera por haber impartido sus enseñanzas de forma gradual, y no solo conocimientos teóricos, sino, estilos de vida.

A mis amigos Andrés, Katherine, José y Johnny, por ser parte de IRIS, donde no solo se demostró que podemos superar nuestros propios límites, sino que, mediante la cooperación, cualquier meta se puede lograr.

Alex J. Villao

DEDICATORIA

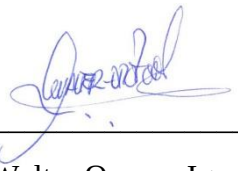
El presente trabajo va dedicado principalmente a mis padres y mi hermana por ser mis pilares fundamentales a lo largo de mi carrera estudiantil y demostrarme que, a pesar de las circunstancias, se debe seguir adelante. También, a mi familia en general.

Alex J. Villao

APROBACIÓN DEL TUTOR

En mi calidad de Tutor del trabajo de titulación denominado: **"Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.'"**, elaborado por el estudiante **Villao Balón Alex Joao**, de la carrera de **Tecnologías de la información** de la Universidad Estatal Península de Santa Elena, me permito declarar que luego de haber orientado, estudiado y revisado, lo apruebo en todas sus partes y autorizo al estudiante para que inicie los trámites legales correspondientes.

La libertad, 18 de agosto del 2021.



Ing. Walter Orozco Iguasnia

TRIBUNAL DE GRADO

Ing. Washington Torres, Mgt.
DIRECTOR (e) DE LA CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN

Ing. Shendry Rosero
DOCENTE ESPECIALISTA

Ing. Walter Orozco, MSc.
DOCENTE TUTOR

Ing. Marjorie Coronel, MGTI.
DOCENTE GUÍA

RESUMEN

Una de las metas que posee la escuela 'Lic. Angélica Villón L.' es el desarrollo estratégico, lo que conlleva una ampliación de infraestructura, y también, reconocimientos por su calidad y excelencia académica. Para cumplir esto, se priorizan elementos como la población estudiantil, el cual, se evalúa mediante indicadores como el rendimiento académico. No obstante, la evaluación de este indicador se encuentra limitada debido a factores como: el escaso conocimiento que impide una óptima toma de decisiones sobre el rendimiento académico, y, la carencia de una gestión y organización de la información adecuada. La institución es privada, sin embargo, al existir una falta de información estructurada, la parcialidad y subjetividad generan resultados pocos convincentes. Para solucionar el problema mencionado, se propuso aplicar técnicas de minería de datos con el objetivo de predecir el rendimiento académico en la institución. El procedimiento se basó en la metodología Descubrimiento de Conocimiento de Base de datos (KDD). La metodología KDD se compone principalmente de cinco fases. Estas van desde la recolección e integración de datos hasta la minería de datos, con la finalidad de obtener conocimiento que será útil a los administradores de la institución. Entre las tecnologías empleadas se encuentra Pentaho Data Integration. Esta herramienta se caracteriza por la generación de procesos de extracción, transformación y carga de datos, así, se permitirá la respectiva integración de los mismos. Para la creación de los modelos de minería de datos, resalta la herramienta Jupyter Notebook. Entre los resultados logrados se encuentran: el empleo de herramientas inteligencia de negocios creando un almacén de datos para la resolución de problemas de gestión y análisis empresarial; y, la aplicación de técnicas de minería de datos como redes neuronales y árboles de decisión de regresión, mediante la creación de modelos de regresión, para predecir el rendimiento académico de la institución.

Palabras claves: KDD, minería de datos, almacén de datos, Inteligencia de Negocios.

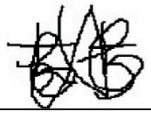
ABSTRACT

One of the goals of the ‘Lic. Angélica Villón L. 'is strategic development, which entails an expansion of infrastructure, and also, recognition for its quality and academic excellence. To achieve this, elements such as the student population are prioritized, which is evaluated through indicators such as academic performance. However, the evaluation of this indicator is limited due to factors such as: the scarce knowledge that prevents optimal decision-making on academic performance, and the lack of adequate information management and organization. The institution is private, however, as there is a lack of structured information, bias and subjectivity generate unconvincing results. To solve the aforementioned problem, it was proposed to apply data mining techniques in order to predict academic performance in the institution. The procedure was based on the methodology Knowledge Discovery Database (KDD). The KDD methodology is mainly composed of five phases. These range from data collection and integration to data mining, in order to obtain knowledge that will be useful to the administrators of the institution. Among the technologies used is Pentaho Data Integration. This tool is characterized by the generation of data extraction, transformation and loading processes, thus, their respective integration will be allowed. For the creation of the data mining models, highlight the Jupyter Notebook tool. Among the results achieved are: the use of business intelligence tools creating a data warehouse for solving management problems and business analysis; and, the application of data mining techniques such as neural networks and regression decision trees, through the creation of regression models, to predict the academic performance of the institution.

Keywords: KDD, data mining, data warehouse, Business Intelligence.

DECLARACIÓN

El contenido del presente Trabajo de Graduación es de mi responsabilidad; el patrimonio intelectual del mismo pertenece a la Universidad Estatal Península de Santa Elena.



Villao Balón Alex Joao

TABLA DE CONTENIDOS

ITEM **PÁGINA**

ÍNDICE DE FIGURAS

ÍTEM DESCRIPCIÓN **PÁGINA**

ÍNDICE DE TABLAS

ÍTEM DESCRIPCIÓN **PÁGINA**

LISTA DE ANEXOS

ÍTEM DESCRIPCIÓN **PÁGINA**

INTRODUCCIÓN

Para un progreso óptimo, las instituciones buscan priorizar y establecer estrategias efectivas que solventen sus necesidades. En este aspecto, la información toma un papel primordial, y es, a través de la misma, que las instituciones desarrollan una mentalidad basada en la mejora continua, donde el análisis de los datos se vuelve una característica fundamental. Así, resalta la escuela "Lic. Angélica Villón Lindao" ubicada en el cantón de Santa Elena de la provincia del mismo nombre.

Esta institución, para evaluar asuntos internos de una manera más detallada como el rendimiento académico, tiene la necesidad de poseer información que facilite la toma de decisiones posteriores relacionada a este proceso. Por ende, la aplicación de metodologías de minería de datos llega a facilitar el análisis extenso del flujo de información, favoreciendo así, no solo la toma de decisiones, sino la comunicación y la administración de datos. Por esta razón, se propone aplicar técnicas de minería de datos para predecir el rendimiento académico de los estudiantes dentro de la institución.

Entre los trabajos previos, relacionados a esta temática, se encuentra "Técnicas de minería de datos para mejorar la precisión de las predicciones del rendimiento académico de los estudiantes: un estudio de caso con Xorro-Q". En este sobresale el uso de modelos clasificadores, sin embargo, existen restricciones como la limitación a un solo curso o el desequilibrio de los datos durante su evaluación. Otro trabajo es "Uso de Aprendizaje Supervisado para predecir el rendimiento estudiantil" desarrollado por Murat Pojon perteneciente a University of Tampere cuyo estudio se basa en fuentes de datos públicas.

El trabajo de titulación se estructura en dos capítulos detallados a continuación:

En el capítulo I se realiza la descripción de los antecedentes abarcando las principales problemáticas de las que es partícipe la institución, resaltando, sobre todo, una falta de información estructurada que permita agilizar procesos, y, de manera adecuada, evaluar indicadores como el rendimiento académico de los

estudiantes. Así, se plantearon los objetivos y la respectiva metodología a aplicar, siendo esta, el Descubrimiento de Conocimiento en Bases de datos (KDD).

En el capítulo II se definen las herramientas empleadas para la elaboración del trabajo, y también, se desarrollan cada una de las etapas de la metodología antes planteada. Esta metodología se compone de cinco etapas, siendo la primera la recopilación e integración de los datos. La segunda abarca la creación de un almacén de datos o data warehouse, mientras que la tercera, corresponde a la aplicación de tres técnicas de minería de datos, las cuales son: árboles de decisión, redes neuronales y vectores de soporte de regresión, una variante de Máquinas de Vectores de Soporte (SVM). La cuarta etapa se basa en la evaluación de las técnicas ya mencionadas, para lo cual, se emplean tres métricas de regresión: error absoluto medio (MAE), error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). La quinta se refiere al difusión y posterior uso del conocimiento obtenido. Finalmente, se exponen los resultados obtenidos a lo largo del trabajo.

CAPÍTULO I

1. Fundamentación

1.1. Antecedentes

La falta de información estructurada, es un grave problema que puede existir en el ámbito de los negocios (empresas, organizaciones o instituciones), pues, afecta a la toma de decisiones [1]. Cuando las instituciones académicas priorizan estrategias, involucrando elementos como la población estudiantil, uno de los factores que se evalúa es el rendimiento académico estudiantil [2]. Así, al buscar un desarrollo estratégico, se genera un análisis de datos ineficaz que proporciona malos resultados para la institución debido al escaso conocimiento del que se dispone, evitando así, una toma de decisiones adecuada sobre el rendimiento académico.

Según Óscar Erazo, en su artículo "El rendimiento académico, un fenómeno de múltiples relaciones y complejidades" para la revista Vanguardia Psicológica, este "es reconocido por su capacidad clasificatoria y su vinculación a la promoción y evaluación de estudiante […] Sin embargo, esta condición no es válida, en tanto existen factores subjetivos y sociales que lo atraviesan, convirtiéndolo en una condición fenomenológica" [3]. De esta manera, al no haber información corroborada por medio de datos, la parcialidad y subjetividad se convierten en opciones pocos convincentes.

Bajo ese contexto, la Escuela de Educación Básica "Angélica Villón Lindao" se convierte en el núcleo de esta propuesta tecnológica. La escuela se encuentra ubicada en el cantón Santa Elena de la provincia del mismo nombre. Se fundó el 12 de septiembre de 2000 y debe su nombre a una educadora santaelenense caracterizada por su "ayuda a los más necesitados". En sí, los cambios que ha tenido repercutieron de forma gradual en el proceso de toma de decisiones desde el ámbito administrativo. La información se corrobora mediante las palabras de uno de sus administradores, el arquitecto Mauricio Yunda, quien manifestó en una entrevista

(ver Anexo 1) que, “el cambio de directiva, con el paso de los años, ha generado inconvenientes en el desarrollo estratégico de la misma, debido a diferencias en su ideología”.

Otra problemática a considerar es que, debido a la carencia de una gestión y organización de la información adecuada, muchas veces se origina una dificultosa colaboración en el ahorro de costos y tiempo para los procesos de la escuela. Este aspecto se convierte en un factor clave debido a que la institución es privada, pues, tampoco cuenta con una infraestructura de TI para solventar necesidades tecnológicas. Por ende, al haber uniformidad de información para realizar predicciones sobe el rendimiento académico, se genera una baja verosimilitud en los análisis e interpretaciones que se realizan, más, si la institución quiere llegar ser el “pináculo” de renombre en toda la provincia.

Además, cabe recalcar que, la escuela tiene como objetivo a largo plazo, extender su infraestructura. Esto conlleva a tener una cantidad de estudiantes mayor que en años anteriores, así, la evaluación del rendimiento académico debe volverse más rígida.

Los métodos de recolección de datos para llevar a cabo esta propuesta se centran en entrevistas con los directores administrativos, además, para realizar los estudios necesarios, y así, determinar un análisis predictivo congruente, se necesitarán de registros de datos de estudiantes que proporcionen la información necesaria sobre estos.

Con respecto al ámbito mundial sobresale el trabajo “Técnicas de minería de datos para mejorar la precisión de las predicciones del rendimiento académico de los estudiantes: un estudio de caso con Xorro-Q” desarrollado por Gomathy Suganya, un tesista de Massey University [4]. Los resultados de este trabajo demostraron que la predicción del rendimiento académico de los estudiantes se puede realizar con las funciones de Xorro-Q, una herramienta educativa usada en ingeniería, con una precisión del 86%, donde los algoritmos de Bosques Aleatorios, o Random Forest,

superan todos los demás clasificadores. Al ser la precisión una métrica importante para evaluar la eficacia del algoritmo, esta no fue tan alta, lo que produjo inconvenientes al tratar de diferenciar el desempeño de los clasificadores. La razón es que los datos disponibles están extremadamente desequilibrados y la elección de las mejores funciones es engorrosa, llegando a afectar a la precisión. El presente estudio utilizó solo los datos de Xorro-Q recopilados en un semestre y para un curso en particular; por lo tanto, el análisis se restringió.

Otro trabajo a nivel mundial es "Uso de Aprendizaje Supervisado para predecir el rendimiento estudiantil" desarrollado por Murat Pojon perteneciente a University of Tampere [5]. El objetivo de esta tesis fue comparar la selección de métodos y la ingeniería de características, en términos de su capacidad para mejorar los resultados de predicción. Se analizaron dos conjuntos de datos diferentes con tres métodos de aprendizaje automático diferentes, y sus resultados se compararon utilizando cuatro medidas de evaluación. Esta investigación tiene ciertas limitaciones que deben tenerse en cuenta. No hubo acceso a un conjunto de datos de estudiantes dedicado, y el estudio se basa en fuentes de datos públicas. Además, no hubo variedad de métodos de aprendizaje automático.

Desde una perspectiva local, sobresale el trabajo "Minería de datos aplicada a la detección de patrones para el análisis de rendimiento académico de los estudiantes de la carrera de Ingeniería en Sistemas Computacionales de la Universidad Católica Santiago de Guayaquil" de Juan José Solines Bernardino. Este trabajo consistió en "elaborar un modelo predictivo en beneficio para los estudiantes, al momento que se inscriban en el semestre, para indicarle cuál es su probabilidad de éxito, alerta o fracaso basado en su rendimiento y en las materias que vaya a tomar" [6]. La limitación de este modelo está en que no se usan otras técnicas de minería de datos, como redes neuronales, para validar la información.

En sí, existe una limitante enorme al no poseer una información organizada que permita, a través de un adecuado análisis, predecir cómo serán los patrones futuros del rendimiento académico, para poder así, tomar las decisiones adecuadas que

salvaguarden las metas institucionales. Además, una vez analizados los trabajos anteriores se determina que, dos limitantes que aparecen en este proceso son: la cantidad de estudiantes evaluados, o población, y la variedad de técnicas empleadas para realizar una correcta minería de datos.

Cabe recalcar que la minería de datos es una etapa del proceso denominado KDD, o Descubrimiento de Conocimiento en Bases de Datos. Este trabajo busca, por medio de una recopilación exhaustiva de datos estudiantiles de varios años, realizar un correcto análisis predictivo aplicando tres técnicas de minería de datos. Posteriormente, para verificar la eficacia de estas, y saber cuál es la mejor, se aplicará una evaluación por medio de métricas de regresión. Además, previamente se creará un data warehouse para integrar los datos y mejorar su análisis, lo que contribuirá de una mejor manera a la toma de decisiones de la Escuela de Educación Básica "Angélica Villón Lindao".

1.2. Descripción del proyecto

Ante la necesidad de mejorar el proceso de toma de decisiones, y que, a su vez, se permita establecer medidas que contribuyan a un rendimiento académico eficaz de los estudiantes, se propone aplicar minería de datos para predecir el rendimiento académico de los estudiantes mediante técnicas de aprendizaje supervisado.

Son cinco etapas las que conforman la propuesta presentada, estas se detallan a continuación:

- Primera etapa: Recolección de información.
- Segunda etapa: Creación de un data warehouse.
- Tercera etapa: Aplicación de minería de datos.
- Cuarta etapa: Evaluación de modelos.
- Quinta etapa: Difusión de conocimiento.

La primera etapa consiste en la extracción de datos a través de fuentes como hojas de cálculos en Excel, y también, de una base de datos estudiantil realizada en

Microsoft Access que posee información entre los años 2015 y 2019. Esta base de datos inicial contiene información elemental de los alumnos como: nombres, apellidos, dirección, entre otros. Por otro lado, en las hojas de cálculo se encuentran datos más específicos del estudiante, además, información con respecto a los padres. De esta manera, para realizar una correcta minería de datos, se emplearán más variables que permitan obtener información novedosa.

Esta etapa también se centra en la creación de una base de datos adicional cuya finalidad es identificar cuáles son las fuentes de los datos en un solo lugar, conteniendo, tanto los campos de la base de datos inicial, como los datos que se encuentran en las hojas de cálculo. Este proceso se lo realizará mediante herramientas de extracción, transformación y carga de datos. Así, al ya existir el nuevo origen de datos, se procederá a generar un dataset en la siguiente etapa que permita, mediante la aplicación de minería de datos, identificar patrones para predecir el rendimiento académico estudiantil.

La segunda etapa tiene como objetivo de crear un data warehouse, es decir, un almacén o repositorio de datos que, una vez integrados los datos, permita una flexibilidad al momento de tomar decisiones sobre temas relacionados al rendimiento académico estudiantil. De esta manera, se obtendrá un conjunto de datos objetivo para analizar. Esto se realizará por medio del proceso ETL (Extracción, transformación y carga) de datos, lo que posibilitará la creación del conjunto de datos objetivo para aplicar las respectivas técnicas de minería de datos.

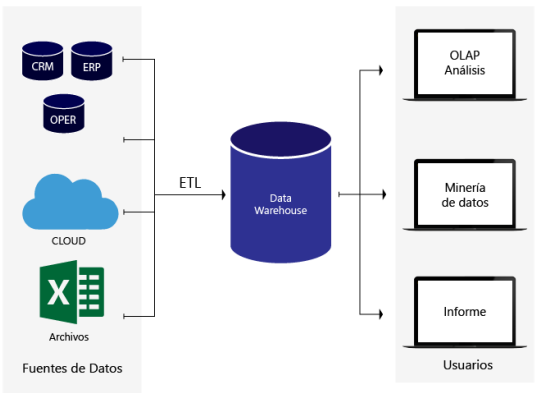


Figura 1. Estructura de un data warehouse.

Fuente: https://gravitar.biz/datawarehouse/data-warehouse-tipos/

La tercera etapa se trata de la aplicación de la minería de datos con la finalidad de realizar un análisis predictivo sobre el rendimiento académico de los estudiantes. Para esto, se realizará un análisis exploratorio de datos para determinar qué variables, del conjunto de datos obtenido, pueden estar más asociadas a los patrones de los datos.

Luego de obtener las variables correspondientes, se procederá a realizar una transformación de los datos para poder aplicar el respectivo proceso de minería. De esta manera, se emplearán las siguientes técnicas supervisadas de aprendizaje automático:

a) Árboles de decisión.
b) Redes neuronales.
c) Máquinas de Vector de Soporte (SVM, por sus siglas en inglés)

En la cuarta etapa se analizarán los resultados obtenidos para determinar cuál es el modelo que posee mejor funcionamiento. Los métodos de evaluación corresponderán a métricas de regresión como el error absoluto medio (MAE), error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), precisando así, qué algoritmo posee menor error; además, se evaluará el coeficiente de determinación para conocer el ajuste de la predicción. En el caso de que haya inconsistencias, será necesario retroceder y analizar nuevamente las etapas anteriores.

La quinta etapa se centra en cómo la información obtenida se transforma en conocimiento. De esta manera, con los resultados obtenidos, se transmitirá el conocimiento a los administradores de la escuela mediante el empleo de una capacitación, así, existirá conocimiento a disposición que servirá de base a los administradores de la escuela para tomar decisiones que posibiliten mejoras en el rendimiento académico de los estudiantes.

Esta propuesta tecnológica se implementará en la escuela "Lic. Angélica Villón Lindao" ubicada en el cantón de Santa Elena de la provincia del mismo nombre. Se consideran datos comprendidos entre los años 2015 y 2019. Las fuentes de datos a emplear consisten en información relacionada a estudiantes y docentes.

Para la realización de este trabajo se han considerado herramientas detalladas a continuación:

- ✓ **Pentaho Business Intelligence:** Para la realización de procesos de Extracción Transformación y Carga (ETL) de datos.

- ✓ **Draw.io:** Su finalidad es representar, mediante diagramas, las relaciones existentes en una base de datos.

- ✓ **Python:** El lenguaje base para la realización de los modelos de minería de datos.

- ✓ **Jupyter Notebook:** El entorno de ejecución de los modelos de minería de datos.

- ✓ **PostgreSQL:** El gestor de base de datos relacional.

- ✓ **Microsoft Access:** El gestor donde se encuentra la base de datos original de la escuela.

- ✓ **Keras:** La biblioteca de Python para redes neuronales.

- ✓ **Scikit-learn:** La biblioteca de Python para la ejecución de los modelos y sus métricas respectivas.

- ✓ **Matplotlib:** La biblioteca de Python para gráficos.

- ✓ **Seaborn:** Una biblioteca de Python usada también para realizar gráficos.

El análisis predictivo del desempeño académico en la escuela "Lic. Angélica Villón L." mediante la evaluación de técnicas de minería de datos contribuirá a la línea de investigación Inteligencia de Negocios (minería de datos) con la finalidad de dar soporte a las decisiones en tiempo real a las empresas [7].

1.3. Objetivos

1.3.1. General

Aplicar técnicas de minería de datos para la determinación de patrones en el rendimiento académico estudiantil mediante la exploración y validación de modelos predictivos.

1.3.2. Específicos

- Recopilar información para la identificación de orígenes de datos de calidad.
- Diseñar un data warehouse por medio de una plataforma que abarque procesos ETL (Extracción, Transformación y Carga) de datos.
- Aplicar técnicas de minería de datos de regresión mediante herramientas de aprendizaje automático.

1.4. Justificación

Es destacable, en el ámbito de los negocios, la importancia que tiene toda institución de contar con información estructurada, pues, esto facilita priorizar y establecer estrategias efectivas gracias a la excelente toma de decisiones. Ante tal situación, es en las organizaciones o empresas donde este factor desempeña un papel importante para tareas diarias. Así, “la toma de decisiones debe basarse en datos y

hechos. Los datos y la información deben ser precisos y confiables, y deben analizarse utilizando métodos válidos" [8].

Por otro lado, poseer una información estructurada resulta muy relevante en la sociedad de la información y, sobre todo, aprender a gestionarla representa "una herramienta clave para poder sobrevivir en un mercado cambiante, dinámico y global. Aprender a competir con esta información es fundamental para la toma de decisiones, el crecimiento y la gestión[…]" [9].

La Escuela de Educación Básica " Lic. Angélica Villón Lindao" tiene la necesidad de poseer información que facilite la toma de decisiones posteriores. Esto permitirá el análisis extenso del flujo de información, favoreciendo así, no solo la toma de decisiones, sino la comunicación y la administración de datos. Por esta razón, se propone realizar una predicción del rendimiento académico de los estudiantes mediante la evaluación de determinadas técnicas de minería de datos. Además, la elaboración de este análisis implica la creación de un conjunto de datos objetivo mediante la recopilación de datos de varias fuentes.

La importancia de que una institución educativa posea información ordenada es fundamental porque esto proporcionará, en el futuro, beneficios como un desarrollo estratégico al evaluar asuntos internos de una manera más detallada, en este caso, el rendimiento académico. Esto contribuirá en sí a un notable fortalecimiento institucional. Además, facilitará a la escuela: ahorro de tiempo y costes, así, los administradores podrán realizar un sinnúmero de consultas de datos por sí solos y mejoras en la toma de decisiones, ya que se proporcionan conocimientos óptimos que conllevan a análisis con precisión y fiabilidad para la creación de determinados informes; también, se proporciona una calidad y coherencia de datos. En sí, el beneficio de un almacén de datos es la mejora continua.

Los reportes generados, al tratar el rendimiento académico de los estudiantes, permitirán una alta verosimilitud en los análisis e interpretaciones que se realizarán. La capacidad de mirar rápidamente hacia atrás en las primeras tendencias y tener

los datos precisos, con el formato adecuado, es esencial para una buena toma de decisiones con respecto a datos relacionados al rendimiento académico de los estudiantes. Así, mediante la minería de datos, se logra extraer información útil que ayuda en las decisiones del negocio, en este caso, una institución educativa.

El tema propuesto está alineado a los objetivos del Plan Nacional de Desarrollo específicamente al siguiente eje:

Eje 2.- Economía al servicio de la sociedad.

Objetivo 5.- Impulsar la productividad y competitividad para el crecimiento económico sostenible de manera redistributiva y solidaria [10].

Política 5.6.- Promover la investigación, la formación, la capacitación, el desarrollo y la transferencia tecnológica, la innovación y el emprendimiento, la protección de la propiedad intelectual, para impulsar el cambio de la matriz productiva mediante la vinculación entre el sector público, productivo y las universidades [10].

1.5. Metodología

1.5.1. Metodología de la investigación

Debido a la falta de información sobre procesos que conllevan a un análisis exhaustivo de datos, como la minería de datos, se procede a utilizar la metodología de investigación de carácter exploratorio. Este tipo de investigación, se realiza cuando "el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes" [11]. Para esto, se investigarán y analizarán trabajos cuyas características se relacionen a esta línea de desarrollo, realizando una comparación donde se determinen semejanzas y diferencias, determinando la relevancia de esta propuesta de investigación con respecto a esos trabajos.

Para obtener información básica sobre la escuela, así como, indagar sobre la organización de la misma, se ha procedido a emplear una metodología de tipo diagnóstica [11] mediante técnicas de recolección de información que posibiliten la posterior toma de decisiones eficaz.

1.5.2. Grupo poblacional involucrado

La población que se ha escogida para aplicar esta propuesta tecnológica está formada por beneficiarios directos e indirectos. Los beneficiarios directos son los administrados de la unidad educativa, mientras que los indirectos están compuestos por estudiantes y docentes de la institución.

Según el Ministerio de Educación en el Archivo Maestro de Instituciones Educativas (AMIE) [12], el cual, recopila información de las instituciones del país en la apertura y clausuras de los años escolares, durante el inicio del periodo 2020-2021, se evidenciaron los siguientes datos:

Beneficiarios	**Cantidad**
Directos	
Administrativos	5
Indirectos	
Docentes	12
Estudiantes	160
Total	**177**

Tabla 1. Grupo poblacional involucrado.

1.5.3. Metodología de recolección de información

El desarrollo de este proyecto involucra el empleo de dos técnicas de recolección de información: la observación y la entrevista.

Observación: Esta técnica "sugiere y motiva los problemas y conduce a la necesidad de la sistematización de los datos" [13].

Entrevista: Esta técnica "es una pesquisa que consiste en el acopio de testimonios orales con la finalidad de obtener información sobre el objeto de estudio, interpretar y plantear soluciones" [14].

1.5.4. Análisis de las técnicas de recolección de información empleadas

La técnica de observación estructurada se limitó a los datos que se obtienen a partir de la información relacionada los estudiantes, como las notas de diferentes años lectivos. Esta información se recopiló mediante una base de datos estudiantil y registros en Excel; también, se reconoció cuál es el patrón estudiantil y el número de estudiantes que ha tenido cada año. Según el AMIE, en los periodos comprendidos entre 2015 y 2019 se registró:

Periodo académico	Cantidad de estudiantes
2015-2016	157
2016-2017	188
2017-2018	187
2018-2019	246
2019-2020	298

Tabla 2. Cantidad de estudiantes por periodo académico.

Con el paso de los años se puede apreciar un aumento notable de la cantidad de estudiantes. Además, se evidenció que no toda la información se encuentra almacenada en un mismo lugar, así, el uso de estadísticas para generar y analizar posteriormente este tipo de datos se ve limitada, de modo que los datos cuantitativos se consideran en general más fiables y objetivos.

Con respecto a la entrevista, esta se realizó de forma estructurada (ver Anexo 1) al arquitecto Mauricio Yunda Villao, uno de los principales administradores de la institución. Además, se formularon ciertas preguntas a la directora de la institución, MSc. Ibelice Tomalá V. relacionadas a su rol administrativo y proceso de toma de decisiones del rendimiento académico estudiantil. Así, se logró evidenciar cómo es el proceso administrativo dentro de la institución.

Entre los hallazgos obtenidos por medio de esta técnica de recolección de información están:

1. Cada dos años se elige una nueva directiva, por ende, esta situación genera la existencia de diferencias ideológicas en torno a los procesos administrativos.
2. Los cambios administrativos recurrente generan problemas debido al poco tiempo que se tiene para analizar los procesos de la escuela relacionados a la falta de información estructurada, sobre todo, datos relacionados al rendimiento académico.
3. No se ha establecido una evaluación acorde a los lineamientos del rendimiento académico, pues, la institución ha estado expandiéndose en los últimos años.
4. El nivel académico de la mayoría de los estudiantes es óptimo, sin embargo, al no haber una forma de interpretar este tipo de información pueden generarse inconvenientes en el futuro.
5. Aunque no exista una manera de evaluar e interpretar el rendimiento académico, el desempeño de los estudiantes se evalúa mediante diagnósticas, lecciones, entre otros.

1.5.5. Variable del proyecto

Lo que se buscar mejorar, en base a la propuesta sugerida, es disminuir el tiempo de evaluación del rendimiento académico estudiantil por parte del personal administrativo de la escuela "Lic. Angélica Villón Lindao". De esta manera, será el tiempo de obtención de los reportes para la toma de decisiones (reportes relacionados a la productividad estudiantil y rendimiento académico) la variable fundamental en esta investigación.

1.5.6. Metodología de desarrollo

Dado que este proyecto se refiere al análisis predictivo del rendimiento académico estudiantil, tiene como base la minería de datos, por lo tanto, para esta propuesta, la metodología que se empleará será Descubrimiento de Conocimiento en Bases de Datos (KDD) [15]. Esta es una de las principales metodologías para la minería de datos y se compone de cinco fases:

- Primera etapa: Recolección de información.
- Segunda etapa: Creación de un data warehouse.
- Tercera etapa: Aplicación de minería de datos.
- Cuarta etapa: Evaluación de los modelos.
- Quinta etapa: Difusión del conocimiento.

Primera etapa. – Se recopilará una cantidad exhaustiva de datos, como hojas de cálculos en Excel, con la finalidad de identificar las fuentes de orígenes de datos de calidad, sobre todo, los datos relacionados a los estudiantes de la institución.

Segunda etapa. – Posteriormente, se creará un data warehouse con la información recopilada de los estudiantes. Este almacén de datos contendrá sus respectivos hechos y dimensiones, mejorando la síntesis y posterior análisis de los datos. El enfoque a emplear será Kimball, generando dos datamarts, uno de estudiantes y otro de profesores. De esta manera, se genera el conjunto de datos objetivo con el que se procederá a aplicar la minería de datos.

Tercera etapa. – Se desarrollará el data wrangling de los datos que consiste en limpiar, transformar y enriquecer el dataset para los objetivos posteriores. Así, se crea la vista minable para eliminar o corregir datos incorrectos que afecten al análisis.

Además, se buscará hacer una predicción de patrones en base a los criterios e información obtenido mediante los pasos anteriores. Las técnicas que se emplearán serán: árboles de decisión, redes neuronales y máquina de vector soporte. Para los

árboles de decisión se aplicará el algoritmo de regresión, para las redes neuronales el perceptrón multicapa y, en cuanto a Máquinas de Vectores de Soporte, se empleará la variante Vectores de Soporte de Regresión (SVR).

Cuarta etapa. – Se analizarán los resultados obtenidos, para evitar así, inconsistencias en el patrón de los datos. Para poder evaluar el rendimiento del modelo y poder determinar cuál es el mejor, se emplearán métricas como el error cuadrático medio (MAE). De esta manera, el modelo no podrá estar sujeto a problemas como el sobreajuste, que carezcan de objetividad al analizar nuevos conjuntos de datos.

Quinta etapa. - Se difundirá el conocimiento a los administradores de la escuela por medio de una capacitación donde se detalle el proceso y los resultados obtenidos, para poder así, tomar las decisiones correctas que posibiliten mejoras en el rendimiento académico próximo de los estudiantes.

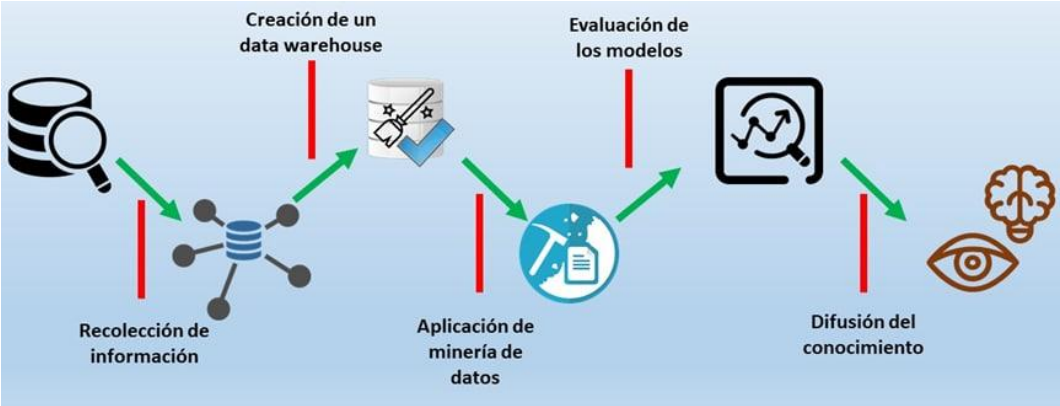


Figura 2. Metodología KDD.

CAPÍTULO II

2. PROPUESTA

2.1. Marco contextual

2.1.1. Institución “Lic. Angélica Villón Lindao”

La institución “Lic. Angélica Villón Lindao” se fundó el 12 de septiembre del año 2000 (Ver Anexo 1). Es una institución con un sostenimiento particular, cuyos niveles de educación son Inicial y Educación General Básica (EGB). Se encuentra ubicada en el cantón Santa Elena de la provincia del mismo nombre.

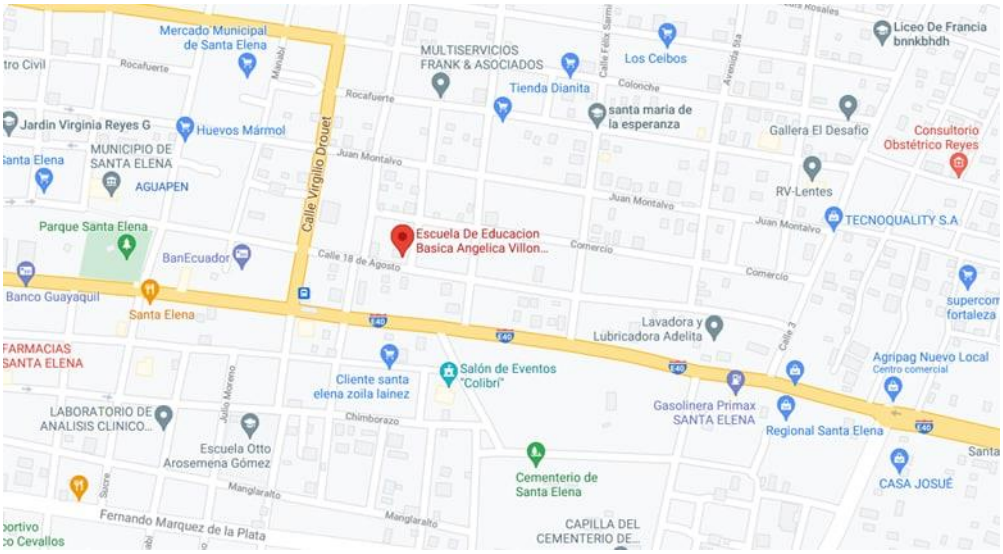


Figura 3. Ubicación de la institución.

Fuente: Google Maps.

2.1.2. Misión de la institución

La Escuela de Educación Básica "Lcda. Angélica Villón Lindao" ofrece un conjunto de talentos humanos técnicos y pedagógicos al servicio de la comunidad, desarrollando procesos educativos, para los estudiantes con o sin necesidades

educativas especiales atendiendo con absoluta responsabilidad y equidad, sin discriminación alguna, sirviendo a la comunidad con un enfoque en derechos (Ver Anexo 3).

2.1.3. Visión de la institución

La Escuela de Educación Básica "Lcda. Angélica Villón Lindao" anhela brindar una educación actualizada acorde a las exigencias que la sociedad actual demanda, con la atención y diversidad, con docentes capacitados comprometidos al cambio para enrumbar a la formación de talentos humanos que se inserten con facilidad de estudios y a la sociedad que adquieren un pensamiento crítico y reflexivo a través de la investigación científica con una práctica permanente de valores éticos y morales por el medio de la inclusión permitiendo a las personas con necesidades educativas especiales o sin discapacidad ser parte comunidad educativa (Ver Anexo 3).

2.1.4. Evaluación del rendimiento académico en la institución y marco legal

En la institución, para medir el rendimiento académico, se realizan actividades como cuestionarios para describir el nivel determinado de los estudiantes. Además, la modalidad de evaluación también se basa en términos de resultados de pruebas y exámenes y la capacidad del estudiante para aplicar lo aprendido y la velocidad a la que los estudiantes avanzan en cuanto a conocimiento (Ver Anexo 1). Esto se rige de acuerdo al artículo 68 de la Ley Orgánica de Educación Intercultural que establece que el desempeño del rendimiento académico es un componente de evaluación continua [16].

2.2. Marco conceptual

2.2.1. Bases de datos

Una base de datos relacional guarda datos en registros. Poseen un flujo de datos que se relacionan a diversos temas como: datos de poblaciones, países, enfermedades, entre otros. La tecnología de base de datos ayuda a resumir este volumen de datos en información necesaria para la toma de decisiones [17].

2.2.2. Pentaho Business Intelligence

Es un conjunto de herramientas que consta de ETL (extracción, transformación y carga de datos), capacidades de generación de informes y cuadros de mando [18]. Dentro de su estructura sobresalen las transformaciones que poseen múltiples formatos para las entradas y salidas en la integración de los datos [18].

2.2.3. Draw.io

Es un software de diagrama en línea gratuito para hacer diagramas de flujo, diagramas de procesos, organigramas, diagramas UML y diagramas de red [19].

2.2.4. Python

Es un lenguaje de scripting potente, interpretado, de código abierto, de uso general y gratuito para aplicaciones web. Es un lenguaje de programación fácil pero poderoso que proporciona estructura y soporte para aplicaciones grandes [20].

2.2.5. Jupyter Notebook

Es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto

narrativo. Los usos incluyen: limpieza y transformación de datos, aprendizaje automático, visualización de datos, entre otros [21].

2.2.6. PostgreSQL

Es un sistema de gestión de bases de datos relacionales de código abierto que permite escribir funciones y procedimientos almacenados en varios lenguajes de programación, y la arquitectura permite flexibilidad de admitir más lenguajes. Tiene características de clase empresarial como funciones de ventanas SQL [22].

2.2.7. Microsoft Access

Posee sólidas funcionalidades de análisis de datos que son fáciles de aprender y ciertamente aplicables a muchos tipos de organizaciones y sistemas de datos, además, puede ayudar a optimizar sus procesos analíticos aumentando su productividad [23].

2.2.8. Keras

Es una biblioteca frontend de código abierto para redes neuronales. Funciona como columna vertebral de la red neuronal, ya que tiene muy buenas capacidades para formar funciones de activación. Keras puede ejecutar diferentes marcos de aprendizaje profundo como backend [24].

2.2.9. Scikit-learn

Es una biblioteca de código abierto de populares algoritmos de aprendizaje automático que permite construir este tipo de sistemas [25]. Además, se compone de herramientas sencillas y eficientes para el análisis predictivo de datos, accesibles para todos y reutilizables en diversos contextos [26].

2.2.10. Matplotlib

Es un paquete de Python para trazado 2D que genera gráficos con calidad de producción. Admite el trazado interactivo y no interactivo, y puede guardar imágenes en varios formatos de salida (PNG, PS y otros). Puede utilizar varios conjuntos de herramientas de ventana (GTK +, wxWidgets, Qt, etc.) y proporciona una amplia variedad de tipos de gráficos (líneas, barras, gráficos circulares, histogramas y muchos más) [27].

2.2.11. Seaborn

Es una librería construida sobre matplotlib y e integrada con pandas que permite realizar visualizaciones con un enfoque técnico y estético. Entre los gráficos que permite hacer destacan diagramas de caja, los cuales, sirven para determinar y analizar la presencia de valores atípicos dentro del conjunto de datos [28].

2.2.12. Inteligencia de Negocios

Los datos se producen tan rápido y en volúmenes extensos que es imposible analizarlos y usarlos de manera efectiva cuando se utilizan métodos manuales tradicionales como hojas de cálculo. Bajo este concepto, surge la Inteligencia de Negocios, que, reúne datos en forma utilizable para su análisis pertinente. La Inteligencia de Negocios apoya la toma de decisiones basada en hechos utilizando datos históricos en lugar de suposiciones carentes de objetividad [29].

2.2.13. Árboles de decisión

Son modelos caracterizados por la precisión mostrando un rendimiento eficaz en la predicción de patrones en datos complejos [30]. Existen árboles de clasificación,

usados para realizar predicciones de variables categóricas, y árboles de regresión, donde se obtienen predicciones con valores continuos.

2.2.14. Redes neuronales

Es un algoritmo cuya finalidad es reconocer la existencia de relaciones en un determinado conjunto de datos por medio de procesos que imitan el funcionamiento del cerebro humano [31]. La ventaja de la red neuronal es que tiene el potencial de detectar todas las interacciones posibles entre las variables predictoras. La red neuronal también podría hacer una detección completa sin tener ninguna duda incluso en relaciones complejas no lineales entre variables dependientes e independientes.

2.2.15. Máquina de vectores de soporte

Es un algoritmo que se lleva a cabo mediante la búsqueda de un hiperplano que separa entre un conjunto de objetos que tienen diferentes clases. Este hiperplano se elige maximizando el margen entre las dos clases para reducir el ruido y aumentar la precisión de los resultados [32]. Cuando la finalidad no es aplicar una clasificación, sino, una regresión, surge una variante al SVM, siendo esta la Regresión de Soporte Vectorial (SVR).

2.2.16. Métricas de rendimiento

Dentro el aprendizaje automático, en el caso de predicciones destacan dos tipos: la clasificación y regresión. Para cada uno de estos tipos resaltan métricas que permitirán evaluar al modelo realizado. En el caso de la clasificación, las métricas que sobresalen están enfocadas en el conteo de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos; entre las cuales destacan: precisión, exactitud, exhaustividad, puntuación f1 y la matriz de confusión. Con respecto a la

regresión, las métricas que miden el rendimiento del algoritmo, son otras, entre las que destacan: error absoluto medio, error cuadrático medio, raíz del error cuadrático medio y el coeficiente de determinación [26].

2.3. Marco teórico

2.3.1. Importancia de la inteligencia de negocios en la actualidad

En la actualidad se han suscitado enormes cambios enfocados en la gestión de los negocios y en los métodos para fundamentar la toma de decisiones [33]. Así, la aplicación de la Inteligencia de Negocios trae consigo un sinnúmero de beneficios dentro de las instituciones entre los que destacan:

- Permite una visión del pasado, el presente, y el futuro al que puede aspirar una empresa.
- BI se acompaña estrictamente del monitoreo con reglas del negocio o métricas que permiten mantener el control de las metas fundamentales de la empresa.
- Aporta información actualizada.

Bajo este contexto, las estrategias en BI se pueden interpretar como la coordinación de forma efectiva de las tecnologías para el análisis adecuado de los datos, cuyo fin es alinearse a las metas y objetivos de una organización.

2.3.2. Estado del arte de la minería de datos

El desarrollo de la tecnología de la información ha generado una gran cantidad de bases de datos y enormes datos en diversas áreas [34]. La investigación en bases de datos y tecnología de la información ha dado lugar a un enfoque para almacenar y manipular estos valiosos datos para una mayor toma de decisiones. Este enfoque es

la minería de datos que se centra en acciones como el descubrimiento de conocimientos por medio de la extracción y el análisis de datos.

En cuanto a metodologías para la aplicación de minería de datos se distinguen varias. Según Gironés en el libro “Minería de datos: Modelos y algoritmos”, se destaca la metodología CRISP-DM compuesta por las fases: comprensión de negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, y finalmente, despliegue [35].

Por otro lado, Hernández, en el libro “Introducción a la Minería de Datos”, resalta la metodología KDD (Descubrimiento de Conocimiento en Bases de Datos), la cual, se componen de cinco fases: integración y recopilación, selección, limpieza y transformación, minería de datos, evaluación e interpretación, y finalmente, difusión y uso [15].

Este proyecto se basa en la metodología KDD, componiéndose de cinco etapas:

1. Recolección de información.
2. Creación de un data warehouse.
3. Aplicación de minería de datos.
4. Evaluación de modelos.
5. Difusión de conocimiento.

Dentro de la amplia gama de algoritmos de minería de datos, resaltan algoritmos que se adaptan a las condiciones del entorno, y de esto depende su clasificación. Por ejemplo, si se desea examinar imágenes de video para ver si alguien en el estacionamiento de una determinada empresa está actuando de una manera inusual se necesitarían de algoritmos para detectar anomalías, por ende, los algoritmos a emplear corresponderían a aprendizaje no supervisado. No obstante, si lo que se quiere es examinar la imagen de una persona que ingresa a su tienda minorista para corroborar si hombre o mujer, se emplea aprendizaje supervisado.

2.4. Componentes de la propuesta

2.4.1. Primera etapa: Recolección de información

Recopilación de información

Para el proyecto de investigación se emplearon dos fuentes de origen, siendo estas: una base de datos en Microsoft Access que contiene la información básica de los estudiantes, así como, los representantes. Este base de datos almacena información correspondiente entre los años 2015 y 2019. El esquema original de la base de datos que posee la institución está compuesto por cuatro tablas, las cuales son: Representante, Estudiante, Estudiante_Dirección y Dirección.

En este, se pudo apreciar que la cardinalidad de las entidades (tablas) era de uno a muchos entre: Representante y Estudiante, Estudiante y Estudiante_Dirección; y, Dirección y Estudiante_Dirección; respectivamente.

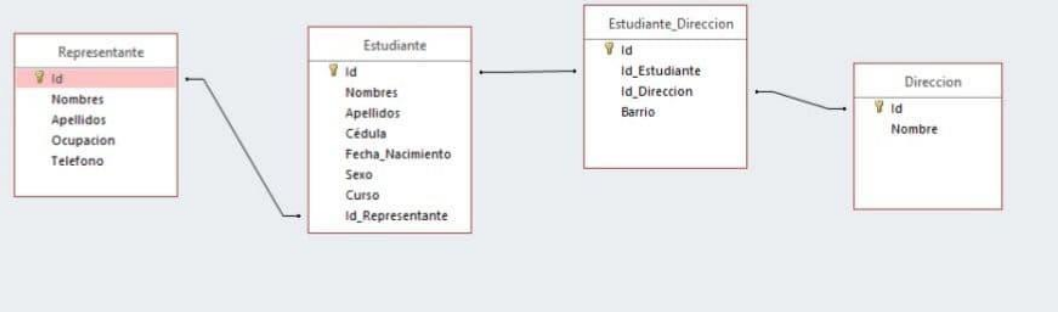


Figura 4. Base de datos original de la institución.

La información contenida en los archivos Excel está relacionada con los docentes de la institución. También, hay hojas de cálculo que contienen datos específicos de los estudiantes, los cuales, han sido realizados por la misma institución con el paso de los años, mediante fichas estudiantes que contienen un formato específico (Ver Anexo 4).

Esquema de la base de datos

A partir de estos datos, la nueva base de datos que se propone contiene información más específica de los estudiantes, además, de datos relacionados con los representantes de los estudiantes, e incluso, los docentes. El esquema se muestra a continuación:

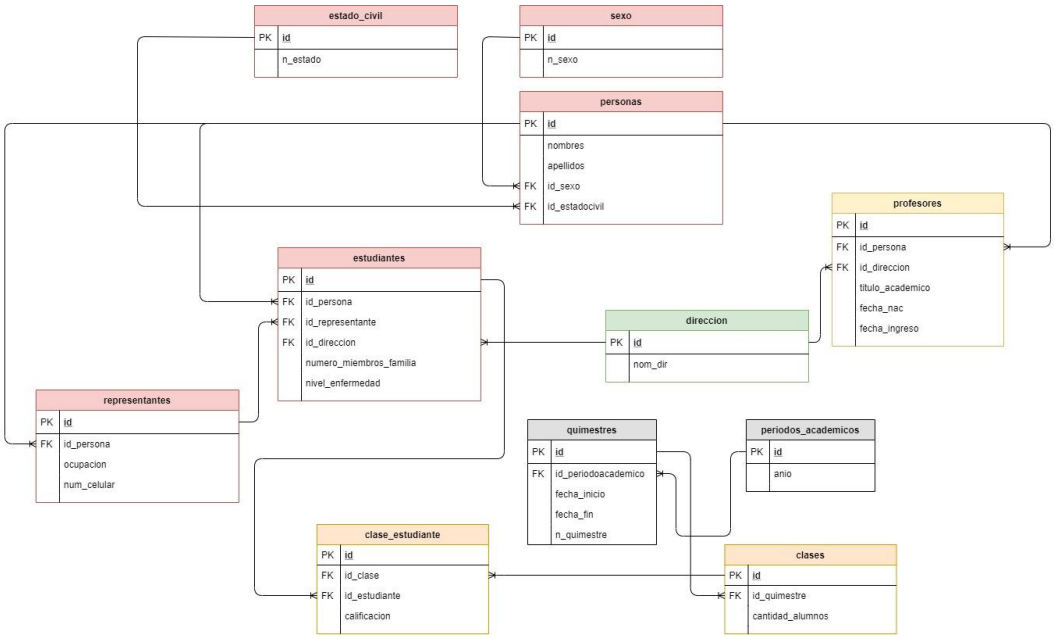


Figura 5. Base de datos propuesta para la institución.

Este nuevo modelo se encuentra formado por once tablas:

- clase_estudiante
- clases
- direccion
- estado_civil
- estudiantes
- periodos_academicos
- personas
- profesores
- quimestres
- representantes

- sexo

A continuación, se describen las columnas que conforman las principales tablas creadas:

Columna	Tipo	Descripción
id	int	Identificador de persona.
nombres	varchar (60)	Nombres de persona.
apellidos	varchar (60)	Apellidos de persona.
id_sexo	int	Identificador del sexo de la persona.
id_estadocivil	int	Identificador del estado civil de la persona.

Tabla 3. Tabla "persona".

Columna	Tipo	Descripción
id	int	Identificador de estudiante.
id_persona	int	Identificador de la persona estudiante.
id_representante	int	Identificador del representante del estudiante.
id_direccion	int	Identificador de la dirección del estudiante.
numero_miembros_familia	int	Número de miembros de la familia.
nivel_enfermedad	int	Nivel de enfermedad.

Tabla 4. Tabla "estudiante".

Columna	Tipo	Descripción
id	int	Identificador del representante.
id_persona	int	Identificador de la persona representante.
ocupacion	varchar (50)	Ocupación del representante.
num_celular	varchar (10)	Número celular del representante.

Tabla 5. Tabla "representantes".

Columna	Tipo	Descripción
id	int	Identificador del profesor.
id_persona	int	Identificador de la persona profesor.
id_direccion	int	Identificador de la dirección del profesor.
titulo_academico	varchar (50)	Título académico del profesor.
fecha_nac	date	Fecha de nacimiento del profesor.
fecha_ingreso	date	Fecha de ingreso a la institución.

Tabla 6. Tabla "profesores".

Columna	Tipo	Descripción
id	int	Identificador de la clase estudiante.
id_clase	int	Identificador de la clase.
id_estudiante	int	Identificador del estudiante.
calificacion	int	Calificación del estudiante en la clase.

Tabla 7. Tabla "clase_estudiante".

Columna	Tipo	Descripción
id	int	Identificador del quimestre.
id_periodoacademico	int	Identificador del periodo académico.
fecha_inicio	date	Fecha de inicio del quimestre.
fecha_fin	date	Fecha de finalización del quimestre.
n_quimestre	int	Número de quimestre.

Tabla 8. Tabla "quimestres".

Creación de base de datos

Luego de haber realizado el esquema, se procedió a crear la base de datos mediante el motor de base de datos PostgreSQL con el nombre 'basedatosEEB'.

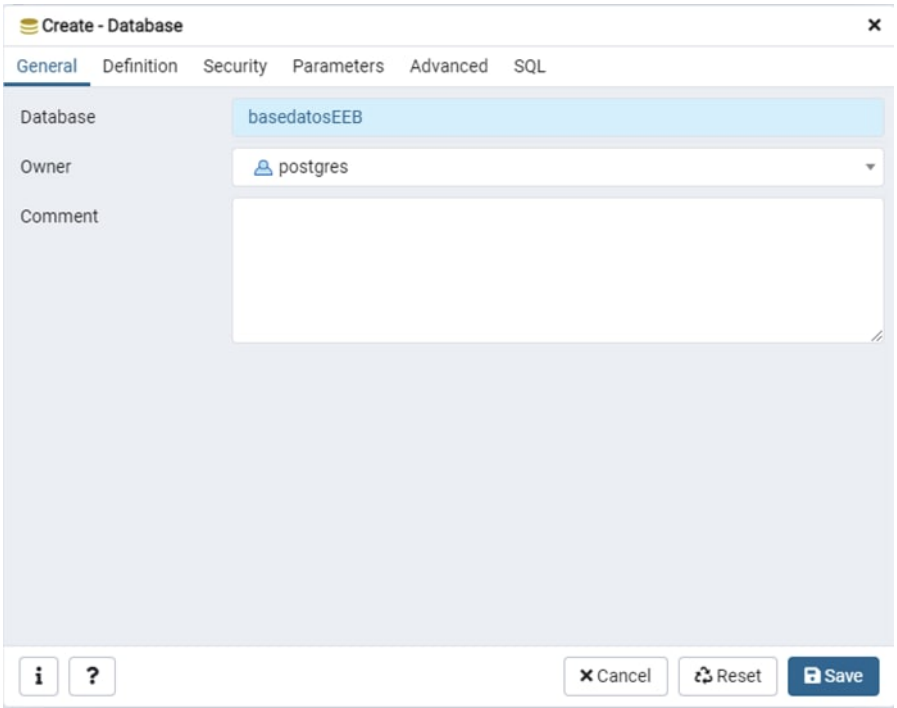


Figura 6. Creación de la base de datos 'basedatosEEB'.

En la sección de esquemas, se pudieron crear las tablas que componen la base de datos.

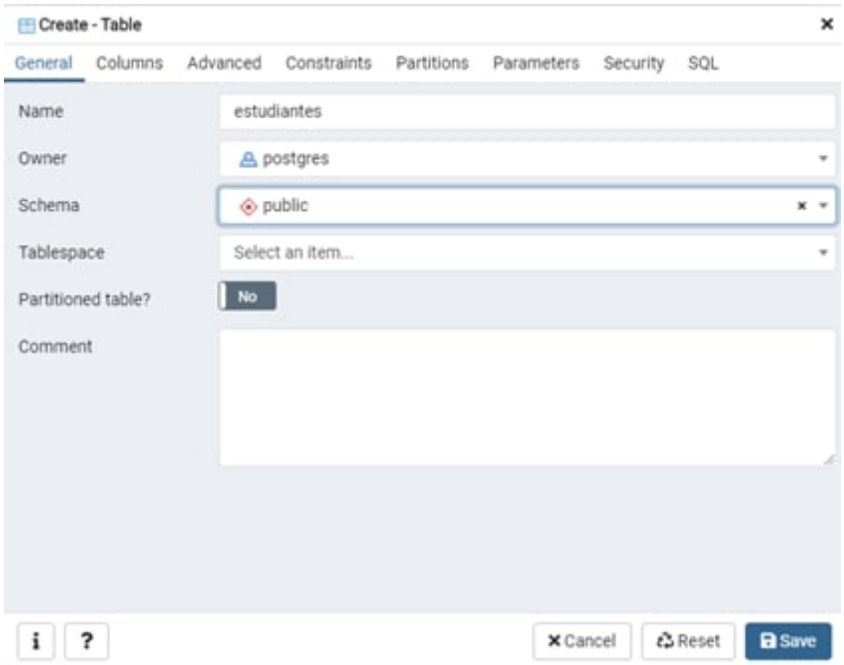


Figura 7. Creación de la tabla 'estudiantes'.

Integración de datos mediante Pentaho Data Integration

Una vez creada la base de datos, se procedió a almacenar los registros dentro de las tablas correspondientes. Para este paso, se usó la herramienta Pentaho Data Integration, la cual, permite realizar procesos de Extracción, Transformación y Carga de datos.

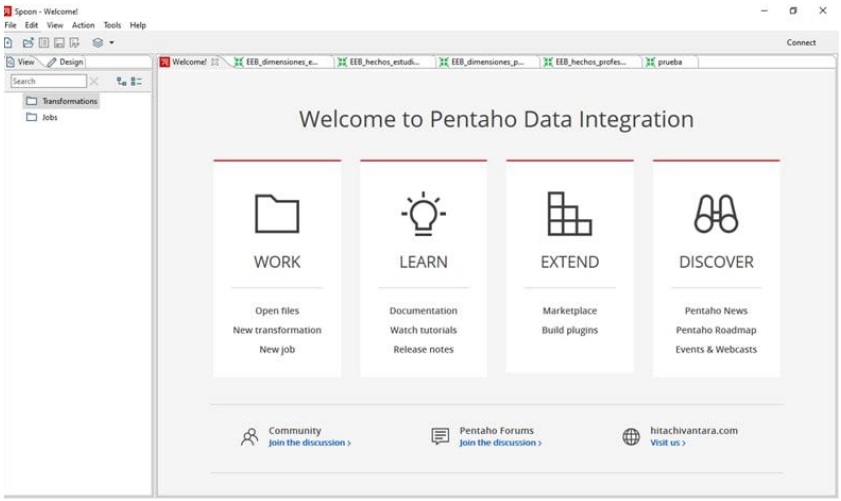


Figura 8. Interfaz de Pentaho Data Integration.

Esta herramienta está caracterizada por tener una modalidad “drag & drop” para la elaboración de sus procesos, entre los cuales destacan: transformaciones, trabajos, pasos y saltos. En cuanto a la base de datos, se creó una nueva transformación:

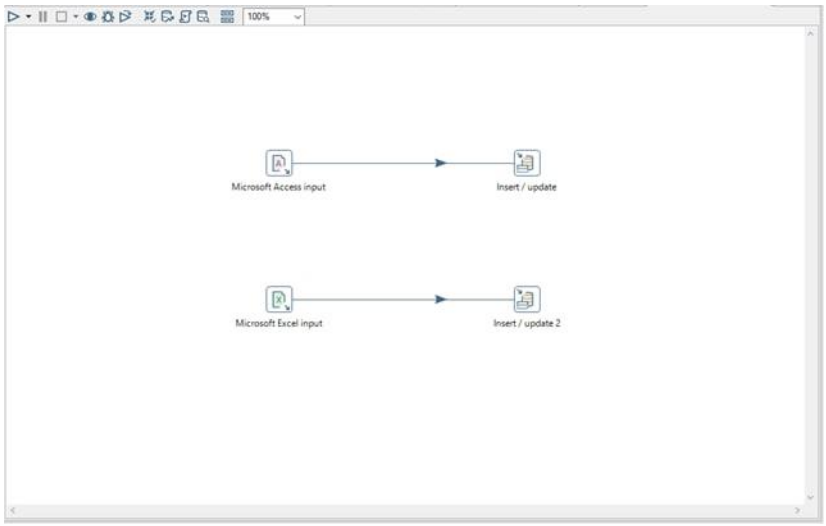


Figura 9. Proceso ETL para los registros de la base de datos 'basedatosEEB'.

En la parte izquierda se encuentran las entradas ‘Microsoft Access input’ y ‘Microsoft Excel input’. La primera permitirá ingresar los datos contenidos en la base de datos original, mientras que la segunda, toda la información que se encuentra alojada en las hojas de cálculo. Las salidas ‘Insert / update’, 1 y 2, se encargan de establecer la conexión pertinente a la base de datos localizada en el administrador ‘pgAdmin’ de PostgreSQL.

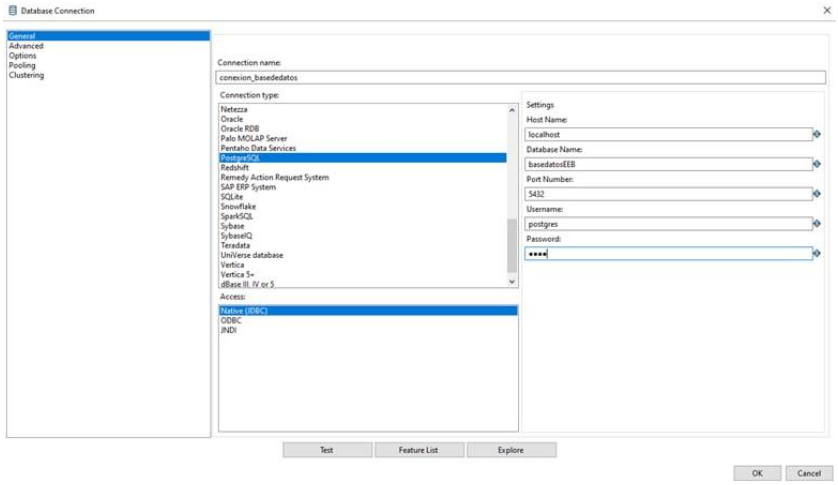


Figura 10. Conexión a base de datos.

En cuanto a los datos de entrada, se procedió a recolectar la información pertinente.

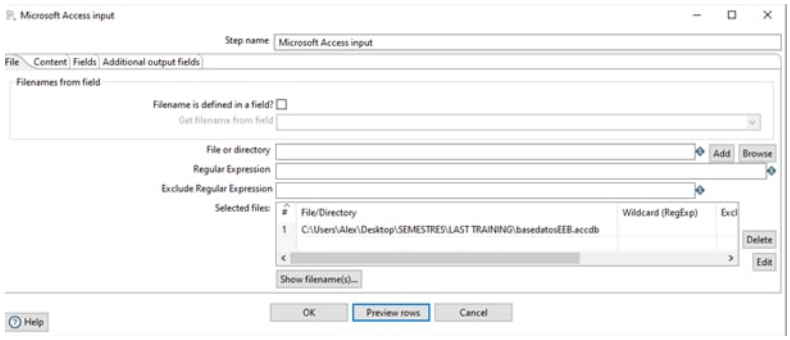


Figura 11. Proceso de entrada base de datos en Microsoft Access.

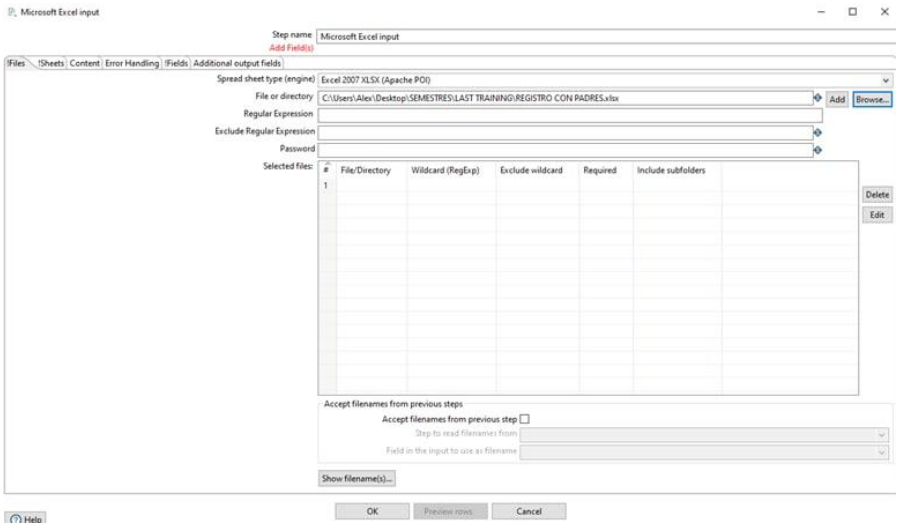


Figura 12. Proceso de entrada de datos de fichas estudiantiles.

Para los procesos de salida, se estableció una conexión con la base de datos creada anteriormente.

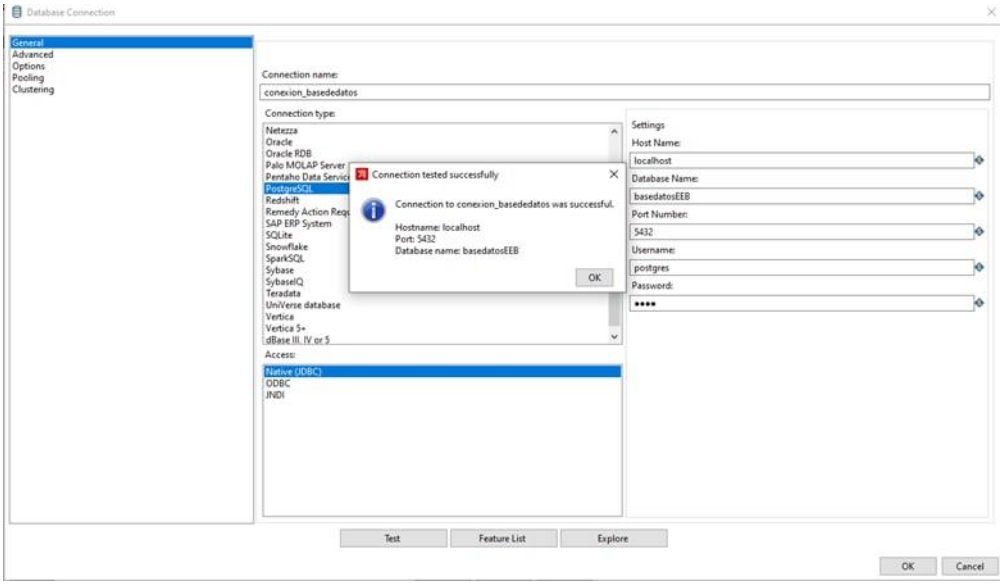


Figura 13. Conexión a base de datos 'basedatosEEB'.

De esta manera, el proceso ETL pudo realizarse de una forma exitosa y los datos se alojaron en la base de datos ‘basedatosEEB’.

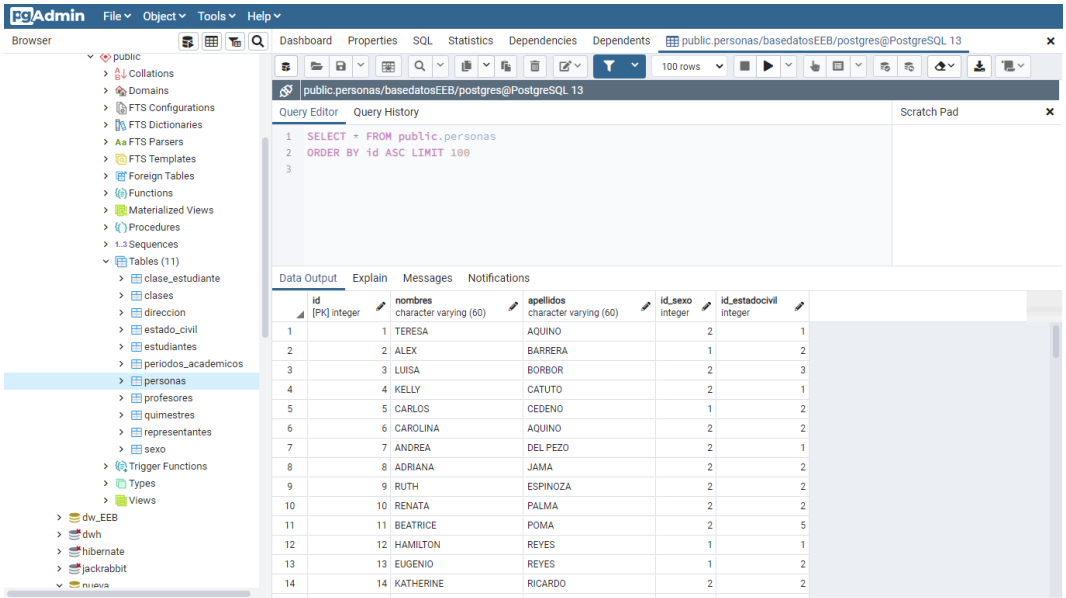


Figura 14. Base de datos 'basedatosEEB'.

2.4.2. Segunda etapa: Creación de un data warehouse

Esquema del data warehouse

Ya creada la base de datos en PostgreSQL, se procede a crear el almacén de datos, el cual, será la base para obtener el conjunto de datos objetivo para la realización de la minería de datos.

Existen dos enfoques para crear un almacén de datos o "data warehouse". El primero, enfoque Inmon, se caracteriza por la creación de un almacén de datos general para el posterior establecimiento de datamarts que centralizan la información a un departamento en específico, y, el enfoque Kimball que describe un proceso contrario, partiendo de la creación de datamarts para luego generar un almacén de datos [36].

Para este proyecto, se escogió el segundo enfoque, Kimball, creándose dos datamarts, uno relacionado a los estudiantes y otro a los profesores.

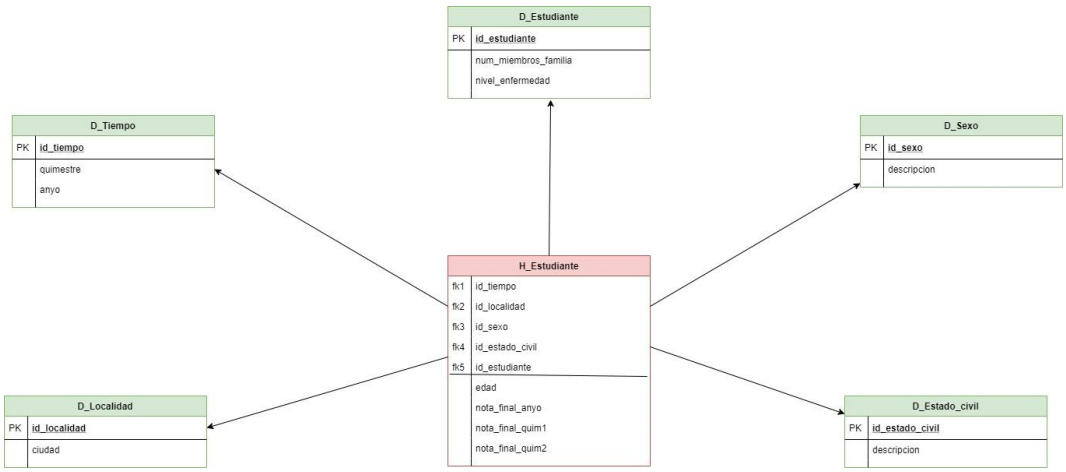


Figura 15. Datamart Estudiante.

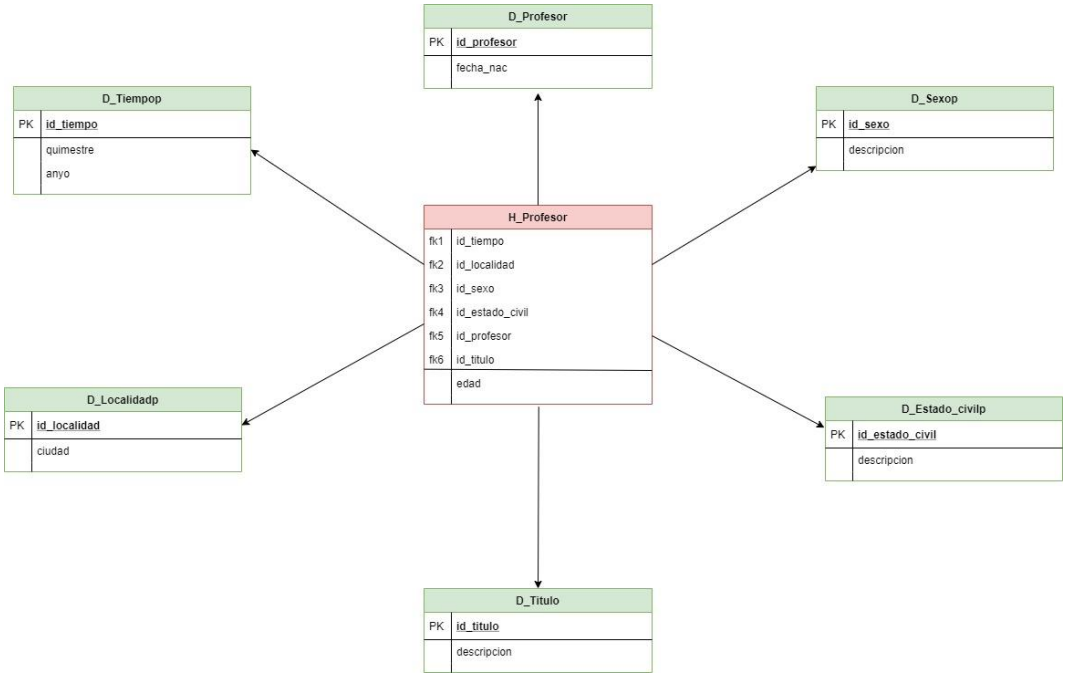


Figura 16. Datamart Profesor.

Creación del data warehouse

Para llenar el data warehouse primero se crea la base de datos donde se alojará el data warehouse.

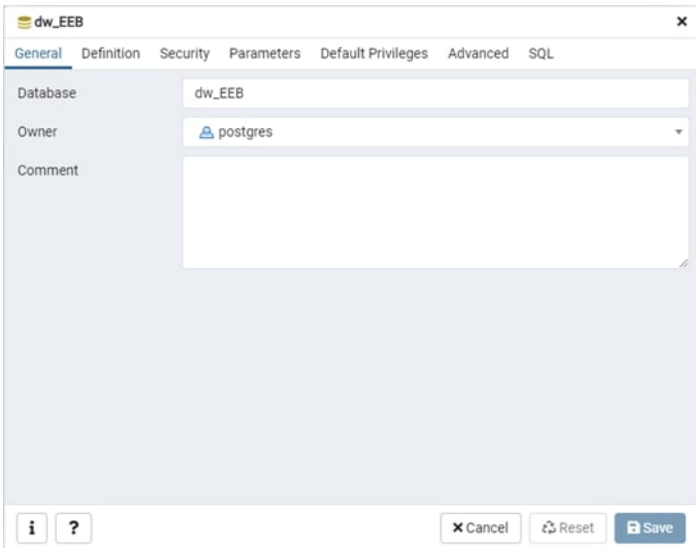


Figura 17. Creación de data warehouse.

Procesos ETL en Pentaho Data Integration

Los procesos de Extracción, Transformación y Carga de datos se realizaron nuevamente en Pentaho Data Integration (Kettle), albergando cuatro transformaciones:

- **EEB_dimensiones_estudiante:** Transformación correspondiente a las dimensiones del datamart Estudiante.
- **EEB_dimensiones_profesor:** Transformación correspondiente a las dimensiones del datamart Profesor.
- **EEB_hechos_estudiante:** Transformación correspondiente a la tabla de hechos del datamart Estudiante.
- **EEB_hechos_profesor:** Transformación correspondiente a la tabla de hechos del datamart Profesor.

Dimensiones de Estudiante

A continuación, se pueden ver los procesos ETL relacionados a las dimensiones estudiantiles, donde los orígenes de datos pertenecen a la base de datos rediseñada y a archivos Excel.

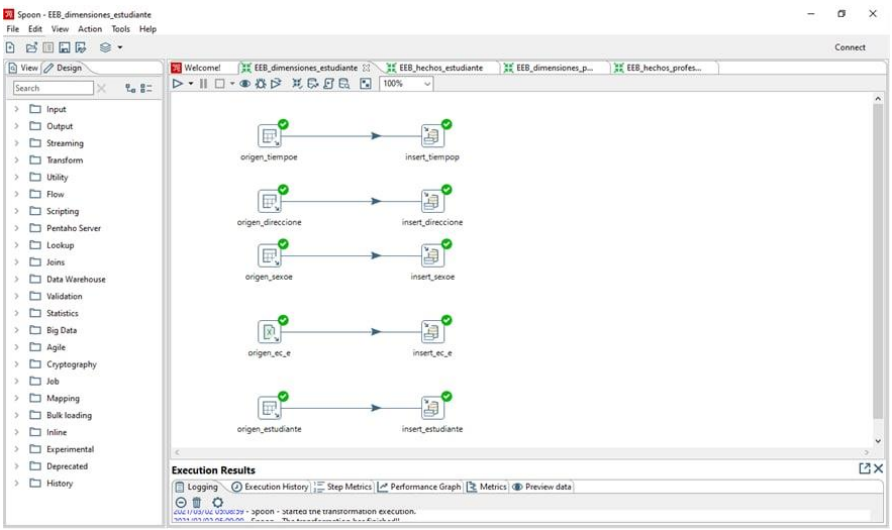


Figura 18. Transformación 'EEB_dimensiones_estudiante'.

Existen cinco entradas, a partir de las cuales se llena el data warehouse, por ejemplo, se muestra la entrada “origen_tiempoe” la cual extrae los datos directamente de la base de datos rediseñada.

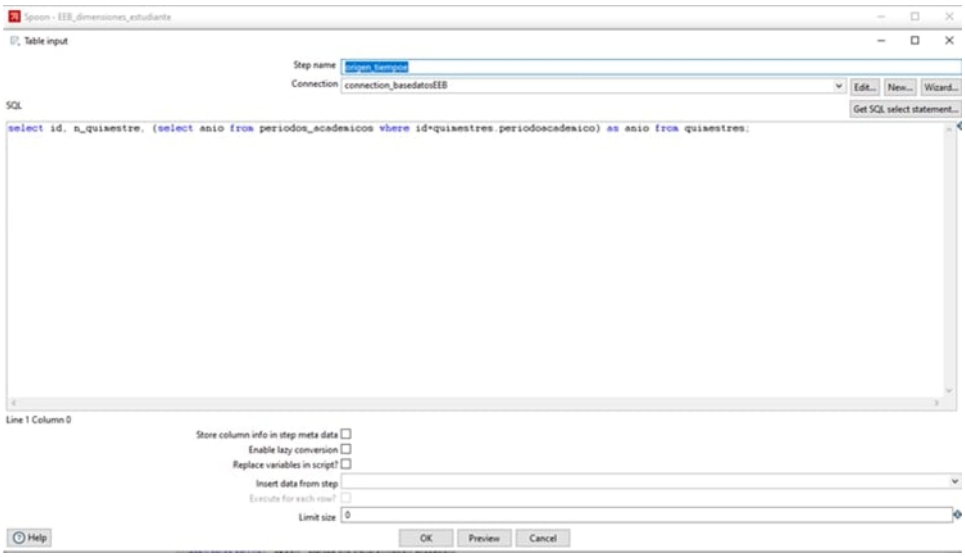


Figura 19. Entrada 'origen_tiempoe'.

Dimensiones de Profesor

A continuación, se muestran los procesos ETL correspondientes a las dimensiones del profesor.

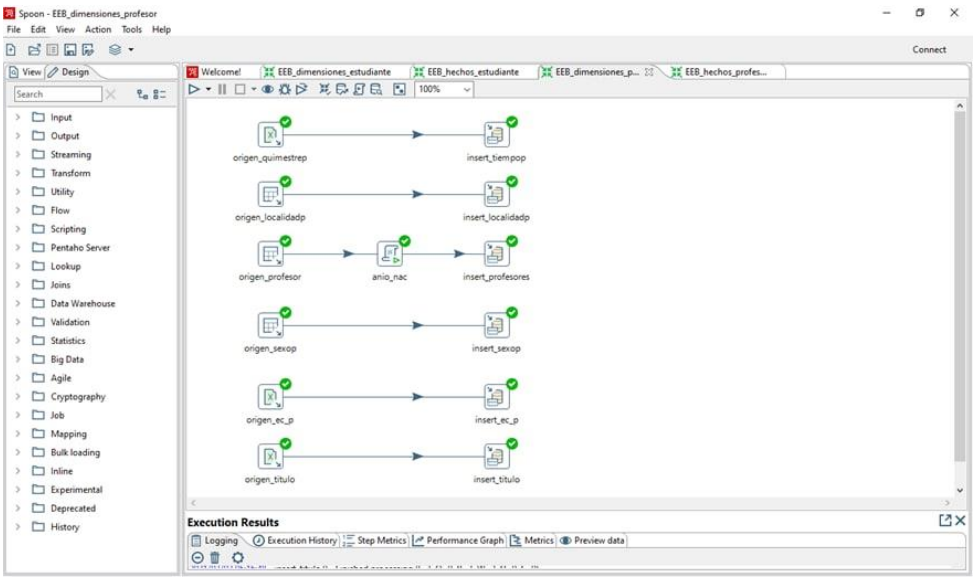


Figura 20. Transformación 'EEB_dimensiones_profesor'.

Tabla de hecho de estudiantes

Luego de llenar las tablas correspondientes a las dimensiones del Estudiante, se procede a realizar el llenado de la tabla de hechos de los estudiantes.

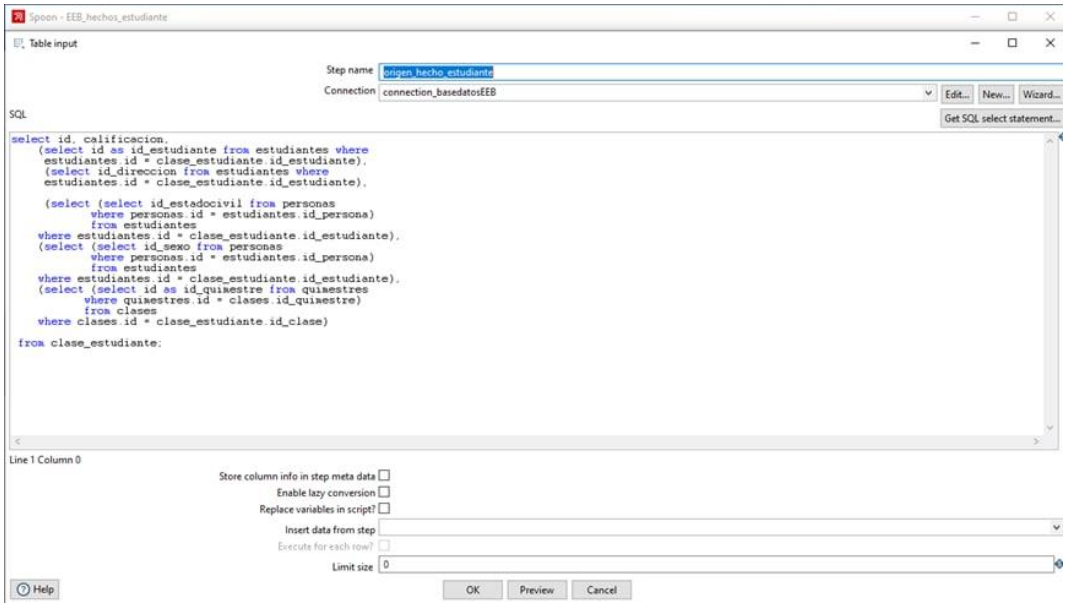


Figura 21. Entrada 'origen_hecho_estudiante'.

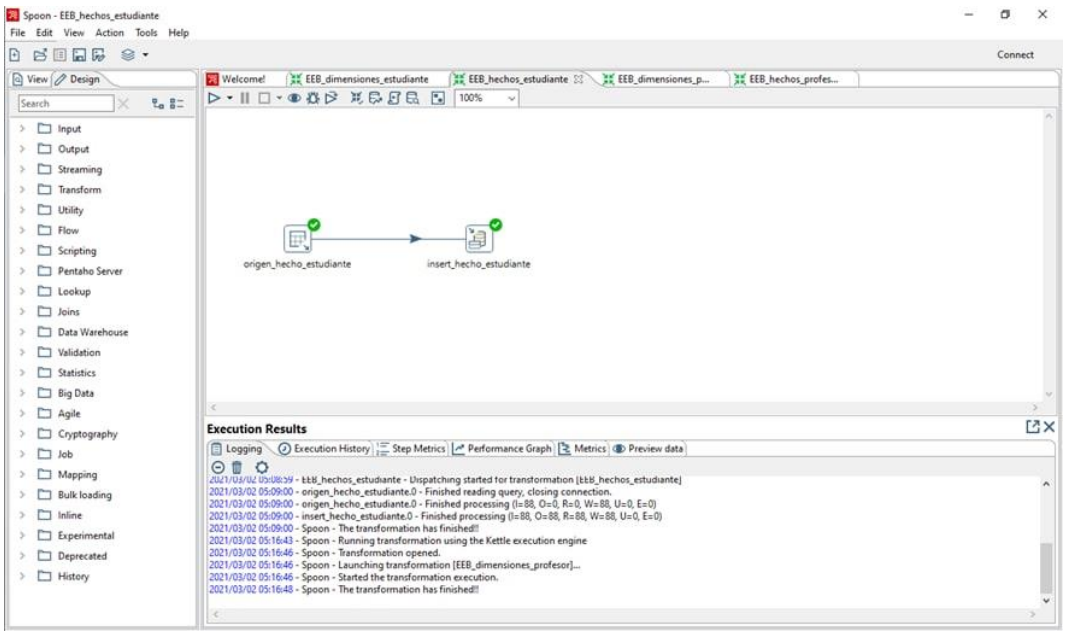


Figura 22. Transformación 'EEB_hechos_estudiante'.

Tabla de hecho de profesores

Posteriormente, se realizó el proceso correspondiente a la tabla de hechos de los profesores.

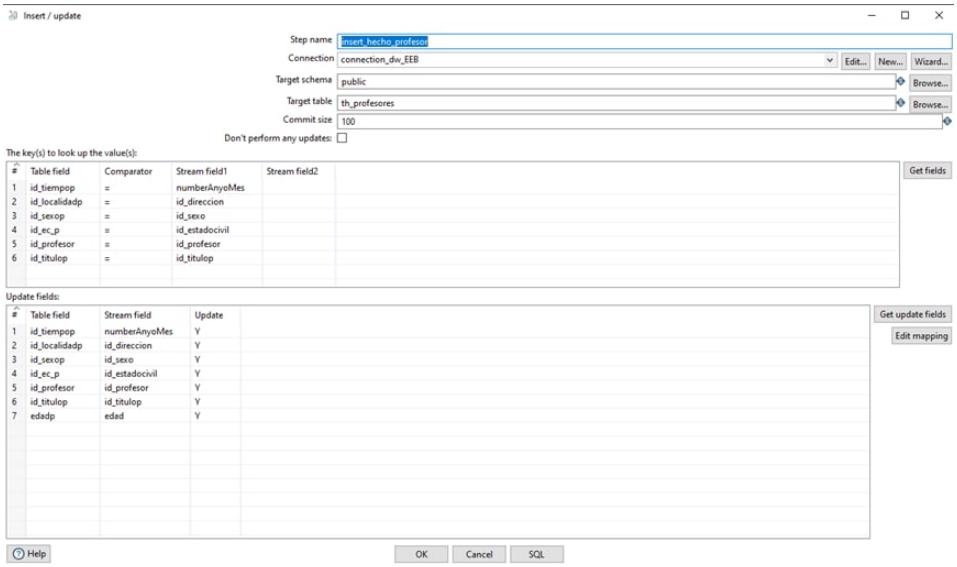


Figura 23. Entrada 'origen_hechos_profesor'.

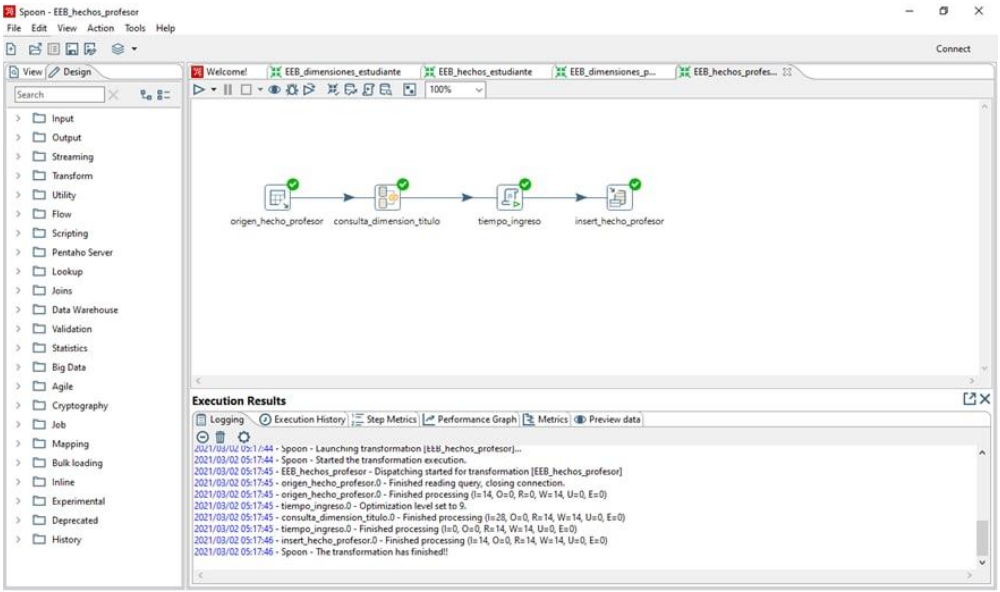


Figura 24. Transformación 'EEB_hechos_profesor'.

De esta manera, se pueden apreciar cómo los datos ya están almacenados en el almacén de datos o data warehouse.

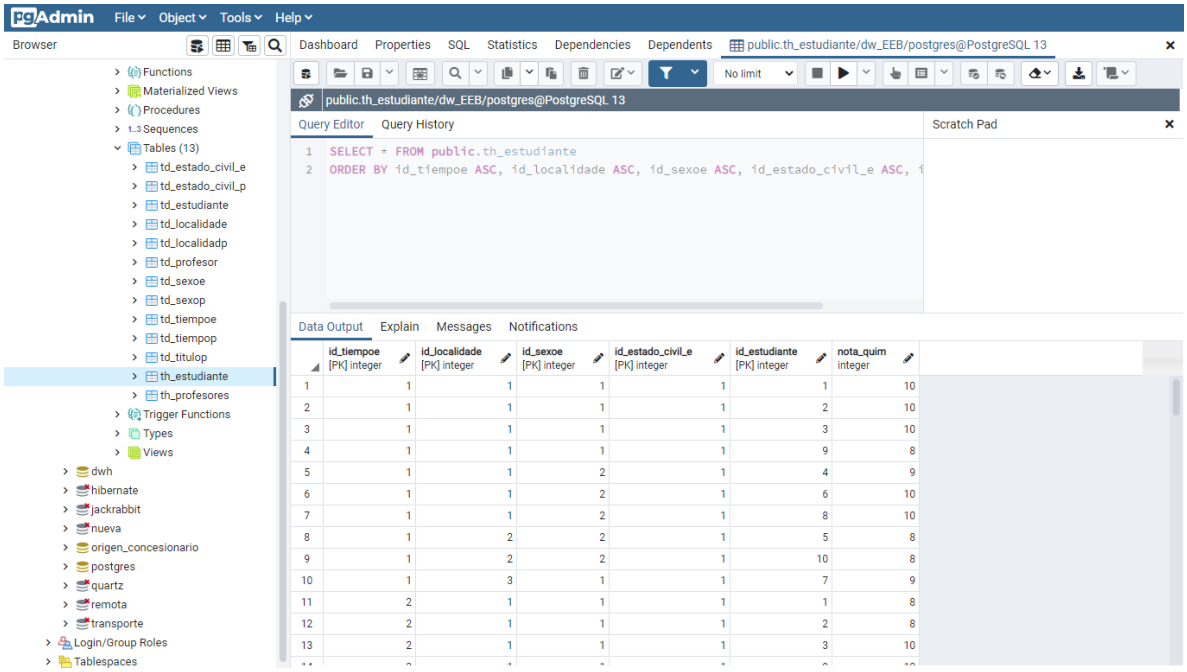


Figura 25. Data warehouse.

Luego de la integración respectiva de los datos, y los procedimientos realizados en la base de datos y en el data warehouse, se procedió a crear el conjunto de datos, o dataset, objetivo, generando un archivo .csv (comma-separated values).

2.4.3. Tercera etapa: Aplicación de minería de datos

Esta etapa se compone de tres subetapas, las cuales son:

- Análisis exploratorio de los datos.
- Transformación de datos.
- Minería de datos

Análisis exploratorio de los datos

El propósito de esta subetapa es, a partir del conjunto de datos generado en la etapa anterior, determinar qué variables serán apropiadas para el proceso de minería de datos, es decir, se busca establecer correlaciones entre las variables independientes y la variable objetivo a predecir.

Las librerías principales que se necesitaron para realizar los procedimientos respectivos son las siguientes:

- Pandas
- Numpy
- Matplotlib
- Seaborn

Una vez importadas las librerías, se procedió a leer el archivo .csv, obtenido en la fase previa, por medio de pandas, empleando el método “read.csv” junto con el parámetro “sep” cuyo valor es “;”. La finalidad es realizar una separación de las variables alojadas en el archivo para evitar errores en el procesamiento de las mismas. Así, se evidencia que la cantidad total de variables a analizar es de 19, de las cuales, 7 corresponden a variables categóricas y 12 a variables numéricas.

La finalidad de realizar el análisis exploratorio de datos es determinar qué tanto están relacionadas las variables independientes con respecto a la variable dependiente. Con respecto a las variables numéricas, se procede a aplicar el coeficiente de correlación de Pearson, y, mediante un gráfico de calor mediante la librería seaborn, se logra apreciar la correlación existente.

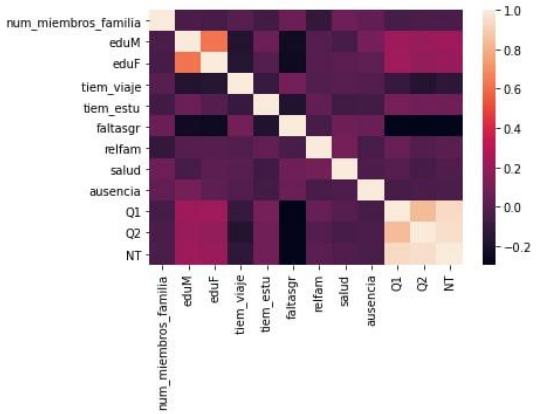


Figura 26. Correlación entre variables numéricas.

Los resultados indican que las variables que más se asocian linealmente a la variable dependiente NT son Q1 y Q2 respectivamente, no obstante, esta asociación es lineal, por ende, se toman otras variables con correlaciones bajas que igual pueden influir en el patrón de NT.

Luego, se procede a analizar las variables categóricas con el propósito de determinar qué variables presentan datos atípicos. Para esto, se procede a realizar diagramas de caja para poder visualizar la ubicación de los cuartiles en relación a la variable NT.

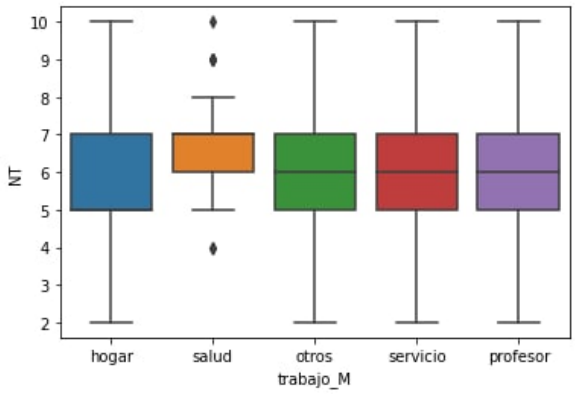


Figura 27. Diagrama de caja de la variable trabajo_M.

Este análisis determina que las variables que presentan valores atípicos están relacionadas al trabajo del representante del estudiante, sea hombre o mujer.

Una vez analizadas las variables, tanto categóricas como numéricas, el nuevo conjunto de datos que se empleará para la siguiente subetapa está compuesto por: sexo, localidad, representante, enfermedades, eduM, eduF, tiem_viaje, faltasgr, Q1, Q2, NT.

También, se procede a revisar cómo está distribuida la variable NT para las variables seleccionadas durante esta subetapa mediante un gráfico de densidad.

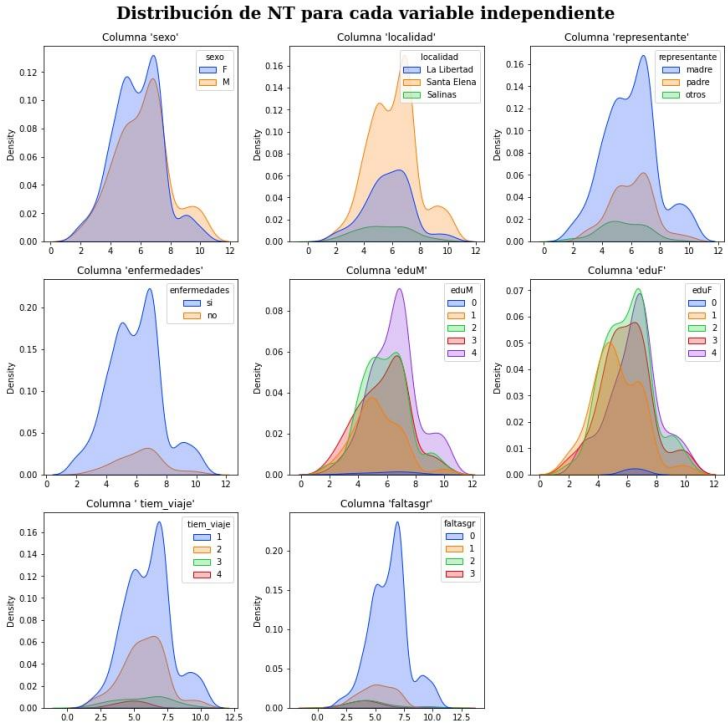


Figura 28. Distribución de NT para las variables seleccionadas.

El resultado de esta subetapa es un nuevo conjunto de datos que contiene 11 variables.

Transformación de datos

Obtenido el conjunto de datos, se corroboró de qué tipo de variables está compuesto, visualizándose la cantidad de variables por tipo de datos que existía.

```
int64     7
object    4
dtype: int64
```

Figura 29. Tipos de datos de los campos del conjunto de datos.

Además, un paso fundamental es analizar el data wrangling, o domado de datos, para corroborar la existencia de valores nulos, valores repetidos o valores que representen outliers, aunque estos ya hayan sido analizados en la subetapa anterior.

Para realizar la correcta aplicación de las técnicas de minería de datos, las variables categóricas deben convertirse a numéricas, puesto que, la predicción a realizar se enfocó en realizar regresiones, y, por ende, aplicar algoritmos concernientes a este caso como Árboles de Decisión de Regresión y Vectores de Soporte de Regresión (SVR).

Una solución empleada para convertir las variables categóricas a variables numéricas fue el empleo del proceso One Hot Encoding. De esta manera, por cada categoría se creó una serie de columnas por medio del método “get.dummies()” de la librería de Pandas.

	eduM	eduF	tiem_viaje	faltasgr	Q1	Q2	NT	F	M	La Libertad	Salinas	Santa Elena	madre	padre	no	si
0	4	4	2	0	10	8	9	1	0	1	0	0	1	0	0	1
1	1	1	1	0	10	9	10	1	0	0	0	1	0	1	0	1
2	1	1	1	3	10	10	10	1	0	0	0	1	1	0	0	1

Figura 30. Proceso One Hot Encoding.

Así, una vez obtenidas las nuevas variables, se procedió a dividir el conjunto de campos en dos variables, una que contenga la variable objetivo y otra que contenga los demás campos para poder aplicar la técnica de minería de datos determinada.

Aplicación de árboles de decisión

Luego del establecimiento de la variable objetivo, el siguiente paso que se realizó fue la importación de "train_test_aplit" por medio de la selección de modelos en la librería "sklearn".

Así, los parámetros que se establecieron fueron cuatro:

- Conjunto de variables a evaluar.
- Variable objetivo.
- Porcentaje del conjunto de pruebas, siendo este, un 20% de los datos totales.
- Estado aleatorio.

Posteriormente, se realizó la importación de "DecisionTreeRegressor" para establecer el regresor. Es en este punto donde se seleccionan los parámetros para la construcción del modelo de regresión, de los cuales dependerá si el modelo posee un buen rendimiento o está sujeto a un sobreajuste. Entre los parámetros establecidos para el modelo, están:

- La profundidad del árbol.
- El número mínimo de muestras necesarias para la división de cada nodo interno.

El primer parámetro se lo estableció en 5, y el segundo, también tuvo un valor de 5. Otros parámetros como "criterion" se mantuvieron con los valores por defecto, en este caso, "mse", o error cuadrático medio para poder ajustar el modelo con los datos correspondientes a entrenamiento.

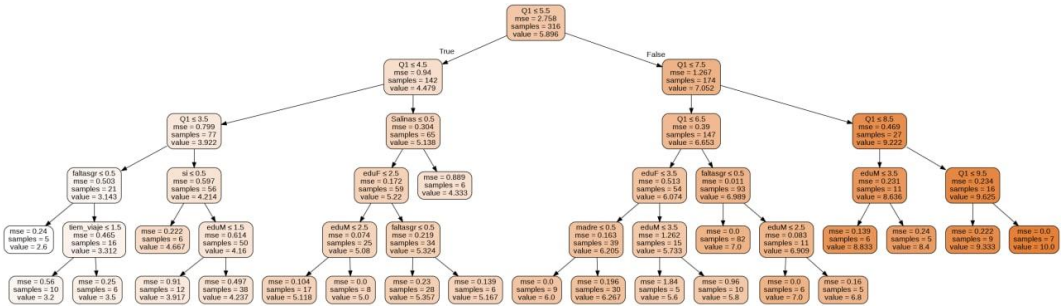


Figura 31. Árbol de decisión de regresión.

Aplicación de redes neuronales

Para este caso, no solo evaluó el conjunto de entrenamiento y prueba, sino que, se separó un porcentaje determinado para el conjunto de validación. Los parámetros fueron los siguientes:

- Conjunto de variables a evaluar.
- Variable objetivo.
- Porcentaje de prueba de 20%.
- Estado aleatorio, cuyo valor fue de 42.

Los mismos parámetros se establecieron al dividir los datos en entrenamiento y validación. A continuación, se muestra la cantidad de datos perteneciente a cada conjunto de datos:

```
Shape of x_train: (284, 14)
Shape of x_test: (79, 14)
Shape of x_val: (32, 14)
Shape of y_train: (284,)
Shape of y_test: (79,)
Shape of y_val: (32,)
```

Figura 32. Conjunto de datos de entrenamiento, prueba y validación.

Debido a las dimensiones de los datos que pertenecen a la variable objetivo, y, para evitar inconsistencias y dar una nueva forma a los arreglos, se procedió a ejecutar la función “reshape”:

```
Shape of x_train: (284, 14)
Shape of x_test: (79, 14)
Shape of x_val: (32, 14)
Shape of y_train: (284, 1)
Shape of y_test: (79, 1)
Shape of y_val: (32, 1)
```

Figura 33. Aplicación de reshape.

Ya aplicados los pasos anteriores, se creó el modelo. Como la base de este modelo de red neuronal fue establecer un perceptrón multicapa, la red se compuso de una capa de entrada, dos capas ocultas y una capa de salida. En cuanto a función de activación se empleó la función Unidad Lineal Rectificada conocida como “ReLU”, mientras que el optimizador para la compilación del modelo fue “Adam” relacionado al momento lineal (momentum) y varianza de la tasa de aprendizaje.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 32)                480
_________________________________________________________________
dense_1 (Dense)              (None, 32)                1056
_________________________________________________________________
dense_2 (Dense)              (None, 16)                528
_________________________________________________________________
dropout (Dropout)            (None, 16)                0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 17
=================================================================
Total params: 2,081
Trainable params: 2,081
Non-trainable params: 0
_________________________________________________________________
None
```

Figura 34. Estructura del modelo de la red neuronal.

A continuación, se muestra la arquitectura de la red neuronal:

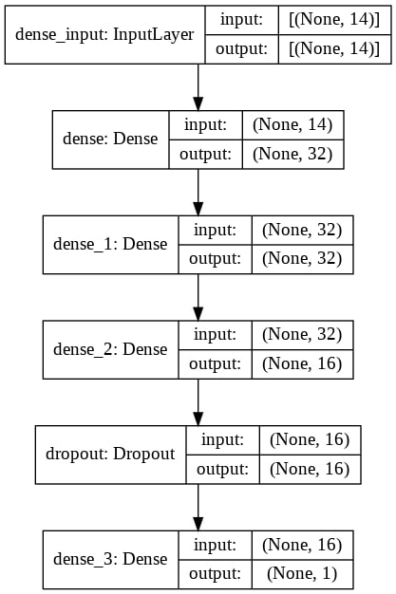


Figura 35. Arquitectura de la red neuronal.

En el ajuste del modelo se establecieron los siguientes parámetros:

- La variable compuesta por la mayoría de campos del conjunto de datos.
- La variable objetivo.
- El conjunto de datos de validación.
- La cantidad de épocas, con un valor de 150 para evitar tanto el subajuste como el sobreajuste.
- El tamaño del lote o batch size, una porción del conjunto de datos de entrenamiento el cual se propagará por toda la red en una iteración.

Aplicación de Máquinas de Vectores de Soporte (SVM)

Con el objetivo de emplear un modelo enfocado en la regresión, se empleó la variante de la Máquina de Vectores de Soporte (SVM), siendo esta, los Vectores de Soporte de Regresión (SVR).

Una vez creada la variable objetivo, se separó los datos empleando los parámetros de los anteriores modelos:

- Conjunto de variables a evaluar.
- Variable objetivo.
- Porcentaje del conjunto de pruebas del 20%.

Posterior a esto, se importó el modelo "SVR" de la librería "sklearn" dentro del método "svm", para lo cual, se establecieron los siguientes parámetros:

- Un kernel.
- Una constante o parámetro de regularización C.
- Un valor gamma.
- Un valor épsilon.

Después del empleo de los parámetros determinados, se ajustó el modelo y creó la variable predictiva por medio del conjunto de datos de prueba.

2.4.4. Cuarta etapa: Evaluación de modelos

Las métricas empleadas para evaluar el funcionamiento de los algoritmos fueron las siguientes:

- Error absoluto medio (MAE).
- Error cuadrático medio (MSE).
- Raíz del error cuadrático medio (RMSE).

Además, otro parámetro para determinar el rendimiento del modelo fue el empleo del parámetro estadístico R^2 o coeficiente de determinación.

El enfoque de estas métricas de regresión se centra en determinar en qué algoritmo existe una menor distancia entre los valores reales y los valores predichos. Por ende, el algoritmo que posea un menor error será el de mejor rendimiento. MAE se encarga de revisar el promedio total del error y sirve como una métrica de precisión tanto para regresión como para pronósticos, mientras que, RMSE, cumple el papel de la desviación estándar de los errores respecto a la función de aproximación, corroborando cuánta variación existe en una predicción.

Para esto, se importaron las métricas por medio de la librería "sklearn".

Métricas en árbol de decisión de regresión

Los resultados para el árbol de decisión de regresión fueron los siguientes:

Métrica	Resultado aproximado
MAE	0.41
MSE	0.38
RMSE	0.62
R^2	0.89

Tabla 9. Métricas del modelo de árboles de decisión de regresión.

Métricas en redes neuronales

Se corroboró el rendimiento del modelo por medio de los valores obtenidos por las métricas de rendimiento, los cuales fueron:

Métrica	Resultado aproximado
MAE	0.58
MSE	0.56
RMSE	0.75
R^2	0.84

Tabla 10. Métricas del modelo de redes neuronales.

Métricas en Vector de Soporte de Regresión (SVR)

Para la respectiva evaluación del rendimiento del modelo SVR se obtuvieron los siguientes datos:

Métrica	Resultado aproximado
MAE	0.62
MSE	0.78
RMSE	0.88
R^2	0.78

Tabla 11. . Métricas del modelo de vectores de soporte de regresión.

2.4.5. Quinta etapa: Difusión de conocimiento

Se realizó la capacitación a los administradores de la institución por medio de una reunión virtual (Ver Anexos 5 y 6), donde se detallaron los siguientes puntos clave:

1. Introducción a la problemática.
2. Objetivos de la propuesta.

3. Metodología aplicada.
4. Resultados.

En cuanto a la introducción a la problemática, se explicó el contexto principal en el que se encuentra la institución y cómo las tecnologías actuales pueden brindar ventajas en el ámbito administrativo, así como, los principales problemas que se han encontrado y que repercuten en el desarrollo estratégico. Además, se especificó cuál fue la técnica de recolección empleada para llevar a cabo la propuesta.

Los objetivos de la propuesta fueron aclarados mediante la explicación de la metodología, detallando cada una de las etapas que conformaba la solución. Antes de llegar a los resultados, se hizo empleo de la visualización de datos para apreciar cuáles eran los datos que se habían obtenido para la realización de la etapa de minería de datos.

Para los resultados, se expuso qué métricas se emplearon y cómo sus valores determinaban el rendimiento del modelo escogido. Además, se realizó la respectiva comparación de los valores reales con los valores predichos a través del modelo de minería de datos escogido.

2.4.6. Requerimientos

De acuerdo a las funciones y etapas del proyecto, los requerimientos a mencionar son los siguientes:

- Se deben conocer las características del uso de los sistemas de Inteligencia de Negocios como almacenes de datos para la resolución de problemas de gestión y análisis empresarial.

- Es necesario entender la problemática de la situación, y así, analizar de una manera detallada para determinar cuál será el punto de partida del proyecto, evitando especulaciones.

- Conocer de las distintas metodologías que pueden emplearse, hará comprensible las fases necesarias e imprescindibles que se deben llevar a cabo durante la ejecución del proyecto.

- Es fundamental conocer cuáles son los roles que participan dentro de las etapas de la investigación, estableciendo así, cuál es la importancia de cada uno y su respectiva función durante la ejecución del proyecto. Entre los roles destacan el Ingeniero de Datos que recolecta la información, y el Analista de datos.

- Se debe proporcionar suficiente información, para que así, se pueda llegar a crear un almacén de datos que contenga numerables campos y datos específicos a emplear para aplicar el proceso de minería de datos.

- Si se desea aplicar un análisis predictivo o descriptivo, se deben conocer las diferentes técnicas de minería de datos y métodos estadísticos para el análisis y procesamiento de datos a aplicar.

- Para crear el conjunto de datos objetivo a minar, se deben aplicar todos los procesos necesarios para su respectiva limpieza, tales como: eliminación de valores duplicados, nulos, e incluso, outliers, los cuales corresponden a datos atípicos. Esto evitará interpretaciones engañosas.

- Uno de los factores que se deben evitar para llegar a conclusiones erróneas es el sesgo, debido al alto riesgo de generar conclusiones subjetivas en la investigación.

- Dependiendo del tipo de minería de datos, clasificación o regresión, se deben seleccionar las métricas adecuadas para evaluar el rendimiento del modelo.

2.5. Diseño de la propuesta

2.5.1. Arquitectura de la solución

La arquitectura planteada en la solución se basa en la metodología KDD, siendo el almacén de datos, la base para realizar la etapa de minería de datos.

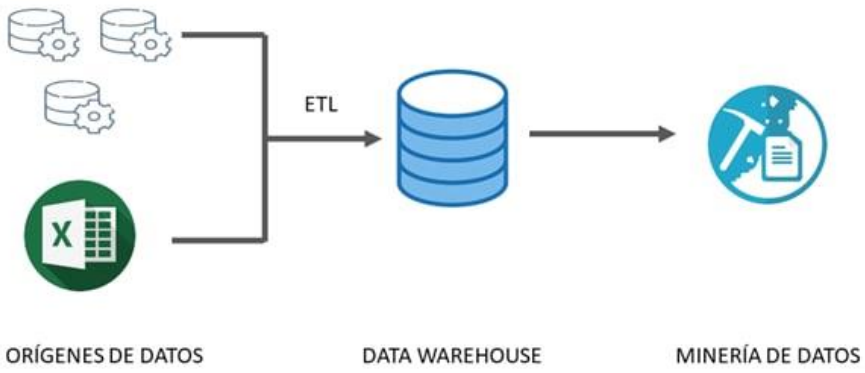


Figura 36. Arquitectura de la solución.

2.5.2. Diagrama físico de datos

Se aplica un modelo físico de datos, el cual se detalla en la primera etapa de la metodología planteada correspondiente a la recopilación y, posteriormente, integración de los datos.

2.5.3. Diccionario de datos

El diccionario de datos se encuentra descrito desde la tabla 3 hasta la tabla 8, mostrando los campos pertenecientes a las tablas más significativas de la base de datos propuesta en la primera etapa de la metodología planteada.

2.6. Presupuesto de la solución

El desarrollo de la solución está delimitado de aspectos como el software, hardware, personal, entre otros, actores que intervienen durante la ejecución del mismo.

Componente	Cantidad	Costo ($)	Detalle	Total ($)
Laptop	1	450.00	Procesador Core i3	450.00

Tabla 12. Presupuesto de Hardware.

Componente
PostgreSQL 13.1
Python 3.8.6
Jupyter Notebook
Pentaho Data Integration
Draw.io

Tabla 13. Presupuesto de software.

Roles	Cantidad	Cantidad de meses	Costo por mes ($)	Total ($)
Ingeniero de datos	1	5	430.00	2150.00
Analista de datos	1	5	430.00	2150.00
Científico de datos	1	5	430.00	2150.00
Total				**6450.00**

Tabla 14. Presupuesto de personal.

Los valores anteriores fueron obtenidos mediante la tabla de salarios mínimos sectoriales 2021 del Ministerio de Trabajo [37].

Componente	Cantidad	Cantidad de meses	Costo por mes ($)	Total ($)
Energía eléctrica	1	5	10.00	50.00
Internet	1	5	24.00	120.00
Total				**170.00**

Tabla 15. Presupuesto de gastos varios.

Presupuesto del proyecto	
El tiempo de elaboración del proyecto es de 5 meses.	
Hardware	$450.00
Software	$0.00
Personal	$6450.00
Gastos varios	$170.00
Total	**$7070.00**

Tabla 16. Presupuesto total de la solución.

De esta manera, presupuesto total aproximado para el desarrollo del proyecto es de $7070, no obstante, en este proyecto el autor asume sus componentes, por ende, el costo total es de $0.00.

2.7. Resultados

2.7.1. Resultados de la evaluación de los modelos

Luego de realizar la cuarta etapa correspondiente a la evaluación de los modelos, se obtuvieron los siguientes resultados:

Técnicas de minería de datos	Métricas de rendimiento			R^2
	MAE	MSE	RMSE	
Árboles de decisión	0.41	0.38	0.62	0.89
Redes neuronales	0.58	0.56	0.75	0.84
SVR	0.62	0.78	0.88	0.78

Tabla 17. Resultados de las métricas de rendimiento.

Según los resultados obtenidos, el modelo que mejor rendimiento posee, debido a un menor valor del error en sus métricas, es el modelo de árboles de decisión de regresión, el cual posee un MAE de 0.41, un MSE de 0.38, y, un RMSE de 0.62. Se emplearon las métricas mostradas anteriormente debido a que los modelos predictivos correspondían a regresión y no a clasificación, por ende, aunque en técnicas como las redes neuronales se pueda emplear métricas como la precisión, esta no es aplicable a las demás técnicas. Además, al evaluar el coeficiente de determinación (R^2), se obtuvo un valor de 0.89 puedo constatar que el modelo obtenido era óptimo.

Además, para corroborar la eficacia del modelo, se procede a realizar un análisis entre los valores reales del conjunto de datos y los valores obtenidos al realizar una predicción.

	Predicción	Valor real	Diferencia (%)
0	4.236842	4	5.921053
1	7.000000	7	0.000000
2	3.200000	3	6.666667
3	4.236842	5	15.263158
4	5.000000	5	0.000000
5	6.266667	7	10.476190
6	8.400000	8	5.000000
7	5.000000	5	0.000000
8	4.666667	4	16.666667
9	5.800000	7	17.142857
10	8.833333	9	1.851852

Figura 37. Valores reales y valores generados mediante la predicción.

La columna “Diferencia (%)” indica el porcentaje de la diferencia entre los valores predichos y los reales. Al revisar la información estadística de esta columna, luego de aplicarse el método “describe” se obtuvo un error aproximado de 8.27%, siendo este, un valor aceptable. Finalmente, se muestran predicciones obtenidas mediante el conjunto de datos, así como el estado del estudiante.

	Predicción	Estado
0	4.2	Reprobado
1	7.0	Aprobado
2	3.2	Reprobado
3	4.2	Reprobado
4	5.0	Reprobado
5	6.3	Reprobado
6	8.4	Aprobado
7	5.0	Reprobado
8	4.7	Reprobado
9	5.8	Reprobado

Figura 38. Predicciones de la nota final de los estudiantes.

2.7.2. Patrones obtenidos

La figura 31, mostrada en la tercera etapa de la metodología al aplicar la técnica de árboles de decisión de regresión, permite determinar cuáles son los patrones determinantes para que los estudiantes de la institución posean un rendimiento adecuado.

- Si la calificación del estudiante durante el primer ciclo (Q1) es mayor a 9.5, el estudiante aprobará con un promedio mayor a 9.
- Si la calificación del estudiante durante el primer ciclo (Q1) es menor o igual a 8.5, el estudiante aprobará con un promedio mayor a 8. No obstante, este patrón se encuentra relacionado al nivel de educación de la madre, pues, si la ponderación equivale a un valor menor o igual a 3.5, el estudiante tendrá un promedio aproximado de 8.8, caso contrario, no superará el 8.4.
- Si la calificación del estudiante durante el primer ciclo (Q1) es menor o igual a 7.5 pero mayor a 6.5, el factor determinante serán las faltas graves que haya tenido durante su periodo escolar, pues, si estas son mínimos o nulas, podrá aprobar con un promedio de 7.0. No obstante, si estas faltas son mayores el patrón vuelve a estar influenciado por el nivel de educación de

la madre, ya que, si este tiene una ponderación menor o igual a 2.5, el estudiante también podrá aprobar con un promedio de 7.0

2.7.3. Resultados de la variable

La variable a evaluar correspondía al tiempo de obtención de los reportes para la toma de decisiones relacionados al rendimiento académico. Según el administrador, en la entrevista realizada previamente (Ver Anexo 1), se determinó que, al no existir una manera de interpretar, de forma directa, indicadores como el rendimiento académico estudiantil, se recurría a un análisis de pruebas como diagnósticos y lecciones, siendo esta, una situación que suele generar demoras en los procedimientos. La solución planteada a través de esta propuesta, permitió conocer, mediante la quinta etapa, difusión de conocimiento (Ver Anexo 5), que estos procesos, ya no involucrarían múltiples sesiones que llegan a abarcar de cinco a seis días, sino, una cantidad máxima de dos días.

Tiempo de obtención de reportes	
Antes	**Después**
5 a 6 días	2 días

Tabla 18. Tiempo de obtención de reportes.

CONCLUSIONES

- Se desarrolló una propuesta tecnológica mediante la aplicación de la metodología Descubrimiento de Conocimiento en Base de Datos (KDD), la misma que consistió de cinco fases. Mediante esta metodología, se establecieron etapas que no solo permitieron conocer el funcionamiento del negocio, sino que, se involucró un estudio de los datos para su respectivo análisis posterior. Además, se pudo comunicar los resultados de una forma óptima a los administradores de la institución.

- Se procedió a la recolección de datos, y posterior integración, en una base de datos que contenía información relacionada a estudiantes, representantes, profesores, entre otros. El resultado de este procedimiento permitió conocer cuáles eran los orígenes de datos, y también, cómo se encontraban relacionados. Además, al aplicar el enfoque Kimball, se logró diseñar un almacén de datos constituido por dos datamarts.

- Se diseñó un almacén de datos, mediante la herramienta Pentaho Data Integration, caracterizada por los procesos de extracción, transformación y carga de datos (ETL). La finalidad de esta etapa se centró en la integración de los datos, siendo el análisis de los mismos, lo que fueron determinantes para emplear la metodología KDD de una forma adecuada.

- La generación del almacén de datos permitió realizar satisfactoriamente la minería de datos, con la finalidad de encontrar conocimiento. Las técnicas de minería fueron: árboles de decisión de regresión, redes neuronales y vectores de soporte de regresión (SVR). Estas se centraron en realizar un análisis predictivo mediante modelos de regresión, por ende, para evaluar su respectivo rendimiento, se emplearon tres métricas, siendo estas: error absoluto medio (MAE), error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE); en donde, el modelo que poseía el menor error, sería el más óptimo a aplicar.

- El método que mejor rendimiento tuvo correspondió a los árboles de decisión de regresión. Los errores obtenidos mediante sus métricas fueron de: un MAE de 0.41, un MSE de 0.38, un RMSE de 0.62, además, al evaluar el coeficiente de determinación, este valor fue de 0.89, lo que indicia que el modelo generado fue óptimo al realizar comparaciones entre los valores reales y valores de la predicción.

- Los patrones obtenidos permitieron conocer que, factores como la educación de la madre y las faltas graves, además de una alta nota del primer ciclo, eran determinantes para que el estudiante pueda aprobar al final de su respectivo periodo académico.

- La propuesta permitió corroborar que, indicadores como el rendimiento académico pueden evaluarse de una forma metódica y evitando cuestiones como el sesgo en la información, esto debido a los resultados obtenidos con la variable, los cuales hacen referencia al tiempo de obtención de los reportes para la toma de decisiones del mismo indicador.

RECOMENDACIONES

- La parte administrativa la institución debe profundizar temas que abarquen el empleo de datos de forma idónea con la finalidad de lograr un desarrollo estratégico futuro, y también, entender que la aplicación de Inteligencia de Negocios es clave para la optimización de procesos de la escuela relacionados a indicadores como el rendimiento académico.

- Emplear una metodología diferente, orientada también a la minería de datos, para determinar si sus respectivas fases pueden cubrir necesidades extras de la institución.

- Para la creación de modelos, aplicar minería de datos predictiva mediante la creación de modelos de clasificación. De esta manera, se podrían evaluar otras métricas de rendimiento, tales como: precisión, exhaustividad, exactitud, entre otras, que también pueden determinar la presencia de irregularidades como el sobreajuste.

- Emplear otros modelos de minería de datos, con el propósito de determinar y ajustar los parámetros de aplicación en situaciones asociadas a la variabilidad de los datos.

BIBILIOGRAFÍA

[1] D. Cohen Karen y E. Asín Lares, Tecnologías de información en los negocios, Quinta ed., México, D.F., México: McGraw-Hill, 2005.

[2] J. Ponce Jarrín, Políticas educativas y desempeño: Una evaluación de impacto de programas educativos focalizados en Ecuador, Primera ed., Quito, Pichincha: FLACSO, 2010.

[3] Ó. A. Erazo, «EL RENDIMIENTO ACADÉMICO, UN FENÓMENO DE MÚLTIPLES RELACIONES Y COMPLEJIDADES,» *Vanguardia Psicológica,* vol. 2, nº 2, p. 30, 2012.

[4] G. Suganya, «Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q,» 2018. [En línea]. Available: https://mro.massey.ac.nz/bitstream/handle/10179/14655/02_whole.pdf?sequence=2&isAllowed=y. [Último acceso: 10 Diciembre 2020].

[5] M. Pojon, «Using Machine Learning to Predict Student Performance,» Junio 2017. [En línea]. Available: https://trepo.tuni.fi/bitstream/handle/10024/101646/GRADU-1498472565.pdf?sequence=1. [Último acceso: 10 Diciembre 2020].

[6] J. J. Solines Bernardino, «Minería de datos aplicada a la detección de patrones para el análisis de rendimiento académico de los estudiantes de la carrera de Ingeniería en Sistemas Computacionales de la Universidad Católica Santiago de Guayaquil,» 24 Septiembre 2018. [En línea]. Available: http://repositorio.ucsg.edu.ec/bitstream/3317/11380/1/T-UCSG-PRE-ING-CIS-202.pdf. [Último acceso: 10 Diciembre 2020].

[7] FACSISTEL, «LÍNEAS DE INVESTIGACIÓN,» 22 Octubre 2020. [En línea]. Available: http://facsistel.upse.edu.ec/index.php?option=com_content&view=article&id=58&Itemid=463. [Último acceso: 17 Diciembre 2020].

[8] J. T. Marchewka, Information Technology Project Management: Providing Measurable Organizational Value, Primera ed., Db Jwo, 2003.

[9] J. L. Cano Giner, Business intelligence: competir con información, Madrid: Fundación Cultural Banesto, 2007.

[10] Secretaría Técnica Planifica Ecuador , «Plan Nacional de Desarrollo 2017 – 2021 Toda una Vida,» [En línea]. Available: https://www.planificacion.gob.ec/plan-nacional-de-desarrollo-2017-2021-toda-una-vida/. [Último acceso: 26 Diciembre 2020].

[11] R. Hernández Sampieri, C. Fernández Collado y M. d. P. Baptista Lucio, Metodología de la Investigación, Sexta ed., México D.F.: McGraw Hill, 2014, p. 91.

[12] Ministerio de Educación, «AMIE (Estadísticas educativas a partir de 2009-2010),» 24 Junio 2021. [En línea]. Available: https://educacion.gob.ec/amie/. [Último acceso: 30 Junio 2021].

[13] M. I. Baas Chable, M. G. Barceló Méndez y G. R. d. F. Herrera Garnica, Metodología de la Investigación, Primera ed., Naucalpan de Juárez, México: Pearson Educación, 2012.

[14] N. Quezada Lucio, Metodología de la investigación: estadística aplicada en la investigación, Primera ed., Lima: Macro, 2010.

[15] J. Hernández Orallo, M. J. Ramírez Quintana y C. Ferri Ramírez, Introducción a la Minería de Datos, D. Fayerman Aragón, Ed., Madrid: Pearson, 2005, p. 680.

[16] Ministerio de Educación, «LEY ORGÁNICA DE EDUCACIÓN INTERCULTURAL,» 14 Marzo 2018. [En línea]. Available: https://educacion.gob.ec/wp-content/uploads/downloads/2020/06/LOEI.pdf. [Último acceso: 18 Julio 2021].

[17] M. V. Mannino, Administración de bases de datos: Diseño y desarrollo de aplicaciones, Tercera ed., Ciudad de México, México: McGraw Hill, 2010.

[18] Nevpro, «Pentaho Business Intelligence Tool,» 23 Septiembre 2020. [En línea]. Available: https://www.nevprobusinesssolutions.com/pentaho-business-intelligence/. [Último acceso: 17 Diciembre 2020].

[19] Draw.io, «Draw.io,» 10 Febrero 2021. [En línea]. Available: https://drawio-app.com/. [Último acceso: 10 Febrero 2021].

[20] R. Gupta, Making use of Python, B. Ryan, Ed., New York: Wiley Publishing, 2002, p. 416.

[21] Jupyter, «Jupyter,» 15 Diciembre 2020. [En línea]. Available: https://jupyter.org/. [Último acceso: 17 Diciembre 2020].

[22] L. Hsu y R. Obe, PostgreSQL: Up and Running, Primera ed., M. Blanchette, Ed., San Francisco, California: O'Reilly Media, 2012.

[23] M. Alexander, Microsoft Access 2007 data analysis, Indiniapolis, Indiana: Wiley Publishing, 2007.

[24] M. Biswas y A. Nandy, Reinforcement Learning : With Open AI, TensorFlow and Keras Using Python, 1 ed., Calcuta, Bengala Occidental: Apress, 2018.

[25] G. Moncecchi y R. Garreta, Learning scikit-learn: Machine Learning in Python, Birmingham, Midlands Occidentales: Packt Publishing, 2013.

[26] Scikit-learn, «Scikit-learn,» 28 Junio 2021. [En línea]. Available: https://scikit-learn.org/stable/. [Último acceso: 19 Julio 2021].

[27] S. Tosi, Matplotlib for Python Developers, Birmingham, Midlands Occidentales: Packt Publishing, 2009.

[28] Seaborn, «Seaborn,» 15 Agosto 2021. [En línea]. Available: https://seaborn.pydata.org/. [Último acceso: 17 Agosto 2021].

[29] J. Curto Díaz, Introducción al Business Intelligence, Primera ed., Barcelona: UOC, 2010.

[30] Z. Díaz Martínez, Predicción de crisis empresariales en seguros no vida mediante árboles de decisión y reglas de clasificación, Primera ed., Madrid, Madrid: Complutense, 2007.

[31] J. M. Fernández Fernández y R. Flórez López, Las Redes Neuronales Artificiales: Fundamentos téoricos y aplicaciones prácticas, Primera ed., La Coruña: Netbiblo, 2008.

[32] J. Bobadilla Sancho, Machine Learning y Deep Learning: Usando Python, Scikit y Keras, Primera ed., Bogotá, Cundinamarca: Ra-Ma, 2020.

[33] J. M. Rodríguez Parrilla, Cómo hacer inteligente su negocio: Business Intelligence a su alcance, Primera ed., Ciudad de México, México: Patria, 2014.

[34] E. Pulido Romero, Ó. Escobar Domínguez y J. Á. Núñez Pérez, Base de datos, Primera ed., Ciudad de México, México: Patria, 2019.

[35] J. Gironés Roig, J. Casas Roma, J. Minguillón Alfonso y R. Caihuelas Quiles, Minería de datos: Modelos y algoritmos, Primera ed., Barcelona: UOC, 2017.

[36] J. Pardillo Vela, J. C. Trujillo Mondejar y N. Mazón López, Diseño y explotación de almacenes de datos: Conceptos básicos de modelado multidimensional, Primera ed., San Vicente, Alicante: ECU, 2011.

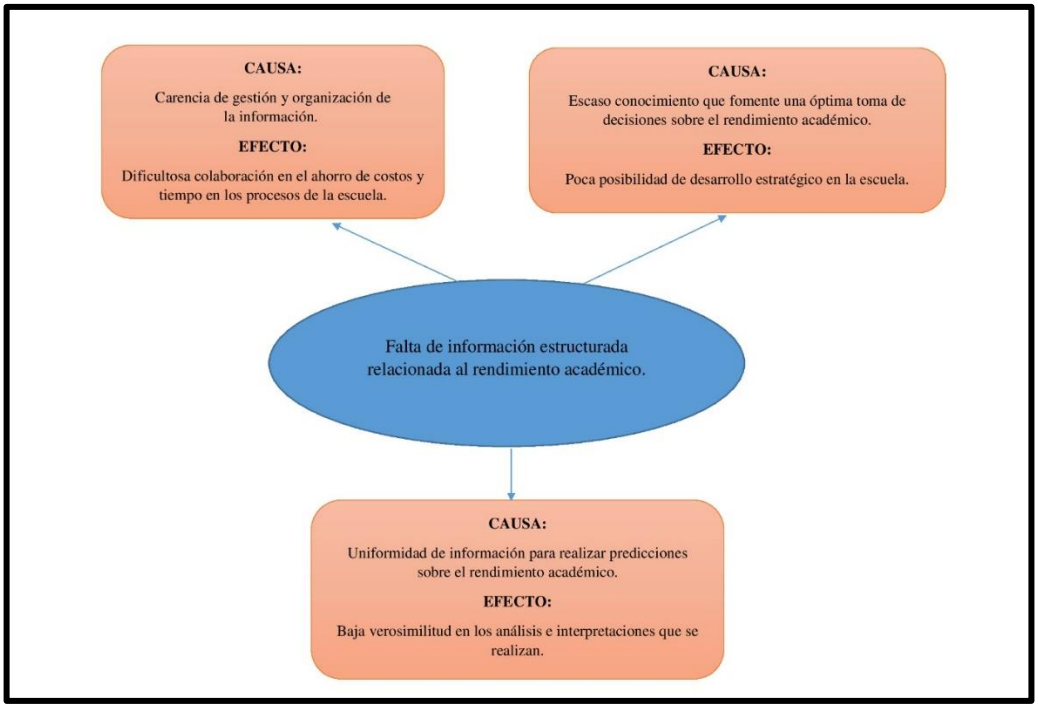
[37] Ministerio del Trabajo, «SALARIOS MÍNIMOS SECTORIALES 2021,» 9 Julio 2021. [En línea]. Available: https://www.trabajo.gob.ec/wp-content/uploads/2020/12/ANEXO-1%E2%80%9CEstructuras-ocupacionales-%E2%80%93-salarios-m%C3%ADnimos-sectoriales-y-tarifas-sa.pdf?x42051. [Último acceso: 2 Agosto 2021].

ANEXOS

ENTREVISTA	
Entrevistador:	Alex Villao Balón
Entrevistado:	Arq. Mauricio Yunda Villao
Objetivo:	Obtener información básica sobre la escuela, así como indagar sobre la organización de la misma.
Fecha:	27/04/2021
Ubicación:	Calle 18 de agosto entre Paquisha y, Fausto Fajardo, Santa Elena

1. ¿En qué año se fundó la escuela?
2. ¿Cómo se encuentra organizada la escuela?
3. ¿Cómo ha ido progresando el rendimiento académico conforme han pasado los años?
4. ¿Existen inconvenientes en el proceso de toma de decisiones de la escuela?
5. ¿Qué opina sobre el cambio de administración en la escuela?
6. ¿Cuál es el nivel actual de los estudiantes de la escuela?
7. ¿Cómo evalúan el rendimiento académico?
8. ¿Cuánto tarda el análisis de resultados con respecto al rendimiento académico?
9. ¿Qué expectativas futuras tiene usted sobre la escuela?

Anexo 1. Formato de la entrevista.



Anexo 2. Árbol de problemas.

Anexo 3. Misión y visión de la institución.

MINISTERIO DE EDUCACIÓN

ESCUELA GENERAL BÁSICA
"LCDA. ANGÉLICA VILLÓN LINDAO"
Fundada el 12 de Septiembre del 2000
2020 -2021
Telf.: 02942-097 0992158416

JARDÍN ESCUELA
SANTA ELENA

FICHA DE MATRÍCULA # .....

DATOS DEL ESTUDIANTE:			
APELLIDOS:	NOMBRES:		
NÚMERO DE CÉDULA:	NACIONALIDAD:		
EDAD:	FECHA DE NACIMIENTO : / /		
DIRECCIÓN DOMICILIARIA:			
CANTÓN:	CIUDAD:	BARRIO:	
Nº GRUPO FAMILIAR:	Nº HERMANOS:	LUGAR ENTRE HERMANOS:	
EL ESTUDIANTE VIVE CON:	PADRES	MADRE	ABUELOS
AÑO DE ESCOLARIDAD:	RELIGIÓN:		
ENFERMEDADES O ALERGIAS :			
OBSERVACIONES :			
DATOS DE LA MADRE :			
NOMBRES:			
APELLIDOS:			
FECHA DE NACIMIENTO:			
NÚMERO DE CÉDULA:			
INDIQUE SU NIVEL EDUCACIONAL:			
NIVEL BÁSICA IMCOMPLETA ----	E. MEDIA ----		
NIVEL BÁSICA COMPLETA ----	E. UNIVERITARIA ----		
SIN ESCOLARIDAD	OTROS.....		
PROFESIÓN :	OCUPACIÓN:		
CELULAR:	TELÉFONO DEL TRABAJO:		
LUGAR DE TRABAJO:	RELIGIÓN:		
DATOS DEL PADRE :			
NOMBRES:			
APELLIDOS:			
FECHA DE NACIMIENTO:			
NÚMERO DE CÉDULA:			
INDIQUE SU NIVEL EDUCACIONAL:			
NIVEL BÁSICA IMCOMPLETA ----	E. MEDIA ----		
NIVEL BÁSICA COMPLETA ----	E. UNIVERITARIA ----		
SIN ESCOLARIDAD	OTROS.....		
PROFESIÓN :	OCUPACIÓN:		
CELULAR:	TELÉFONO DEL TRABAJO:		
LUGAR DE TRABAJO:	RELIGIÓN:		
EMPODERAMIENTO DE PARTE DE PADRES			
YO,	padre de familia del estudiante,		
Delego a	CON C.I	pueda retirar a mi representado.	

Lcda. Ibelice Tomalá Villón MSc.
DIRECTORA ADMINISTRATIVA

PADRE DE FAMILIA

Anexo 4. Ficha estudiantil de la institución.

Anexo 5. Etapa de difusión del proyecto.

Anexo 6. Difusión de resultados de los modelos de minería de datos.

Anexo 7. Cronograma de la propuesta tecnológica.