Recibido 28 sep. 2021 *ReCIBE, Año 11 No. 1, may. 2022*

Aceptado 22 jun. 2022

Implementación de un almacén de datos en el Centro de Enseñanza del Laboratorio Nacional de Informática Avanzada (LANIA) aplicado al estudio de perfiles de deserción.

Implementation of a data Warehouse in the Centro de Enseñanza del Laboratorio Nacional deInformática Avanzada (LANIA) applied to the study of desertion profiles.

Jesús Alberto Torres Sosa
jtorres@lania.edu.mx

Juan Manuel Gutiérrez Méndez

Idalì Nieto Jiménez

Resumen: En la actualidad las organizaciones generan mucha información día a día, misma que puede ser de utilidad para adquirir conocimiento y para la toma de decisiones. El proyecto se enfoca en la construcción de un almacén de datos, que contiene las variables correspondientes para poder aplicar estadísticas descriptivas y así visualizar patrones de comportamiento en la deserción académica de los alumnos de posgrado del Centro de Enseñanza LANIA; dichas variables se obtuvieron de un estudio del estado del arte. El trabajo brinda al lector un panorama de la construcción del almacén de datos, donde se destaca principalmente el proceso ETL, el cual se refiere a la extracción, transformación y carga de los datos. También se puede destacar la aplicación de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), donde se llevó a cabo cuatro fases de seis que contiene. Se pudo observar en los resultados que el 57.14% de los estudiantes que desertaron de la Maestría en Computación Aplicada (MCA) está en el rango de 23 a 25 años y el 40.47% de los estudiantes que desertaron de la Maestría en Redes y Sistemas Integrados (MRySI) está en el rango de 27 a 31 años.

Palabras claves: Deserción académica, almacén de datos, metodología CRISP-DM, proceso ETL.

Abstract: Today, organizations generate a lot of information every day, which can be useful to acquire knowledge and for decision making. The project focuses on the construction of a data warehouse, which contains the corresponding variables to be able to apply descriptive statistics and thus visualize behavior patterns in the academic desertion of postgraduate students of the Centro de Enseñanza LANIA; these variables were obtained from a study of the state of the art. The work provides the reader with an overview of the construction of the data warehouse, where the ETL process stands out, which refers to the extraction, transformation and loading of data. The application of the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology can also be highlighted, where four phases of the six it contains were carried out. It was observed in the results that 57.14% of the students who dropped out of the Maestría en Computación Aplicada (MCA) are in the range of 23 to 25 years old and 40.47% of the students who dropped out of the Maestría en Redes y Sistemas integrados (MRySI) is in the range of 27 to 31 years.

Keywords: Academic desertion, Warehouse, CRISP-DM methodology, ETL process

1. Introducción.

En América Latina las Instituciones de Educación Superior (IES), representan una alta tasa de deserción estudiantil, especialmente en los primeros semestres, según (Pereira et al., 2013). Por otra parte, la deserción académica es un fenómeno que no solo se presenta en las IES, sino que es un problema en todos los niveles educativos (Díaz Peralta, citado por (Eckert & Suénaga, 2015)). Por lo tanto, este tipo de problema conlleva a varios tipos de efectos como lo son financieros, académicos y sociales, como para las instituciones educativas, para el alumno y hasta para los países (Pereira et al., 2013). Actualmente existen varias definiciones de lo que es deserción; la deserción académica: "es el abandono de las actividades estudiantiles antes de terminar algún grado o nivel educativo" esta es una definición de la Secretaría de Educación Pública citada por (Romo et al., 2015). La deserción se entiende como "la interrupción que realizan los alumnos en los estudios emprendidos antes de culminar el total de número de las asignaturas del programa académico en el cual se encuentra matriculado" (Romo et al., 2015).

En respuesta a este tipo de problemática, varias instituciones de América Latina han estudiado este problema con ayuda de la tecnología y se han utilizado técnicas de minería de datos. La minería de datos, también conocida como Descubrimiento de Conocimiento en Bases de Datos por sus siglas en inglés ("KDD; Knowledge Discovery in Databases"), es el campo el cual permite descubrir información en grandes volúmenes de datos (Galindo & García, 2010). También han analizado los datos de los alumnos a través del tiempo que pasa en la institución, esto gracias a los almacenes de datos (data Warehouse). Los repositorios de almacenes de datos se crean con el propósito de generar reportes y de analizar datos históricos de una organización de forma eficaz, y así encontrar tendencias de comportamiento en el tiempo (Matamala et al., 2011).

En este artículo se presenta la construcción de una data Warehouse, como un almacén de datos que guarde información de los alumnos que hayan estudiado y se encuentren cursando actualmente en LANIA, el cual se construyó de la revisión literaria que han abordado problemas similares a la deserción académica, el cual sirvió para determinar las variables o atributos que llevó la data Warehouse.

1.1 Problemática.

Actualmente existen registro donde hay alumnos de las maestrías que se han dado de baja del programa académico quienes argumentan como causas: el rendimiento académico, problemas económicos, problemas de salud entre otros aspectos personales; pero dicha información no es clara en las fuentes de información con las que cuenta LANIA para gestionar los datos de los alumnos, las cuales incluyen entre otras fuentes de información como Moodle ya que se encarga de llevar el registro de las actividades y tareas de los alumnos al interior de cada materias del mapa curricular; SIPAL que se encarga del control y seguimiento de los pagos de las inscripciones de los alumnos y Vfront que se encarga del acceso a la información de la trayectoria académica de los alumnos. Estas fuentes de información muestran datos en vistas parciales y no es posible identificar indicios de que existe un problema con el alumno, lo que origina que la detección de riesgo se lleve a cabo de forma tardía y no permita atender el problema a tiempo ya que no se cuenta con una herramienta que se alimente de las diferentes fuentes de información de LANIA, donde se guarde los datos de los alumnos para que se puedan analizar, procesar y encontrar patrones o indicios de que alumnos son candidatos a desertar.

2 Trabajos relacionados.

2.1 Estudios sociales sobre deserción académica.

(Ander-Egg, 1980) menciona en su libro que un estudio social es un proceso que usa la metodología científica para obtener nuevos conocimientos mediante las entrevistas o las encuestas, ya que son unas de las técnicas más utilizadas por los investigadores (Cerón & Cerâon, 2006). En este tipo de estudio se encontró variables que influyen en la toma de decisión al momento de desertar, pero dichas variables se encuentran categorizadas.

(Romo et al., 2015) en su trabajo analiza las características que involucran a los alumnos que desertan de los programas de posgrado, donde identificaron cinco categorías con sus respectivas variables que influyen en la deserción académica a nivel posgrado durante su investigación del estado del arte. Las clasificaciones de las categorías se muestran en la Tabla 1. Una vez identificadas las categorías realizaron un caso práctico de los posgrados del Centro Universitario de Ciencias Económico-Administrativas de la Universidad de Guadalajara donde obtuvieron como resultados que las categorías que intervienen en la deserción de los alumnos son los sociológicos y los socioeconómicos.

Categorías	Variables
Individuales	● Problemas de salud. ● Problema Psicológico. ● Acontecimiento biológico. ● Expectativas personales. ● Hábito de estudio.
Socioeconómicos	● Situación económica familiar. ● Situación económica personal. ● Expectativa de ingreso. ● Necesidad económica.
Sociológicos	● Situación laboral. ● Problema laboral. ● Nivel educativo de los padres. ● Familia disfuncional. ● Problema familiar.
Institucionales	● Calidad del programa. ● Método de estudio. ● Estructura curricular. ● Modalidad de estudio. ● Apoyo institucional.
Académicos	● Integración a la comunidadestudiantil. ● Calidad de los profesores. ● Acompañamiento estudiantil. ● Desempeño académico anterior. ● Compromiso con el programa.

Tabla 1. Categorías y variables identificadas por Romo et al. (2015).

Otro de los trabajos revisado fue el de (Agudelo & Angulo, 2015) realizado en Colombia donde utilizaron un enfoque cualitativo haciendo uso de las entrevistas para conocer los motivos que se involucran en la deserción de los alumnos, donde en su investigación logró identificar tres categorías; en la Tabla 2 se muestra las categorías encontradas con sus respectivas variables. Al finalizar su estudio identificó que la categoría con más presencia al momento de la deserción es el Académico por lo que sugiere monitorear las plataformas académicas que se utilizan para el programa, así como también tener sesiones de chat e incorporar foros para los alumnos.

Categorías	Variables
Proceso de admisión	• Información incompleta.
Institucionales	• Plan de estudio. • Calidad del programa. • Método de estudio.
Académicos	• Mal servicio del tutor académico. • Integración al trabajo de equipo. • Nivel educativo de los padres. • Falta de atención al alumno. • Dificultad con la plataforma de apoyo.

Tabla 2. Categorías y variables identificadas por Agudelo & Angulo (2015).

(Barrientos & Umaña, 2010) realizó una investigación en Costa Rica a nivel posgrado, donde la información para dicho trabajo la obtuvo consultando bases de datos, entrevistas estructuradas, entrevistas libres a alumnos y, por último, registros internos del posgrado; logrando identificar seis categorías con sus respectivas variables que afecta en la deserción académica; las categorías y sus variables; se muestra en la Tabla 3.

Categorías	Variables
Psicológicas	• Egresado no titulado. • Modalidad del programa. • Estado civil. • Tener hijos. • Sexo. • Edad. • Apoyo económico.
Académicos	• Educación previa. • Grado académico. • Promedio de hora a la semana. • Promedio de calificación.
Sociológicos	• Grado académico del núcleo familiar. • Situación laboral.
Socioeconómicos	• Casa propia. • Vive con los padres. • Cuenta con ingreso familiar.
Organizacionales	• Capacidad del docente. • Perfil de salida del programa. • Cambio en el proyecto de tesis.
Institucionales	• Calidad de la institución. • Programas menos interesantes. • Cambio de modalidad de estudio.

Tabla 3. Categorías y variables identificadas por Barrientos & Umaña (2010).

2.2 Estudios técnicos sobre deserción académica.

En el estudio técnico se hace uso de las técnicas computacionales como es la minería de datos, la inteligencia artificial, entre otras; con el propósito de crear modelos predictivos. Para llevar a cabo la creación de modelo predictivo es necesario extraer información existente con el propósito de predecir tendencia y encontrar patrones de comportamientos (Espino Timón, 2017).

(Eckert & Suénaga, 2015) en su trabajo analizó información académica para identificar factores que influyen en la deserción de los alumnos de la carrera en Informática. Gastón Dachary en Argentina, analizó variables relacionadas con los resultados académicos del alumno como los siguientes: promedio general del alumno en el primer año, número de materias cursadas en el primer año, localización geográfica, entre otras. Utilizo tres algoritmos en la herramienta de WEKA (Waikato Environment for Knowledge Analysis), los cuales fueron: árbol de decisión (J48), clasificador Naive Bayes y el algoritmo clasificador OneR. Teniendo como resultado que el algoritmo J48 fue el que mayor porcentaje de clasificación obtuvo correctamente y el clasificador OneR fue el que menor porcentaje de clasificación obtuvo.

(Azoumana, 2014) realizó un análisis sobre la deserción de los alumnos de la Universidad Simón Bolívar donde seleccionó una población de 707 alumnos del programa de Ingeniería de Sistemas, tomando las variables siguientes: pérdida de semestre, dificultad financiera, ingreso al mercado laboral, otros intereses y por último indeterminado. Para procesar la información utilizó la herramienta de WEKA donde aplicó el algoritmo de clasificación y como resultado obtuvo un margen de confianza de 94% y la variable que se pudo observar que era determinante para que el alumno tomara la decisión de desertar fueron: Indeterminado y Dificultad financiera.

(Matamala et al., 2011) en su trabajo realizó un almacén de datos con el objetivo de analizar el desempeño académico de los estudiantes, dicho almacén de datos se utilizó con el enfoque de Procesamiento Analítico OnLine Relacional (ROLAP por sus siglas en inglés) ya que les permitió la acción de explorar a través de operaciones definidas el análisis y la creación de reportes bajo el modelo relacional, uno de los procesos que se llevó a cabo en la creación del almacén de datos fue el ETL (Extraction, Transformation, Load), el cual les permitió obtener la información de las bases de datos originales para posteriormente pasar a la limpieza o transformación de los datos y por último cargarla al almacén de datos. También implementaron una red neuronal con el propósito de predecir el comportamiento del alumno en el siguiente semestre el cual se alimentaba de la información del almacén de datos.

(Pereira et al., 2013) en su trabajo detectó patrones de deserción de los alumnos utilizando técnicas de minería de datos tomando como referencia datos socioeconómicos, académicos, disciplinares e institucionales de los alumnos de programas de pregrado de la Universidad de Nariño y la Institución Universitaria de la ciudad de Pasto Colombia. También optaron por la implementación de un repositorio de datos, utilizando PostgreSQL como su sistema gestor de base de datos, el cual alimentaron con datos de los alumnos que ingresaron en los años 2004, 2005 y 2006 con el propósito de realizar un seguimiento completo de los alumnos. En su trabajo pudieron observar patrones generales en la deserción de los alumnos al momento de aplicar técnicas de clasificación y de agrupamiento sobre los datos. Las variables que encontraron al momento de la deserción fueron: un promedio bajo, materias reprobadas en los primeros semestres y alto costo de la colegiatura.

(Pereira & Toledo, 2014) en su trabajo detectó patrones de deserción de alumnos de la Universidad de Nariño e Institución Universitaria, donde aplicó la metodología CRISP-DM para su proyecto de minería de datos, construyeron un repositorio con el objetivo de tener los datos centralizados para encontrar patrones que inciden en la deserción de los alumnos, la información utilizada fueron: datos socioeconómicos y datos académicos. Para la técnica de minería de datos utilizaron el algoritmo de árboles de decisión (J48), donde obtuvieron como resultado las siguientes variables involucradas en la deserción: promedio bajó, materias reprobadas en los primeros semestres y puntaje bajo en la prueba de admisión.

(Moreno Serrano, 2017) en su trabajo muestra la aplicación de la metodología CRISP-DM, donde aplicó técnicas de minería de datos con los alumnos de los diferentes diplomados impartidos por el Laboratorio de Informática Avanzada, utilizando los algoritmos de árbol de decisión (J48), red bayesiana y redes neuronales con el objetivo de tener una predicción semanal del riesgo de deserción que tiene un alumno. Obteniendo mejores resultados con el algoritmo de red bayesiana y la red neuronal.

En la Tabla 4; se muestra los nombres de los estudios revisados así como el número de variables y la técnica utilizada en la investigación, el cual es un resumen general de lo encontrado en el estudio técnico.

Autores	Estudios	Número de variables	Técnicas
(Eckert & Suénaga, 2015)	Análisis de deserción permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos.	10	● Árbol de decisión ● Naive Bayes ● Regla OneR
(Azoumana, 2014)	Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnica de minería de datos.	5	● Clasificaciones
(Matamala et al., 2011)	Análisis de rendimiento académico estudiantil usando almacén de datos y redes neuronales.	15	● Análisis Rolap ● Redes neuronales
(Pereira et al., 2013)	Descubrimiento de perfiles de deserción estudiantil con técnica de minería de datos.	13	● Árbol de decisión ● K-Means
(Pereira & Toledo, 2014)	Detección de patrones de deserción estudiantil en programas de pregrado de Institución Superior con CRISP-DM.	30	● Árbol de decisión ● K-Means ● Apriori
(Moreno Serrano, 2017)	A2: MP Prototipo del módulo de predicción del asesor de asesores.	35	● Redes neuronales ● Naive Bayes ● Árbol de decisión

Tabla 4. Resumen de los estudios técnicos.

2.3 Metodología CRISP-DM.

(Parra et al., 2010) la metodología CRISP-DM por sus siglas en inglés (Cross Industry Standard Process for Data Mining) es una de las principales metodologías seguida por los analistas en la inteligencia de negocios, donde se puede rescatar primordialmente el almacén de datos y la minería de datos, ya que está sustentada en estándares internacionales por lo que facilitan la unificación de sus fases. (Chapman et al., 2000) y (Galán Cortina, 2016) la metodología cuenta con seis fases, las cuales se describen a continuación:

- **Fase 1: Comprensión del negocio.** Se centra en comprender los objetivos del proyecto y los requisitos desde una perspectiva de negocio.
- **Fase 2: Comprensión de los datos.** Se comienza con la recolección inicial de los datos y posteriormente identificar problemas de calidad de los datos y descubrir los primeros conocimientos.
- **Fase 3: Preparación de los datos.** Abarca todas las actividades necesarias para construir el conjunto de datos finales a partir de los datos iniciales, es probable que esta tarea se realice varias veces.
- **Fase 4: Modelado.** Se centra en seleccionar y aplicar técnicas de modelado, también en calibrar los parámetros para la optimización de los datos y construir modelos de pruebas y finales.
- **Fase 5: Evaluación.** Antes de proceder a la última fase, es importante evaluar y revisar los pasos ejecutados, para estar seguro de que el modelo logra adecuadamente los objetivos.
- **Fase 6: Despliegue.** Dependiendo de los requisitos, el cliente puede realizar la implementación y no el analista ya que en esta fase se generan los reportes finales.

De las seis fases que tiene esta metodología, solamente se emplearon las primeras cuatro para la creación del almacén de datos, como se muestra en la Figura 1 de color amarillo claro.

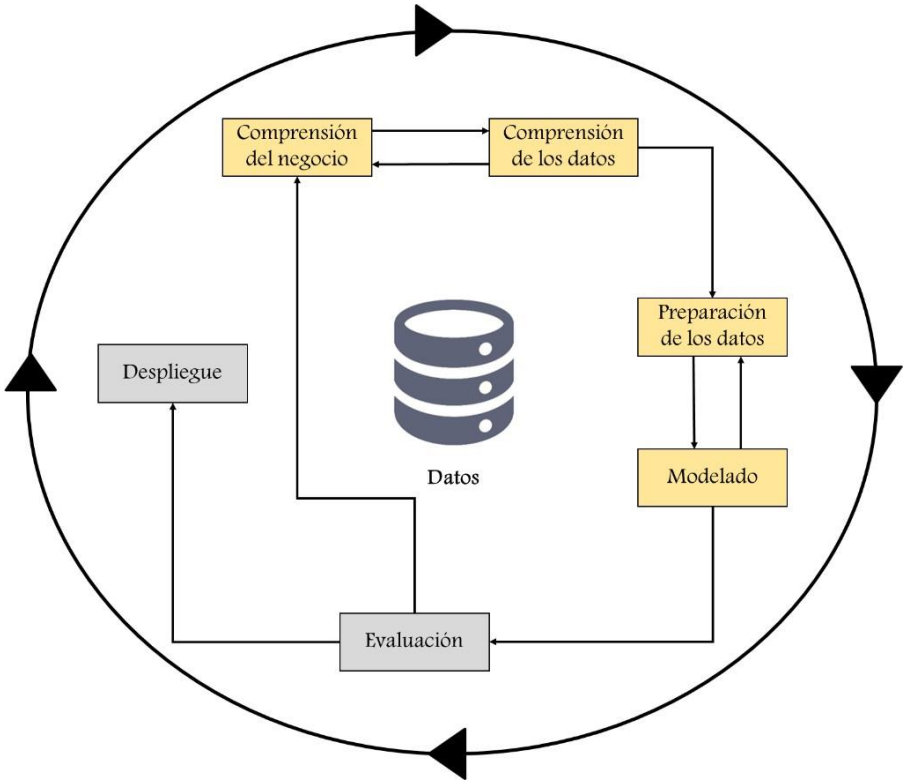


Figura *1*. Fases de la metodología CRISP-DM.
Fuente: Elaboración propia.

3 Metodología.

La metodología que se utilizó para llevar a cabo la investigación estuvo compuesta de las siguientes fases:

Fase 1: Estudio de la literatura. (El cual se realizó en el apartado 2.)
Fase 2: Selección de variables y categorías.
Fase 3: Modelado e implementación del almacén de datos.
- Aplicación de la metodología CRISP-DM

Fase 4: Proceso ETL.
Fase 5: Validación de los datos
Fase 6: Aplicación de estadística descriptiva.

3.1 Selección de variables y categorías.

En la Fase 2; se seleccionó las categorías y variables para la construcción del almacén de datos a partir de la revisión de la literatura que se realizó en la Fase 1, se determinaron seis categorías y cuarenta y seis variables que son cualitativa y cuantitativa y por último dos variables que son de control. Estas últimas dos variables no sirven para un modelo predictivo, la única función es la de identificar al alumno. Las categorías y variables seleccionadas se muestran en la Tabla 5.

Categorías	Variables
Personales	• Nombre. • Id del usuario. • Fecha de nacimiento. • Dirección. • Estado civil. • Procedencia. • Nacionalidad.
Individuales	• Problemas de salud. • Problema psicológico. • Acontecimiento biológico. • Expectativas personales.
Sociológicos	• Situación laboral. • Problema laboral. • Nivel educativo del padre. • Nivel educativo de la madre. • Familia disfuncional. • Tiempo invertido en el trabajo. • Problemas personales.

Socioeconómicos	● Necesidades económicas. ● Casa propia. ● Cantidad que paga en el estudio. ● Vive con sus padres. ● Dependientes económicos. ● Cuenta con deudas.
Académicos	● Integración estudiantil. ● Acompañamiento estudiantil. ● Desempeño académico anterior. ● Compromiso con el programa. ● Trabajo de equipo. ● Plataforma de apoyo. ● Grado académico actual. ● Horas invertidas. ● Promedio actual. ● Número de materias cursadas. ● Número de materias aprobadas. ● Número de materias reprobadas. ● Calificación de la prueba de admisión. ● Tipo de colegio previo. ● Generación. ● Estatus académico. ● Motivo de deserción. ● Programa al que pertenece. ● Fecha del periodo. ● Número de periodo.
Psicológicas	● Tiene hijos. ● Sexo. ● Edad. ● Apoyo económico.

Tabla 5. Categorías y variables finales.

3.2 Modelado e implementación del almacén de datos.

Para la Fase 3; se llevó a cabo el modelado y la implementación del almacén de datos, donde se utilizó la metodología CRISP-DM en sus primeras 4 etapas.

Etapa 1; comprensión del negocio, se realizó una evaluación de los recursos de LANIA, en dicha avaluación se determinaron las fuentes de información con las que se cuenta, con el propósito de saber de dónde se podría obtener la información de los alumnos, también se identificó si son fuentes internas o externas y por último si las fuentes se encontraban de manera On-line u Off-line. A continuación se muestran las fuentes de información que tiene LANIA.

- SICEES - On-line / Externo: Es un sistema de la secretaría de educación para uso de control escolar en donde se realizan inscripciones/reinscripciones también se gestiona los planes de estudios, entre otras cosas.
- SISBEC - On-line / Externo: Es un sistema de la secretaría de educación del estado de Veracruz para instituciones particulares donde pueden postular a los alumnos para becas particulares.
- Moodle - On-line / Interno: Es el sistema que se encarga de gestionar las actividades del alumno como son las tareas.
- CONACyT - On-line / Externo: Es el sistema donde se sube un reporte general del alumno de lo que realizó durante el periodo escolar.
- Vfront - Off-line / Interno
- SIPAL - On-line / Interno: Es el sistema encargado de gestionar los pagos de colegiatura e inscripción del alumno durante su trayectoria académica.
- SharePoint - On-line / Interno
- Documentos de Excel - Off-line / Interno

Etapa 2; comprensión de los datos; se dividieron en dos conjuntos las variables que se muestran en la Tabla 6; el primer conjunto hace referencia a las variables, con las que cuentan LANIA que es un total de veinticinco variables incluyendo una variable para identificar al alumno que es el **Nombre** y son las primeras veinticinco variables y el segundo conjunto de datos son con las que no cuenta actualmente LANIA que es un total de veintitrés variables las cuales empiezan en la veintiséis. En esta fase también se identificaron variables con diferentes formatos uno de ellos fue el sexo que se encontró como (H y M) y (Masculino y Femenino), también se encontraron datos nulos; las diferencias de formato en esta variable se deben a que la información revisada viene de dos departamentos diferentes y los datos nulos se debe a que los alumnos no llenan el formulario de inscripción o reinscripción completo.

Variables		
1. Materias cursadas.	17. Dirección.	33. Endeudamiento.
2. Materias aprobadas.	18. Fecha de nacimiento.	34. Familia disfuncional.
3. Promedio general actual.	19. Nombre completo.	35. Examen aprobado.
4. Materias reprobadas.	20. Nacionalidad.	36. Promedio de admisión.
5. Tipo de colegio previo.	21. Sexo.	37. Integración estudiantil.
6. Promedio académico anterior.	22. Edad.	38. Trabajo en equipo.
7. Acompañamiento estudiantil.	23. Apoyo económico.	39. Dificultad con la plataforma de apoyo.
8. Compromiso con el programa.	24. Estatus.	40. Tiene hijo.
9. Grado académico actual.	25. Periodo.	41. Dependientes económicos.

10. Hora semanal en la escuela.	26. Factor económico.	42. Cuenta con casa propia.

11. Estado de procedencia.	27. Expectativas personales.	43. Vive con sus padres.
12. Estado civil.	28. Acontecimiento.	44. Tipo de residencia.
13. Programa.	29. Problemas psicológicos.	45. Problema laboral.
14. Generación	30. Problemas de salud.	46. Educación del padre.
15. Fecha fin del periodo.	31. Situación laboral.	47. Educación de la madre.
16. Motivo de deserción.	32. Hora semanal al trabajo.	48. Situación personal.

Tabla 6. Variables consideradas para la construcción del almacén de datos.

A continuación, se describen algunas variables que su función no es sencilla de interpretar de la tabla anterior.
Materia cursada: almacena el número de materias cursadas por el alumno.
Materia aprobada: almacena el número de materias que aprobó durante el trimestre o cuatrimestre.
Materias reprobadas: almacena el número de materias que reprobó durante del trimestre o cuatrimestre.
Tipo de colegio previo: almacena si el colegio previo era privada o pública.
Periodo: almacena el número de trimestre o cuatrimestre que se encuentra el alumno.
Fecha fin del periodo: almacena la fecha en que termina el trimestre o cuatrimestre según el calendario escolar.
Factor económico: almacena un número, el cual se obtiene entre la división de ingreso del alumno y los gastos de este.
Acontecimiento: almacena si el alumno recientemente está pasando por una pérdida familiar o si tiene un familiar enfermo.
Endeudamiento: almacena si el alumno tiene deudas externas a LANIA.
Examen aprobado: almacena el total de exámenes que el alumno aprobó.
Tipo de residencia: almacena si el lugar donde vive es prestada, rentada o propia.

En la preparación de los datos que es la Etapa 3; se seleccionaron las variables con las que se trabajó las cuales son veintiséis; dichas variables se pueden observar en la Tabla 6, ya que son las primeras veinticinco variables y se agregó una nueva variable que no se encuentra en la tabla anterior que es la Matrícula del alumno. La preparación de los datos de estas variables se llevó a cabo en archivos de Excel, ya que las fuentes de información primarias fueron otros archivos de Excel y documentos en físico de los alumnos como por ejemplo actas de nacimiento y certificado de licenciatura.

En la Etapa 4; para el modelado del almacén de datos se utilizó el esquema copo de nieve, porque tiene la ventaja de representar las dimensiones de manera normalizadas (Tamayo & Moreno, 2006); el cual permite a partir de un esquema en estrella expandir la jerarquía de cada dimensión del almacén de datos (Hernández del Razo, 2009). En la Figura 2, se muestra el diseño que se realizó para el almacén, en ella se puede observar la tabla central o la tabla de hecho, sus seis dimensiones principales y por último una dimensión secundaria que es una dimensión normalizada, de la dimensión *dPersonales.*

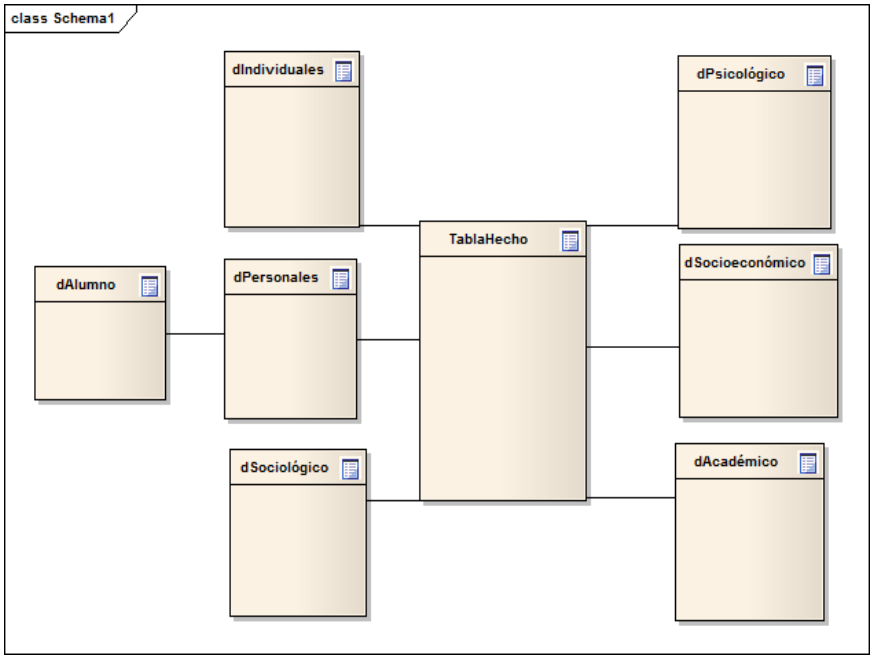


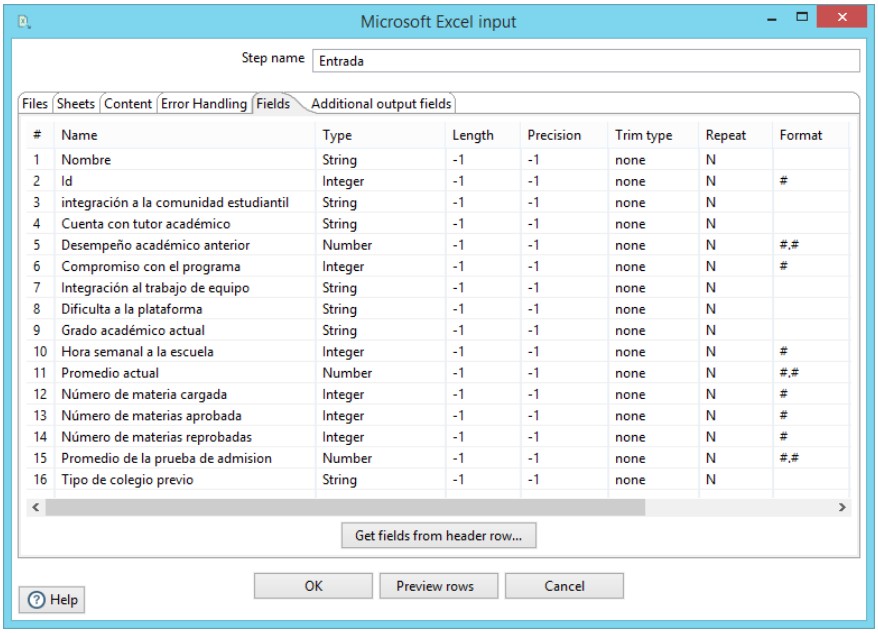
Figura 2. Modelo del almacén de datos.
Fuente: Elaboración propia.

Una vez que se obtuvo el diseño, se creó el script correspondiente del almacén de datos utilizando PostgreSQL, ya que es un gestor de software libre y es multiplataforma (Camps et al., 2005).

3.3 Proceso ETL.

Una vez creado el almacén de datos, se dio por concluida la aplicación de la metodología CRISP-DM y se prosiguió con la Fase 4 de la metodología general; que es el proceso ETL el cual es el encargado de extraer los datos de las fuentes de información originales, transformándola y por último realizar la carga al almacén de datos (Matamala et al., 2011); dicho proceso se realizó en la herramienta llamada Kettle que es una herramienta de la suite de Pentaho que ofrece análisis de negocios (Sekar, 2017).

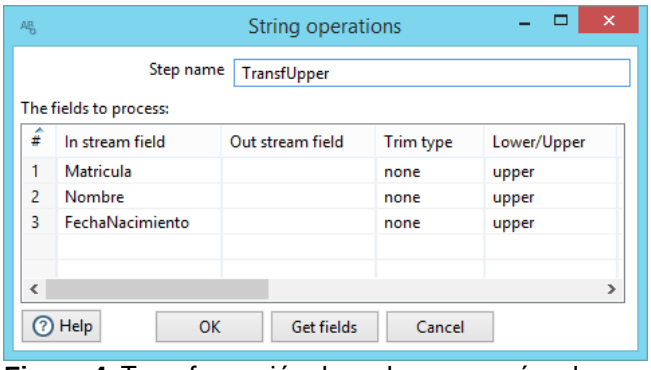
En el proceso de Extracción, fue uno de los procesos que más tiempo se le invirtió debido a que se prepararon los archivos en Excel para realizar la carga masiva de información al almacén de datos, los datos que se cargaron se tuvieron que capturar en Excel debido a que las fuentes de información originales se encontraban de manera física, es decir, en papel, por ejemplos: acta de nacimiento, copia del INE, copia de la CURP, formato de inscripción o reinscripción, entre otros, pero también se contó con archivos de Excel para obtener datos. Una vez concluido el preparado de los archivos se procedió a configurar la extracción de los datos en la herramienta de Kettle, como se muestra en la Figura 3.



#	Name	Type	Length	Precision	Trim type	Repeat	Format
1	Nombre	String	-1	-1	none	N	
2	Id	Integer	-1	-1	none	N	#
3	integración a la comunidad estudiantil	String	-1	-1	none	N	
4	Cuenta con tutor académico	String	-1	-1	none	N	
5	Desempeño académico anterior	Number	-1	-1	none	N	#.#
6	Compromiso con el programa	Integer	-1	-1	none	N	#
7	Integración al trabajo de equipo	String	-1	-1	none	N	
8	Dificulta a la plataforma	String	-1	-1	none	N	
9	Grado académico actual	String	-1	-1	none	N	
10	Hora semanal a la escuela	Integer	-1	-1	none	N	#
11	Promedio actual	Number	-1	-1	none	N	#.#
12	Número de materia cargada	Integer	-1	-1	none	N	#
13	Número de materias aprobada	Integer	-1	-1	none	N	#
14	Número de materias reprobadas	Integer	-1	-1	none	N	#
15	Promedio de la prueba de admision	Number	-1	-1	none	N	#.#
16	Tipo de colegio previo	String	-1	-1	none	N	

Figura 3. Configuración de la extracción de los datos.
Fuente: Elaboración propia.

Para el proceso de Transformación, se realizó dos tipos de transformaciones las cuales fueron: 1, reducción de cadena de la variable fecha_nacimiento y la variable fecha_fin_periodo, ya que se extrajo como tipo date, pero se necesitó guardar como tipo cadena, por ejemplo, dd/MM/yyyy:HH:MI:SS→dd/MM/yyyy; y 2, fue convertir los datos de tipo cadena a mayúsculas, ya que había datos con minúscula y mayúscula por los diferentes departamentos de donde se obtuvieron dichos datos, por ejemplo, Omar→OMAR; en la Figura 4 se muestra el proceso de la segunda transformación.



#	In stream field	Out stream field	Trim type	Lower/Upper
1	Matricula		none	upper
2	Nombre		none	upper
3	FechaNacimiento		none	upper

Figura 4. Transformación de cadena a mayúscula.
Fuente: Elaboración propia.

El último proceso en llevarse a cabo fue la carga de los datos al almacén de datos, la carga se realizó al motor de bases de datos de PostgreSQL, se configuró la herramienta de Kettle para realizar el proceso, donde se creó una conexión a la base de datos como se muestra en la Figura 5.

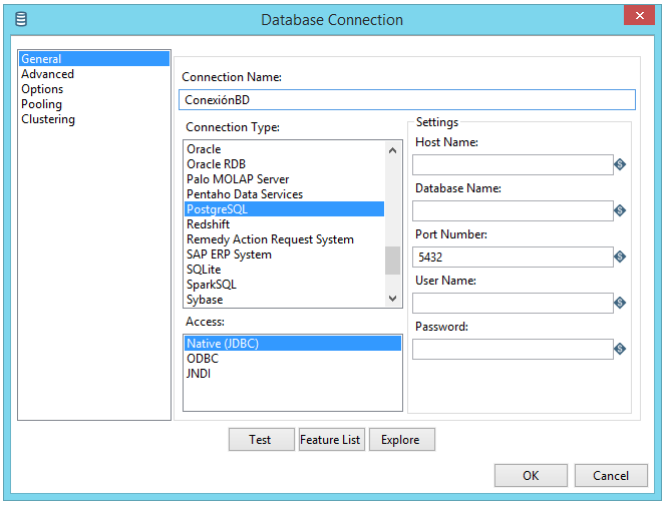


Figura 5. Creación de la conexión a PostgreSQL.
Fuente: Elaboración propia.

Para finalizar el proceso fue preciso mapear las variables que se estaban mandando con las variables contenidas en la base de datos como se muestra en la Figura 6, esto con el objetivo de que no hubiera errores al momento de ejecutar la carga y que los datos se guardaran en los campos correspondientes.

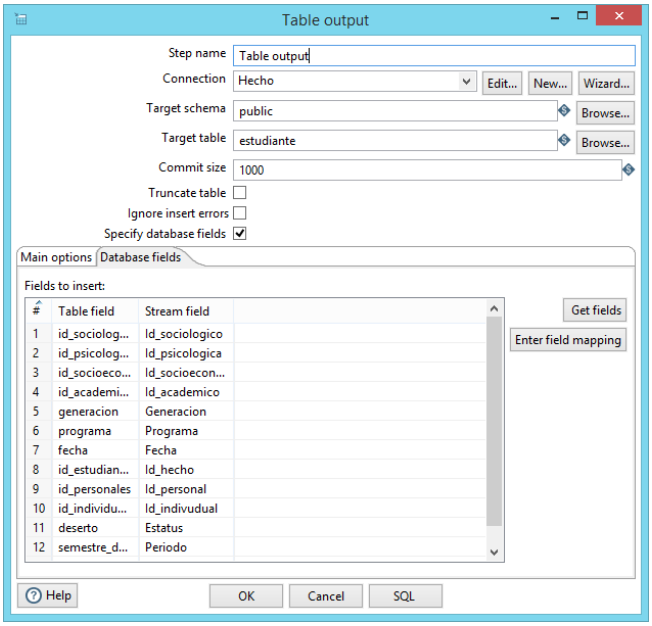


Figura 6. Mapeo de las variables.
Fuente: Elaboración propia.

3.4 Validación de los datos.

En la fase 5 se llevó a cabo una validación de la información que se cargó al almacén de datos, con la finalidad de comprobar que la información que se encontraba en el departamento de control escolar coincidiera con la información contenida en el almacén de datos. Para esta validación se realizaron dos pruebas, en la primera prueba se tomó el número de alumnos que existen por generaciones y por maestrías de control escolar y se comparó con la información que existen en el almacén de datos, los resultados de esta prueba se pueden observar en la Tabla 7, donde se puede apreciar que la generación 2010 para la maestría MRySI no coincide el número de datos, eso se debió a que un alumno realizó un cambio de la especialidad de redes a la MRySI.

Generación	Control escolar	Almacén de datos	Maestría
2010	11	12	MRySI
2011	9	9	MCA
2012	17	17	MRySI
2013	21	21	MCA
2014	19	19	MRySI
2015	17	17	MCA
2016	9	9	MRySI
2017	20	20	MCA

Tabla 7. Validación de números de alumnos.

La segunda prueba consistió en tomar una muestra de la población, ya que es rápido en obtener los resultados (Ciro, 2016), el tamaño de la muestra fue tom ar cincuenta y cuatro fechas de nacimiento totalmente al azar para realizar la comparación de los datos, donde el tamaño de la muestra salió aplicando la siguiente fórmula (Caparó, 2017):

$$Tamaño\ muestral = \frac{Z^2 * P(1-P) * N}{Z^2 * P * (1-P) + E^2 * (N-1)} \quad (1)$$

Donde:
N = tamaño de la población total.
E = margen de error.
Z = puntuación de la desviación estándar en porcentaje.
P = probabilidad de que suceda el fenómeno.
6
Valores de las variables:
N = 250
E = 10% = 0.1
Z = 1.65
P = 0.5

De las cincuenta y cuatro fechas de nacimiento que se compararon, dos de las fechas no coinciden con los datos de control escolar, variaron por un día. Se cree que eso se debe que al principio de la extracción la fecha de nacimiento se extrajo en formato POSIX y a momento de realizar la conversión al formato fecha la conversión no fue exacta.

4 Resultados.

Una vez finalizado el proceso ETL, se tiene un almacén de datos, donde cada dimensión principal y la tabla de hecho cuenta con un total de 1,114 registros y la dimensión secundaria cuenta con un total de 250 registros los cuales están comprendidos del año 2010 hasta el 2017. También se cuenta con una implementación desarrollada con el framework Yii 2.0, con el fin de visualizar la información que se encuentra en el almacén de datos. Se utilizó esta herramienta porque genera una estructura básica del código fuente lo que ahorra tiempo a la hora de programar (Enríquez, 2019). En la Figura 7 se muestra parte de la información que se encuentra en la tabla de hecho, se puede observar celdas con la leyenda (not set) eso se debe a que al momento de buscar la información en las fuentes primarias, no se encontraron datos por lo que se decidió dejarlo vacío.

Alumnos | Personales | Individuales | Académico | Psicológicas | Sociológicos | Socioeconómicos | **Estudiantes** | Graficas

Estado	Nacionalidad	Estado Civil	Edad	Sexo	Beca	Promedio Licenciatura	Grado Actual	Promedio Actual	Tipo De Escuela	Generacion	Deserto	Semestre Deserto	Programa	Fecha
VERACRUZ	MEXICANA	SOLTERO	28	MASCULINO	NO	9.0	LICENCIATURA	6.3	PÚBLICA	2013	IRREGULAR	8	MRYSI	15/08/2015
VERACRUZ	MEXICANA	SOLTERO	27	MASCULINO	NO	9.8	LICENCIATURA	9.5	PÚBLICA	2013	EGRESADO	8	MRYSI	15/08/2015
VERACRUZ	MEXICANA	CASADO	39	MASCULINO	NO	9.0	LICENCIATURA	6.8	PÚBLICA	2013	IRREGULAR	8	MRYSI	15/08/2015
VERACRUZ	MEXICANA	SOLTERO	35	MASCULINO	NO	9.7	LICENCIATURA	9.5	PÚBLICA	2013	EGRESADO	8	MRYSI	15/08/2015
VERACRUZ	MEXICANA	SOLTERO	36	FEMENINO	NO	9.2	LICENCIATURA	9.6	PÚBLICA	2013	EGRESADO	8	MRYSI	15/08/2015
VERACRUZ	MEXICANA	CASADO	28	FEMENINO	NO	9.0	LICENCIATURA	9.6	PÚBLICA	2013	EGRESADO	8	MRYSI	21/11/2005
VERACRUZ	MEXICANA	SOLTERO	(not set)	MASCULINO	NO	8.7	LICENCIATURA	6.7	PÚBLICA	2013	IRREGULAR	8	MRYSI	21/11/2005
CHIAPAS	MEXICANA	SOLTERO	25	MASCULINO	(not set)	8.8	LICENCIATURA	9.3	PÚBLICA	2012	REGULAR	2	MCA	26/04/2013
VERACRUZ	MEXICANA	SOLTERO	24	FEMENINO	(not set)	8.1	LICENCIATURA	8.3	PÚBLICA	2012	REGULAR	2	MCA	26/04/2013
VERACRUZ	MEXICANA	SOLTERO	23	MASCULINO	(not set)	9.2	LICENCIATURA	9.1	PÚBLICA	2012	REGULAR	2	MCA	26/04/2013
VERACRUZ	MEXICANA	SOLTERO	25	MASCULINO	SI	8.5	LICENCIATURA	8.1	PÚBLICA	2013	REGULAR	2	MCA	25/04/2014
VERACRUZ	MEXICANA	SOLTERO	(not set)	FEMENINO	NO	9.2	LICENCIATURA	9.7	PRIVADA	2013	EGRESADO	8	MRYSI	21/11/2005
VERACRUZ	MEXICANA	CASADO	42	MASCULINO	NO	9.2	LICENCIATURA	9.7	PÚBLICA	2014	EGRESADO	8	MRYSI	

Figura 7. Vista de la tabla de hecho con Yii 2.0.
Fuente: Elaboración propia.

Por último, se realizó estadística descriptiva el cual permitió visualizar la información del almacén de datos de manera de gráficos donde se utilizó la herramienta de Highcharts, ya que tiene una gran cantidad de gráficas, tiene una velocidad de respuesta muy rápida y por último se le puede agregar contenido JavaScript (Wang & Wang, 2015).

Una de las gráficas realizada fue para observar la comparación de los alumnos que se dieron de alta en las diferentes generaciones y el número de alumnos que se dieron de baja de su respectiva generación. En la Figura 8 se puede observar la gráfica que le corresponde a la MCA donde se puede ver que en la generación 2015 fue la generación con un mayor número de alumno dado de baja, también se puede observar un total de 28 alumnos que se dieron de baja, lo que representa un 21.53% de las 8 generaciones.

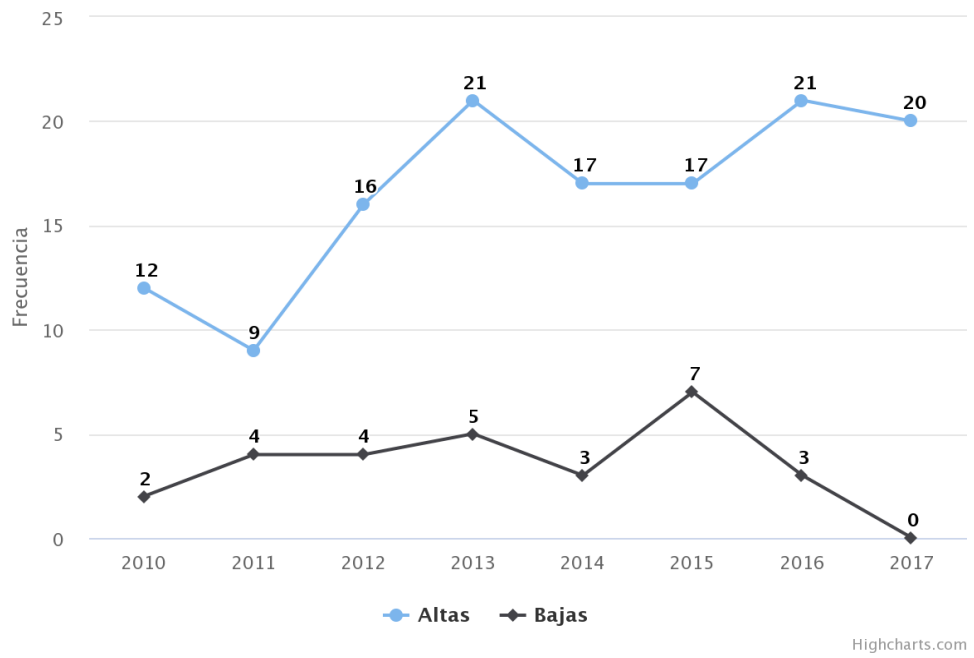


Figura 8. Altas y bajas de la MCA.
Fuente: Elaboración propia.

En la Figura 9 se muestra la gráfica que le corresponde a la MRySI donde se observa que la generación 2014 tiene el mayor número de alumno dado de baja, también se puede observar un total de 46 alumnos que se dieron de baja lo que representa el 39.65% de las 8 generaciones.

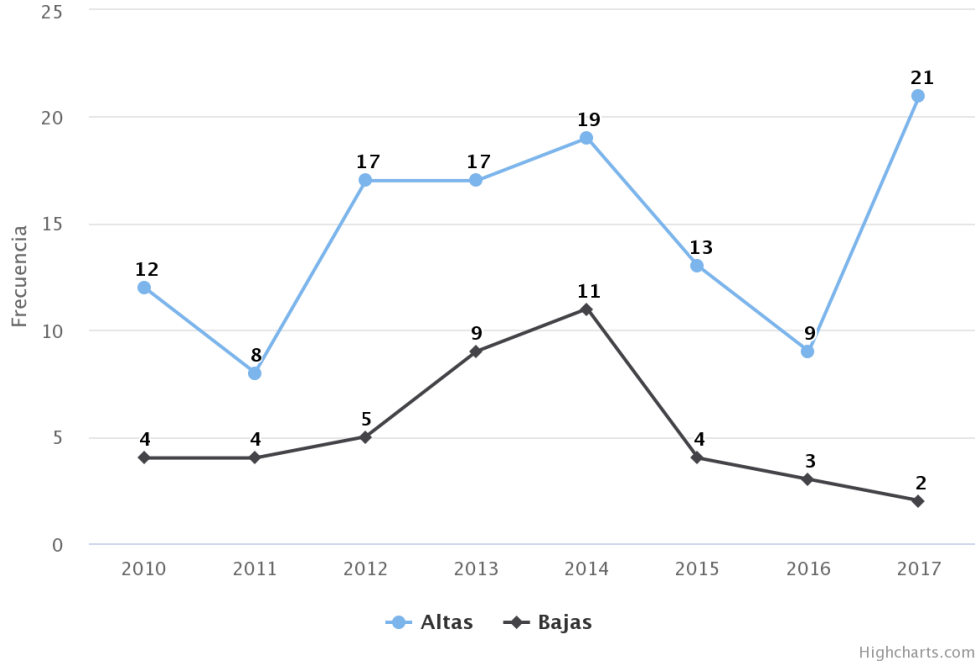


Figura 9. Altas y bajas de la MRySI.
Fuente: Elaboración propia.

Para las gráficas anteriores la línea azul representa el número de alumnos que se dieron de alta y la línea negra representa el número de alumnos que se dieron de baja, el eje de la X representa la generación; por ejemplo, el 2010 hace referencia a la Generación 2010, es decir, los alumnos que ingresaron en ese año o en su caso los alumnos que se dieron de baja.

En la Figura 10 se muestra una gráfica de caja y bigote, donde se puede observar la distribución de los datos de las bajas con referencia a la mediana, también se puede observar que la MRySI ha tenido un mayor número de deserción en una de sus generaciones eso se observa a través de su bigote máximo, para ambas maestrías la mediana es igual, ya que en ambas maestrías el número de baja en una generación fue cuatro y eso se puede observar a través de la línea que divide la caja y por último también para ambas maestrías el número mínimo de deserción por generación es dos y se puede observar en el bigote mínimo. Para la elaboración de la gráfica de la MCA no se incluyó la generación 2017.

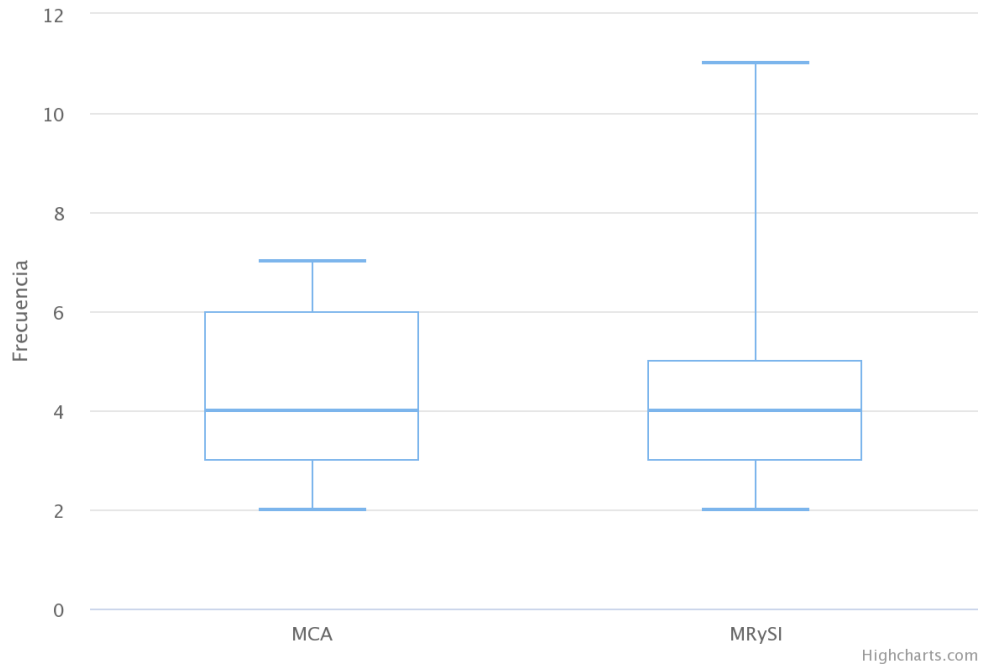


Figura 10. Distribución de baja de las maestrías.
Fuente: Elaboración propia.

También se realizaron gráficas de intervalos que muestran las edades de los alumnos al momento de solicitar la baja de las diferentes maestrías, en la Figura 11 se muestra la gráfica correspondiente a la MCA donde se observa que los alumnos entre 23 a 25 años son los que más solicitan su baja y representa el 57.14% del total de los alumnos que ha desertado. En la Figura 12 la gráfica correspondiente a la MRySI donde se observa que los alumnos entre 27 a 31 años son los que más solicitan su baja y representa el 40.47% del total de los alumnos que ha desertado. Existen reglas empíricas que se utilizan para calcular el número de intervalos o clases, la más empleada es la regla de Sturges (Pereyra, 2021). La regla de Sturges (k) es:

$$K = 1 + 3.3220 * Log(N) \quad (2)$$

Donde:
N = total de frecuencia
K = número de clases
Log N= logaritmo decimal o base 10 de N

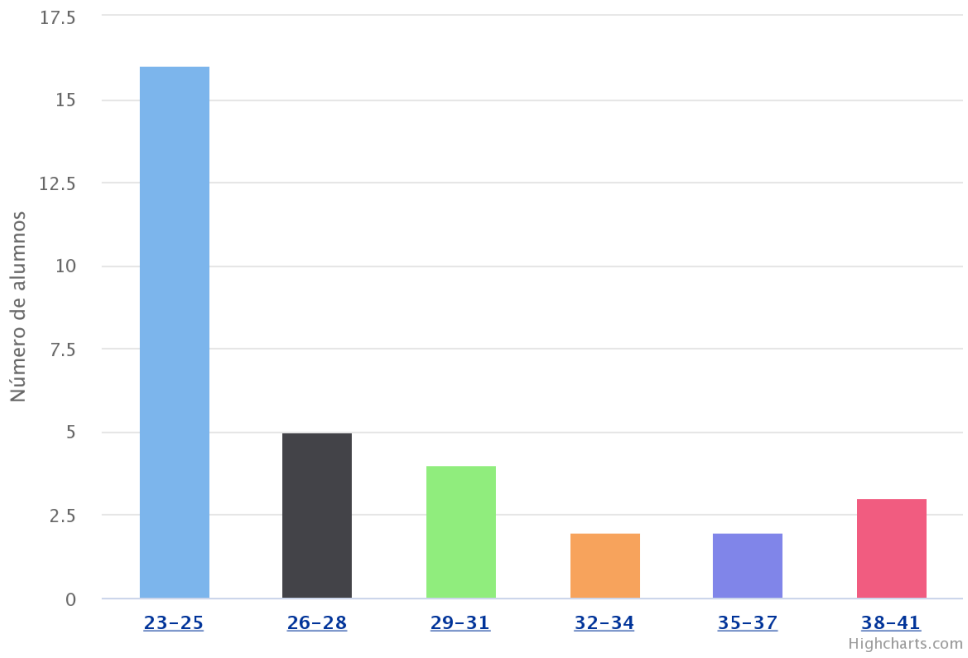


Figura 11. Edad de los alumnos al momento de desertar de la MCA.
Fuente: Elaboración propia.

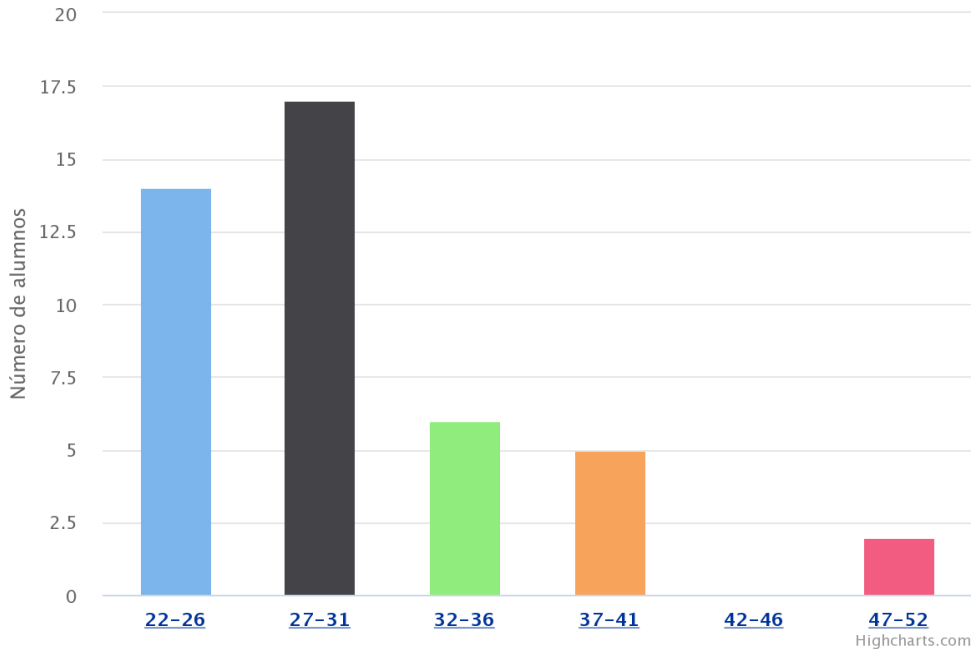


Figura 12. Edad de los alumnos al momento de desertar de la MRySI.
Fuente: Elaboración propia.

5 Conclusiones.

El problema de la deserción académica, como se puede evidenciar en el presente trabajo, no es solo una problemática propia de México, si no que podemos encontrarlo en otros países, quienes ya se encuentran trabajando en el tema a fin de frenar la deserción académica en los diferentes niveles educativos con ayuda de las herramientas tecnológicas. Son muchos los factores que lleva a que el alumno tome la decisión de desertar del programa académico y uno de los factores detectados en el presente trabajo y del que es, el más reiterativo, versa sobre la situación económica del alumno

Haciendo uso de las herramientas tecnológicas, se desarrolló un almacén de datos optando por el modelado copo de nieve, se decidió hacer uso de este modelo porque permite la normalización de dimensiones. Al final se obtuvo un almacén de datos con una tabla de hecho, seis dimensiones primarias y una dimensión secundaria que salió de la normalización de una dimensión primaria. Con la implementación del almacén de datos se propone tener las pautas a la hora de tomar decisiones con ayuda de la creación de reportes gráficos, para visualizar el progreso de los alumnos, también se tiene una base sólida con información histórica para la creación de un sistema de predicción de deserción académica dentro del Centro de Enseñanza LANIA.

Por otro lado, se realizó estadística descriptiva con la información que contiene el almacén de datos donde se tuvo como resultado diferentes reportes gráficos, como reporte de frecuencia que muestra las edades de los alumnos al momento de desertar de las maestrías, otro de los reportes fue saber la distribución de las bajas respecto a la mediana en las generaciones, teniendo como resultado una mediana de cuatro bajas en una de sus generaciones para ambas maestrías.

La carga del almacén de datos se llevó a cabo gracias al proceso ETL, ya que permitió realizar carga masiva, pero también fue la actividad donde se invirtió más tiempo en el proyecto al momento de extraer los datos ya que se tuvieron que construir los archivos con los datos a extraer.

Por último, el optar por una metodología como es CRISP-MD en la construcción de este proyecto proporcionó mucha ayuda para llevar el control de las actividades ya que facilitó la planificación durante todo el proyecto y tener un proyecto estandarizado, de tal manera que facilita llevar a cabo las últimas dos fases que no se utilizaron qué es la Evaluación y el Despliegue por otra persona.

Referencias.

Agudelo, N. E. M. and Angulo, P. J. R. (2015). Motivos de deserción estudiantil en programas virtuales de posgrado: revisión de caso y consideraciones desde el mercadeo educativo y el mercadeo relacional para los programas de retención. Revista de Educación a Distancia, (45).

Ander-Egg, E. (1980). *Técnicas de investigación social* (Vol. 14). Buenos Aires: El Cid Editor.

Azoumana, K. (2013). Analysis of student desertion at Universidad Simon Bolívar, faculty of systems engineering, with data mining techniques. REVISTA PENSAMIENTO AMERICANO, 6(10):41-51.

Barrientos, Z. and Umaña, R. (2010). Deserción estudiantil en posgrados semi-presenciales de la universidad estatal a distancia (UNED), costa rica: ¿deserción o retraso? UNED Research Journal/Cuadernos de Investigación UNED, 1(2).

Camps, R., Casillas, L., Costal, D., Gibert, M., Martín, C., & Pérez, O. (2005). Bases de datos. *España. Editorial Eureka Media SL*.

Carrillo, J. A. B., Villalobos, R. S., Madera, C. G. A., and Ramos, A. (2016). Estimación de porosidad en areniscas a partir de micrografía digitales utilizando r-studio.

Caparó, E. V. (2017). El tamaño muestral para la tesis.¿ cuántas personas debo encuestar?. *Odontología Activa Revista Científica*, *2*(1), 59-62.

Cerón, M. C. and Cerâon, M. C. (2006). Metodología de la investigación social. LOM ediciones.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000).Crisp-dm 1.0 step-by-step data mining guide.

Ciro, M. B. (2016). *Estadística básica aplicada*. Ecoe Ediciones.

Eckert, K. B. and Suénaga, R. (2015). Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. Formación universitaria, 8(5):03-12.

Enríquez Tarapues, L. G. (2018). *Sistema web utilizando framework YII 2.0 para la evaluación de desempeño docente en la unidad educativa "La Paz"* (Bachelor's thesis).

Espino Timón, C. (2017). Análisis predictivo: Técnicas y modelos utilizados y aplicaciones del mismo-herramientas open source que permiten su uso.

Galán Cortina, V. (2016). Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario. B.S. thesis.

Galindo, A. and García, H. (2010). Minería de datos en la educación. Universidad Carlos III, pages 1-8.

Hernández del Razo, J. A. (2009). Optimización de la ejecución de escenarios ETL para almacén de datos. M.S. thesis, LANIA.

Matamala, C. Z., Díaz, D. R., Cuello, K. C., and Leiva, G. A. (2011). Análisis de rendimiento académico estudiantil usando almacén de datos y redes neuronales/analysis of students' academic performance using almacén de datos and neural networks. Ingeniare: Revista Chilena de Ingeniería, 19(3):369.

Moreno Serrano, A. K. (2017). A2: MP prototipo del módulo de predicción del asesor de asesores.

M.S. thesis, LANIA.

Parra, O. J. S., Galeano, R. M., and Rodríguez, L. G. (2010). Metodología crisp para la implementación almacén de datos. Tecnura, 14(26):35-48.

Pereira, R. T., Romero, A. C., and Toledo, J. J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. Revista vínculos, 10(1):373-383.

Pereira, R. T. and Toledo, J. J. (2014). Detección de patrones de deserción estudiantil en programas de pregrado de instituciones de educación superior con crisp-dm. Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación.

Pereyra, L. E. (Ed.). (2021). *Probabilidad y estadística*. Klik.

Romo, O. K. H., Mora, R. P., and Estévez, G. G. (2015). La deserción en los posgrados, un problema no menor. Diálogos sobre educación, (8).

Sekar, V. (2017). *Restructuring the Internals of kettle/Pentaho platform* (Doctoral dissertation).

Tamayo, M. and Moreno, F. J. (2006). Comparing the molap the rolap storage models. Ingeniería e Investigación, 26(3):135-142.

Wang, Y., & Wang, J. (2015, April). Application of Highcharts in the Analysis of the Multi-source Track Inspection Data. In *2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-15)*. Atlantis Press.

Esta obra está bajo una licencia de Creative Commons
Atribución-NoComercial-CompartirIgual 4.0 Internacional.