UNIVERSIDAD TÉCNICA DEL NORTE

Facultad de Ingeniería en Ciencias Aplicadas

Carrera de Ingeniería en Sistemas Computacionales

DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE.

Trabajo de grado previo a la obtención del título de Ingeniero en Sistemas Computacionales

Autor:

Dayana Patricia Vila Espinosa

Tutor:

PhD. Iván Danilo García Santillán

Ibarra – Ecuador

Abril 2019

UNIVERSIDAD TÉCNICA DEL NORTE
BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1004028245		
APELLIDOS Y NOMBRES:	VILA ESPINOSA DAYANA PATRICIA		
DIRECCIÓN:	ZUMBA 17-80 E ISLA FERNANDINA		
EMAIL:	dpvila@utn.edu.ec		
TELÉFONO FIJO:	2-950-703	**TELÉFONO MÓVIL:**	0987952630

DATOS DE LA OBRA	
TÍTULO:	DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE
AUTOR (ES):	DAYANA PATRICIA VILA ESPINOSA
FECHA: DD/MM/AAAA	11/04/2019
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	[X] **PREGRADO** [ ] **POSGRADO**
TITULO POR EL QUE OPTA:	INGENIERA EN SISTEMAS COMPUTACIONALES
ASESOR /DIRECTOR:	PhD. IVÁN GARCÍA

2. CONSTANCIAS

El autor (es) manifiesta (n) que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto la obra es original y que es (son) el (los) titular (es) de los derechos patrimoniales, por lo que asume (n) la responsabilidad sobre el contenido de la misma y saldrá (n) en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 11 días del mes de Abril de 2019

EL AUTOR:

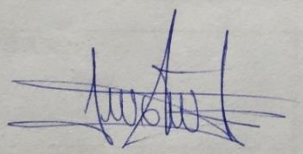
(Firma)..... Patricia Vila
Nombre: Dayana Patricia Vila Espinosa

II

CERTIFICADO TUTOR

En mi calidad de tutor del Trabajo de Grado presentado por la egresada **DAYANA PATRICIA VILA ESPINOSA** para optar por el Título de Ingeniera en Sistemas Computacionales cuyo tema es: **DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE.** Considero que el presente trabajo reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte del tribunal examinador que se designe.

En la ciudad de Ibarra, a los 11 días del mes de Abril del 2019.



PhD. Iván García Santillán
TUTOR TRABAJO DE GRADO

III

UNIVERSIDAD TECNICA DEL NORTE

Universidad Acreditada resolución 002-CONEA-2010-129-DC
Resolución No. 001-073-CEAACES-2013-13

DIRECCION DE DESARROLLO TECNOLOGICO E INFORMATICO

DIRECTOR DE LA DIRECCIÓN DE DESARROLLO TECNOLÓGICO E INFORMÁTICO

CERTIFICA

QUE: La señorita DAYANA PATRICIA VILA ESPINOSA con cédula identidad 1004028245 estudiante de la Facultad de Ingeniería en Ciencias Aplicadas – de la Carrera de Ingeniería en Sistemas Computacionales, ha desarrollado con los datos entregados de la Dirección de Desarrollo Tecnológico e Informático, el Proyecto de Tesis **"DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE".**

QUE: El estudio del proyecto fue entregado a la Dirección de Desarrollo Tecnológico e Informático el 10 de abril del 2019.

Es todo cuanto puedo certificar, facultando a la interesada hacer uso de este certificado como estime conveniente, excepto para trámites judiciales.

.

Ibarra, 10 de abril del 2019

Atentamente
CIENCIA Y TECNICA AL SERVICIO DEL PUEBLO

Ing. Juan Carlos García
DIRECTOR

Av. 17 de Julio 5 – 21 y José María Córdova
Ciudadela Universitaria Barrio El Olivo
Teléfono: (06) 2997800 ext. 7040 Casilla 199
www.utn.edu.ec
Ibarra - Ecuador

AUTORÍA

Yo, **DAYANA PATRICIA VILA ESPINOSA**, portadora de la cédula de ciudadanía número **100402824-5**, declaro bajo juramento que el trabajo aquí descrito es de mi autoría, **DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE** que no ha sido previamente presentada para ningún grado, ni calificación profesional, y que se han respetado las diferentes fuentes y referencias.

Patricia Vila

DAYANA PATRICIA VILA ESPINOSA

C.I.**100402824-5**

V

Dedicatoria

Por haberme dado la vida, amor, esfuerzo, constancia, paciencia y por todo lo que hoy soy, este proyecto de titulación va dedicado a mi madre.

Agradecimientos

A mi familia, por haberme brindado el apoyo y motivación incondicional durante mi carrera universitaria y creer en mi en todo momento.

A todos los profesionales que han formado parte del proceso de investigación y análisis del presente proyecto, a quienes les debo gran parte de mis conocimientos, de manera muy especial a mi tutor de tesis, por haberme guiado no solo en el desarrollo del presente trabajo de titulación sino a lo largo de mi formación profesional, a mis compañeros de carrera con quienes hemos atravesado este gran reto, por su complicidad y amistad sincera, y finalmente un eterno agradecimiento a esta prestigiosa universidad, por haber abierto sus puertas a jóvenes ávidos de saber, prepararnos para un futuro profesional competitivo y formarnos como personas de bien.

Tabla de Contenido

Índice de Figuras

Índice de Cuadros

Resumen

La deserción estudiantil constituye un problema que afecta a las instituciones de educación superior y por ende a sus estándares de calidad; las causas probables que ocasionan esta problemática pueden ser personales, académicas o su situación socioeconómica. Esta investigación tiene como objetivo principal investigar patrones de deserción estudiantil y los principales factores que contribuyen a esta problemática en la Universidad Técnica del Norte (Ecuador), aplicando técnicas predictivas de minería de datos (clasificación y regresión), para procesar datos históricos de los estudiantes desde del año 2017 a 2018. El proceso KDD (Proceso de descubrimiento de conocimiento en bases de datos) sirvió para obtener una vista minable con 11200 registros, para aplicar técnicas bayesianas, árboles de decisión y regresión en el software Weka. Para definir el mejor algoritmo se evaluaron cuantitativamente cada uno de ellos, mediante la matriz de confusión y medidas estadísticas. Los principales resultados demostraron que los mejores algoritmos fueron RandomTree y Logistic, para obtener el conocimiento se tomó en cuenta la intersección de los resultados obtenidos de ambos algoritmos.

Palabras claves: deserción estudiantil, descubrimiento de patrones, minería de datos, técnicas predictivas

Abstract

Student desertion is a problem that affects higher education institutions and therefore their quality standards; The probable causes that cause this problem can be personal, academic or socioeconomic situation. The main objective of this research is to investigate patterns of student desertion and the main factors that contribute to this problem in the Universidad Técnica del Norte (Ecuador), applying predictive techniques of data mining (classification and regression), to process historical data of students from 2017 to 2018. The KDD process (knowledge discovery process in databases) was used to obtain a viewable minable with 11200 records, to apply Bayesian techniques, decision trees and regression in the Weka software. To define the best algorithm, each of them was evaluated quantitatively, using the confusion matrix and statistical measures. The main results showed that the best algorithms were RandomTree and Logistic, to obtain the knowledge the intersection of the results obtained from both algorithms was considered.

Keywords: student desertion, pattern discovery, data mining, predictive techniques

Introducción

Antecedentes

La deserción estudiantil constituye, por su magnitud, un problema importante en los sistemas educativos en toda Latinoamérica. Las altas tasas de abandono de los estudios que se producen en todos los niveles educativos afectan negativamente los procesos económicos, sociales y culturales en el desarrollo de los países. Por lo anterior, naciones como Costa Rica, Argentina, Colombia, entre otros, han comenzado a diseñar profundos procesos de mejoramiento para aumentar la retención en los primeros años de estudios(UNESCO, 2004).

Nuevos datos revelados por el Instituto de Estadística de la UNESCO (IEU) revelan que, en 2010, aproximadamente 32,2 millones de estudiantes de educación primaria repitieron el grado en el que se encontraban y 31,2 millones abandonaron la escuela y, probablemente, nunca más regresen a las aulas (Instituto de Estadística de la UNESCO, 2012). La última edición del Compendio Mundial de Educación destaca la urgente necesidad de abordar el problema que representa el alto número de niños y niñas que repiten grados y dejan la escuela antes de concluir la educación primaria o el primer ciclo de secundaria. El informe refleja que baja la cantidad de alumnos que repiten curso, pero no la de quienes la abandonan.

Actualmente en el Ecuador ocho de cada diez estudiantes que ingresaron a una universidad pública en el año 2012 continuaron con sus estudios en el año 2013, y siete de cada diez continuaron en el 2014 (SENESCYT, 2015).

Existen cuatro tipos de deserción estudiantil (Bazantes et al., 2017): (i) deserción o mortalidad estudiantil absoluta, la que corresponde a retiros del estudiante por motivos académicos o de otra índole; (ii) deserción o mortalidad estudiantil relativa, referida a la proporción entre los estudiantes que se retiran y el total de matriculados; (iii) la deserción académica absoluta, que sería el número de estudiantes que no aprueban el semestre académico siguiente en el cual están matriculados, porque se retiraron de la universidad o perdieron cursos y no alcanzaron el total de puntos requeridos para avanzar al siguiente semestre, y por último, (iv) la deserción académica relativa, que viene a ser la relación entre el número de estudiantes que no pasan al semestre académico siguiente, respecto del total de matriculados en cualquier semestre académico.

Algunos trabajos relevantes que abordan la deserción estudiantil son los siguientes:

- (Lehr et al., 2016), utilizaron técnicas de minería de datos en la Universidad Aeronáutica (ERAU), en específico la tarea de predicción, para prever la retención de los estudiantes universitarios. Para ello, han empleado la clasificación de los datos de 972 estudiantes matriculados en ERAU en el 2008, los cuales fueron objetos de entrenamiento para los modelos predictivos, tomando en cuenta la preparación previa de los estudiantes, el rendimiento de estos en el primer año, datos personales y financieros. Las tareas de clasificación que emplearon son: Regresión Logística, Naive Bayes, K Vecinos Más Próximos, Random Forest, Perceptrón Multicapa y Árboles de Decisión.
- (Gonzalez et al., 2016), desarrollaron un sistema predictivo que permita detectar al alumno con alta probabilidad de deserción y más aún proporcionar el perfil de los alumnos desertores, por medio de técnicas de minería de datos tales como regresión logística, clustering, árboles de decisión y redes neuronales, tomando en cuenta la información almacenada en el sistema escolar del instituto, únicamente del programa educativo de Ingeniería en Tecnologías de la Información un Comunicaciones, los resultados que se obtuvieron mediante la regresión logística son los que mejor se acoplaron al estudio realizado. Los resultados obtenidos fueron comparados con los resultados reales que tiene el sistema de información del instituto.
- (Sadiq et al., 2018) recopilaron información socioeconómica, demográfica y académica de 300 estudiantes de tres instituciones de Assam, India. Formaron una vista minable con 24 atributos, para posteriormente aplicar 4 algoritmos de clasificación, J48, PART, Random Forest y Bayes Network con la herramienta de minería de datos Weka (Waikato, 2018). Los resultados de evaluación de los clasificadores mostraron que el algoritmo Random Forest supera al resto de clasificadores en función de la tase de error y la precisión. Adicionalmente emplearon el algoritmo descriptivo A priori para encontrar reglas de asociación que también dio buenos resultados.
- (Hernández et al., 2018) realizaron un análisis sobre la deserción estudiantil en la Universidad Nacional de Colombia con sede en Manizales para identificar las características o factores más relevantes que influyen a esta problemática. Emplearon los datos del periodo de tiempo de 2009 a 2014 para ejecutar el proceso de descubrimiento del conocimiento en bases de datos, y para obtener la información emplearon las técnicas de árboles de decisión y reglas de inducción, los cuales fueron evaluados para determinar con qué modelo se obtienen los mejores resultados.

- (Noboa et al., 2018) argumentan que la detección temprana de abandono escolar es importante para mejorar los índices de graduación y calidad de la educación. Por estos motivos mediante el uso de técnicas básicas de aprendizaje supervisado, predijeron el abandono escolar en la Escuela Superior Politécnica del Litoral en Guayaquil, Ecuador con una muestra total de 4294 estudiantes, por medio de árboles de decisión y regresión logística, obtuvieron resultados tales como que casi el 22% de los estudiantes son potenciales desertores y por otro lado, llegaron a la conclusión de que en caso de aprobar más de 12 asignaturas y tener acceso a la biblioteca puede asegurar la estancia de los estudiantes en la universidad.
- (Vila et al., 2018) argumentan que la deserción estudiantil es un fenómeno que afecta a los estándares de calidad de la educación superior, su objetivo principal fue detectar patrones de deserción estudiantil basándose en información personal y académica de los estudiantes de la Universidad Técnica del Norte por medio de técnicas predictivas de minería de datos, bayesianas y árboles de decisión, realizaron una comparativa y la técnica que dio mejores resultados fue árboles de decisión.
- (Palacios-Pacheco et al., 2018) manifiestan que la minería de datos es uno de los problemas más abordados en las universidades de Latinoamérica, varios países han optado por elaborar estrategias de retención de estudiantes de los niveles bajos, por este motivo realizaron una minería de datos con la finalidad de descubrir patrones en los alumnos para dar respuesta al efecto que se presenta. Para ello emplearon la herramienta Weka para procesar la información, mientras que los resultados hasta la fecha de presentación de este trabajo se encuentran en etapa de validación.
- (Espinoza & Gallegos, 2019) manifiestan que en la actualidad las actividades comerciales de servicios y productos han ido acumulando un sinnúmero de información a lo largo del tiempo, por esta razón argumentan la importancia de un profesional para analizar dicha información y obtener beneficios de ella. Presentan una revisión de la literatura para identificar el perfil del profesional llamado Data Scientist, ya que no existen perfiles concretos, llegando a la conclusión de que las principales características de este profesional es estar instruido en materias como las finanzas, inversión, producción, recursos humanos, entre otros, de igual forma menciona que los rasgos personales innatos deben ser el liderazgo y autoeducación para alcanzar sus objetivos.

Problema

El abandono estudiantil en la educación superior es un fenómeno que se ha venido presentando a lo largo del tiempo, causando una problemática social que afecta a la economía nacional. Dicho fenómeno se produce por diferentes causas que pueden ser académicas, socioeconómicas, psicológicas o psíquicas (Gonzalez et al., 2016). Por ello, atacar este suceso a tiempo es indispensable. En la UTN, se ha observado que no existe un método óptimo que permita evidenciar o registrar la permanencia de los estudiantes en el nivel de grado, ya que el seguimiento académico a posibles desertores se lo realiza cuando ya están a punto de abandonar sus carreras universitarias, y es un proceso manual; de igual forma no cuenta con personal administrativo capacitado para analizar los datos almacenados en el sistema académico, provocando de alguna manera la pobreza en el país y baja calidad en la casona universitaria lo cual se refleja en el presupuesto asignado a la institución por parte de los organismos competentes.

Objetivos

Objetivo General

Detectar patrones de deserción estudiantil utilizando técnicas predictivas de clasificación y regresión en minería de datos para la gestión académica de la Universidad Técnica del Norte.

Objetivos Específicos

- Elaborar un marco teórico que sustente las técnicas predictivas de minería de datos y el proceso de descubrimiento de conocimiento en base de datos (KDD).
- Construir un data warehouse a partir de datos académicos y socio económicos de los estudiantes de pregrado de los últimos 5 años de la UTN utilizando la Suite de Pentaho
- Aplicar técnicas predictivas de clasificación y regresión a la vista minable utilizando el software Weka.
- Obtener patrones de deserción estudiantil utilizando los mejores modelos de clasificación y regresión basado en métricas de calidad y la característica de Consistencia de la ISO/IEC 25012.

Justificación

El desarrollo de un estudio que permita predecir los estudiantes que tienen alta posibilidad de abandonar sus estudios en la Universidad Técnica del Norte, permitirá agilizar de manera significativa la detección de este grupo vulnerable y prever la retención de los estudiantes universitarios en el Departamento de Bienestar Universitario, y así atacar de forma temprana esta problemática, que contribuye al aumento del desempleo y pobreza del país(Gonzalez et al., 2016).

En la actualidad existe una gran variedad de técnicas empleadas en la minería de datos en el ámbito educativo, por este motivo en el presente proyecto se plantea el uso de técnicas predictivas, con el objetivo de emplear la información existente en las plataformas educativas para comprender las características de los desertores académicos (Bazantes et al., 2017), con el fin de elevar los estándares de categorización de la casona universitaria y brindar una educación de calidad.

Los beneficiarios directos del presente trabajo serían los estudiantes, ya que con la información que se obtendrá se podrá tomar decisiones ajustadas a la realidad y poner en marcha planes de contingencia para atacar desde la raíz este problema, de igual forma entre los principales beneficiarios se encuentra la UTN ya que mejoraría significativamente sus estándares de calidad, ya que en la evaluación del CEAACES se considera la "eficiencia académica" donde lo que cuenta es la tasa de retención de estudiantes y el número de graduados (Roig, 2014).

Alcance

Mediante la presente investigación se obtendrán patrones de deserción estudiantil que permitirán identificar a tiempo, de forma eficiente y eficaz los candidatos a abandonar su carrera de pregrado. Motivo por el cual el estudio consiste en realizar un análisis de los datos académicos almacenados en los repositorios de las bases de datos Oracle del nivel de pregrado de todas las facultades de la UTN (licenciaturas e ingenierías) mediante el proceso de descubrimiento de conocimiento KDD, que consiste en realizar la extracción, transformación y limpieza de los datos con la Suite de Pentaho 7, para posteriormente almacenarlos en un data warehouse en la base de datos Oracle, que constituye la vista minable, la cual servirá para ejecutar los algoritmos predictivos de clasificación y de regresión en la herramienta Weka 9.3; para posteriormente, realizar una comparativa de los datos obtenidos mediante las métricas cuantitativas para evaluar los algoritmos de minería de datos, como la matriz de confusión, tasa de error, coeficiente Kappa, curvas ROCC, tasa de verdaderos positivos y falsos positivos, entre otros, para así deducir cuál es el que mejor se acopla al estudio y que permita tener datos de mayor calidad, de acuerdo a los datos de entrada, luego de un estudio se determinará el método estadístico que permitirá validar los resultados.

CAPÍTULO 1

Marco teórico

1.1. Introducción a la minería de datos

En la actualidad, con los avances tecnológicos la mayoría de las instituciones, públicas o privadas, cuentan con sistemas que almacenan información que es vital para el giro del negocio; por este motivo el análisis de una gran cantidad de información (datos masivos) se muestra como uno de los más grandes desafíos actuales de la informática. Frecuentemente para el análisis de grandes volúmenes de datos no se puede realizar por los métodos tradicionales, esto se da por el tamaño masivo de información (Tan et al., 2005).

1.1.1. La minería de datos

La minería es la actividad económica que comprende el proceso de extracción, explotación y aprovechamiento de minerales que se hallan en la superficie terrestre (BCE, 2018), por ejemplo, el oro o la plata. El dato por su parte es el valor que toma una variable, parámetro, atributo, característica, etc. (Lara, 2014). Entonces, la minería de datos consiste en aplicar técnicas especiales para analizar datos con el objetivo de obtener información que se encuentra oculta y que es esencial para analizar el giro/negocio de la institución a la cual pertenecen los datos.

La minería de datos hace referencia al proceso de extracción de información en forma de modelo, que se puede tomar como una característica extrema de los datos (Leskovec et al., 2014).

Hay quienes consideran que la minería de datos es sinónimo de aprendizaje automático. No hay duda de que una minería de datos utiliza apropiadamente algoritmos de aprendizaje automático. Los practicantes de aprendizaje automático usan los datos como un conjunto de entrenamiento, para entrenar un algoritmo, de los muchos existentes, como redes de Bayes, máquinas de vectores de soporte (SVM), árboles de decisión, redes neuronales artificiales (ANN), ocultos Modelos de Markov, y muchos otros (Leskovec et al., 2014, p. 2).

En la actualidad se puede aseverar que la minería de datos ha demostrado contar con una primera generación de algoritmos válidos para diferentes aplicaciones del mundo real, por ello el desarrollo de tecnología de minería de datos es un área importante de estudio por su afán de buscar sacar provecho de la información que se encuentra almacenada (Riquelme, Ruiz, & Gilbert, 2006).

La minería de datos abarca técnicas que pretenden descubrir el conocimiento que se encuentra oculto en grandes volúmenes de datos, con el objetivo de obtener patrones, perfiles o tendencias por medio del estudio de datos (Pérez & Santín, 2007).

1.1.2. Relación de la minería de datos con otras áreas

La minería de datos es una simbiosis de varias disciplinas diferentes, entre las áreas que se relacionan se encuentran:

a. Estadística

Gran parte de las técnicas que se emplean en la minería de datos son o tienen su raíz en la estadística (Lara, 2014). De una u otra forma muchos de los conceptos y técnicas que se emplean en la estadística, tales como modelos matemáticos o inferencias basadas en probabilidades, se utilizan también en minería de datos.

b. Bases de datos

Puesto que como se ha mencionado anteriormente, la información que se analiza en la minería de datos proviene de repositorios o bases de datos (por ejemplo, Oracle, SQL Server, PostgreSQL, MySQL, etc.). Estas dos disciplinas tienen una estrecha relación entre sí.

c. Visualización

La meta de la minería de datos es obtener conocimiento que sea relevante y útil, por ello es indispensable que dicho conocimiento pueda ser plasmado visualmente, ya sea en diagramas, gráficos, resúmenes, etc., para posteriormente ser analizado por el experto de cada área (Lara, 2014).

d. Aprendizaje automático

Las diferentes técnicas de reconocimiento de formas han adquirido una gran popularidad debido a su gran eficiencia y eficacia al momento de solucionar problemas de ámbitos de la vida real, por ello se relaciona estrechamente con la minería de datos, ya que solucionan problemas mediante mecanismos automáticos (Sierra, 2006).

e. Otras

De igual forma, se relaciona con varias áreas tales como sistemas de apoyo a la toma de decisiones, tratamiento y procesamiento de señales (incluyendo análisis de imágenes), así como recuperación de información.

A continuación, en la Fig. 1 se puede apreciar cómo se relaciona la minería de datos con las áreas antes mencionadas.

Fig. 1. Áreas con las que se relaciona la Minería de Datos (Lara, 2014)

1.2. Tipos de datos

Los datos son hechos que describen sucesos y entidades, pueden ser comunicados por varios tipos de símbolos, por ejemplo, letras del alfabeto, números, movimientos, señales, etc. Una característica importante de estos símbolos es que se pueden ordenar de tal forma que se convierten en información, y que por sí mismos no tienen capacidad de comunicar ningún significado y por consiguiente no pueden afectar el comportamiento de quien los observa (D'ambrosio, 2008). En la Tabla 1.1 se muestra un ejemplo de 3 datos para una ciudad de Ecuador, el dato "Ibarra" que representa al atributo "NombreCiudad", el dato "181175" representa al atributo "NúmeroHabitantes" y el dato "C" que hace referencia al atributo Capital.

TABLA 1.1
DATOS ASOCIADOS A CIUDAD

Atributo	Dato
NombreCiudad	Ibarra
NúmeroHabitantes	181175
Capital	C

Fuente: Propia

Los datos generalmente se clasifican dependiendo de la naturaleza del atributo al cual representan (Pérez & Santín, 2006), así tenemos:

1.2.1. Cuantitativos

Realizan representaciones numéricas o de magnitudes:

a. **Discretos.** Son los datos que pueden tomar un número limitado de valores diferentes. **Ej.** Número de empleados

b. **Continuos.** Son aquellos datos para los que se efectúa que, para cualquier par de valores, siempre se puede encontrar un valor intermedio.
   Ej. Sueldo, peso, altura de una persona

1.2.2. Cualitativos

Realizan representaciones de categorías:

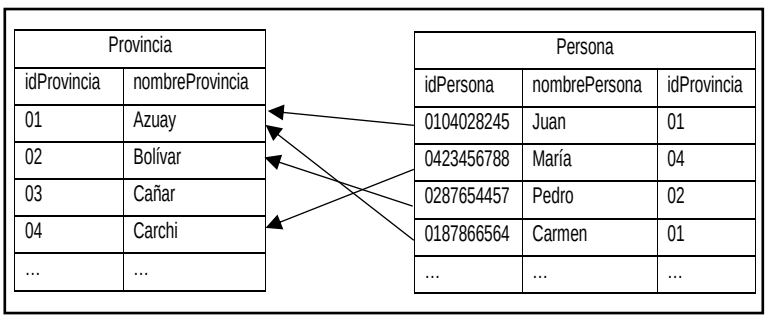
a. **Nominales.** Aquellos para los cuales existe una asignación parcial de elementos de una categoría.
   Ej. Estado civil de las personas, colores.
b. **Ordinales.** Aquellos para los que existe una relación directa con el orden entre las categorías.
   Ej. Alto, medio o bajo, escala de Likert (3, 5, 7 puntos)

1.2.3. Bases de datos

Bases de datos relacionales

Actualmente, todas las instituciones poseen repositorios de información que se denominan bases de datos. Generalmente para almacenar información del negocio o información transaccional son las bases de datos relacionales, que contienen estas bases de datos se representan en tablas (Lara, 2014).

Las tablas de una base de datos relacional se encuentran formada por un conjunto de atributos, formados por las columnas o campos, y pueden contener un sinnúmero de tuplas, que también se conocen como filas o registros. Cada registro representa un objeto, que se encuentra representado por cada valor que contiene cada atributo, de igual forma se caracteriza por tener una clave única que lo representa (Hernández et al., 2004). En la Fig. 2 se muestra un ejemplo de una base de datos relacional.



Provincia	
idProvincia	nombreProvincia
01	Azuay
02	Bolívar
03	Cañar
04	Carchi
...	...

Persona		
idPersona	nombrePersona	idProvincia
0104028245	Juan	01
0423456788	María	04
0287654457	Pedro	02
0187866564	Carmen	01
...	...	...

Fig. 2. Ejemplo de base de datos relacional

Esquemas desnormalizados

Para realizar minería de datos, no se necesita de forma indispensable contar con un esquema relacional, sino es necesario contar con información redundante o duplicada (Lara, 2014). A este tipo de esquemas se los conoce generalmente como esquemas desnormalizados, en la Fig. 3 se muestra de forma gráfica este tipo de esquema en donde la última columna contiene datos redundantes.

Persona		
idPersona	nombrePersona	nombreProvincia
0104028245	Juan	Azuay
0423456788	María	Carchi
0287654457	Pedro	Bolívar
0187866564	Carmen	Azuay
…	…	…

Fig. 3. Ejemplo de esquema desnormalizado

1.3. Proceso de descubrimiento del conocimiento (KDD)

La minería de datos es una parte importante del proceso de descubrimiento del conocimiento, KDD por sus siglas en inglés Knowledge Discovery in Databases. Este proceso integra varias etapas hasta llegar a obtener el conocimiento para la toma de decisiones. En la Fig. 4 se muestran las fases principales del proceso KDD:

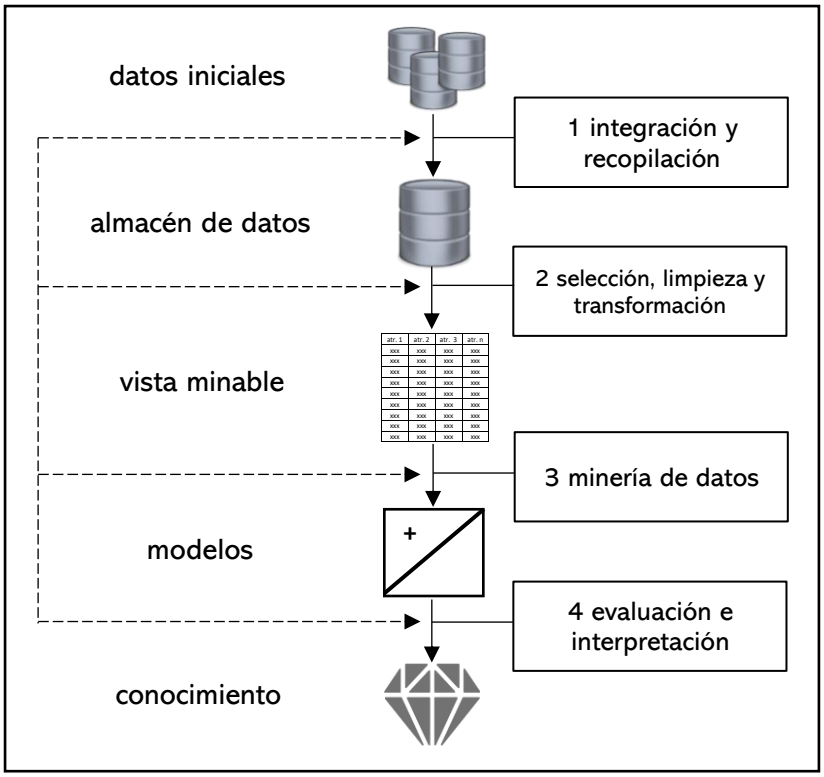


Fig. 4.Proceso KDD. (Hernández Orallo et al., 2004)

El proceso KDD inicia con la recopilación de la información de los diferentes orígenes de datos (data warehouse), estos pueden ser de bases de datos relacionales, temporales, multimedia, texto, etc., esta etapa es mejor conocida como la etapa de selección de datos y es el pilar para que el proceso completo sea exitoso (Pérez & Santín, 2007). La siguiente fase es el procesamiento de datos, donde el objetivo es seleccionar, limpiar y transformar los datos, para obtener una vista minable que permita aplicar las diferentes técnicas de minería de datos.

La etapa más importante del proceso KDD es la minería de datos, ya que, aplicando las diferentes tareas de minería de datos, predicción y clasificación, en este caso, se pueden obtener los diferentes modelos que representarán a los datos analizados, para llegar a la etapa final que consiste en evaluar e interpretar la información obtenida para obtener el conocimiento.

1.4. Etapas del proceso KDD

1.4.1. Fase de integración y recopilación

Como se mencionó anteriormente, las instituciones emplean bases de datos relacionales para efectuar sus transacciones, ya que son suficientes para cubrir sus necesidades diarias, no obstante, no son suficientes para tomar decisiones estratégicas para la organización. Sin embargo, para este tipo de situaciones se debe llevar a cabo el proceso KDD. Asimismo, puede suscitarse el caso de que los datos necesarios para llevar a cabo el análisis sean ajenos a la institución, siendo de bases de datos externas como censos, datos demográficos o climatológicos (Hernández et al., 2004).

El análisis posterior de la información se tornará más sencillo, si la fuente de datos a ser analizados es unificada, accesible y sobre todo que se encuentre separada del trabajo transaccional (Pérez & Santín, 2007). La idea de integrar múltiples bases de datos, con sus respectivos formatos, identificadores, etc., es un reto significativo que ha dado lugar a los conocidos almacenes de datos o data warehouse, que según (PowerData, 2018) son los que permiten crear un repositorio de bases de datos transaccionales provenientes de diferentes fuentes para la toma de decisiones. Para crear un almacén de datos se pueden aplicar un sinnúmero de técnicas, una de las más comunes es integrar y almacenar la información en un nuevo esquema como se indica en la Fig. 5.

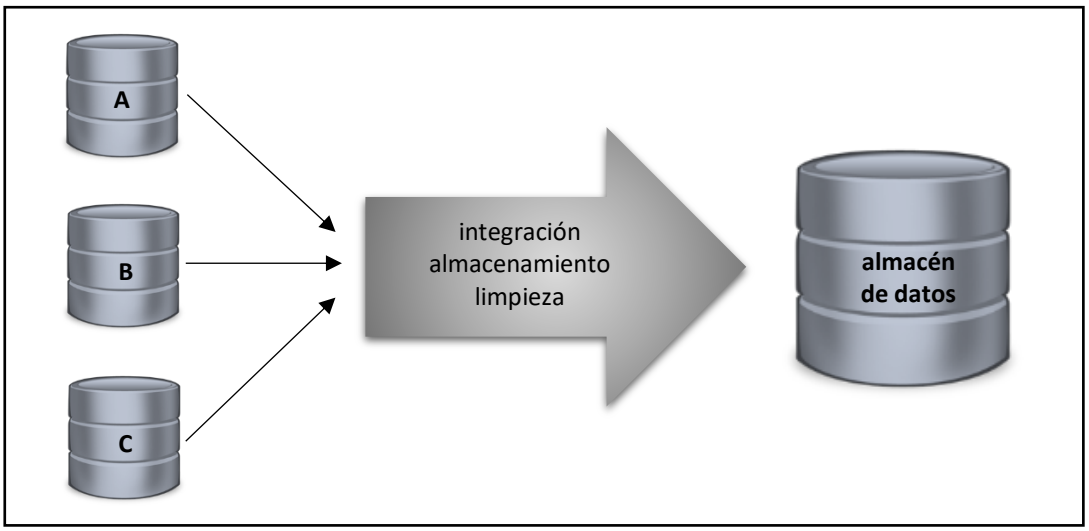


Fig. 5. Fase de integración y recopilación – Fuente: (Hernández Orallo et al., 2004)

Los almacenes de datos se emplean para integrar información de manera sofisticada, por esta razón los datos se modelan con una estructura de bases de datos multidimensional, en el cual cada dimensión corresponde a un atributo o grupo de atributos en el esquema en torno a hechos que almacenan el valor de acuerdo con una medida agregada, por ejemplo, la cantidad de estudiantes que aprobaron una materia en un año en concreto de una carrera. De esta forma se da paso a que los almacenes de datos sean adecuados para el procesamiento analítico en línea (on-line analytical processing, OLAP), que es un análisis de datos superior al clásico SQL (Structured Query Language), ya que permite presentar la información a diferentes niveles de abstracción, dependiendo de las necesidades del usuario (Hernández et al., 2004).

Para el presente trabajo de titulación, se recopilará la información proveniente de las bases de datos transaccionales en Oracle (Oracle, 2018) en las cuales la Universidad Técnica del Norte almacena su información, mientras que para integrar la información y consolidar el data warehouse se emplearán las herramientas de Pentaho Data Integration (Hitachi Vantara, 2018).

1.4.2. Fase de selección, limpieza y transformación

Para asegurar la calidad del conocimiento que se pretende obtener, no solo se deben aplicar las técnicas de minería de datos adecuadas, sino también que los datos a minar deben ser de calidad; por este motivo después de la fase de integración y recopilación de la información es indispensable seleccionar, limpiar y transformar la información, de esta forma se obtiene una vista minable (Pérez & Santín, 2006).

a. Selección

Para realizar la selección de datos se realiza el proceso de filtrado que se puede efectuar a varios niveles (Lara, 2014):

- **Filtrado de atributos**

La selección de los atributos es uno de los pasos más relevantes del proceso KDD, ya que los atributos que se van a considerar deben ser relevantes para el estudio (Hernández Orallo et al., 2004), por ejemplo para identificar los desertores estudiantiles de una universidad uno de los principales datos irrelevantes es el nombre del estudiante y su pasaporte o cédula de ciudadanía, ya que no aportan al estudio.

- **Filtrado de registros**

La selección de registros en muchas ocasiones depende de la naturaleza del problema que se pretende solucionar, por ejemplo, es posible que en el caso anterior se desee tomar en cuenta únicamente a los estudiantes que se encuentran cursando una carrera afín a la informática y por ende se eliminarían el resto de los registros.

b. Limpieza

La limpieza de datos consiste en eliminar el mayor número posible de datos erróneos o inconsistentes y ausencia de valores. En esta etapa del proceso, generalmente, se emplean herramientas de consulta de información y herramientas estadísticas (Pérez López & Santín González, 2007).

En el caso de que exista ausencia de valores, por ejemplo, puede ser que en el data warehouse exista un campo que haga referencia al lugar de nacimiento de la persona y dicho campo en un registro se encuentre vacío, es indispensable que caso de faltantes se realice un análisis minucioso, ya que puede ser que arrojen información interesante, tal como cuando una persona no quiere dar a conocer su información o se reserva el derecho a divulgarla, o dado el caso se puede pasar por alto la ausencia y continuar con el análisis o filtrar el registro (Lara, 2014).

Cuando se habla de datos inconsistentes o erróneos, se dice que pueden representar ruidos o excepciones, sin embargo, otros son muy relevantes y el resultado se puede alterar por ello, por ejemplo, se puede dar el caso en el cual el estudiante tenga como año de nacimiento 2203, lo cual sería un dato erróneo. En muchos casos no es conveniente eliminarlos, ya que en ciertos casos como detecciones de fraudes pueden ser más interesantes que los datos regulares (Hernández et al., 2004).

c. Transformación

La transformación de datos es la etapa en la cual se preparan los datos para facilitar el uso de las diferentes técnicas de minería de datos que requieren los diferentes datos; existen varias técnicas de transformación de datos, entre las principales se encuentran (Lara, 2014):

- **Numerización**

La numerización consiste en transformar un atributo de tipo cualitativo a cuantitativo (Hernández et al., 2004), por ejemplo al momento de almacenar si un estudiante aprobó o no, se puede colocar 1 o 0respectivamente.

- **Discretización**

Es el proceso inverso a la numerización, en el cual los valores numéricos son transformados en discretos o nominales (Lara, 2014). Un ejemplo claro es el peso de una persona que puede pasar de > de 18.5 a deficiente, de 18.5 a 24.9 a normal, de 25.0 a 29.9 a sobrepeso y de <30 a obesidad («Calculadora del índice de masa corporal (IMC)», s. f.).

- **Creación de características**

Consiste en crear nuevos atributos en función a atributos existentes, que son las variaciones de los mismos (Hernández et al., 2004). En el caso del estudiante se puede crear el promedio general sumando el promedio de cada semestre cursado y dividiéndolo para el número de semestres.

- **Normalización**

La normalización consiste en transformar el rango de valores que toma un determinado atributo. Generalmente se emplea la normalización lineal uniforme, que transforma los valores de un atributo a una escala entre 0 y 1 mediante la Ecuación 1 (Lara, 2014):

$$ValorNormalizado = \frac{ValorInicial - ValorMínimo}{ValorMáximo - ValorMínimo} \quad \text{EC. 1}$$

Para seleccionar, limpiar y transformar la información se emplearán las herramientas estadísticas que nos brinda la herramienta ofimática de Microsoft Excel (Microsoft, 2018), y la suite de Pentaho (Hitachi Vantara, 2018). En la Fig. 6 se muestra el proceso mediante el cual se obtiene una vista minable, insumo para la siguiente fase del proceso KDD:

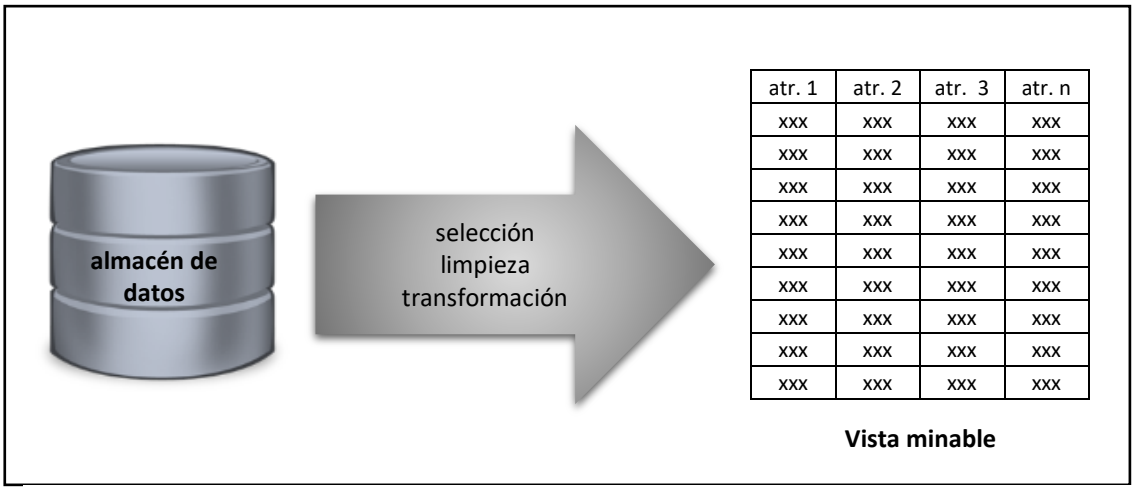


Fig. 6. Fase de selección, limpieza y transformación

1.4.3. Fase de minería de datos

Cuando ya se tiene consolidada la vista minable, el siguiente paso es aplicar las diferentes técnicas de minería de datos para obtener los modelos que representarán a dicha vista minable. Para poder iniciar con la fase de minería de datos es importante tomar decisiones que afectarán a la calidad del conocimiento que se pretende obtener (Hernández et al., 2004), entre ellas:

- Determinar la tarea de minería de datos que es la más apropiada para el análisis, por ejemplo, si se desea predecir cierta información.
- Escoger el modelo de acuerdo con la forma que se pretende que la información sea presentada, por ejemplo, en reglas de decisión.
- Elegir el algoritmo más eficiente al momento de resolver la tarea y que devuelva el modelo que se está buscando.

1.4.3.1. Tareas de minería de datos

En la etapa de minería de datos, se aplican diferentes técnicas para resolver los diferentes problemas. A los diferentes tipos de problemas que se pueden resolver por medio de las técnicas de minería de datos se conocen como tareas. Las tareas de minería de datos pueden ser predictivas o descriptivas (Lara, 2014).

a. Tareas predictivas

Las tareas predictivas de minería de datos sirven para predecir un valor desconocido de uno o varios atributos para varios registros de la vista minable, entre las principales tareas predictivas se encuentran la clasificación y la regresión (Lara, 2014).

- **Clasificación**

La tarea de clasificación es la más utilizada, consiste en utilizar un conjunto de entrenamiento para construir un modelo que a futuro se empleará para clasificar elementos o individuos desconocidos en base a una variable (atributo) de clase de tipo cualitativo (Hernández et al., 2004).

- **Regresión**

La regresión es similar a la clasificación, sin embargo, en este caso el atributo a predecir no es cualitativo, sino más bien cuantitativo que es la variable de clase (Lara, 2014).

b. Tareas descriptivas

Las tareas descriptivas generan modelos que de una forma u otra describen un grupo de datos.

- **Agrupamiento**

El objetivo principal del agrupamiento es obtener grupos homogéneos a partir de los datos heterogéneos, ya que en este caso se habla de grupos y no de clases, puesto que no analiza los datos a partir de una etiqueta conocida, sino que analiza los datos para obtener dicha etiqueta (Hernández et al., 2004).

- **Asociación**

El objetivo de la tarea de asociación es buscar las relaciones que no se encuentran explícitas entre los atributos, por medio de reglas de asociación (Lara, 2014).

- **Detección de atípicos**

La detección de atípicos consiste en encontrar objetos que presentan un comportamiento que se diferencia de manera notable en medio de un grupo de registros dentro de una vista minable (Pérez & Santín, 2007).

En la Fig. 7 se muestra el proceso mediante el cual se obtienen los modelos.

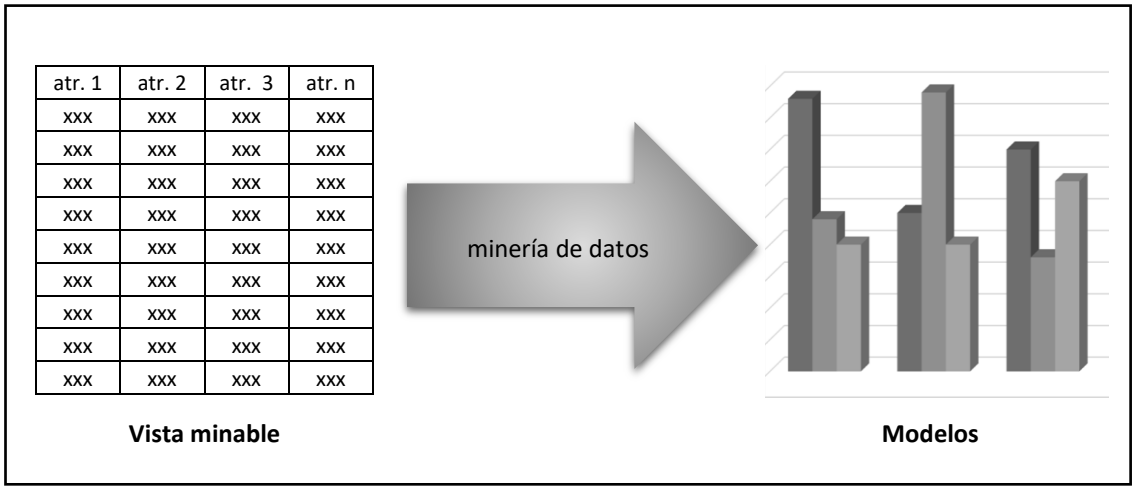


Fig. 7. Fase de minería de datos

En la fase de minería de datos se empleará la herramienta Weka (Waikato, 2018), ya que ofrece un sinnúmero de herramientas para aplicar tareas de minería de datos.

1.4.4. Fase de evaluación e interpretación

Finalmente, después de obtener los modelos de minería de datos el siguiente paso es evaluar la calidad de los modelos e interpretarlos, con la finalidad de obtener el conocimiento deseado (Lara, 2014).

Evaluación

Validar la bondad de un modelo predictivo es lo que se conoce como evaluar y sirve principalmente para medir su capacidad de predicción de nuevas instancias. Habitualmente la validación se realiza en base a las siguientes consideraciones (Sierra, 2006):

- **Matriz de confusión:**

La matriz de confusión permite apreciar, mediante una tabla de contingencia la repartición de los errores que se suscitaron por un clasificador, en esta tabla se cruza la variable de la clasificación predicha por el modelo con la variable que conserva la verdadera clasificación (ground truth). Para el caso de dos clases se tiene la matriz de confusión que se aprecia en la Tabla 1.2.

TABLA 1.2
MATRIZ DE CONFUSIÓN CASO DE ESTUDIO DE DOS CLASES

		Clase predicha		
		A(+)	B(-)	
Clase verdadera	A(+)	*TP*	*FN*	*πA*
	B(-)	*FP*	*TN*	*πB*
		pA	*pB*	*N*

Fuente: (Sierra Araujo, 2006)

Cada fila simbolizará los valores reales por cada clase, mientras que cada columna representará el número de predicciones para cada clase realizadas por el clasificador, de esta forma cada valor en la matriz de confusión quedan divididos de la siguiente forma (Pina, 2018):

a. True Positives (TP)

Los TP son el número de verdaderos positivos, es decir, las predicciones correctas para la clase A(+).

b. False Positives (FP)

Los FP son el número de falsos positivos, es decir que la predicción es la clase A(+) cuando realmente debía ser la B(-); a este tipo de casos se los conoce como errores tipo I.

c. False Negatives (FN)

Los FN son el número de falsos negativos, en donde la predicción en la clase B(-) cuando realmente debían ser en la clase A(+), a este tipo de casos se los denomina errores tipo II y son peores que los errores tipo I.

d. True Negatives (TN)

Los TN son el número de verdaderos negativos, es decir, el número de predicciones correctas para la clase B(-).

En la Tabla 1.2 se tiene que (Sierra, 2006):

a. *πA* denota la probabilidad a priori de la clase A
b. *πB* denota la probabilidad a priori de la clase B; donde *πB = 1 - πA*
c. *pA* indica la proporción de casos que el clasificador predice en la clase A
d. *pB* indica la proporción de casos que el clasificador predice en la clase B; pB = 1 – pA
e. *N = TP + FP + FN + TN*

De la matriz de confusión, también se puede extraer información tal como la sensibilidad, especialidad, proporción de falsos positivos o proporción de falsos negativos, así (Pina, 2018):

a. Tasa de error:

La tasa de error del clasificador es la medida más habitual para medir el éxito de un clasificador mediante la Ecuación 2. Se entiende el error como la clasificación incorrecta:

$$Tasa\ de\ error = \frac{número\ de\ errores}{número\ total\ de\ casos} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{FP+FN}{N} \quad EC.\ 2$$

b. Sensibilidad

También conocida como recall, proporciona la probabilidad de que, dada una observación realmente positiva el clasificador lo de así, siempre y cuando *TP* + *FN* sea mayor que 0, de lo contrario se define recall como 1. La sensibilidad se calcula mediante la Ecuación 3.

$$Recall = \frac{TP}{TP+FN} \quad EC.\ 3$$

c. Especificidad

La especificidad también se conoce como ratio de verdaderos negativos; que proporciona la probabilidad de que dada una clasificación en la clase B(-), el valor real también sea B(-), la especificidad se obtiene mediante la Ecuación 4.

$$Especifidad = \frac{TN}{TN+FP} \quad EC.\ 4$$

d. Precisión

La precisión también conocida como valor de predicción positiva, da la probabilidad de que dada una clasificación en la case A(+) el valor real también sea en la clase A(+); siempre y cuando *TP* + *FP* sea mayor que 0; caso contrario se define la precisión como 1, la fórmula se la encuentra en la Ecuación 5.

$$Precisión = \frac{TP}{TP+FP} \quad EC.\ 5$$

e. Accuracy

La medida accuracy ofrece el porcentaje de los aciertos de nuestro modelo, para calcular esta medida se emplea la Ecuación 6.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N} \quad EC.\ 6$$

f. F1-measure

La medida F_1 o F_1-measure, es una de las métricas de rendimiento más populares, ya que es la medida armónica de la precisión, ya que al crear un clasificador generalmente se hace una predicción en función a la recuperación y la clasificación, ese torna complicado comparar un modelo con alta capacidad de recuperación y baja precisión contra un modelo de alta precisión y baja recuperación, es aquí donde F1-measure puede comparar ambos modelos (Strandjev, 2017). En la Ecuación 7 se aprecia la fórmula de F_1-measure.

$$F_1 = \frac{2 \times precisión \times recall}{precisión \times recall} \qquad EC.\ 7$$

g. Curvas Receiver Operating Characteristic (ROC)

Las curvas ROC proporcionan una representación de la sensibilidad y especificidad para cada valor en el umbral, que permite comparar dos o más clasificadores en función a su capacidad discriminante, de acuerdo al área bajo la curva (AUC por sus siglas en inglés, area under curve), que comprende valores entre 0.5 y 1, donde 1 es el diagnóstico perfecto y 0.5 es un diagnóstico sin valor, es decir, que la clasificación se realiza al azar (Del Valle, s. f.).

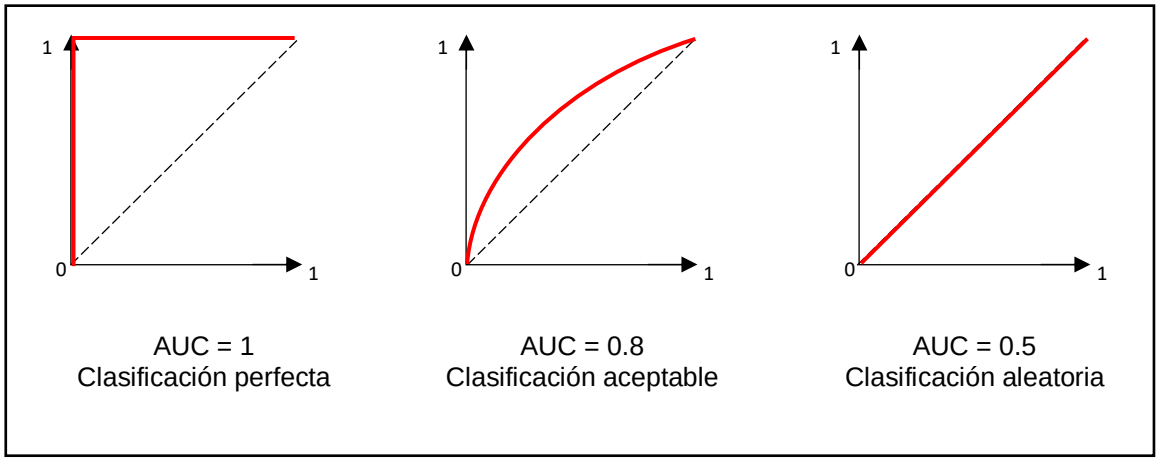


Fig. 8. Curva ROC

h. Coeficiente Kappa

El coeficiente Kappa es la medida que puede ser calculada en tablas de cualquier dimensión, puede tomar valores entre -1 y +1; mientras más se acerque a +1 mayor será el grado de concordancia de la clasificación, mientras que al contrario cuando más se acerque a -1 mayor es la discordancia de predicción. Cuando k = 0 se refleja la relación observada que se esperaba a causa del azar (Cerda & Villarroel, 2008). La fórmula del coeficiente Kappa se aprecia en la Ecuación 8. Para ver la forma de cálculo visitar el siguiente link: http://bit.ly/2RBCgBe.

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \qquad \textit{EC. 8}$$

Adicionalmente, existen dos conceptos importantes, la tasa de error verdadera y la tasa de error aparente. La tasa de error verdadera es la probabilidad de que el modelo clasifique incorrectamente nuevos casos que aún no se emplean. El objetivo de dicha tasa es obtener una estimación más apegada a la realidad. Mientras que la tasa de error aparente se obtiene al clasificar las instancias que ya se han empleado. A diferencia de la anterior esta tasa no se apega a la realidad, puesto que las instancias utilizadas para entrenar el modelo suelen adaptarse mejor a él que las nuevas instancias. Después de tener claros los conceptos anteriormente expuestos, los principales métodos de validación de los clasificadores son los siguientes (Sierra, 2006):

a. Método Holdout

Este método es uno de los más sencillos de validación, inicialmente se particiona el conjunto de casos en dos grupos, el de entrenamiento y el de testeo; el grupo de entrenamiento generalmente son dos terceras partes del conjunto total de casos y es usado para entrenar un modelo clasificatorio, mientras que el resto empleado para el grupo de testeo que sirve para estimar la tasa de error verdadera.

b. Método de remuestreo

Las técnicas de remuestreo consiste en tomar muestras de forma repetida del conjunto completo de muestras que se posee (*Remuestreo.pdf*, s. f.), para aplicar el método H en cada una de las muestras, de esta manera la estimación del error se obtiene a partir de la media de las diferentes tasas de error que se obtuvieron (Sierra , 2006).

a. Método de validación cruzada (Cross validation)

La validación proviene de la generalización, de dividir la muestra total en K grupos de aproximadamente el mismo tamaño, en el cual K -1 constituye el grupo de entrenamiento y el resto el grupo de testeo(Lara, 2014), un caso particular de validación cruzada es el método de ‘dejar uno fuera’, en donde K es igual al número de instancias que son empleadas para entrenar el modelo, ya que el proceso se repite varias veces para obtener el promedio de error cometido (Sierra, 2006).

Cuando ya se ha evaluado el modelo, es indispensable expresar el conocimiento en términos que conoce el usuario final; por ello es importante que se relacione la minería de datos con técnicas de visualización, para que dichos modelos sean comprendidos e interpretados por los expertos de cada área; así los expertos podrán comparar el

conocimiento obtenido con la realidad que ellos perciben para emplearlo en la toma de decisiones estratégicas (Lara, 2014), tal como se aprecia en la Fig. 9.

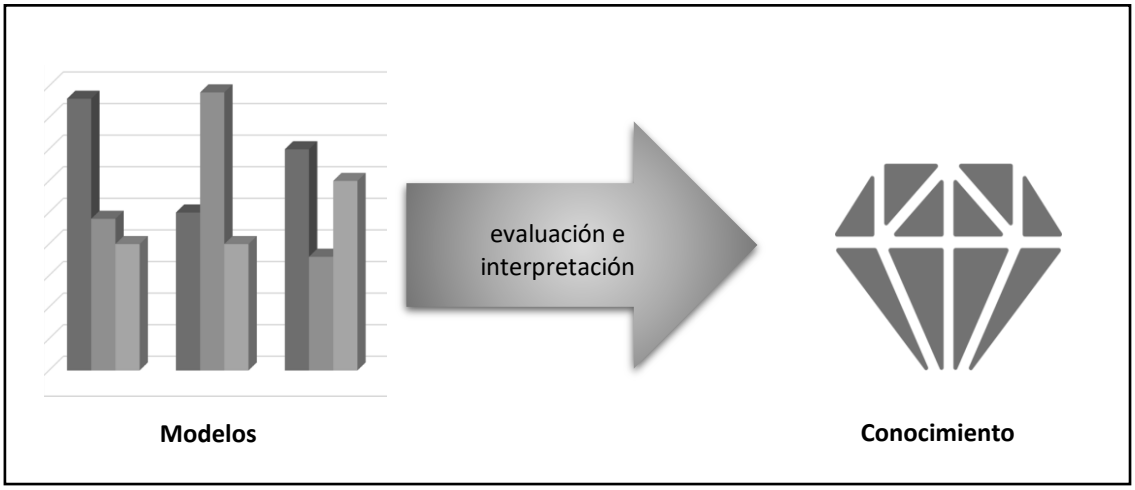


Fig. 9. Fase de evaluación e interpretación

Para visualizar la información se emplearán las herramientas de visualización de información que ofrece la herramienta de Microsoft Excel (Microsoft, 2018) y Weka.

1.5. Tareas y modelos predictivos

Las tareas predictivas de minería de datos nos proporcionan diferentes modelos predictivos, que tienen por objetivo obtener un modelo válido para tratar futuros casos (Sierra, 2006), los cuales se dividen en clasificación y regresión.

1.5.1. Clasificación

Entre las principales técnicas de clasificación se encuentran:

Algoritmos de clasificación por vecindad

Los algoritmos de clasificación por vecindad, como su nombre lo dice se basa en aproximaciones conocidos en criterios de vecindad (Dasarathy, 1991). Los algoritmos de vecindad exigen una definición de una cierta medida de distancia entre los elementos del espacio en representación. Una de las principales ventajas de esta técnica, es la simplicidad de su concepto, ya que la clasificación de un nuevo punto del espacio de representación se calcula en función a las clases, de los puntos más próximos a él (Sierra, 2006). Por ejemplo, el algoritmo K-NN.

Árboles de clasificación

Los arboles de decisiones, son un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la predicción se puede realizar siguiendo el camino de las condiciones hasta una de sus hojas. Entre las principales ventajas de los árboles de decisión es que se pueden expresar en grafo o en reglas de decisiones (Hernández et al., 2004). Por ejemplo, el algoritmo Random Tree, Random Forest, entre otros. El algoritmo Random Tree considera un número dado de características aleatorias en cada nodo sin realizar ninguna poda, mientras que Random Forest construye bosques aleatorios mediante el empaquetamiento de conjuntos de Random Tree (Frank et al., 2016).

Redes bayesianas

Las redes bayesianas modelan un hecho mediante un conjunto de variables y relaciones entre ellas, basadas en el teorema de Bayes; esta técnica predictiva permite estimar la probabilidad futura de las variables desconocidas en base a las conocidas (Sierra Araujo, 2006). Thomas Bayes expresa que, el teorema de Bayes es un resultado que expresa la probabilidad condicionada de un evento aleatorio dado otro evento (Lara, 2014). Por ejemplo, el algoritmo Naive Bayes, que implementa la teoría bayesiana para generar las probabilidades, también puede usar estimaciones de densidad del kernel, que mejora el rendimiento del algoritmo en caso de que el supuesto de normalidad sea muy incorrecto (Frank et al., 2016).

Redes neuronales artificiales

Las redes neuronales se basan en el aprendizaje humano, las neuronas cerebrales. Gracias a ellas un valor de entrada se transforma en salida mediante una función no lineal. Las redes neuronales poseen las siguientes características (Lara, 2014):

- La exactitud usualmente es muy alta
- Trabajan de manera correcta incluso con datos erróneos
- La salida puede ser un valor real o un conjunto de reales
- Evolucionan rápidamente de la función de entrenamiento

Por ejemplo, el algoritmo perceptrón multicapa.

Máquinas de vectores de soporte (SVM)

El desarrollo del aprendizaje supervisado ha dado paso a la creación de algoritmos denominados métodos de kernel, que han tomado un gran éxito, por su método denominado

Máquinas de Vectores Soporte(support vector machines). Entre sus principales ventajas se encuentra la aplicabilidad a cualquier tipo de datos, ya que las funciones kernel sirven como mecanismo de transformación y representación de información de entrada al algoritmo,

1.5.2. Regresión

Entre las principales técnicas de regresión se tienen las siguientes:

Regresión lineal

La regresión lineal es la forma más sencilla de regresión, puesto que los datos se modelan usando una línea recta; mediante una variable aleatoria que se denomina variable respuesta, que es la función lineal de otras variables, a_i $(0 \leq i \geq k)$, denominadas variables predictoras, como se aprecia en la Ecuación 9 (Lara, 2014):

$$y = w_0\ a_0 + w_1\, a_1 + \cdots + w_k\ a_k \qquad \textit{EC. 9}$$

El principal objetivo de la regresión es obtener el valor de una serie de pesos, a partir de los datos de un conjunto de entrenamiento. Los pesos se calculan mediante la técnica de los mínimos cuadrados, con el objetivo de minimizar la expresión de la Ecuación 10, donde y^i representa a la variable de respuesta y para el objeto *i* (Lara, 2014).

$$\sum (y^i - \sum w_j {a_j}^i)^2,\ (1 \leq i \geq n),\ (0 \leq j \geq K) \qquad \textit{EC. 10}$$

Regresión logística

La regresión logística se denomina así puesto que es un modelo más generalizado de regresión, denominada también discriminación logística; este tipo de tarea predictiva obtiene una estimación de probabilidades para variables categóricas, es decir una variable que puede adoptar un número limitado de categorías, y cuando se habla de dos variables de clase se habla de ranking (Hernández et al., 2004).

La regresión logística es un tipo de análisis discriminante predictivo, ya que su objetivo principal es brindar procedimientos sistemáticos para clasificar una observación cuyo origen se desconoce empleando los valores que toman las variables clasificadoras, que generalmente son variables de tipo cuantitativo. El objetivo de esta técnica es estimar probabilidades a posteriori $\{P(G_i|X); i = 1, \ldots, M\}$, en este estudio el número de grupos a discriminar es $M = 2$, para la clase desertó que toma los valores Si o No. Constituyen modelos de la forma que se presenta en la Ecuación 11 (Sierra, 2006):

$$P(G_1|X) = F(X'\beta);\ P(G_2|X) = 1 - P(G_1|X) \qquad \textit{EC. 11}$$

Donde *F* es la función de distribución de probabilidad acumulada, denominada función de enlace, particularmente si $F(x) = \varphi x$, la función es estándar y por ende el modelo es de Regresión Binomial tal como se aprecia en la Ecuación 12, para ampliar información ver el siguiente enlace http://bit.ly/2Xzyooy :

$$F(x) = \frac{exp(x)}{1+exp(x)} \qquad \textit{EC. 12}$$

1.6. ISO/IEC 25012:2008

La Organización Internacional de Normalización (ISO) en conjunto con la Comisión Electrotécnica Internacional (IEC) forman el sistema especializado para la normalización mundial, en el campo de las tecnologías de la información estos entes han determinado un comité técnico conjunto denominado conjunto ISO/IEC JTC 1 (Comité técnico). La norma ISO/IEC 25012 forma parte de las Normas Internacionales SQuaRE que se compone por las siguientes divisiones, que son referentes a Ingeniería de Software, Requerimientos y Evaluación de Calidad del Producto de Software (SQuaRE) (ISO & IEC, 2014):

- División de Gestión de Calidad (ISO/IEC 2500n)
- División de Modelo de Calidad (ISO/IEC 2501n)
- División de Medición de Calidad (SO/IEC 2502n)
- División de Requerimientos de calidad (ISO/IEC 2503n)
- División de Evaluación de Calidad (ISO/IEC 2504n)

Actualmente la cantidad de datos e información que es manejada por sistemas computacionales está en aumento, por ello verificar la calidad de esta es fundamental para varios procesos de negocios. En la Norma Internacional ISO/IEC 25012 la calidad de datos se encuentra descrita utilizando un modelo de datos previamente definido, que categoriza los atributos de calidad en quince características consideradas desde los puntos de vista inherente y dependiente del sistema (ISO & IEC, 2014):

1.6.1. Calidad inherente de datos

La calidad inherente de datos hace referencia al grado en el que las características de calidad tienen el potencial específico para satisfacer necesidades cuando la información se encuentre bajo condiciones específicas.

1.6.2. Calidad de datos de un sistema dependiente

La calidad de datos de un sistema dependiente se refiere al grado en el que la calidad se alcanza y se mantiene dentro de un sistema computacional cuando la información se encuentre bajo condiciones específicas.

A continuación, en la Fig. 10 se explica la división de las características de Calidad de Datos Inherente y Calidad de Datos Dependiente del Sistema.

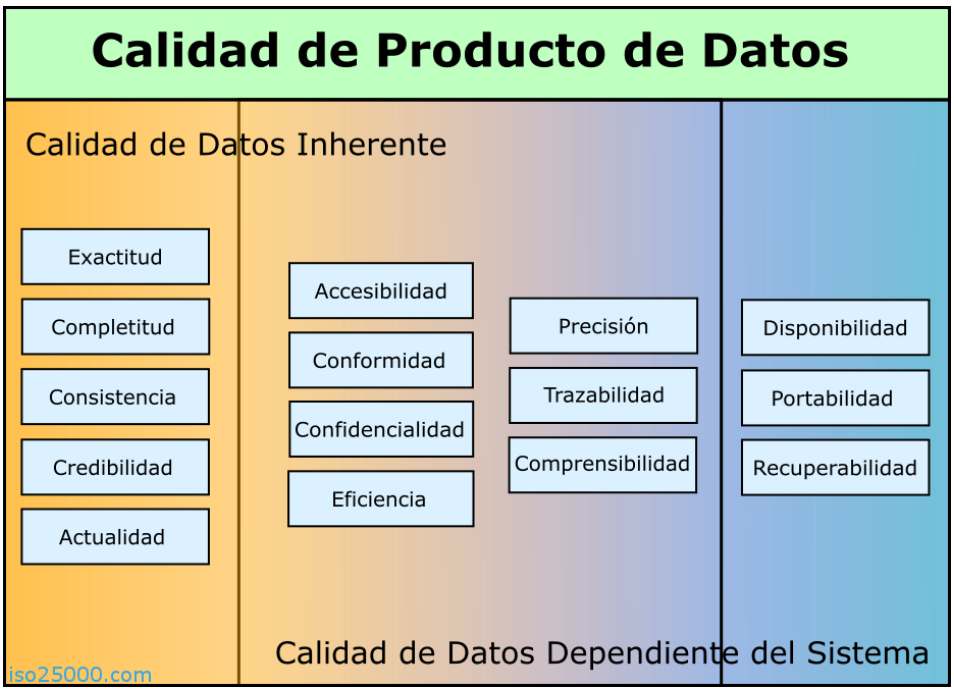


Fig. 10. Distribución ISO/IEC 25012:2008 - Fuente: Normas ISO 25000

Las características de la normativa que son aplicables a la naturaleza del presente proyecto son inherentes ya que los datos de análisis no se ligan o dependen de ningún sistema. A continuación, se detallan las características según IEC & ISO (2014):

- **Exactitud**

La exactitud cuenta con dos características la exactitud sintáctica y semántica. La exactitud sintáctica se define como la proximidad de los valores de los datos a un conjunto de valores definidos en un dominio considerado sintácticamente correcto, por ejemplo, cuando la palabra debería ser “abril” se encuentra como “abrjl”. Mientras que la exactitud semántica se define como la proximidad de los valores de los datos a un conjunto de valores definidos en un dominio considerado semánticamente correcto, por ejemplo, el mes se registra como “abril” cuando realmente debería ser “marzo”.

Para aplicar la característica sintáctica de exactitud se usa la Ecuación 12, donde A es el número de registros en el campo especificado sintácticamente exacto y B es el número total de registros:

$$exactitud\ sintáctica\ del\ registro = \frac{A}{B} \qquad EC.\ 13$$

- **Integridad**

La integridad es el grado en el que los datos del registro en cuestión a una entidad tienen valores para todos los atributos que se esperan, y casos de la entidad relacionados en un contexto de uso específico. Por ejemplo, para la base de datos de un estudiante, la integridad reduce si los registros del estudiante no contienen la información acerca del nivel en el que él se encuentra cursando. Para calcular la integridad de un registro en específico se aplica la Ecuación 13, donde A es el número de datos requeridos para un contexto particular en el archivo de información y B es el número de datos en un contexto particular especificado de uso previsto:

$$integridad\ de\ datos\ en\ un\ archivo = \frac{A}{B} \qquad EC.\ 14$$

- **Consistencia**

La consistencia mide el grado en el cual los datos tienen atributos que no tiene argumentación y son coherentes con otros datos en un contexto específico de uso. Por ejemplo, un estudiante no puede nacer en el año en el cual se matricula a su semestre académico. La Ecuación 14 especifica la forma de cálculo de la consistencia, donde A es el número de datos consistentes en el archivo y B es el número de datos guardado en el archivo.

$$integridad\ de\ datos\ en\ un\ archivo = \frac{A}{B} \qquad EC.\ 15$$

- **Credibilidad**

La credibilidad es el grado en el cual los datos poseen atributos que se relacionan como verdaderos y ciertos por los usuarios en un contexto específico de uso, así como también incluye el concepto de autenticidad. Para calcular la credibilidad se usa la Ecuación 15, donde A es el número de datos certificados por una auditoría interna y B es el número de datos usados para obtener información de riesgo de crédito:

$$credibilidad\ de\ datos = \frac{A}{B} \qquad EC.\ 16$$

CAPÍTULO 1

Desarrollo del Proceso KDD

La implementación del presente proyecto de minería de datos proporcionará la información necesaria para que los encargados de velar por el bienestar y permanencia de los estudiantes de la UTN tomen decisiones estratégicas en base a los datos históricos (personales, académicos, socioeconómicos, demográficos, psicológicos, etc.) que se encuentran almacenados en los repositorios de las bases de datos institucionales. Este capítulo fue desarrollado juntamente con el trabajo de titulación denominado “Detección de patrones de deserción estudiantil utilizando técnicas descriptivas de agrupamiento, asociación y atípicos de minería de datos, para la gestión académica de la Universidad Técnica del Norte” elaborado por Saúl Andrés Cisneros Buitrón, siendo un tronco en común de ambos trabajos por lo que ciertas tablas y figuras son similares.

A continuación, se procederá a explicar los pasos desarrollados para la implementación de las técnicas predictivas de minería de datos en base al Proceso de Descubrimiento del Conocimiento en Bases de Datos (KDD).

Para el desarrollo del presente proyecto de titulación se han considerado los siguientes aspectos:

2.1. Generalidades

- **Suposiciones**

Para la realización de la minería de datos se dispone de información histórica de la base de datos transaccional de la UTN, que será la fuente de datos en bruto para el análisis. La información resultante se expresará de forma que las personas encargadas de analizar por qué los estudiantes abandonan su carrera universitaria puedan interpretarla con mayor facilidad.

Se ha establecido una jerarquía que permitirá el desarrollo del proyecto, conformada por el jefe del proyecto, personas especializadas en minería de datos y personas especializadas en la problemática de deserción estudiantil; quienes cumplirán funciones tales como determinar roles, aportar conocimientos técnicos del área y brindar asesoría sobre factores clave para la problemática, respectivamente. Asimismo, el jefe del proyecto debe asignar un patrocinador que cuente con influencia entre los participantes del proyecto.

Al momento del desarrollo del proyecto se deben cumplir con los tiempos establecidos, considerando que cada fase del proceso KDD es la entrada (insumo) a la siguiente fase.

- **Restricciones**

El tiempo de desarrollo del proyecto no debe ser mayor a 6 meses, por ello la distribución del tiempo se determinó tomando en cuenta la naturaleza de las actividades que se han establecido, por ejemplo, para el desarrollo de recopilación de la información se estableció un tiempo prudente de 3 semanas para constituir el data warehouse que es la base del análisis.

El proyecto debe irse desarrollando de acuerdo con la disponibilidad de tiempo de las personas especializadas en cada área, con el objetivo de recibir la asesoría e insumos para cada fase del proyecto.

Los permisos de uso de las herramientas de software deben estar disponibles, de igual forma se debe contar con un especialista en el área de aprendizaje supervisado.

2.2. Entregables del Proyecto

A continuación, se detallan los artefactos que serán generados y utilizados por el proyecto, mientras dura el proceso los artefactos van variando hasta llegar a una vista minable de acuerdo con el proceso KDD.

- Check list de acuerdo con la ISO/IEC 25012:2008
- Data Warehouse: Correspondiente a la Fase de Integración de Datos
- Vista Minable: Correspondiente a la Fase de Selección, Limpieza y Transformación de datos.
- Modelos predictivos (clasificación y regresión): Correspondientes a la Fase de Minería de Datos
- Conocimiento (clasificación y regresión): Correspondientes a la Fase de Evaluación e Interpretación

2.3. Organización del Proyecto

2.3.1. Participantes del Proyecto

A continuación, en la TABLA 2.1 se especifican los directores de las áreas implicadas.

TABLA 2.1
DIRECTIVOS DE LAS ÁREAS IMPLICADAS

DEPENDENCIA	ENCARGADO	FUNCIÓN
Coordinación Carrera de Ingeniería en Sistemas Computacionales	Mgs. Pedro Granda	Asignar especialista en aprendizaje supervisado
Dirección de Desarrollo Tecnológico e Informático	Mgs. Juan Carlos García	Asignar especialista en Bases de Datos
Departamento de Bienestar Universitario	Dra. Eugenia Orbes	Asignar especialista en la problemática de deserción estudiantil

Fuente: Propia

En la TABLA 2.2 se detallan los participantes directos en el proyecto

TABLA 2.2
PARTICIPANTES DIRECTOS DEL PROYECTO

ROL	DEPENDENCIA	NOMBRE
Jefe de Proyecto	Carrera de Ingeniería en Sistemas Computacionales	Dr. Iván García
Administrador de bases de datos	Dirección de Desarrollo Tecnológico e Informático	Ing. Evelyn Enríquez Ing. Fernanda Rivera
Analista de Sistemas	Carrera de Ingeniería en Sistemas Computacionales	Dayana Vila

Fuente: Propia

2.3.2. Roles y Responsabilidades

En la TABLA 2.3 se describen los roles y las responsabilidades de cada uno de los participantes directos del proyecto.

TABLA 2.3
ROLES Y RESPONSABILIDADES

ROL	DEPENDENCIA
Jefe de Proyecto	La principal responsabilidad del jefe de proyecto es tomar decisiones que tiendan a cumplir los objetivos. Es el encargado de establecer comunicación con el usuario final de los entregables y el patrocinador, así como de gestionar los recursos empleados durante el proyecto, toma las decisiones necesarias para conocer en todo momento la situación actual en relación con los objetivos establecidos (Gutiérrez, 2017).
Administrador de bases de datos	El administrador de base de datos cumple con la función de proporcionar los datos almacenados en la base de datos transaccional de la UTN provenientes de las tablas Persona, Matrículas, Dependencias, Localidades, Notas, Ficha Socioeconómica, entre otras.
Analista de Sistemas	Aplicación de la ISO/IEC 25012:2008, desarrollo del proceso KDD, validación de resultados, documentación y análisis de impacto.

Fuente: Propia

2.4. Gestión del Proceso

2.4.1. Estimaciones

En las Tabla 2.4, Tabla 2.5 y Tabla 2.6 se detalla el presupuesto estimado y recursos involucrados. Para ello el método de estimación del costo se realizó considerando el número de horas empleadas por el costo por hora.

TABLA 2.4
TALENTO HUMANO

DESCRIPCIÓN	N. DE HORAS	COSTO POR HORA ($)	COSTO TOTAL ($)
Horas de investigación del proyecto	200	20.00	4000.00
Horas de desarrollo del proyecto	200	20.00	4000.00
		TOTAL	8000.00

Fuente: Propia

TABLA 2.5
RECURSOS MATERIALES

DESCRIPCIÓN	COSTO REAL ($)	COSTO ACTUAL ($)
Hardware		
Computadora portátil	750.00	00.00
Impresora	200.00	00.00
Software		
Microsoft Excel	00.00	00.00
Microsoft Word	00.00	00.00
Zotero	00.00	00.00
Pentaho Data Integration	00.00	00.00
Weka	00.00	00.00
Materiales de Oficina		
Tinta de impresora	50.00	30.00
Hojas A4	03.50	03.50
Esferos	01.00	01.00
Internet	120.00	120.00
Flash Memory	20.00	00.00
Investigación		
Textos	40.00	00.00
ISO IEC 25012:2008	8.00	00.00
TOTAL	1192.50	159.00

Fuente: Propia

TABLA 2.6
COSTO TOTAL DEL PROYECTO

DESCRIPCIÓN	COSTO ($)
Talento humano	8000.00
Recursos materiales	1192.50
TOTAL	9192.50

Fuente: Propia

2.4.2. Plan del Proyecto

El desarrollo de la minería de datos consta de varias fases del proceso KDD. En la Tabla 2.7 se encuentra detallada la duración por hora de cada fase y la aplicación de la normativa.

TABLA 2.7
DISTRIBUCIÓN DE HORAS

FASE	DURACIÓN EN HORAS
Fase de Integración y Recopilación	30
Implementación ISO/IEC 25012:2008	10
Fase de Selección, Limpieza y Transformación de Datos	30
Fase de Minería de Datos	30
Fase de Evaluación e Interpretación	50
Documentación e investigación	170
Análisis de Resultados	40
Análisis de Impactos	40
TOTAL	400

Fuente: Propia

En la TABLA 2.8 se detallan los hechos que determinan que una fase ha concluido.

TABLA 2.8
HECHOS IMPORTANTES

HECHO	DESCRIPCIÓN
Obtención del conocimiento bibliográfico	Se obtiene el conocimiento necesario para iniciar con el análisis de los datos, para finalmente determinar en qué consiste el proceso KDD y qué modelos se emplearán en la etapa de minería de datos de acuerdo con la naturaleza del conocimiento que se pretende obtener.
Obtención de los datos a analizar	Este hecho es determinante para iniciar con el análisis de los datos, concluye cuando ya se tienen los datos necesarios para realizar las fases del proceso KDD.
Implementación ISO/IEC 25012:2008	Se obtiene el grado en el cuales los datos son de calidad bajo condiciones específicas, de acuerdo con características determinadas en la norma.
Obtención del data warehouse	Se define el almacén de datos provenientes de los diferentes repositorios de datos de la UTN.

obtención de la vista minable	Se construye la vista minable a la cual se pueden aplicar técnicas predictivas de minería de datos, en la cual los datos se encuentran categorizados o numerizados.
obtención de los modelos predictivos	Se obtienen los modelos predictivos de minería de datos que permitirán el posterior análisis o interpretación del conocimiento.
obtención de métricas de calidad y conocimiento	Se determina si el modelo predijo bien dependiendo del atributo de clase y determinar si se ajusta al tipo de conocimiento que se pretende obtener.
Obtención de documentación	Se obtiene la documentación de cada una de las fases y actividades adicionales que se desarrollaron durante el proyecto.

Fuente: Propia

2.5. Fase de integración y recopilación

2.5.1. Tipos de datos base

Para el presente proyecto de titulación se emplearon los datos académicos, socioeconómicos, demográficos y personales de los estudiantes de la UTN que se encuentran almacenados en los repositorios de la base de datos Oracle 11g de la institución. Los tipos de datos que se encontraron son enteros, reales, fechas y cadenas de caracteres, sin embargo, se encontraron inconsistencias en los datos. Para aplicar las técnicas de minería de datos predictivas es indispensable clasificarlos en dos tipos de datos que son numéricos y categóricos o discretos (Hasperué, 2013). A continuación, en las Tablas 2.9- 2.17 se especifican los tipos de datos a analizar.

- **Tipos de datos de la tabla CICLO_ACADEMICOS_102018**

TABLA 2.9
ESTRUCTURA TABLA CICLO_ACADEMICOS_102018

ATRIBUTO	TIPO DE DATO
CODIGO	Cadena de caracteres
PER_ACAD_CODIGO	Cadena de caracteres
DESCRIPCION	Cadena de caracteres
FECHA_INICIO	Fecha
FECHA_FIN	Fecha
ESTADO	Caracter
ORDEN	Entero
TCICLOACAD_CODIGO	Entero
OBSERVACION	Cadena de caracteres
ANIO	Fecha
PORCENTAJE_PRIMERA_MATRICULA	Real
PORCENTAJE_SEGUNDA_MATRICULA	Real

PORCENTAJE_TERCERA_MATRICULA	Real
PORCENTAJE_GASTOS_ADM	Real
PORCENTAJE_SEGUNDA_PRORROGA	Real

Fuente: DDTI-UTN

- **Tipos de datos de la tabla DEPENDENCIAS_102018**

TABLA 2.10
ESTRUCTURA TABLA DEPENDENCIAS_102018

ATRIBUTO	TIPO DE DATO
CODIGO	Cadena de caracteres
NOMBRE	Cadena de caracteres
FUNCION	Caracter
DEPEN_CODIGO	Cadena de caracteres
DESCRIPCION	Cadena de caracteres
SIGLAS	Cadena de caracteres
OBSERVACION	Cadena de caracteres
ESTADO	Caracter
COD_SUBAREA_UNESCO	Cadena de caracteres
SECTOR	Entero
DEPEN_ANTIGUA	Cadena de caracteres

Fuente: DDTI-UTN

- **Tipos de datos de la tabla DETALLE_MATRICULAS_102018**

TABLA 2.11
ESTRUCTURA TABLA DETALLE_MATRICULAS_102018

ATRIBUTO	TIPO DE DATO
PARALELO_CODIGO	Cadena de caracteres
MATERIA_CODIGO	Cadena de caracteres
DOCENTE_CEDULA	Cadena de caracteres
INST_CODIGO	Cadena de caracteres
MODA_ESTUD_CODIGO	Cadena de caracteres
SIST_ESTUD_CODIGO	Cadena de caracteres
TCICLOACAD_CODIGO	Cadena de caracteres
TFINANCIA_CODIGO	Cadena de caracteres
DEPEN_CODIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
NIVEL_CODIGO	Cadena de caracteres
MATRICULA_CODIGO	Entero
ESTUDIANTE_CEDULA	Cadena de caracteres
ESTADO	Caracter
NUMERO_MATRICULA	Entero
ANULACION	Caracter

FECHA_ANULACION	Fecha
PENSUM_CODIGO	Caracter
ESTADO_EVAL_DOC	Caracter

Fuente: DDTI-UTN

- **Tipos de datos de la tabla ESTUDIANTE_CARRERA_102018**

TABLA 2.12
ESTRUCTURA TABLA ESTUDIANTE_CARRERA_102018

ATRIBUTO	TIPO DE DATO
ESTUDIANTE_CEDULA	Cadena de caracteres
DEPEN_CARRERA	Cadena de caracteres
NUMERO_CARRERA	Entero
FECHA_INGRESO	Fecha
GRATUIDAD	Caracter
MOTIVO_CODIGO	Cadena de caracteres
ESTADO	Caracter
FECHA_ULTIMA_MATRICULA	Fecha
TERMINA_CARRERA	Caracter
PIERDE_TERCERA	Caracter
FECHA_PIERDE_TERCERA	Fecha
USUARIO	Cadena de caracteres
OBSERVACION	Cadena de caracteres
NUMERO_CAMBIO	Entero
PRIMER_CICLO	Cadena de caracteres
ULTIMO_CICLO	Cadena de caracteres
INST_CODIGO	Cadena de caracteres
MOTIVO_SALE	Cadena de caracteres
FECHA_FINALIZACION	Fecha
MODA_ESTUD_CODIGO	Cadena de caracteres
PENSUM_CODIGO	Caracter
PENSUM_CICLO_ACAD_CODIGO	Cadena de caracteres
PENSUM_MODA_ESTUD_CODIGO	Cadena de caracteres
PENSUM_SIST_ESTUD_CODIGO	Cadena de caracteres
PENSUM_RESOLUCION	Cadena de caracteres
USUARIO_ACTUALIZA_PENSUM	Cadena de caracteres
CAMBIO_MALLA	Caracter

Fuente: DDTI-UTN

- **Tipos de datos de la tabla LOCALIDADES_102018**

TABLA 2.13
ESTRUCTURA TABLA LOCALIDADES_102018

ATRIBUTO	TIPO DE DATO
CODIGO	Cadena de caracteres
TLOCALIDAD_CODIGO	Cadena de caracteres
DESCRIPCION	Cadena de caracteres

ESTADO	Caracter
LOCALIDAD_CODIGO	Cadena de caracteres
GENTILICIO	Cadena de caracteres
FUNCION	Caracter
OBSERVACION	Cadena de caracteres
ZONA_PLANIFICACION	Entero

Fuente: DDTI-UTN

- **Tipos de datos de la tabla MATRICULAS_102018**

TABLA 2.14
ESTRUCTURA TABLA MATRICULAS_102018

ATRIBUTO	TIPO DE DATO
ESTUDIANTE_CEDULA	Cadena de caracteres
CODIGO	Cadena de caracteres
INST_CODIGO	Cadena de caracteres
MODA_ESTUD_CODIGO	Cadena de caracteres
SIST_ESTUD_CODIGO	Cadena de caracteres
TCICLO_ACAD_CODIGO	Cadena de caracteres
TFINANCIA_CODIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
DEPEN_CODIGO	Cadena de caracteres
TMATRICULA_CODIGO	Entero
USUARIOS_CUENTA	Cadena de caracteres
ESTADO	Caracter
NUMERO_MATRICULA	Entero
FECHA_INSCRIPCION	Fecha
FECHA_MATRICULA	Fecha
NIVEL_CODIGO	Cadena de caracteres
TRAN_NRO_TRANSACCION	Cadena de caracteres
EXONERADO	Caracter
ARRASTRES	Entero
LEGALIZADO	Caracter
FECHA_LEGALIZACION	Fecha
CARNETIZADO	Caracter
CONTINGENCIA	Caracter

Fuente: DDTI-UTN

- **Tipos de datos de la tabla NOTAS_102018**

TABLA 2.15
ESTRUCTURA TABLA NOTAS_102018

ATRIBUTO	TIPO DE DATO
MATERIA_CODIGO	Cadena de caracteres
PARALELO_CODIGO	Caracter
MATRICULA_CODIGO	Entero
DOCENTE_CEDULA	Cadena de caracteres
INST_CODIGO	Cadena de caracteres
MODA_ESTUD_CODIGO	Cadena de caracteres
SIST_ESTUD_CODIGO	Cadena de caracteres
TCICLOACAD_CODIGO	Cadena de caracteres
TFINANCIA_CODIGO	Cadena de caracteres
DEPEN_CODIGO	Cadena de caracteres
CICLO_ACAD_CODIGO	Cadena de caracteres
NIVEL_CODIGO	Cadena de caracteres
ESTUDIANTE_CEDULA	Cadena de caracteres
APROBO	Caracter
NOTA1	Entero
NOTA2	Entero
NOTA3	Entero
NOTA4	Entero
NOTA5	Entero
RESULTADO1	Entero
RESULTADO2	Entero
RESULTADO3	Entero
FINAL1	Real
FINAL2	Real
FINAL3	Real
NOTA_FINAL	Real
FECHA_REGISTRO	Fecha
RESOLUCION	Cadena de caracteres
OBSERVACION	Cadena de caracteres
PORCENTAJE_FALTAS	Real
OBSERVACION_AULA	Cadena de caracteres
PIERDE_POR_FALTAS	Caracter

Fuente: DDTI-UTN

- **Tipos de datos de la tabla PERSONAS_102018**

TABLA 2.16
ESTRUCTURA TABLA PERSONAS_102018

ATRIBUTO	TIPO DE DATO
CEDULA	Cadena de caracteres
LUGAR_NACIMIENTO	Cadena de caracteres
LUGAR_RESIDENCIA	Cadena de caracteres
NACIONALIDAD	Cadena de caracteres
LUGAR_PROCEDENCIA	Cadena de caracteres
TIPO_IDENTIFICACION	Caracter
FECHA_NACIMIENTO	Fecha
GENERO	Caracter
ESTADO_CIVIL	Caracter
ESTADO	Caracter
TIPO_SANGRE	Cadena de caracteres
LIBRETA_MILITAR	Cadena de caracteres
ID_SUBGRUPO_DISCAPACIDAD	Entero
CARNET_CONADIS	Cadena de caracteres
PORCENTAJE_DISCAPACIDAD	Real
COD_ETNIA	Cadena de caracteres
IDENTIFICACION	Cadena de caracteres

Fuente: DDTI-UTN

- **Tipos de datos de la tabla FICHA_112018**

TABLA 2.17
ESTRUCTURA TABLA FICHA_112018

ATRIBUTO	TIPO DE DATO
ESTUDIANTE_CEDULA	Cadena de caracteres
CODIGO_MATRICULA	Entero
CONVIVIENTE	Cadena de caracteres
TIPO_VIVIENDA	Cadena de caracteres
FINANCIAMIENTO	Cadena de caracteres
INGRESO_MENSUAL	Real
ACTIVIDAD_PADRE	Cadena de caracteres
AREA_PADRE	Cadena de caracteres
ACTIVIDAD_MADRE	Cadena de caracteres
AREA_MADRE	Cadena de caracteres
ACTIVIDAD_ESTUDIANTE	Cadena de caracteres
AREA_ESTUDIANTE	Cadena de caracteres
EMPLEO_ESTUDIANTE	Cadena de caracteres
INGRESOS_ESTUDIANTE	Real

Fuente: DDTI-UTN

2.5.2. Implementación de la norma ISO/IEC 25012:2008

La calidad de los datos es un factor clave, ya que el acierto de las decisiones que toma una organización depende en gran medida de la calidad de la información en que dichas decisiones se basan. Para iniciar con el proceso KDD se tomó en cuenta varias características de calidad de los datos inherentes de la ISO/IEC 25012. Puesto que la minería de datos se realiza independientemente del sistema que registra los datos, se aplicaron las características que se ajustan a la naturaleza del presente estudio, como se aprecia en la Tabla 2.18:

Indicar como se obtuvieron los resultados.

TABLA 2.18
EVALUACIÓN ISO/IEC:25012

MEDIDA / TABLA	EXACTITUD		INTEGRIDAD	CONSISTENCIA	CREDIBILIDAD
	SINTÁCTICA	SEMÁNTICA			
CICLO_ACADEMICOS_102018	94.11	100	60	100	100
DEPENDENCIAS_102018	100	100	100	100	100
DETALLE_MATRICULAS_102018	100	100	100	100	100
ESTUDIANTE_CARRERA_102018	100	100	79.16	100	100
LOCALIDADES_102018	100	100	85.71	100	100
MATRICULAS_102018	100	100	90.90	100	100
NOTAS_102018	100	100	75	100	100
PERSONAS_102018	100	100	88.23	99.75	100
FICHA_102018	99.82	100	64..28	98.95	100

Fuente: Propia

2.5.3. Construcción del data warehouse

Para construir el data warehouse se empleó la herramienta Pentaho Data Integration (PDI) la cual permite realizar procesos de Extracción, Transformación y Carga de información (ETL, por sus siglas en inglés Extract, Transform and Load), que serán útiles a lo largo del análisis.

Para realizar el data warehouse se ejecutaron las transformaciones en diferentes equipos con las características que se especifican en la Tabla 2.19:

TABLA 2.19
CARACTERÍSTICAS DE LOS EQUIPOS

DESCRIPCIÓN	SISTEMA OPERATIVO	MEMORIA RAM	PROCESADOR
EQUIPO 1	Windows 10	8GB	CORE i5 6ta Generación
EQUIPO 2	Windows 10	16GB	CORE i7 7ma Generación
EQUIPO 3	Windows 7	4GB	CORE i3 5ta Generación
EQUIPO 4	Windows 7	8GB	CORE i3 5ta Generación

Fuente: Propia

- **Dimensión LOCALIDADES**

En la Fig. 11 se muestra la transformación en PDI para obtener la dimensión LOCALIDADES.

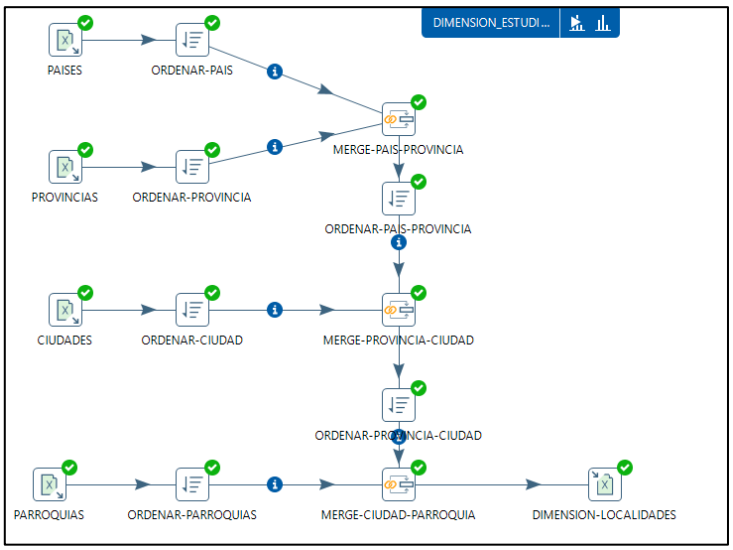


Fig. 11. Transformación PDI para Dimensión LOCALIDADES

En la Tabla 2.20 se detallan los tiempos de procesamiento en diferentes ordenadores con 1225 filas:

TABLA 2.20
TIEMPOS DE RESPUESTA DIMENSIÓN LOCALIDADES

EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
7.7 s	6.5 s	10.3 s	12 s

Fuente: Propia

- **Dimensión DEPENDENCIAS**

En la Fig. 12 se muestra la transformación en PDI para obtener la dimensión DEPENDENCIAS

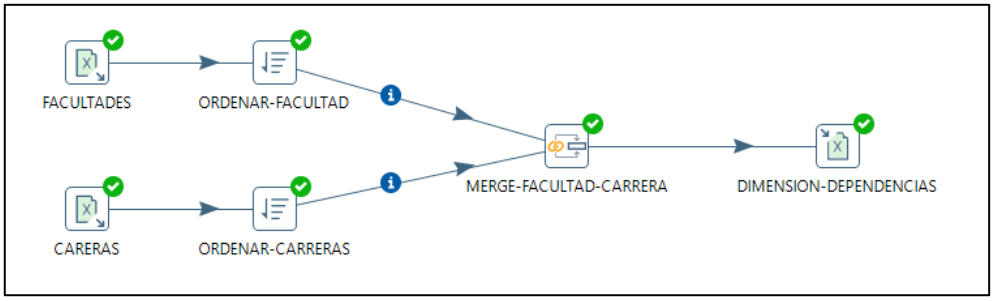


Fig. 12. Transformación PDI para Dimensión DEPENDENCIAS

En la Tabla 2.21 se detallan los tiempos de procesamiento en diferentes ordenadores con 80 filas:

TABLA 2.21
TIEMPOS DE RESPUESTA DIMENSIÓN DEPENDENCIAS

EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
6.4 s	5.4 s	8 s	12.5 s

Fuente: Propia

- **Dimensión ESTUDIANTE_CARRERA**

En la Fig. 13 se muestra la transformación en PDI para obtener la dimensión ESTUDIANTE_CARRERA

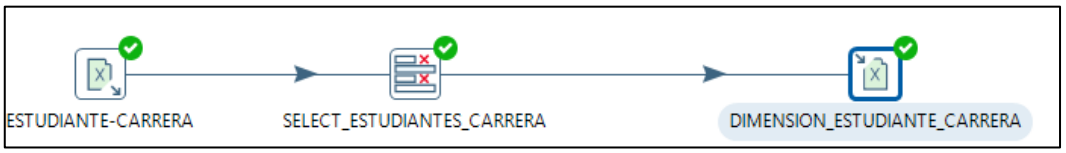


Fig. 13. Transformación PDI para Dimensión ESTUDIANTE_CARRERA

En la Tabla 2.22 se detallan los tiempos de procesamiento en diferentes ordenadores con 13675 filas:

TABLA 2.22
TIEMPOS DE RESPUESTA DIMENSIÓN ESTUDIANTE_CARRERA

EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
32.8 s	15.9 s	40.6 s	32,5 s

Fuente: Propia

- **Dimensión PERSONAS**

En la Fig. 14 se muestra la primera transformación en PDI para obtener la dimensión PERSONAS.

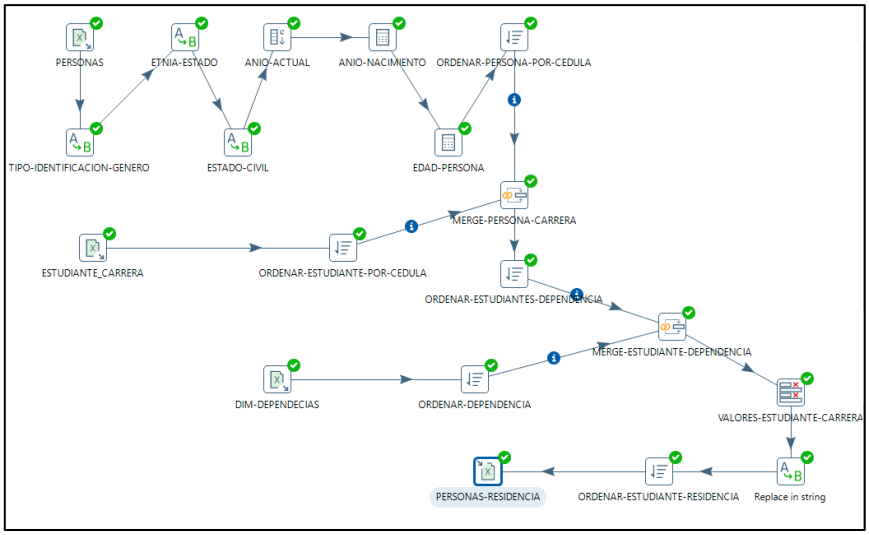


Fig. 14. Primera transformación PDI para Dimensión PERSONA

En la Fig. 15 se muestra la segunda transformación en PDI para obtener la dimensión PERSONAS.

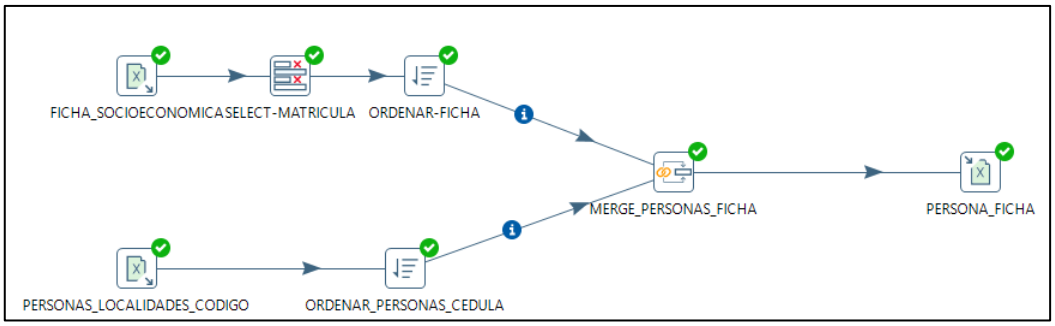


Fig. 15. Segunda transformación PDI para Dimensión PERSONA

En la Fig. 16 se muestra la tercera transformación en PDI para obtener la dimensión PERSONAS.

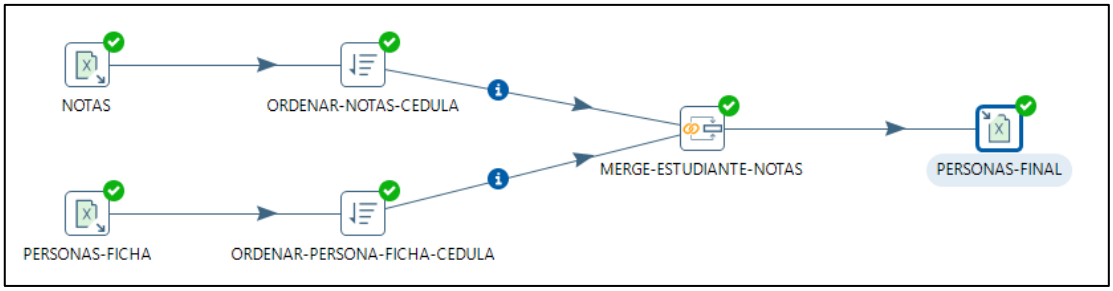


Fig. 16. Tercera transformación PDI para Dimensión PERSONA

En la Tabla 2.23 se detallan los tiempos de procesamiento en diferentes ordenadores con 13677 filas:

TABLA 2.23
TIEMPOS DE RESPUESTA DIMENSIÓN PERSONA

N. TRANSFORM.	EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
1	1m 7s	40.6 s	1m 23s	1m 14s
2	1m 39s	59.1 s	1m 57s	1m 42s
3	1m 24s	48.1 s	1m 32s	1m 30s

Fuente: Propia

- **Dimensión del DATA_WAREHOUSE**

En la Fig. 17 se muestra la transformación en PDI para obtener el DATA_WAREHOUSE.

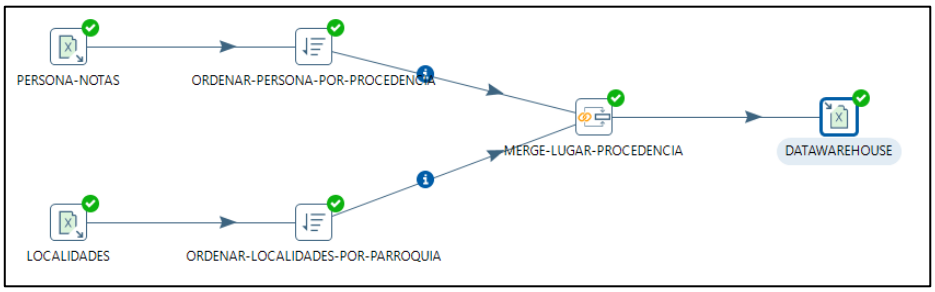


Fig. 17. Transformación PDI para el Data Warehouse

En la Tabla 2.24 se muestra la estructura del DATA_WAREHOUSE que cuenta con 11201 filas:

TABLA 2.24
ESTRUCTURA DATA_WAREHOUSE

ATRIBUTO	TIPO DE DATO
ESTUDIANTE_CEDULA	Cadena de caracteres
CONVIVIENTE	Cadena de caracteres
TIPO_VIVIENDA	Cadena de caracteres
FINANCIAMIENTO	Cadena de caracteres
INGRESO_MENSUAL	Real
ACTIVIDAD_PADRE	Cadena de caracteres
AREA_PADRE	Cadena de caracteres
ACTIVIDAD_MADRE	Cadena de caracteres
AREA_MADRE	Cadena de caracteres
ACTIVIDAD_ESTUDIANTE	Cadena de caracteres
AREA_ESTUDIANTE	Cadena de caracteres
EMPLEO_ESTUDIANTE	Cadena de caracteres

INGRESOS_ESTUDIANTE	Real
SIGLAS_FACULTAD	Cadena de caracteres
NOMBRE_CARRERA	Cadena de caracteres
CEDULA	Cadena de caracteres
NACIONALIDAD	Cadena de caracteres
TIPO_IDENTIFICACION	Cadena de caracteres
GENERO	Cadena de caracteres
ESTADO_CIVIL	Cadena de caracteres
ESTADO	Cadena de caracteres
TIPO_SANGRE	Cadena de caracteres
PORCENTAJE_DISCAPACIDAD	Real
COD_ETNIA	Cadena de caracteres
EDAD_PERSONA	Cadena de caracteres
NUMERO_CARRERA	Entero
ESTADO_CARRERA	Cadena de caracteres
MOTIVO_SALE	Cadena de caracteres
PROMEDIO_DE_NOTA_FINAL	Real
DESCRIPCION_PROVIENCIA	Cadena de caracteres

Fuente: Propia

En la Tabla 2.25 se detallan los tiempos de procesamiento en diferentes ordenadores con 11200 filas:

TABLA 2.25
TIEMPOS DE RESPUESTA DATA_WAREHOUSE

EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
1m 3s	42.3 s	1m 34s	1m 10s

Fuente: Propia

2.6. Fase de selección, limpieza y transformación

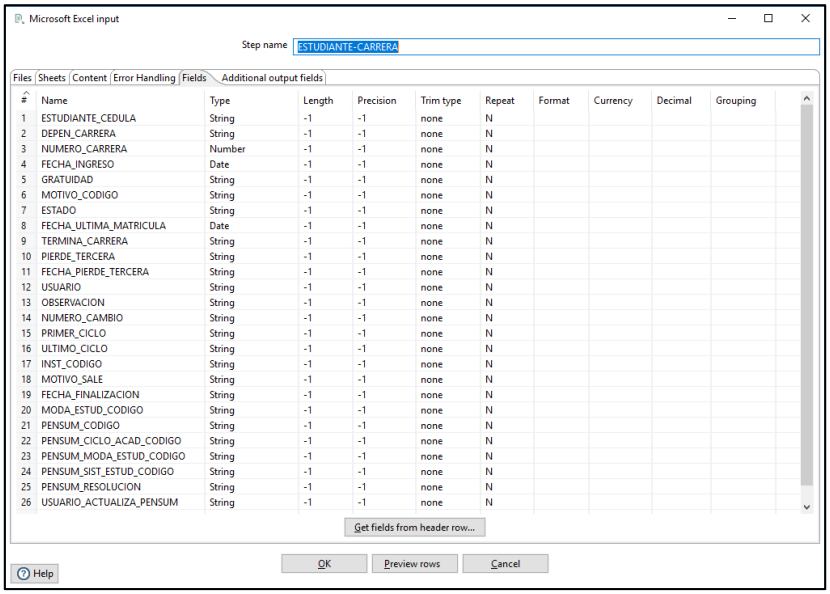
Una vez consolidado el data warehouse se procede a seleccionar, limpiar y transformar los datos, para lo cual se empleó la herramienta PDI.

2.6.1. Selección

- **Filtrado de atributos**

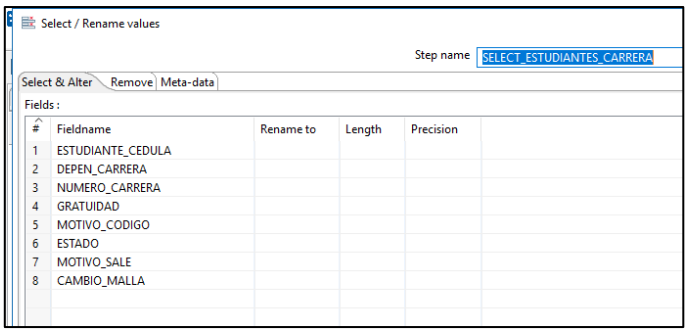
La etapa de selección y limpieza se inició eliminando los atributos de las tablas que no son relevantes para nuestro estudio, por ejemplo, en la tabla ESTUDIANTE_CARRERA

inicialmente se contaba inicialmente con 26 atributos, después de la selección quedaron 8 atributos relevantes para este estudio, tal como se muestra en las Fig. 18 y Fig. 19.



#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	ESTUDIANTE_CEDULA	String	-1	-1	none	N				
2	DEPEN_CARRERA	String	-1	-1	none	N				
3	NUMERO_CARRERA	Number	-1	-1	none	N				
4	FECHA_INGRESO	Date	-1	-1	none	N				
5	GRATUIDAD	String	-1	-1	none	N				
6	MOTIVO_CODIGO	String	-1	-1	none	N				
7	ESTADO	String	-1	-1	none	N				
8	FECHA_ULTIMA_MATRICULA	Date	-1	-1	none	N				
9	TERMINA_CARRERA	String	-1	-1	none	N				
10	PIERDE_TERCERA	String	-1	-1	none	N				
11	FECHA_PIERDE_TERCERA	String	-1	-1	none	N				
12	USUARIO	String	-1	-1	none	N				
13	OBSERVACION	String	-1	-1	none	N				
14	NUMERO_CAMBIO	String	-1	-1	none	N				
15	PRIMER_CICLO	String	-1	-1	none	N				
16	ULTIMO_CICLO	String	-1	-1	none	N				
17	INST_CODIGO	String	-1	-1	none	N				
18	MOTIVO_SALE	String	-1	-1	none	N				
19	FECHA_FINALIZACION	String	-1	-1	none	N				
20	MODA_ESTUD_CODIGO	String	-1	-1	none	N				
21	PENSUM_CODIGO	String	-1	-1	none	N				
22	PENSUM_CICLO_ACAD_CODIGO	String	-1	-1	none	N				
23	PENSUM_MODA_ESTUD_CODIGO	String	-1	-1	none	N				
24	PENSUM_SIST_ESTUD_CODIGO	String	-1	-1	none	N				
25	PENSUM_RESOLUCION	String	-1	-1	none	N				
26	USUARIO_ACTUALIZA_PENSUM	String	-1	-1	none	N				

Fig. 18. Atributos de la tabla ESTUDIANTE_CARRERA



#	Fieldname	Rename to	Length	Precision
1	ESTUDIANTE_CEDULA			
2	DEPEN_CARRERA			
3	NUMERO_CARRERA			
4	GRATUIDAD			
5	MOTIVO_CODIGO			
6	ESTADO			
7	MOTIVO_SALE			
8	CAMBIO_MALLA			

Fig. 19. Atributos relevantes al estudio

En el filtrado final de los atributos se consideraron datos personales, académicos, demográficos y socioeconómicos de los estudiantes, como se observa en la Fig. 20.

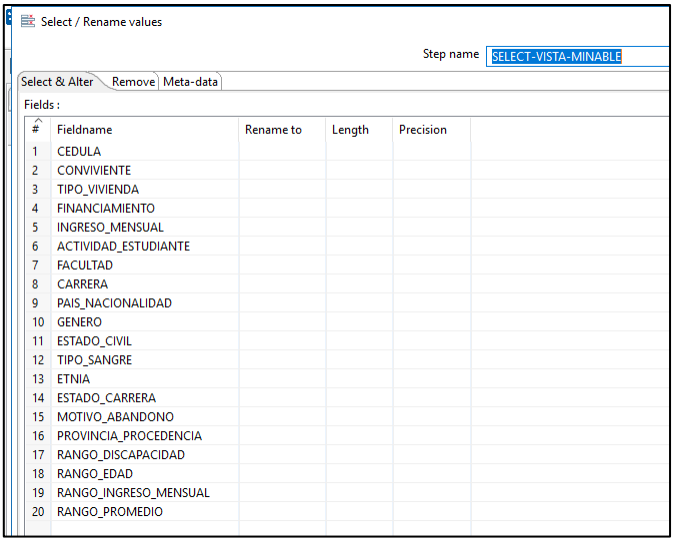


Fig. 20. Atributos seleccionados para realizar el análisis

- **Filtrado de registros**

Al igual que en el caso anterior se procedió a eliminar los registros que no se disponen completamente, tales como los datos del año 2013 al primer periodo académico del año 2017, ya que al realizar las diferentes transformaciones y cruces con los datos socioeconómicos (tipo de vivienda, financiamiento, ingresos económicos mensuales, actividad del estudiante y conviviente) se eliminan al no estar habilitada la ficha socioeconómica en dicho periodo. Por lo que quedaron los datos del primer periodo académico del 2017 hasta el primer periodo académico de 2018.

2.6.2. Transformación

Para realizar las transformaciones se categorizaron las clases que se ajustaban a criterios específicos, de acuerdo con lo siguiente:

- **Clase CONVIVIENTE:**

Se categorizó la clase CONVIVIENTE ya que las opciones que se almacenan en los repositorios son muy amplias y los resultados se podrían dispersar demasiado, como se muestra en la Tabla 2.26:

TABLA 2.26
CATEGORIZACIÓN CLASE CONVIVIENTE

CATEGORÍA	VALORES
FAMILIARES	Abuela
	Abuelo
	Abuelos
	Padrinos
	Primos

	Tíos
	Hermanos
	Hijos
	Familiar
PAREJA	Cónyuge
	Pareja
SOLO	Sólo
	Financiamiento propio
MADRE	Madre
PADRE	Padre
PADRES	Padres
OTROS	Otros
	Amigos
	No Asignado

Fuente: Propia

- **Clase FINANCIAMIENTO:**

De igual forma que en la clase CONVIVIENTE se categorizaron los datos para reducir la gama de opciones, como se detalla en la Tabla 2.27:

TABLA 2.27
CATEGORIZACIÓN CLASE FINANCIAMIENTO

CATEGORÍA	VALORES
FAMILIARES	Abuela
	Abuelo
	Abuelos
	Padrinos
	Primos
	Tíos
	Hermanos
	Hijos
	Familiar
PAREJA	Cónyuge
	Pareja
SOLO	Sólo
	Financiamiento propio
MADRE	Madre
PADRE	Padre
PADRES	Padres
BECA	Beca
CRÉDITO	Crédito
OTROS	Otros
	Amigos
	No Asignado

Fuente: Propia

- **Clase INGRESO_MENSUAL**

La clase ingreso mensual se categorizó tomando en cuenta el Salario Básico Unificado (RMU) del Ecuador del año 2018 que es $386.00 (Telégrafo, 2017), el canasta básica familiar $ 720.53 (INEC, 2018) y el promedio entre los valores más altos almacenados en el data warehouse, de acuerdo con la Tabla 2.28.

TABLA 2.28
CATEGORIZACIÓN CLASE INGRESO_MENSUAL

CATEGORÍA	VALORES ($)
MUY BAJO	0 a 386.00
BAJO	386.01 a 720.53
MEDIO	720.54 a 1000.00
ALTO	1000.01 a 2000.00
MUY ALTO	2000.01 a 16000.00

Fuente: Propia

- **Clase CARRERA**

Para categorizar la clase CARRERA se clasificó por cada una de las facultades de la UTN: Facultad de Ingeniería en Ciencias Aplicadas (FICA), Facultad de Ciencias Administrativas y Económicas (FACAE), Facultad de Ciencias de la Salud (FCCSS), Facultad de Educación, Ciencia y Tecnología (FECYT) y Facultad de Ingeniería en Ciencias Agropecuarias y Ambientales (FICAYA). Posteriormente se aplicó el nombre de la carrera del rediseño curricular de cada carrera.

En la Tabla 2.29 se muestra la categorización de la FACAE

TABLA 2.29
CATEGORIZACIÓN CLASE CARRERA FACAE

CATEGORÍA	VALORES
ADMINISTRACIÓN DE EMPRESAS	Administración de Empresas (Rediseño)
	Administración de Empresas
	Administración Pública de Gobiernos Seccionales
CONTABILIDAD SUPERIOR Y AUDITORÍA	Contabilidad y Auditoría
	Contabilidad y Auditoría CPA
ECONOMÍA	Economía (Rediseño)
	Economía Mención Finanzas
MERCADOTECNIA	Mercadotecnia (Rediseño)
	Mercadotecnia
GASTRONOMÍA	Gastronomía (Rediseño)
	Gastronomía
TURISMO	Turismo (Rediseño)
	Turismo

DERECHO	Derecho

Fuente: Propia

En la Tabla 2.30 se muestra la categorización de la FCCSS

TABLA 2.30
CATEGORIZACIÓN CLASE CARRERA FCCSS

CATEGORÍA	VALORES
ENFERMERÍA	Enfermería (Rediseño)
	Enfermería
TERAPIA FÍSICA MÉDICA	Fisioterapia
	Terapia Física Médica
NUTRICIÓN Y DIETÉTICA	Nutrición y Salud Comunitaria
	Nutrición y Dietética
MEDICINA	Medicina

Fuente: Propia

En la Tabla 2.31 se muestra la categorización de la FECYT

TABLA 2.31
CATEGORIZACIÓN CLASE CARRERA FECYT

CATEGORÍA	VALORES
PSICOLOGÍA EDUCATIVA Y O. V.	Psicología Educativa y O. V.
PSICOLOGÍA GENERAL	Psicología
	Psicología (Rediseño)
	Psicopedagogía
ARTES PLÁSTICAS	Artes Plásticas (Rediseño)
	Artes Plásticas
	Pedagogía de las Artes y Humanidades
DISEÑO Y PUBLICIDAD	Diseño y Publicidad
	Publicidad
DISEÑO GRÁFICO	Diseño Gráfico (Rediseño)
	Diseño Gráfico
GESTIÓN Y DESARROLLO SOCIAL	Gestión y Desarrollo Social
RELACIONES PÚBLICAS	Relaciones Públicas
SECRETARIADO EJECUTIVO EN ESPAÑOL	Secretariado Ejecutivo en Español
INGLÉS	Inglés
PARVULARIA	Parvularia
	Educación Inicial
ENTRENAMIENTO DEPORTIVO	Entrenamiento Deportivo (Rediseño)
	Educación Física
	Entrenamiento Deportivo
	Pedagogía de la Actividad Física y Deporte
EDUCACIÓN GENERAL BÁSICA	Educación Básica

	Educación Básica Lenguaje y Comunicación Convenio Inst. Pedagógico
FÍSICA Y MATEMÁTICA	Físico Matemático Pedagogía en las Ciencias Experimentales
CONTABILIDAD Y COMPUTACIÓN	Contabilidad y Computación
IDIOMAS NACIONALES Y EXTRANJEROS	Idiomas Nacionales y Extranjeros

Fuente: Propia

En la Tabla 2.32 se muestra la categorización de la FICA

TABLA 2.32
CATEGORIZACIÓN CLASE CARRERA FICA

CATEGORÍA	VALORES
TELECOMUNICACIONES	Electrónica y Redes de comunicación Telecomunicaciones
SOFTWARE	Sistemas Computacionales Software
MECATRÓNICA	Mecatrónica (Rediseño) Mecatrónica
INDUSTRIAL	Industrial (Rediseño) Industrial
TEXTILES	Textil Textiles
AUTOMOTRIZ	Automotriz Mantenimiento Automotriz
ELECTRICIDAD	Electricidad Mantenimiento Eléctrico

Fuente: Propia

En la Tabla 2.33 se muestra la categorización de la FICAYA

TABLA 2.33
CATEGORIZACIÓN CLASE CARRERA FICAYA

CATEGORÍA	VALORES
AGROINDUSTRIAS	Agroindustrias Agroindustrial
AGRONEGOCIOS AVALÚOS Y CATASTROS	Agronegocios Avalúos y Catastros
AGROPECUARIA	Agropecuaria (Rediseño) Agropecuaria
RECURSOS NATURALES RENOVABLES	Recursos Naturales Renovables (Rediseño) Recursos Naturales
FORESTAL	Forestal (Rediseño) Forestal
BIOTECNOLOGÍA	Biotecnología (Rediseño)

	Biotecnología
ENERGÍAS RENOVABLES	Energías Renovables

Fuente: Propia

- **Clases GENERO y TIPO_IDENTIFICACION**

Para categorizar la clase GENERO y TIPO_IDENTIFICACION se tomó en cuenta los datos almacenados en la base de datos de la UTN, de acuerdo con la Tabla 2.34.

TABLA 2.34
CATEGORIZACIÓN CLASES GENERO Y TIPO_IDENTIFICACION

CATEGORÍA	VALORES
MASCULINO	M
FEMENINO	F
CÉDULA	C
PASAPORTE	P

Fuente: Propia

- **Clase ESTADO_CIVIL**

La clase ESTADO_CIVIL se categorizó de acuerdo con la Ley de Registro Civil, Identificación y Cedulación (Asamblea Nacional del Ecuador, 2016), como se muestra en la Tabla 2.35.

TABLA 2.35
CATEGORIZACIÓN CLASE ESTADO_CIVIL

CATEGORÍA	VALORES
CASADO	C
DIVORCIADO	D
SOLTERO	S
VIUDO	V
UNIÓN DE HECHO	U

Fuente: Propia

- **Clase EDAD**

Para la clase EDAD se tomó la fecha de nacimiento del estudiante y se calculó la edad con la que este finalizó el año 2018 como se observa en la Fig. 21, para posteriormente categorizar las edades tomando en cuenta el promedio en que los estudiantes del nivel de grado culminan su carrera que es entre los 25 años, mientras que los estudiantes del nivel de postgrado culminan en promedio entre los 40 años, mientras que en las edades mayores a 40 años son menos frecuentes como se aprecia en la TABLA 2.36 (Vila et al., 2018).

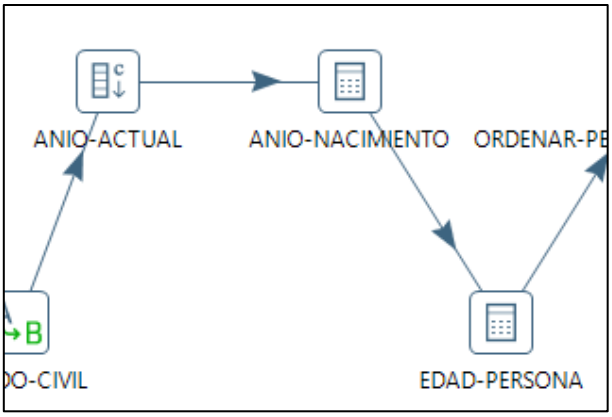


Fig. 21. Parte de la transformación que calcula la edad hasta el año 2018

TABLA 2.36
CATEGORIZACIÓN CLASE EDAD

CATEGORÍA	VALORES
BAJA	18 a 25
MEDIA	26 a 39
ALTA	40 en adelante

Fuente: Propia

- **Clase ETNIA**

La clase ETNIA se categorizó de acuerdo con las etnias que se registraron en el Censo de Población y Vivienda 2001 y 2010 (Villacís & Carrillo, 2012) como se aprecia en la Tabla 2.37.

TABLA 2.37
CATEGORIZACIÓN CLASE ESTADO_CIVIL

CATEGORÍA	VALORES
MESTIZO	ME
AFRODESCENDIENTE	AF/ MU/NE
INDÍGENA	IN
MONTUBIO	MO
NO ASIGNADO	NO

Fuente: Propia

- **Clase PROVINCIA_PROCEDENCIA**

Para la provincia de procedencia se tomó en cuenta las provincias del Ecuador y para los lugares fuera del país se consideró únicamente el nombre del país, con la finalidad de que las opciones se reduzcan, como se observa en la Tabla 2.38. Este proceso se realizó para estudiantes de nacionalidad colombiana y peruana, puesto que para el resto de los estudiantes extranjeros ya se encontraba el país en el nivel de provincia.

TABLA 2.38
CATEGORIZACIÓN CLASE PROVINCIA_PROCEDENCIA

CATEGORÍA	VALORES
COLOMBIA	ANTIOQUIA
	BOYACÁ
	CAUCA
	CUNDINAMARCA
	NARIÑO
	PUTUMAYO
	VALLE DEL CAUCA
PERÚ	HUILCA

Fuente: Propia

- **Clase PORCENTAJE_DISCAPACIDAD**

Para el porcentaje de discapacidad se consideró la base establecida para considerar discapacidad que es el 30% (Pérez, 2017), como se muestra en la Tabla 2.39.

TABLA 2.39
CATEGORIZACIÓN CLASE PORCENTAJE_DISCAPACIDAD

CATEGORÍA	VALORES (%)
NO TIENE	0 a 29
LEVE	30 a 49
MODERADO	50 a 74
SEVERO	75 a 84
MUY SEVERO	85 a 100

Fuente: Propia

- **Clase PROMEDIO**

Para esta clase se calculó el general de las notas del estudiante, por ejemplo, si el estudiante tiene registrado 4 semestres con 7 materias cada uno se calcula el promedio por semestre y luego se realiza el promedio general, para ellos se empleó la herramienta de tablas dinámicas de Microsoft Excel como se muestra en la Fig. 22.

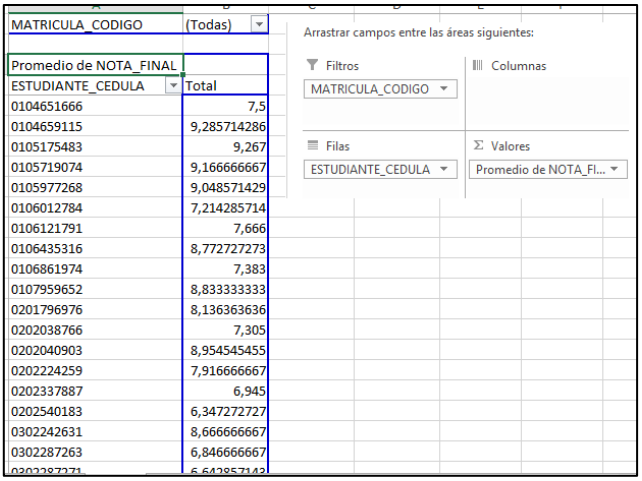


Fig. 22. Cálculo del promedio general de cada estudiante

La categorización de la clase PROMEDIO se realizó mediante los siguientes rangos que se definieron considerando que un promedio menor a 7 es insuficiente para aprobar el semestre, de 7 a 8 puntos es suficiente, de 8 a 9 es bueno y de 9 a 10 es excelente, como se aprecia en la Tabla 2.40.

TABLA 2.40
CATEGORIZACIÓN CLASE PROMEDIO

CATEGORÍA	VALORES
INSUFICIENTE	0 a 6.99
SUFICIENTE	7.00 a 8.00
BUENO	8.01 a 9.00
EXCELENTE	9.01 a 10.00

Fuente: Propia

A continuación, en la Tabla 2.41 se puede apreciar los tiempos de respuesta de la etapa de transformación en diferentes ordenadores.

TABLA 2.41
TIEMPOS DE RESPUESTA ETAPA DE TRANSFORMACIÓN

EQUIPO 1	EQUIPO 2	EQUIPO 3	EQUIPO 4
44.7 s	27.7 s	32.5 s	45.7 s

Fuente: Propia

2.6.3. Limpieza

En la fase de limpieza de datos se eliminaron los datos inconsistentes que quedaron después de la transformación, que generalmente tenían relación con las fechas. De igual forma se corrigieron los datos que quedaron con errores ortográficos o que no se transformaron bien, tal como se aprecia en la Fig. 23 y Fig. 24.

NGO_EDAD	TIPO_SANGRE	PAIS_NACIONALIDAD	GENERO	ETNIA	RANGO_DISCAPACIDAD	ESTADO_CIVIL	CONVIVIENTE	TIPO_VIVIENDA	FINANCIAMIENTO
DIA	A+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	O+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
DIA	O+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
IA	O-	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PRESTADA	CRÉDITO
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	CASADO	FAMILIARs	PROPIA	MADRE
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	A+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	A+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
IA	O+	ECUADOR	FEMENINO	MONTUBI	NO TIENE	SOLTERO	FAMILIARs	PRESTADA	PADRES
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
IA	A+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
IA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
IA	O+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
TA	A+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	FAMILIAR
IA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	FAMILIARs
IA	NO ASIGNADO	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
DIA	O+	ECUADOR	MASCULINC	MESTIZO	NO TIENE	UNION DE HECHO	FAMILIARs	PROPIA	PADRE
IA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
IA	O+	ECUADOR	FEMENINO	NO ASIGN	NO TIENE	SOLTERO	FAMILIARs	PROPIA	FAMILIARs

Fig. 23. Error ortográfico en Clase CONVIVIENTE categoría FAMILIAR

PADRE	NO TRABAJA	MEDIO	FICAYA	RECURSOS NATURALES RENOVABLES Renovables	IMBABURA	SUF
SOLO	NO TRABAJA	MUY BAJO	FECYT	DISEÑO Y DISEÑO Y PUBLICIDAD	CARCHI	BUE
SOLO	NO TRABAJA	MUY BAJO	FECYT	Licenciatura en Ed. B. Lenguaje y Comunicación - Convenio Inst. Pedagógicos	CARCHI	BUE
PADRE	NO TRABAJA	MUY BAJO	FICA	Ingeniería en Mantenimiento Automotriz	CARCHI	FAL
MADRE	NO TRABAJA	BAJO	FICA	MECATRÓNICA	CARCHI	FAL
PADRE	NO TRABAJA	BAJO	FACAE	MERCADOTECNIA	IMBABURA	SUF
SOLO	TRABAJA	ALTO	FECYT	Licenciatura en ENTRENAMIENTO DEPORTIVO	CARCHI	BUE
FAMILIAR	TRABAJA	MUY BAJO	FECYT	Licenciatura en ENTRENAMIENTO DEPORTIVO	CARCHI	BUE
MADRE	NO TRABAJA	MEDIO	FICA	TELECOMUNICACIONES	PICHINCHA	INSU
PADRES	NO TRABAJA	MUY BAJO	FICAYA	AGROPECUARIA	CARCHI	INSU
MADRE	NO TRABAJA	BAJO	FICA	TEXTILES	CARCHI	SUF
MADRE	NO TRABAJA	ALTO	FCCSS	Licenciatura en Enfermería	CARCHI	SUF
MADRE	NO TRABAJA	MUY BAJO	FICAYA	RECURSOS NATURALES RENOVABLES Renovables	CARCHI	BUE
SOLO	TRABAJA	MEDIO	FECYT	PARVULARIA	IMBABURA	EXC
PADRES	NO TRABAJA	BAJO	FICA	TELECOMUNICACIONES	CARCHI	INSU
PADRES	NO TRABAJA	ALTO	FECYT	RELACIONES PÚBLICAS	CARCHI	EXC
SOLO	TRABAJA	MUY ALTO	FECYT	Licenciatura en ENTRENAMIENTO DEPORTIVO	TUNGURAHUA	BUE
MADRE	NO TRABAJA	MUY BAJO	FACAE	CONTABILIDAD SUPERIOR Y AUDITORÍA	IMBABURA	BUE
PADRES	TRABAJA	MUY BAJO	FICA	SOFTWARE	IMBABURA	SUF
MADRE	NO TRABAJA	MUY BAJO	FACAE	ADMINISTRACIÓN DE EMPRESAS	CARCHI	FAL
SOLO	TRABAJA	MUY BAJO	FECYT	DISEÑO GRÁFICO	CARCHI	BUE
MADRE	NO TRABAJA	MUY BAJO	FICA	INGENIERÍA INDUSTRIAL	CARCHI	BUE
PADRES	NO TRABAJA	BAJO	FICA	INGENIERÍA INDUSTRIAL	CARCHI	SUF
PADRE	NO TRABAJA	ALTO	FICA	INGENIERÍA INDUSTRIAL	CARCHI	BUE
PADRES	NO TRABAJA	BAJO	FICA	SOFTWARE	CARCHI	FAL
PADRE	NO TRABAJA	MUY BAJO	FACAE	ADMINISTRACIÓN DE EMPRESAS	CARCHI	INSU

Fig. 24. Error de transformación en la Clase CARRERA

En el campo promedio se encontraron 59 registros que no tienen promedio, por esta razón a los estudiantes que se encuentran inactivos en la carrera se le asignó un promedio INSUFICIENTE, mientras que, al resto, se le asignó la media de la clase que es promedio SUFICIENTE.

2.7. Fase de minería de datos

Una vez consolidada la vista minable, se procede a transformarla en un archivo con extensión *.csv (comma-separated values), ya que en esta fase del proceso KDD se emplea la herramienta Weka. A continuación, en la Fig. 25 se observan los datos en formato *.csv:

```
RANGO_EDAD,TIPO_SANGRE,PAIS_NACIONALIDAD,GENERO,ETNIA,RANGO_DISCAPACIDAD,ESTADO_CIVIL,CONVIVIENTE,TIPO_VIVIENDA,FINANCIAMIENTO,ACTIVIDAD
_ESTUDIANTE,RANGO_INGRESO_MENSUAL,FACULTAD,PROVINCIA_PROCEDENCIA,RANGO_PROMEDIO,MOTIVO_ABANDONO,ESTADO_CARRERA
BAJA,O+,ECUADOR,MASCULINO,AFRODESCENDIENTE,NO TIENE,SOLTERO,PADRES,ARRENDADA,PADRES,NO TRABAJA,MUY ALTO,FICA,AZUAY,SUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,MADRE,ARRENDADA,MADRE,NO TRABAJA,BAJO,FECYT,AZUAY,EXCELENTE,NINGUNO,A
BAJA,A-,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,MADRE,ARRENDADA,MADRE,NO TRABAJA,MUY BAJO,FACAE,IMBABURA,BUENO,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRE,ARRENDADA,PADRES,NO TRABAJA,MUY ALTO,FECYT,AZUAY,BUENO,NINGUNO,A
MEDIA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,FAMILIAR,ARRENDADA,SOLO,TRABAJA,BAJO,FECYT,IMBABURA,EXCELENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,CASADO,PAREJA,ARRENDADA,PAREJA,NO TRABAJA,MUY BAJO,FACAE,AZUAY,EXCELENTE,NINGUNO,A
MEDIA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRES,PROPIA,SOLO,NO TRABAJA,MEDIO,FACAE,IMBABURA,EXCELENTE,NINGUNO,I
BAJA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,PADRE,HIPOTECADA,PADRE,NO TRABAJA,ALTO,FICA,PICHINCHA,SUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRE,PROPIA,PADRE,NO TRABAJA,BAJO,FECYT,IMBABURA,BUENO,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,SOLO,PROPIA,MADRE,NO TRABAJA,MUY BAJO,FCCSS,AZUAY,BUENO,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,LEVE,SOLTERO,SOLO,ARRENDADA,PADRES,NO TRABAJA,MUY BAJO,FECYT,AZUAY,BUENO,NINGUNO,A
BAJA,A+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,MADRE,PROPIA,MADRE,NO TRABAJA,BAJO,FECYT,IMBABURA,BUENO,NINGUNO,A
MEDIA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,PADRES,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICA,AZUAY,SUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRES,PROPIA,PADRES,NO TRABAJA,BAJO,FACAE,AZUAY,BUENO,NINGUNO,A
ALTA,O+,CUBA,FEMENINO,MESTIZO,NO TIENE,DIVORCIADO,MADRE,PROPIA,SOLO,TRABAJA,MUY BAJO,FECYT,CUBA,BUENO,NINGUNO,I
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRES,PROPIA,PADRE,NO TRABAJA,MUY ALTO,FICAYA,IMBABURA,BUENO,NINGUNO,I
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,PADRES,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICAYA,IMBABURA,SUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,SOLO,ARRENDADA,MADRE,NO TRABAJA,BAJO,FECYT,BOLÍVAR,BUENO,NINGUNO,A
BAJA,O+,ECUADOR,MASCULINO,INDÍGENA,NO TIENE,SOLTERO,SOLO,PROPIA,PADRES,NO TRABAJA,BAJO,FICAYA,BOLÍVAR,BUENO,NINGUNO,A
BAJA,A+,ECUADOR,MASCULINO,INDÍGENA,NO TIENE,SOLTERO,FAMILIAR,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICA,BOLÍVAR,SUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,FEMENINO,INDÍGENA,NO TIENE,SOLTERO,SOLO,PROPIA,PADRE,NO TRABAJA,MEDIO,FICA,BOLÍVAR,FALTA,NINGUNO,A
BAJA,A+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,SOLO,ARRENDADA,PADRE,NO TRABAJA,BAJO,FICAYA,BOLÍVAR,INSUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,MADRE,PROPIA,MADRE,NO TRABAJA,ALTO,FACAE,AZUAY,BUENO,NINGUNO,A
BAJA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,FAMILIAR,ARRENDADA,PADRES,NO TRABAJA,MEDIO,FICA,MORONA
SANTIAGO,INSUFICIENTE,NINGUNO,A
BAJA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,FAMILIAR,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICA,MORONA SANTIAGO,INSUFICIENTE,NINGUNO,A
BAJA,A+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,SOLO,ARRENDADA,PADRES,NO TRABAJA,ALTO,FICA,MORONA SANTIAGO,SUFICIENTE,NINGUNO,A
ALTA,A+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,DIVORCIADO,PAREJA,ARRENDADA,SOLO,TRABAJA,MUY BAJO,FACAE,CARCHI,INSUFICIENTE,NINGUNO,A
ALTA,A+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,PADRE,PROPIA,PADRE,NO TRABAJA,BAJO,FECYT,CARCHI,SUFICIENTE,NINGUNO,A
ALTA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,UNION DE HECHO,PAREJA,PROPIA,SOLO,TRABAJA,MUY BAJO,FECYT,SUCUMBÍOS,BUENO,NINGUNO,I
ALTA,O+,ECUADOR,FEMENINO,NO ASIGNADO,NO TIENE,CASADO,PAREJA,PRESTADA,SOLO,TRABAJA,MUY BAJO,FACAE,CARCHI,BUENO,NINGUNO,A
ALTA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,CASADO,PAREJA,ARRENDADA,SOLO,TRABAJA,MUY BAJO,FECYT,PICHINCHA,BUENO,NINGUNO,I
ALTA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,CASADO,PAREJA,ARRENDADA,SOLO,TRABAJA,MUY BAJO,FECYT,IMBABURA,BUENO,NINGUNO,A
ALTA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,CASADO,PAREJA,PROPIA,SOLO,NO TRABAJA,MUY BAJO,FACAE,PICHINCHA,BUENO,NINGUNO,A
ALTA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,SOLO,ARRENDADA,SOLO,TRABAJA,BAJO,FICA,CARCHI,INSUFICIENTE,NINGUNO,I
ALTA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,CASADO,PAREJA,ARRENDADA,SOLO,NO TRABAJA,ALTO,FACAE,CARCHI,EXCELENTE,NINGUNO,A
MEDIA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,PAREJA,PROPIA,SOLO,TRABAJA,ALTO,FACAE,IMBABURA,BUENO,NINGUNO,A
ALTA,O+,ECUADOR,MASCULINO,NO ASIGNADO,NO TIENE,CASADO,PAREJA,ARRENDADA,SOLO,TRABAJA,MEDIO,FACAE,CARCHI,BUENO,NINGUNO,A
MEDIA,O+,ECUADOR,MASCULINO,MESTIZO,NO TIENE,SOLTERO,MADRE,PROPIA,MADRE,NO TRABAJA,MUY BAJO,FACAE,CARCHI,INSUFICIENTE,NINGUNO,A
MEDIA,O+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,CASADO,PAREJA,PROPIA,PAREJA,NO TRABAJA,BAJO,FACAE,IMBABURA,BUENO,NINGUNO,A
MEDIA,A+,ECUADOR,FEMENINO,MESTIZO,NO TIENE,SOLTERO,SOLO,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICA,CARCHI,SUFICIENTE,NINGUNO,A
MEDIA,O+,ECUADOR,MASCULINO,NO ASIGNADO,NO TIENE,SOLTERO,SOLO,ARRENDADA,PADRES,NO TRABAJA,BAJO,FICA,CARCHI,FALTA,NINGUNO,A
ALTA.O+.ECUADOR.MASCULINO.MESTIZO.NO TIENE.CASADO.PAREJA.ANTICRESIS.SOLO.TRABAJA.ALTO.FECYT.CARCHI.BUENO.NINGUNO.I
```

Fig. 25. Datos en formato *.csv

2.7.1. Clasificación

Árboles de decisión

Para realizar la clasificación se consideró la técnica de árboles de decisión con el algoritmo Random Tree, debido a la simplicidad de su modelo, fácil interpretación y velocidad para clasificar nuevos datos (Vila et al., 2018). De igual forma se consideró este algoritmo puesto q varias de los datos personales y académicos de los estudiantes que se abordan en este estudio ya se han considerado en trabajos pasados y se busca analizar cómo influyen los datos socioeconómicos en la predicción de la deserción. De igual forma se consideró el algoritmo RandomForest, ya que se cuenta con antecedentes que reportan buenos resultados al emplear este algoritmo, tales como Lehr (2016) y Sadiq (2018).

Naive Bayes

Al igual que en el caso anterior, se escogió esta técnica, para analizar a futuro cómo se comportan las variables con respecto a más atributos, e identificar si los patrones predichos varían de acuerdo con datos socio económicos. Adicionalmente, por la naturaleza del algoritmo permite interpretar la información de acuerdo a cómo se relacionan las variables de análisis, que en ciertas ocasiones se pueden interpretar como relaciones de causa efecto (Sierra Araujo, 2006).

2.7.2. Regresión

Regresión Logística Binomial

La regresión logística binomial emplea una variable dependiente cualitativa; por ello se ha empleado las técnicas de clasificación basadas en este planteamiento, ya que la variable a discriminar es ESTADO_CARRERA con los valores ACTIVO/ INACTIVO (Sierra Araujo, 2006). Generalmente, el análisis de regresión se emplea para modelar la relación entre variables y realizar predicciones a partir de variables explicativas y variables de análisis, descartando aquellas variables que no aportan información (Madrid, 2017).

CAPÍTULO 2

Validación de resultados

3.1. Fase de evaluación e interpretación

Para evaluar los algoritmos predictivos aplicados (clasificación y regresión) se emplearon métricas cuantitativas de calidad basadas en la matriz de confusión.

3.1.1. Evaluación de Tareas de Clasificación

- **Árboles de decisión**

RandomTree

En las Fig. 26 y Fig. 27 se muestran los resultados del algoritmo RandomTree, con Validación Cruzada de 10, 17 atributos conteniendo 11200 instancias. Parta visualizar las reglas de decisión completas visitar el siguiente enlace: http://bit.ly/2Op5gMo

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     VISTA-MINABLE-FACULTADES-ARBOLES-DE-DECISION
Instances:    11200
Attributes:   17
              RANGO_EDAD
              TIPO_SANGRE
              PAIS_NACIONALIDAD
              GENERO
              ETNIA
              RANGO_DISCAPACIDAD
              ESTADO_CIVIL
              CONVIVIENTE
              TIPO_VIVIENDA
              FINANCIAMIENTO
              ACTIVIDAD_ESTUDIANTE
              RANGO_INGRESO_MENSUAL
              FACULTAD
              PROVINCIA_PROCEDENCIA
              RANGO_PROMEDIO
              MOTIVO_ABANDONO
              ESTADO_CARRERA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomTree
==========

MOTIVO_ABANDONO = NINGUNO
|   FACULTAD = FICA
|   |   RANGO_EDAD = BAJA
|   |   |   RANGO_PROMEDIO = SUFICIENTE
|   |   |   |   FINANCIAMIENTO = PADRES
|   |   |   |   |   CONVIVIENTE = PADRES
|   |   |   |   |   |   TIPO_VIVIENDA = ARRENDADA : A (32/0)
|   |   |   |   |   |   TIPO_VIVIENDA = PROPIA
|   |   |   |   |   |   |   TIPO_SANGRE = O+
|   |   |   |   |   |   |   |   ETNIA = AFRODESCENDIENTE : A (4/0)
|   |   |   |   |   |   |   |   ETNIA = MESTIZO
|   |   |   |   |   |   |   |   |   RANGO_INGRESO_MENSUAL = MUY ALTO : A (7/0)
|   |   |   |   |   |   |   |   |   RANGO_INGRESO_MENSUAL = BAJO
|   |   |   |   |   |   |   |   |   |   GENERO = MASCULINO
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = AZUAY : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = IMBABURA : A (30/1)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = PICHINCHA : A (4/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = CUBA : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = BOLÍVAR : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = MORONA SANTIAGO : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = CARCHI : A (1/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = SUCUMBÍOS : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = TUNGURAHUA : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = NO ASIGNADO : A (1/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = COLOMBIA : A (0/0)
|   |   |   |   |   |   |   |   |   |   |   PROVINCIA_PROCEDENCIA = ESPAÑA : A (0/0)
```

Fig. 26. Primera parte de resultados del algoritmo RandomTree

```
Size of the tree : 13603

Time taken to build model: 0.05 seconds

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

Correctly Classified Instances        9846               87.9107 %
Incorrectly Classified Instances      1354               12.0893 %
Kappa statistic                          0.1799
K&B Relative Info Score            -220260.5761 %
K&B Information Score                -1018.2522 bits     -0.0909 bits/instance
Class complexity | order 0            5176.2623 bits      0.4622 bits/instance
Class complexity | scheme          1220081.3961 bits    108.9358 bits/instance
Complexity improvement     (Sf)    -1214905.1339 bits   -108.4737 bits/instance
Mean absolute error                      0.1469
Root mean squared error                  0.3495
Relative absolute error                 83.1755 %
Root relative squared error            117.6279 %
Total Number of Instances            11200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,953    0,803    0,916      0,953    0,934      0,186    0,599     0,920     A
                 0,197    0,047    0,313      0,197    0,242      0,186    0,599     0,161     I
Weighted Avg.    0,879    0,729    0,857      0,879    0,867      0,186    0,599     0,846

=== Confusion Matrix ===

    a    b   <-- classified as
 9630  474 |    a = A
  880  216 |    b = I
```

Fig. 27. Segunda parte de resultados del algoritmo RandomTree

En la Tabla 3.1 se aprecia la matriz de confusión (Fig. 27) que se obtuvo después de ejecutar el algoritmo RandomTree, con 17 atributos, 11200 instancias y validación cruzada de 10-iteraciones.

TABLA 3.1
MATRIZ DE CONFUSIÓN DEL ALGORITMO RANDOMTREE

		Clase predicha	
		A	I
Clase verdadera	A	9630	474
	I	880	216

Fuente: Resultados del software Weka

Donde:

- TP = 9630
- TN = 216
- FN = 474
- FP =880
- Número total de instancias = 11200

Además de la matriz de confusión existen otras métricas de calidad derivadas de ésta. A continuación, en la Tabla 3.2 se aprecian dichas medidas estadísticas:

TABLA 3.2
MEDIDAS ESTADÍSTICAS DE CALIDAD DE RANDOMTREE

MEDIDA	**VALOR**
Tasa de error	12.0893%

Sensibilidad	95.3088%
Especificidad	19.7080%
Accuracy	87.9107%
Coeficiente Kappa	0.1799
Curva ROC	0.599
Precisión	85.7%
Recall	87.9%
TP Rate	87.9%
FP Rate	72.9%
F – Measure	86.7%

Fuente: Propia

RandomForest

A continuación, en las Fig. 28 y 29 se muestran los resultados obtenidos después de aplicar el algoritmo RandomForest en la herramienta Weka. Para visualizar las reglas de decisión completas visitar el siguiente enlace: http://bit.ly/2HFme8Q

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -print -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -depth 2
Relation:     VISTA-MINABLE-FACULTADES-ARBOLES-DE-DECISION
Instances:    11200
Attributes:   17
              RANGO_EDAD
              TIPO_SANGRE
              PAIS_NACIONALIDAD
              GENERO
              ETNIA
              RANGO_DISCAPACIDAD
              ESTADO_CIVIL
              CONVIVIENTE
              TIPO_VIVIENDA
              FINANCIAMIENTO
              ACTIVIDAD_ESTUDIANTE
              RANGO_INGRESO_MENSUAL
              FACULTAD
              PROVINCIA_PROCEDENCIA
              RANGO_PROMEDIO
              MOTIVO_ABANDONO
              ESTADO_CARRERA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -depth 2 -do-not-check-capabilitiesAll the base classifiers:


RandomTree
==========

MOTIVO_ABANDONO = NINGUNO
|   RANGO_PROMEDIO = SUFICIENTE : A (3764/189)
|   RANGO_PROMEDIO = EXCELENTE : A (673/115)
|   RANGO_PROMEDIO = BUENO : A (4719/456)
|   RANGO_PROMEDIO = FALTA : A (611/134)
|   RANGO_PROMEDIO = INSUFICIENTE : A (1241/83)
MOTIVO_ABANDONO = CAMBIO_C : I (76/0)
MOTIVO_ABANDONO = PIERDE TERCERA MATRÍCULA
|   RANGO_EDAD = BAJA : A (34/11)
|   RANGO_EDAD = MEDIA : I (81/22)
|   RANGO_EDAD = ALTA : I (1/0)

Size of the tree : 12
Max depth of tree: 2
```

Fig. 28. Primera parte de resultados de RandomForest

```
Size of the tree : 27
Max depth of tree: 2



Time taken to build model: 0.28 seconds

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

Correctly Classified Instances       10168               90.7857 %
Incorrectly Classified Instances      1032                9.2143 %
Kappa statistic                          0.1019
K&B Relative Info Score             -42759.3773 %
K&B Information Score                 -197.6742 bits     -0.0176 bits/instance
Class complexity | order 0            5176.2623 bits      0.4622 bits/instance
Class complexity | scheme             4589.6542 bits      0.4098 bits/instance
Complexity improvement     (Sf)        586.6081 bits      0.0524 bits/instance
Mean absolute error                      0.1639
Root mean squared error                  0.2802
Relative absolute error                 92.8109 %
Root relative squared error             94.3092 %
Total Number of Instances            11200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,941    0,907      1,000    0,951      0,230    0,748     0,958     A
                 0,059    0,000    0,985      0,059    0,112      0,230    0,748     0,383     I
Weighted Avg.    0,908    0,849    0,915      0,908    0,869      0,230    0,748     0,901

=== Confusion Matrix ===

     a     b   <-- classified as
 10103     1 |     a = A
  1031    65 |     b = I
```

Fig. 29. Segunda parte de resultados de RandomForest

En la Tabla 3.3 se aprecia la matriz de confusión (Fig. 29) que se obtuvo después de ejecutar el algoritmo RandomForest, con 17 atributos, 11200 instancias y validación cruzada de 10-iteraciones.

TABLA 3.3
MATRIZ DE CONFUSIÓN DEL ALGORITMO RANDOMFOREST

		Clase predicha	
		A	I
Clase verdadera	A	10103	1
	I	1031	65

Fuente: Resultados del software Weka

Donde:

- TP = 10103
- TN = 65
- FN = 1
- FP = 1031
- Número total de instancias = 11200

A continuación, en la Tabla 3.4 se aprecian las medidas estadísticas que se derivan de la matriz de confusión:

TABLA 3.4
MEDIDAS ESTADÍSTICAS DE CALIDAD DE RANDOMFOREST

MEDIDA	**VALOR**

Tasa de error	9.2143%
Sensibilidad	99.9901%
Especificidad	5.9307%
Accuracy	90.7857%
Coeficiente Kappa	0.1019
Curva ROC	0.748
Precisión	91.5%
Recall	90.8%
TP Rate	90.8%
FP Rate	84.9%
F – Measure	86.9%

Fuente: Propia

- **Naive Bayes**

En las Fig. 30 y 31 se observan los resultados después de procesar los datos con el algoritmo NaiveBayes bajo las mismas condiciones que los árboles de decisiones, para ver las probabilidades completas visitar el siguiente enlace: http://bit.ly/2uuGmSm

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     VISTA-MINABLE-FACULTADES-ARBOLES-DE-DECISION
Instances:    11200
Attributes:   17
              RANGO_EDAD
              TIPO_SANGRE
              PAIS_NACIONALIDAD
              GENERO
              ETNIA
              RANGO_DISCAPACIDAD
              ESTADO_CIVIL
              CONVIVIENTE
              TIPO_VIVIENDA
              FINANCIAMIENTO
              ACTIVIDAD_ESTUDIANTE
              RANGO_INGRESO_MENSUAL
              FACULTAD
              PROVINCIA_PROCEDENCIA
              RANGO_PROMEDIO
              MOTIVO_ABANDONO
              ESTADO_CARRERA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                                    Class
Attribute                             A       I
                                    (0.9)   (0.1)
=====================================================
RANGO_EDAD
  BAJA                             7526.0   526.0
  MEDIA                            2505.0   550.0
  ALTA                               76.0    23.0
  [total]                         10107.0  1099.0


TIPO_SANGRE
  O+                               4973.0   583.0
  A-                                 30.0     5.0
  A+                               4136.0   385.0
  B+                                500.0    49.0
  O-                                151.0     9.0
  NO ASIGNADO                       217.0    52.0
  AB+                                76.0    12.0
  AB-                                10.0     4.0
  B-                                 20.0     6.0
  [total]                         10113.0  1105.0
```

Fig. 30.Primera parte de resultados del algoritmo NaiveBayes

```
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances       10156               90.6786 %
Incorrectly Classified Instances      1044                9.3214 %
Kappa statistic                          0.224
K&B Relative Info Score            -172664.7848 %
K&B Information Score                 -798.2195 bits     -0.0713 bits/instance
Class complexity | order 0           5176.2623 bits      0.4622 bits/instance
Class complexity | scheme            4538.0177 bits      0.4052 bits/instance
Complexity improvement     (Sf)       638.2446 bits      0.057  bits/instance
Mean absolute error                      0.1524
Root mean squared error                  0.2796
Relative absolute error                 86.2607 %
Root relative squared error             94.1105 %
Total Number of Instances            11200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,987    0,835    0,916      0,987    0,950      0,276    0,739     0,956     A
                 0,165    0,013    0,584      0,165    0,257      0,276    0,739     0,346     I
Weighted Avg.    0,907    0,754    0,883      0,907    0,882      0,276    0,739     0,896

=== Confusion Matrix ===

    a    b   <-- classified as
 9975  129 |    a = A
  915  181 |    b = I
```

Fig. 31. Segunda parte de resultados del algoritmo NaiveBayes

En la Tabla 3.5 se especifica la matriz de confusión (Fig. 31) que se obtuvo después de ejecutar el algoritmo NaiveBayes, con 17 atributos, 11200 instancias y validación cruzada de 10-iteraciones.

TABLA 3.5
MATRIZ DE CONFUSIÓN DEL ALGORITMO NAIVEBAYES

		Clase predicha	
		A	I
Clase verdadera	A	9975	129
	I	915	181

Fuente: Resultados del software Weka

Donde:

- TP = 9975
- TN = 181
- FN = 129
- FP =915
- Número total de instancias = 11200

Al igual que en el algoritmo anterior, se obtienen varias medidas de calidad, a continuación, en la Tabla 3.6 se aprecian las medidas estadísticas que se derivan de la matriz de confusión:

TABLA 3.6
MEDIDAS ESTADÍSTICAS DE CALIDAD DE NAIVEBAYES

MEDIDA	**VALOR**
Tasa de error	9.3214%
Sensibilidad	98.7233%
Especificidad	16.5146%
Accuracy	90.6786%
Coeficiente Kappa	0.2240
Curva ROC	0.739
Precisión	88.3%
Recall	90.7%
TP Rate	90.7%
FP Rate	75.4%
F - Measure	88.2%

Fuente: Propia

3.1.2. Evaluación de Tareas de Regresión

- **Regresión Logística**

En las Fig. 32 y 33 se aprecia el resultado del algoritmo Logistic, con Validación Cruzada de 10-iteraciones, 17 atributos y 11200 instancias. Parta visualizar los resultados completos visitar el siguiente enlace: http://bit.ly/2FAzsBx

```
=== Run information ===

Scheme:       weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation:     VISTA-MINABLE-FACULTADES-REGRESION-LOGISTICA
Instances:    11200
Attributes:   17
              RANGO_EDAD
              TIPO_SANGRE
              PAIS_NACIONALIDAD
              GENERO
              ETNIA
              RANGO_DISCAPACIDAD
              ESTADO_CIVIL
              CONVIVIENTE
              TIPO_VIVIENDA
              FINANCIAMIENTO
              ACTIVIDAD_ESTUDIANTE
              RANGO_INGRESO_MENSUAL
              FACULTAD
              PROVINCIA_PROCEDENCIA
              RANGO_PROMEDIO
              MOTIVO_ABANDONO
              ESTADO_CARRERA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                                                                      Class
Variable                                                                A
================================================================================
RANGO_EDAD=BAJA                                                       0.5403
RANGO_EDAD=MEDIA                                                     -0.4988
RANGO_EDAD=ALTA                                                      -1.2013
TIPO_SANGRE=O+                                                       -0.0439
TIPO_SANGRE=A-                                                       -0.3936
```

Fig. 32. Primera parte de resultados del algoritmo Logistic

```
Time taken to build model: 6.03 seconds

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

Correctly Classified Instances       10215               91.2054 %
Incorrectly Classified Instances       985                8.7946 %
Kappa statistic                          0.2118
Mean absolute error                      0.1483
Root mean squared error                  0.2732
Relative absolute error                 83.9867 %
Root relative squared error             91.9655 %
Total Number of Instances            11200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,996    0,862    0,914      0,996    0,953      0,307    0,745     0,955     A
                 0,138    0,004    0,791      0,138    0,235      0,307    0,745     0,374     I
Weighted Avg.    0,912    0,778    0,902      0,912    0,883      0,307    0,745     0,898

=== Confusion Matrix ===

     a     b   <-- classified as
 10064    40 |     a = A
   945   151 |     b = I
```

Fig. 33. Segunda parte de resultados del algoritmo Logistic

En la Tabla 3.7 se especifica la matriz de confusión (Fig. 33) que se obtuvo después de ejecutar el algoritmo Logistic, con 17 atributos, 11200 instancias y validación cruzada de 10-iteraciones.

TABLA 3.7
MATRIZ DE CONFUSIÓN DEL ALGORITMO LOGISTIC

		Clase predicha	
		A	I
Clase verdadera	A	10064	40
	I	945	151

Fuente: Resultados del software Weka

Donde:

- TP = 10064
- TN = 151
- FN = 40
- FP =945
- Número total de instancias = 11200

En la Tabla 3.8 se aprecian las medidas estadísticas que se derivan de la matriz de confusión después de ejecutar el algoritmo Logistic:

TABLA 3.8
MEDIDAS ESTADÍSTICAS DE CALIDAD DE LOGISTIC

MEDIDA	**VALOR**
Tasa de error	8.7946%
Sensibilidad	99.6041%
Especificidad	13,7774%
Accuracy	91,2054%
Coeficiente Kappa	0,2118
Curva ROC	0.745
Precisión	90.2%
Recall	91.2%
TP Rate	91.2%
FP Rate	77.8%
F - Measure	88.3%

Fuente: Propia

3.2. Análisis e interpretación de resultados

3.2.1. Análisis e interpretación de resultados de las tareas de clasificación

- **Análisis de resultados**

Se evaluaron cuantitativamente los resultados de los algoritmos de clasificación RandomTree, RandomForest y NaiveBayes, mediante la validación cruzada con valor de 10-iteraciones, precisión, tasa de error, coeficiente Kappa, tasa de TP, tasa de FT, puntaje de F y el área bajo la curva (Curva ROC) (Vila et al., 2018). Adicionalmente, se tomó en cuenta otras métricas de calidad, tales como la sensibilidad y la especificidad.

En la Tabla 3.2, Tabla 3.4 y Tabla 3.6 se observa que los clasificadores cuentan con valores similares, excepto en la tasa de error, este hecho se da gracias a que los datos que fueron procesados son no balanceados (Saito & Rehmsmeier, 2015), es decir, para una categoría de la variable de clase existen más registros que para la otra categoría, en este caso para el atributo de clase ESTADO_CARRERA la categoría predominante es A, que pertenece a los estudiantes activos, existen 10104 registros, mientras que para la categoría I, inactivos, existen 1096 registros. A partir de la Tabla 3.1, Tabla 3.3 y Tabla 3.5 que son las matrices de confusión de los algoritmos RandomTree, RandomForest y NaiveBayes respectivamente, y considerando lo anteriormente expuesto se tiene que:

El algoritmo RandomTree clasificó 9846 instancias de forma correcta, mientras que 1354 instancias son clasificadas incorrectamente, entonces se dice:

- Para la categoría A (activo) el algoritmo clasificó correctamente 9630 registros y de forma errónea 474 registros.
- Para la categoría I (inactivo) el algoritmo clasificó correctamente 216 registros mientras que incorrectamente clasifica 880.

El algoritmo RandomForest clasificó correctamente 10168 y de forma incorrecta 1032 registros, donde:

- Para la categoría A el algoritmo clasificó correctamente 10103 registros y de forma incorrecta 1 registro.
- Para la categoría I el algoritmo clasificó correctamente 65 registros y de forma incorrecta 1031 registros.

El algoritmo NaiveBayes clasificó correctamente 10156 y de forma incorrecta 1044 registros, donde:

- Para la categoría A el algoritmo clasificó correctamente 9975 registros mientras que de forma incorrecta 129 registros.
- Para la categoría I el algoritmo clasificó correctamente 181 registros mientras que incorrectamente 915 registros.

Para los tres casos se aprecia que para la categoría I (inactivo) los datos toman una tendencia a clasificar incorrectamente ya que existen más datos en la categoría A (activo), por lo que se expuso anteriormente acerca de los datos no balanceados. En este caso, que las métricas de calidad son similares para los tres clasificadores, se considera los tipos de errores de los cuales RandomTree tiene 880 errores tipo I y 474 errores tipo II, RandomForest tiene 1031 errores tipo I y un error de tipo II y NaiveBayes tiene 915 errores de tipo I y 129 errores de tipo II. Tomando en cuenta que los errores de tipo II son los peores, el algoritmo RandomForest es el mejor clasificador para este caso, por ello se empleará la información obtenida por medio de este algoritmo que proporcionan un conjunto de reglas de decisión (que son equivalentes al árbol de decisión) que proporcionan un fácil procesamiento e interpretación de la información, mientras que NaiveBayes no genera un modelo sino que realiza la clasificación en el momento que se solicita, es decir en tiempo real (Vila et al., 2018).

- **Interpretación de resultados**

Una vez efectuado el análisis de los resultados y determinado el mejor modelo predictivo, árbol y reglas de decisión con el algoritmo RandomForest. La interpretación de datos se realizó desde la raíz hacia las hojas cuyo atributo de clase sea ESTADO_CARRERA = I (Pajares & De La Cruz, 2010), identificando los siguientes patrones de deserción estudiantil a partir del algoritmo RandomForest en la UTN:

- Estudiantes con cambio de carrera
- Estudiantes que pierden tercera matrícula
- Estudiantes de la FICAYA o FICA con promedio excelente
- Estudiantes con promedio excelente que proceden de la provincia de Bolívar o Zamora Chinchipe
- Estudiantes que provienen de Sucumbíos y que cambiaron su carrera
- Estudiantes que provienen de Tungurahua que viven con su pareja
- Estudiantes de Colombia que viven solos
- Estudiantes de la provincia de Orellana con promedio excelente
- Estudiantes de Zamora Chinchipe que viven solo con su padre
- Estudiantes que proceden de Cuba que viven solos
- Estudiantes de Colombia de edad alta

- Estudiantes de Manabí que son casados
- Estudiantes cuyo promedio es medio y viven con su pareja
- Estudiantes con edad baja por cambio de carrera
- Estudiantes de Cuba que están en la FECYT
- Estudiantes de edad alta en la FECYT
- Estudiantes de la FCCSS con tipo de sangre AB-
- Estudiantes de tipo de sangre A- de Imbabura
- Estudiantes de edad media de Manabí
- Estudiantes de la FICA de edad alta
- Estudiantes con sangre AB- cuya casa es propia o prestada
- Estudiantes cubanas de sexo femenino
- Estudiantes de Carchi de edad alta
- Estudiantes de Tungurahua con ingresos económicos mensuales muy altos
- Estudiantes de Zamora Chinchipe con ingresos económicos mensuales medios
- Estudiantes divorciados que viven solo con su padre
- Estudiantes viudos que viven con sus padres
- Estudiantes de etnia montubio que viven solos
- Estudiantes de Loja que estudian en FCCSS
- Estudiantes de etnia afrodescendiente de la provincia de Orellana
- Estudiantes de Santo Domingo de los Tsáchilas que viven solo con su madre o con un familiar
- Estudiantes de sexo femenino de Zamora Chinchipe
- Estudiantes de Santo Domingo de los Tsáchilas con discapacidad leve
- Estudiantes que viven en unión de hecho de Colombia
- Estudiantes viudos con promedio suficiente
- Estudiantes con ingreso económico mensual muy alto que cambiaron su carrera
- Estudiantes con ingreso económico mensual bajo que cambiaron su carrera
- Estudiantes de edad alta de Cuba, Sucumbíos, Colombia o Esmeraldas
- Estudiantes con promedio medio cuya vivienda es adquirida en anticresis
- Estudiantes que provienen de Loja y viven con sus padres

Como ya se ha mencionado RandomForest realiza un bosque compuesto de varios RandomTree, y varios de los patrones obtenidos son repetitivos; así se formaron los siguientes patrones finales a partir de RandomForest:

- Estudiantes de la FICAYA o FICA con promedio excelente

- Estudiantes divorciados o viudos que viven con sus padres
- Estudiantes con promedio excelente que proceden de la provincia de Bolívar, Orellana o Zamora Chinchipe, que tienen ingresos económicos mensuales suficientes para la canasta básica familiar.
- Estudiantes de etnia mestiza o afrodescendiente que pierden tercera matrícula
- Estudiantes que cambiaron su carrera y que viven con sus padres, solo con su madre o solos
- Estudiantes de Colombia, Carchi, Sucumbíos, o Esmeraldas que viven solos o de edad alta
- Estudiantes de edad media que son casados
- Estudiantes de edad alta en la FICA, FECYT, FCCSS
- Estudiantes cuyo promedio es medio y viven con su pareja
- Estudiantes que pierden tercera matrícula con tipo de sangre A+ o B+

3.2.2. Análisis e interpretación de resultados de las tareas de regresión

- **Análisis de resultados**

Una vez evaluados cuantitativamente el algoritmo de regresión logística, Logistic, mediante la validación cruzada con valor de 10-iteraciones, precisión, tasa de error, coeficiente Kappa, tasa de TP, tasa de FT, puntaje de F y el área bajo la curva (Curva ROC) (Vila et al., 2018). A partir de la Tabla 3.8 que es la matriz de confusión del algoritmo Logistic, y considerando que los errores de tipo II son menores que los de tipo I, para clase ESTADO_CARRERA la categoría “A” (activo) es igual 10104 registros y la categoría “I” (inactivo) es igual a 1096, entonces se tiene que:

El algoritmo Logistic clasificó 10215 instancias de forma correcta, mientras que 985 instancias son clasificadas incorrectamente, entonces se dice:

- Para la categoría A (activo) el algoritmo clasificó correctamente 10064 registros y de forma errónea 40 registros.
- Para la categoría I (inactivo) el algoritmo clasificó correctamente 151 registros mientras que incorrectamente clasifica 945.

- **Interpretación de resultados**

Considerando que la regresión logística trabaja por medio de funciones matemáticas, se tiene que cuando el coeficiente que da como resultado Weka es positivo más probabilidades existen de que la predicción sea en la clase a la que se asignó el valor; por lo que se mencionó

anteriormente acerca de los datos no balanceados, la herramienta únicamente nos dio resultados en la clase “A”. Como se expresó anteriormente, el modelo posee valores independientes que se asumen como una influencia sobre el resultado, y tomando en cuenta que la aproximación típica en la regresión logística son los mínimos cuadrados ordinarios, que en este caso son incorrectos debido a que los errores de regresión en algunas variables son heterocedásticos y no son normales, es decir que la varianza de los errores no es constante en todas las observaciones realizadas, debido a que los datos no se encuentran balanceados, los resultados de probabilidad estimados cuyo coeficiente se encuentra por encima de 1 y por debajo de -1 carecen de sentido, por ello se tomó en cuenta únicamente las variables cuyos coeficientes estén en este rango (Mayra & Mauricio, 2018) que corresponden a las probabilidades de ocurrir una determinada categoría. A continuación, en la Tabla 3.9 se observan las categorías de los atributos cuyos coeficientes se emplearán para la interpretación de la regresión logística:

TABLA 3.9
COEFICIENTES CORRECTAMENTE ESTIMADOS POR EL ALGORITMO LOGISTIC

CATEGORÍA	COEFICIENTE
RANGO_EDAD=BAJA	0.5403
RANGO_EDAD=MEDIA	-0.4988
TIPO_SANGRE=O+	-0.0439
TIPO_SANGRE=A-	-0.3936
TIPO_SANGRE=A+	0.0462
TIPO_SANGRE=B+	0.1357
TIPO_SANGRE=O-	0.4594
TIPO_SANGRE=NO ASIGNADO	-0.3294
TIPO_SANGRE=AB+	-0.0952
TIPO_SANGRE=B-	-0.7839
GENERO=FEMENINO	-0.1112
ETNIA=AFRODESCENDIENTE	0.1185
ETNIA=MESTIZO	-0.3313
ETNIA=INDÍGENA	0.2211
ETNIA=NO ASIGNADO	0.9303
ETNIA=MONTUBIO	-0.1267
ESTADO_CIVIL=SOLTERO	0.036
ESTADO_CIVIL=CASADO	-0.1058
ESTADO_CIVIL=DIVORCIADO	0.0036
ESTADO_CIVIL=UNIÓN DE HECHO	0.1973

ESTADO_CIVIL=VIUDO	0.392
CONVIVIENTE=PADRES	-0.0623
CONVIVIENTE=MADRE	-0.1646
CONVIVIENTE=PADRE	0.0596
CONVIVIENTE=FAMILIAR	0.1679
CONVIVIENTE=PAREJA	0.187
CONVIVIENTE=SOLO	0.1168
CONVIVIENTE=OTROS	0.0723
TIPO_VIVIENDA=ARRENDADA	-0.1185
TIPO_VIVIENDA=PROPIA	-0.0914
TIPO_VIVIENDA=HIPOTECADA	0.5467
TIPO_VIVIENDA=PRESTADA	0.0906
TIPO_VIVIENDA=ANTICRESIS	0.426
TIPO_VIVIENDA=NO ASIGNADO	-0.1211
FINANCIAMIENTO=PADRES	-0.1902
FINANCIAMIENTO=MADRE	0.0252
FINANCIAMIENTO=SOLO	0.2635
FINANCIAMIENTO=PAREJA	0.2713
FINANCIAMIENTO=PADRE	-0.0625
FINANCIAMIENTO=FAMILIAR	0.2367
FINANCIAMIENTO=OTROS	-0.2987
ACTIVIDAD_ESTUDIANTE=TRABAJA	-0.3838
RANGO_INGRESO_MENSUAL=MUY ALTO	0.1605
RANGO_INGRESO_MENSUAL=BAJO	0.0077
RANGO_INGRESO_MENSUAL=MUY BAJO	-0.2144
RANGO_INGRESO_MENSUAL=MEDIO	0.0943
RANGO_INGRESO_MENSUAL=ALTO	0.1751
FACULTAD=FICA	-0.2305
FACULTAD=FECYT	0.0425
FACULTAD=FACAE	0.4389
FACULTAD=FCCSS	-0.76
FACULTAD=FICAYA	0.2125
PROVINCIA_PROCEDENCIA=IMBABURA	-0.8061
PROVINCIA_PROCEDENCIA=PICHINCHA	-0.2117
PROVINCIA_PROCEDENCIA=BOLÍVAR	-0.4112
PROVINCIA_PROCEDENCIA=CARCHI	-0.6392

PROVINCIA_PROCEDENCIA=SUCUMBÍOS	-0.6379
PROVINCIA_PROCEDENCIA=NO ASIGNADO	-0.2016
PROVINCIA_PROCEDENCIA=COLOMBIA	0.4284
PROVINCIA_PROCEDENCIA=LOJA	-0.1868
PROVINCIA_PROCEDENCIA=ESMERALDAS	-0.0472
PROVINCIA_PROCEDENCIA=SANTO DOMINGO DE LOS TSÁCHILAS	-0.9214
RANGO_PROMEDIO=SUFICIENTE	0.4571
RANGO_PROMEDIO=EXCELENTE	-0.8631
RANGO_PROMEDIO=BUENO	-0.2221
RANGO_PROMEDIO=FALTA	-0.7528
RANGO_PROMEDIO=INSUFICIENTE	0.382
MOTIVO_ABANDONO=PIERDE TERCERA MATRÍCULA	-0.7614

Fuente: Weka

De acuerdo con la Tabla 3.9 la interpretación de regresión logística queda de la siguiente manera:

EDAD

Cuando un estudiante tiene edad baja posee más posibilidades de permanecer activo en la carrera, mientras que cuando tiene edad media (o alta) las posibilidades disminuyen.

TIPO_SANGRE

Los estudiantes con los tipos de sangre A+, B+ o O- tienen más posibilidades de permanecer activos en su carrera mientras que los estudiantes con tipo de sangre O+, A-, AB+, B- o los que no registraron su tipo de sangre tienen menos posibilidades de continuar con su carrera universitaria.

GENERO

Las personas de género femenino poseen menos probabilidades de permanecer activas en su carrera universitaria.

ETNIA

Los estudiantes que pertenecen a las etnias afrodescendiente o indígena poseen más probabilidades de permanecer en su carrera, mientras que las personas de etnia mestiza o montubia tienen menos posibilidades de permanecer activos en su carrera universitaria.

ESTADO_CIVIL

Los estudiantes cuyo estado civil es soltero, divorciado, unión de hecho o viudo poseen mayor probabilidad de permanecer activos en su carrera universitaria, mientras que las personas casadas tienen mayor probabilidad de abandonar sus estudios.

CONVIVIENTE

Las personas que conviven con su padre, familiar, pareja, solo o con otras personas poseen mayor probabilidad de permanecer activos en sus estudios, mientras que las personas que viven con sus padres o madre tienden a abandonar sus estudios.

TIPO_VIVIENDA

Las personas que viven en una vivienda hipotecada, prestada o anticresis tienen mayor probabilidad de permanecer en su carrera, mientras que las personas que viven en casas arrendadas, propias o que no registran el tipo de vivienda tienen una probabilidad más baja de permanecer en su carrera universitaria.

FINANCIAMIENTO

Las personas cuyo financiamiento para sus estudios proviene de madre, solo, pareja o familiar tienen mayor probabilidad de permanecer activos en su carrera universitaria, mientras que las personas que sus estudios son financiados por ambos padres, solo padre u otros tienen menos probabilidad de permanecer en su carrera universitaria.

ACTIVIDAD_ESTUDIANTE

Los estudiantes que trabajan tienen menos probabilidades de permanecer activos en su carrera universitaria.

RANGO_INGRESO_MENSUAL

Las personas cuyo rango de ingreso mensual es muy alto, alto, medio o bajo, tienen mayor probabilidades de permanecer en su carrera universitaria, mientras que las personas cuyo rango de ingreso mensual es muy bajo tienen menos probabilidades de permanecer en su carrera universitaria.

FACULTAD

Los estudiantes que pertenecen a la FECYT, FACAE o FICAYA poseen mayor probabilidad de permanecer activos en su carrera universitaria, mientras que los estudiantes que pertenecen a la FICA o FCCSS poseen menos probabilidades de permanecer activos en su carrera universitaria.

PROVINCIA_PROCEDENCIA

Los estudiantes que proceden de Imbabura, Pichincha, Bolívar, Carchi, Sucumbíos, Loja, Esmeraldas, Santo Domingo de los Tsáchilas o las personas que no especifican su lugar de procedencia poseen menos probabilidades de permanecer activos en su carrera universitaria.

RANGO_PROMEDIO

Las personas que tienen un promedio suficiente o insuficiente poseen mayor probabilidades de permanecer en la carrera universitaria, mientras que las personas con promedios excelente o bueno poseen menor probabilidad de permanecer activos en su carrera universitaria.

MOTIVO_ABANDONO

Las personas que tienen perdida de tercera matrícula poseen menos probabilidades de permanecer activos en su carrera universitaria.

3.3. Fase de obtención del conocimiento

Para obtener el conocimiento se realizó un análisis de los resultados obtenidos con los algoritmos RandomForest (clasificación) y Logistic (regresión), verificando que los resultados de ambos algoritmos se intersecan, obteniendo como resultado que los potenciales desertores cumplen con una o más de las siguientes características:

- Estudiantes que proceden de Imbabura, Bolívar, Colombia, Carchi, Sucumbíos, Esmeraldas, Santo Domingo de los Tsáchilas o Loja.
- Estudiantes que cambiaron su carrera
- Estudiantes de la FICA, FCCSS o FECYT.
- Estudiantes divorciados, viudos o casados.
- Estudiantes cuyo promedio es excelente.
- Estudiantes que pierden tercera matrícula.
- Estudiantes de género femenino.
- Estudiantes con promedio medio y viven con su pareja.
- Estudiantes cuyo ingreso mensual es igual o inferior a la canasta básica familiar.

También se pudo obtener que factores o atributos que son determinantes para que un estudiante abandone su carrera universitaria, para ello se consideró lo anteriormente expuesto que la regresión logística selecciona únicamente a las variables que se encuentran vinculadas a la respuesta, descartando aquellas que no aportan mayor conocimiento al estudio (Madrid, 2017), en este caso las variables descartadas en base al coeficiente obtenido por medio de la regresión logística son RANGO_EDAD, TIPO_VIVIENDA, FINANCIAMIENTO, DISCAPACIDAD, PAÍS_NACIONALIDAD y ACTIVIDAD_ESTUDIANTE. Para identificar las variables que más aportan se tomó en cuenta los atributos que predominan en el bosque que arroja el algoritmo RandomForest, obteniendo como resultado que las variables y sus categorías que más influyen en la deserción estudiantil son las que se aprecian en la Tabla 3.10 ordenados por prioridad, de acuerdo con su presencia en el bosque de RandomForest:

TABLA 3.10
VARIABLES Y CATEGORÍAS QUE MÁS INFLUYEN EN LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL

ATRIBUTO	CATEGORÍA
MOTIVO_ABANDONO	PIERDE TERCERA MATRÍCULA
	CAMBIO CARRERA
FACULTAD	FCCSS
	FICA
	FECYT
RANGO_PROMEDIO	EXCELENTE
RANGO_INGRESO_MENSUAL	BAJO
PROVINCIA_PROCEDENCIA	IMBABURA
	BOLÍVAR
	CARCHI
	COLOMBIA
	SUCUMBÍOS
	ESMERALDAS
	SANTO DOMINGO DE LOS TSÁCHILAS
	LOJA
GENERO	FEMENINO
ETNIA	MONTUBIOS
	MESTIZOS
	AFRODESCENDIENTES
TIPO_SANGRE	A+
	B+

	DIVORCIADO
ESTADO_CIVIL	VIUDO
	CASADO
	MADRE
CONVIVIENTE	PADRES
	SOLO

Fuente: Propia

3.4. Análisis de impactos

El análisis de impactos define las posibles consecuencias que se podrían presentar cuando se tomen decisiones estratégicas en base a los patrones de deserción estudiantil obtenidos; por este motivo es indispensable analizar el efecto de las decisiones tomadas cualificando y cuantificando las bondades o defectos de acuerdo con ciertas dimensiones e indicadores (Estévez, 2013).

Para evaluar los impactos del presente estudio se utilizó la matriz de impactos que permite identificar los aspectos positivos y negativos que la ejecución del proyecto provocará en un grupo o área específica; siendo este un análisis de impactos prospectivo, de acuerdo con la Tabla 3.11 (Posso-Yépez, 2013):

TABLA 3.11
NIVELES DE IMPACTOS

Niveles de Impactos	Ponderación
Impacto Alto Positivo	3
Impacto Medio Positivo	2
Impacto Bajo Positivo	1
Punto de Indiferencia	0
Impacto Bajo Negativo	-1
Impacto Medio Negativo	-2
Impacto Alto Negativo	-3

Fuente: Posso, 2013

Para este análisis se tomaron en cuenta el impacto que tendrá en presente trabajo en el ámbito educativo, sociocultural y económico, que se aprecian en las Tablas 3.12 a 3.14 y en la Tabla 3.15 se detalla el impacto general del proyecto.

3.1.1. Impacto Educativo

TABLA 3.12
IMPACTO EDUCATIVO

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Nivel académico del alumno							X
Nivel de desempeño del alumno							X
Fuente de apoyo para otras instituciones				X			
Niveles de deserción							X
Niveles de repitencia						X	
TOTAL				0		2	9

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{11}{5} = 2.2$$

Nivel de Impacto Educativo = Medio positivo

En el ámbito educativo el proyecto tendrá un impacto medio positivo, puesto que el nivel académico del alumno será alto positivo, ya que con la información obtenida los directivos pueden poner en marcha planes de acción para garantizar una educación de calidad del estudiante.

El nivel del desempeño del estudiante se determina como alto positivo, puesto que se llevará a cabo un seguimiento, por lo que este se sentirá respaldado y motivado en sus estudios.

Se considera que el presente estudio no tendrá impacto en otras instituciones de educación superior puesto que los datos son diferentes y por ende los resultados también variarán, sin embargo, podrían tomar como referencia el presente estudio y replicarlo con sus datos para emplear esta información de manera estratégica.

El impacto que tendrá el presente trabajo en los niveles de deserción se considera alto positivo, puesto que esta es la problemática que se pretende atacar, y los patrones obtenidos en el presente estudio abordan esta problemática directamente.

En cuanto a la repitencia se considera que se tendrá un impacto medio positivo ya que el punto central del presente estudio son los potenciales desertores estudiantiles, sin embargo, los estudiantes con tendencia a repetir las materias también se verán beneficiados de las decisiones que se tomen en base al conocimiento obtenido.

3.1.2. Impacto Sociocultural

TABLA 3.13
IMPACTO SOCIOCULTURAL

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Calidad de vida de los alumnos							X
Empleo						X	
TOTAL						2	3

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{5}{2} = 2.5$$

Nivel de Impacto Sociocultural = Alto positivo

El impacto socio cultural del presente proyecto se considera alto positivo, puesto que los alumnos de la UTN podrán mejorar su calidad de vida al poder formarse como profesionales siendo este un impacto alto positivo.

En cuanto al empleo se dice que este trabajo de titulación tendrá un impacto medio positivo, ya que los estudiantes que se formen profesionalmente tendrán mayores posibilidades de conseguir empleo.

3.1.3. Impacto Económico

TABLA 3.14
IMPACTO ECONÓMICO

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Productividad de la UTN							X
Presupuesto universitario							X
Presupuesto del estudiante							X
TOTAL							9

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{9}{3} = 3$$

Nivel de Impacto Económico = Alto positivo

El ámbito económico se considera que tendrá un impacto alto positivo, puesto que la productividad de la universidad aumentará, ya que la información almacenada en los repositorios de Oracle se empleará con el objetivo de aumentar la retención estudiantil, esto se verá reflejado en el presupuesto asignado a la universidad y en los indicadores de calidad.

En cuanto al presupuesto universitario se tiene un impacto alto positivo puesto que los recursos que se asignan para la educación serán aprovechados al máximo si se eleva los niveles de retención estudiantil.

El presupuesto del estudiante no se verá afectado, ya que por medio de los planes que se pondrán en marcha el estudiante no tendrá repitencia y evitará los pagos de los aranceles de matrícula por materia.

Considerando que la SENESCYT anualmente provee los costos óptimos por carrera anuales(SENESCYT, 2016), para el año 2018 se invirtió $2 385 364,00 en los estudiantes que abandonaron su carrera universitaria.

3.1.4. Impacto General

TABLA 3.15
IMPACTO GENERAL

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Impacto educativo						X	
Impacto sociocultural							X
Impacto económico							X
TOTAL						2	6

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{8}{3} = 2.6$$

Nivel de Impacto Tecnológico = Alto positivo

El impacto general del proyecto es alto positivo, lo cual genera altas expectativas para la toma de decisiones en base al conocimiento obtenido, ya que al beneficiarse el estudiante también se beneficia la comunidad universitaria UTN y la sociedad en general, especialmente en la zona de influencia directa de la universidad (Imbabura).

3.5. Discusión

El presente trabajo es continuación de la investigación realizada por (Vila et al., 2018), donde se abordó esta problemática mediante el análisis de datos personales y académicos de los estudiantes de la UTN utilizando 13 variables, basado en técnicas predictivas de clasificación (RandomTree). En el presente estudio se realizó el análisis incorporando datos académicos, personales, socioeconómicos (17variables) y adicionalmente se trabajó con otras tareas de clasificación (RandomForest) y regresión (Logistic).

Los resultados obtenidos en el presente proyecto de titulación coinciden con algunos trabajos existentes, tales como los de (Vila et al., 2018) en el que se obtiene coincidencia en que los potenciales desertores son las personas cuyo país de nacionalidad es Colombia, las personas de género femenino o que viven con su pareja. Considerando que se utilizó como datos base la información de los mismos estudiantes, se tiene como mayor similitud que existe una relación directa entre la deserción estudiantil con las carreras afines a las ciencias de la salud, el arte o la agricultura, siendo un aporte importante de este estudio que las carreras de ingeniería también mantienen relación con la deserción. Como principal diferencia se tiene que en el trabajo realizado por Vila et al. entre las variables independientes más influyentes se encuentra la edad, mientras que en este trabajo no se considera como tal.

En el trabajo realizado por (Hernández-Leal et al., 2018), se tiene que existe una relación directa entre la deserción estudiantil y el motivo de abandono de su carrera, el promedio de notas académicas o la situación económica del estudiante. De igual forma existen ciertos puntos en común con el trabajo realizado por (Gonzalez et al., 2016) en el que se obtuvo como principal factor de abandono escolar el lugar de procedencia del estudiante, y con la investigación realizada por (Merchán & Duarte, 2016) donde se establece que el género del estudiante y el nivel económico del estudiante son factores clave para determinar la permanencia de un estudiante en su carrera universitaria al igual que en los resultados que se obtuvieron en este trabajo.

También se registraron ciertas discrepancias con otras investigaciones, tales como la de (Noboa et al., 2018), en la que ellos afirman que las características personales del estudiante, a excepción de la edad, no inciden en la permanencia de los estudiantes en su carrera universitaria, resultados totalmente contrarios a los del presente estudio. (Cuji et al., 2017) definieron que las variables que no fueron incluidas en su modelo para obtener los patrones de deserción estudiantil son el género de la persona, estado civil, etnia y el lugar de procedencia que a diferencia de este estudio son variables determinantes. (Pérez et al., 2018) determinó que los estudiantes de género masculino tienen mayor probabilidad de abandonar los estudios a diferencia del presente trabajo donde son las mujeres.

Entre las principales limitaciones que se presentaron en este trabajo se tiene que no se tiene una base de datos que almacene los datos psicológicos de los estudiantes, sino únicamente de los datos médicos que se hacen atender por afecciones físicas leves o de aquellos estudiantes que hacen uso de los servicios odontológicos en la UTN.

Finalmente, entre los principales trabajos a futuro que surgieron después de realizar el presente trabajo siguiendo la misma línea de investigación están: (i) realizar un modelo

predictivo en base a las variables obtenidas en el presente trabajo, considerando adicionalmente los datos psicológicos de los estudiantes; (ii) generar otros modelos basados en máquinas de vectores de soporte (SVM) y redes neuronales artificiales (perceptrón multicapa); y finalmente (iii) establecer un modelo predictivo por cada una de las cinco facultades de la UTN.

Conclusiones y recomendaciones

Conclusiones

Se llevó a cabo el proceso de descubrimiento de conocimiento en bases de datos (KDD), con el objetivo de encontrar patrones de deserción estudiantil, desarrollando en base a las etapas y requerimientos que el proceso en sí demanda. El desarrollo del proceso KDD fue dirigido para atacar una de las problemáticas que enfrenta la educación superior hoy en día en base a la realidad.

Se verificó la importancia del análisis de datos que se encuentran almacenados en sistemas transaccionales, con el fin de obtener información que se encuentra oculta en las bases de datos y puede ser vital para analizar el giro del negocio.

Para elaborar la vista minable se emplearon datos históricos académicos, personales y socioeconómicos de los alumnos de las 39 carreras de grado de la UTN que se encuentran almacenados en la base de datos Oracle, información que fue preprocesada en Pentaho Data Integration y Microsoft Excel, con 11200 instancias desde 2017-2018.

Se empleó el software de Weka para aplicar las técnicas predictivas de minería de datos, lo que permitió construir 4 tipos diferentes de modelos: árboles de decisión (2), naive bayes y regresión logística, para garantizar la calidad de los resultados, de igual manera se identificó que el software posee una amplia gama de algoritmos predictivos con su respectiva documentación y que se obtiene buenos resultados con los mismos.

Se aplicaron las características inherentes de la ISO/IEC 25012:2008, sin embargo, no se obtuvieron resultados relevantes (sección 2.5.2), ya que al desarrollar el proceso KDD se verifica que la vista minable sea de calidad y que se realice un trabajo ordenado. Se evaluaron los resultados de los algoritmos y los modelos mediante medidas cuantitativas de calidad como la matriz de confusión, coeficiente Kappa y otras medidas estadísticas como tasa de error, exactitud, sensibilidad, recall, Curva ROC, F1, número de tipo de errores, entre otros; obteniéndose como mejores algoritmos a RandomForest (clasificación) y Logistic (regresión).

Para obtener el conocimiento se interpretó la información de cada uno de los modelos llegando a la conclusión de que se complementan y se realizó un análisis de ambos algoritmos para obtener los patrones finales (sección 3.3) y las variables que se relacionan directamente, en este caso con la deserción estudiantil (Tabla 3.10), lo que contribuye para que los encargados propongan planes de acción efectivos para mitigar esta problemática, que entre otras, impacta los indicadores educativo, social y económico (Tabla 3.1.4).

Recomendaciones

Al momento de definir desde qué punto se pretende combatir la deserción estudiantil, se debería realizar una investigación previa acerca de la información, estructura tecnológica y organización de la institución en la cual se pretende aplicar la minería de datos, con el objetivo de seleccionar correctamente los materiales para el análisis.

Analizar periódicamente la información contenida en el repositorio de bases de datos institucionales, ya que las necesidades de los estudiantes van cambiando, por ello es indispensable contar con información en base a la realidad para la toma de decisiones estratégicas actuales.

Elaborar una vista minable por facultad para poder construir un modelo independiente, y de esta manera poder identificar si la carrera tiene incidencia en el abandono escolar. Asimismo, se recomienda que la información ingresada por los estudiantes sea validada para verificar su veracidad y evitar datos incongruentes en las bases de datos ya que de ello depende la precisión de los resultados.

Para trabajos a futuro de minería de datos emplear otros modelos como máquinas de vectores soporte (SVM) o redes neuronales artificiales, ya que existe la documentación necesaria y trabajos previos para poder realizar cualquier tipo de análisis, y determinar si la naturaleza de los datos se aplica al algoritmo seleccionado, e investigar nuevas herramientas tecnológicas para realizar cada fase del proceso KDD, con el objetivo de ampliar la gama de algoritmos.

Verificar si aplicar cierta normativa aporta al trabajo, para optimizar esfuerzos y costos, identificar si la vista minable posee datos atípicos que puedan sesgar el resultado, y tratarlos de acuerdo con las necesidades del estudio. Analizar todos los resultados de los algoritmos detalladamente para verificar que se interpretarán los resultados del mejor algoritmo.

Utilizar la información obtenida para poner en marcha planes de acción con el objetivo de disminuir los niveles de abandono estudiantil y minimizar el impacto económico que se produce tanto a nivel de la universidad como del estado, de igual forma se puede emplear la información obtenida para realizar trabajos de minería de datos con los atributos que se obtuvieron para identificar las causas específicas relacionadas con esta problemática.

GLOSARIO DE TÉRMINOS

Bayesiana: tipo de inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta.

Cross validation: es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Deserción estudiantil: desde el punto de vista individual, el cual debe referirse a las metas y propósitos que tienen las personas al incorporarse al sistema de educación superior en el que desertar significa el fracaso para completar un determinado curso de acción o alcanzar una meta deseada, considerándose deserción estudiantil al abandono de la carrera, ya sea por cambio de carrera dentro de la misma institución, cambio a otra institución o por otros motivos.

Discretización: modificación de la granularidad de la variable o categoría, como cuando se agregan múltiples variables discretas o se fusionan múltiples categorías discretas.

Heterocedásticos: propiedad de algunos modelos de regresión lineal en los que los errores de estimación son constantes a lo largo de las observaciones.

ISO: International Organization for Standardization

Normalización: es el proceso de elaborar, aplicar y mejorar las normas que se aplican a distintas actividades científicas, industriales o económicas, con el fin de ordenarlas y mejorarlas.

BIBLIOGRAFÍA

Asamblea Nacional del Ecuador. (2016). *Ley orgánica de gestión de la identidad y datos civiles*. 16.

Bazantes, Z., Ruiz, M., & Álvarez, M. (2017, febrero 14). DESERCIÓN ESTUDIANTIL UNIVERSITARIA EN ECUADOR Y SU INFLUENCIA EN LA CALIDAD DEL EGRESADO | Revista Magazine de las Ciencias. ISSN 2528-8091. Recuperado 22 de julio de 2018, de https://revistas.utb.edu.ec/index.php/magazine/article/view/183

BCE. (2018). *Sector Minero* (p. 4). Recuperado de https://contenido.bce.fin.ec/documentos/Estadisticas/Hidrocarburos/cartilla00.pdf

Calculadora del índice de masa corporal (IMC). (s. f.). Recuperado 11 de octubre de 2018, de Texas Heart Institute website: https://www.texasheart.org/heart-health/heart-information-center/topics/calculadora-del-indice-de-masa-corporal-imc/

Cerda L, J., & Villarroel Del P, L. (2008). Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. *Revista chilena de pediatría*, *79*(1), 54-58. https://doi.org/10.4067/S0370-41062008000100008

Cuji, B., Gavilanes, W., & Sánchez, R. (2017). Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Espacios*, *38*, 17-26.

D'ambrosio, S. (2008, febrero 6). El Concepto de Datos - Monografias.com. Recuperado 9 de octubre de 2018, de https://www.monografias.com/trabajos14/datos/datos.shtml

Dasarathy, B. (1991). *Nearest neighbor (NN) norms: nn pattern classification techniques* (ilustrada). Recuperado de https://books.google.com.ec/books/about/Nearest_neighbor_NN_norms.html?id=k2dQAAAAMAAJ&redir_esc=y

Del Valle, A. R. (s. f.). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones.* (Universidad de Sevilla). Recuperado de https://idus.us.es/xmlui/bitstream/handle/11441/63201/Valle%20Benavides%20Ana%20Roc%C3%ADo%20del%20TFG.pdf?sequence=1

Espinoza, M., & Gallegos, D. (2019). Data Scientist: A Systematic Review of the Literature. *Communications in Computer and Information Science*, *895*, 476 a 487. https://doi.org/10.1007/978-3-030-05532-5_35

Estévez, F. (2013). *ESTUDIO DE FACTIBILIDAD PARA LA CREACIÓN DE UNA MICROEMPRESA DEDICADA A LA PRODUCCIÓN Y COMERCIALIZACIÓN DE PASTAS (FIDEOS) CON HARINA DE FRÉJOL Y MAÍZ EN LA PARROQUIA DE SAN ANTONIO, CANTÓN IBARRA, PROVINCIA DE IMBABURA* (Grado). Universidad Técnica del Norte, Ibarra, Ecuador.

Frank, E., Hall, M., & Witten, I. (2016). *Data Mining. Practical Machine Learning Tools and Techniques* (Cuarta). Recuperado de https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Gonzalez, A. G. H., Armenta, R. A. M., Rosales, L. A. M., Barrientos, A. G., Xihuitl, J. L. T., & Algredo, I. (2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-

Mexico. *IEEE Latin America Transactions*, *14*(11), 4573-4578. https://doi.org/10.1109/TLA.2016.7795831

Gutierrez, A. (2017, agosto 15). Funciones del Jefe de Proyectos (II). Recuperado 5 de diciembre de 2018, de Jefe de Proyectos website: https://jefedeproyectos.com/funciones-del-jefe-de-proyectos-ii/

Hasperué, L. W. (2013). *Extracción de Conocimiento en Grandes Bases de Datos Utilizando Estrategias Adaptativas*. 212.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid, España: PEARSON EDUCACIÓN, S.A.

Hernández-Leal, E. J., Quintero-Lorza, D. P., Escobar-Naranjo, J. C., Ramírez-Gómez, J. S., & Duque-Méndez, N. D. (2018). Educational data mining for the analysis of student desertion. *Learning Analytics for Latin America 2018*, *2231*, 51-60.

Hitachi Vantara. (2018). Pentaho Data Integration - Accelerate Data Pipeline. Recuperado 17 de octubre de 2018, de https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html

INEC. (2018). *Informe Ejecutivo de las Canastas Analíticas: Básica y Vital* [Informe Ejecutivo]. Recuperado de Gobierno Nacional de la República del Ecuador website: http://www.ecuadorencifras.gob.ec/documentos/web-inec/Inflacion/canastas/Canastas_2018/Noviembre-2018/1.%20Informe_Ejecutivo_Canastas_Analiticas_nov_2018.pdf

Instituto de Estadística de la UNESCO. (2012). *Oportunidades perdidas: El impacto de la repetición y de la salida prematura de la escuela. Compendio mundial de la educación 2012*. Instituto de Estadística de la UNESCO.

ISO, & IEC. (2014). ISO/IEC 25012:2008(en), Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Recuperado 3 de diciembre de 2018, de https://www.iso.org/obp/ui/#iso:std:iso-iec:25012:ed-1:v1:en

Lara, J. (2014). *Minería de Datos* (CENTRO DE ESTUDIOS FINANCIEROS).

Lehr, S., Liu, H., Kinglesmith, S., Konyha, A., Robaszewska, N., & Medinilla, J. (2016). Use Educational Data Mining to Predict Undergraduate Retention. *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 428-430. https://doi.org/10.1109/ICALT.2016.138

Leskovec, J., Rajaraman, A., & Ullman, J. (2014). *Mining of Massive Datasets* (Segunda). Recuperado de https://www.cambridge.org/es/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB

Madrid, J. (2017). *Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria* (Maestría, Universidad Nacional de Colombia). Recuperado de http://bdigital.unal.edu.co/58059/1/71787491.2017.pdf

Mayra, A., & Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in Ecuador. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 1238-1242. https://doi.org/10.1109/EDUCON.2018.8363371

Merchan, S., & Duarte, J. (2016). Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance. *IEEE Latin America Transactions*, *14*(6), 2783-2788. https://doi.org/10.1109/TLA.2016.7555255

Microsoft. (2018). Microsoft Excel 2016, hojas de cálculo, prueba gratuita. Recuperado 17 de octubre de 2018, de https://products.office.com/es-ww/excel

Noboa, C., Ordóñez, M., & Magallanes, J. (2018). Statistical Learning to Detect Potential Dropouts in Higher Education: A Public University Case Study. *Learning Analytics for Latin America 2018*, *2231*, 12-21.

Oracle. (2018). Base de datos | Base de datos en la nube | Oracle España. Recuperado 17 de octubre de 2018, de https://www.oracle.com/es/database/index.html

Pajares, G., & De La Cruz, J. (2010). *Aprendizaje automático: un enfoque práctico*. Recuperado de http://www.ra-ma.es/libros/APRENDIZAJE-AUTOMATICO-UN-ENFOQUE-PRACTICO/23487/978-84-9964-011-2

Palacios-Pacheco, X., Villegas-Ch, W., & Luján-Mora, S. (2018). Application of Data Mining for the Detection of Variables that Cause University Desertion. *Communications in Computer and Information Science*, *895*, 510 a 520. https://doi.org/10.1007/978-3-030-05532-5_38

Perez, A., Escobar, C., Toledo, M., Gutierrez, L., & Reyes, G. (2018). *Modelo de predicción de la deserción estudiantil de primer año en la Universidad Bernardo O´Higgins*. https://dx.doi.org/10.1590/s1678-4634201844172094

Pérez, B. (2017, julio 7). Corte Constitucional: Se considerarán personas con discapacidad aquellas que posean 30% o más de discapacidad. Recuperado 4 de enero de 2019, de Pérez Bustamante & Ponce website: http://www.pbplaw.com/corte-constitucional-se-consideraran-personas-con-discapacidad-aquellas-que-posean-30-o-mas-de-discapacidad/

Pérez López, C., & Santín González, D. (2006). *Data Mining. Soluciones con Enterprise Miner*. Madrid, España: RA-MA Editorial.

Pérez López, C., & Santín González, D. (2007). *Minería de Datos: Técnicas y Herramientas*. Madrid, España: Clara Ma de la Fuente Rojo.

Pina, K. (2018, abril 22). Matriz de confusión. Recuperado 14 de octubre de 2018, de Koldo Pina - Data Science & Inteligencia Artificial website: https://koldopina.com/matriz-de-confusion/

Posso-Yépez, M. (2013). *Proyectos, Tesis y Marco Lógico: Planes e Información de investigación*. Quito, Ecuador.

PowerData, G. (2018). Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos. Recuperado 10 de octubre de 2018, de https://www.powerdata.es/data-warehouse

presentación-rendición-de-cuentas.pdf. (s. f.). Recuperado de http://www.senescyt.gob.ec/rendicion2015/assets/presentaci%C3%B3n-rendici%C3%B3n-de-cuentas.pdf

Remuestreo.pdf. (s. f.). Recuperado de http://www.cs.us.es/~fsancho/ficheros/IAML/Remuestreo.pdf

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, *10*(29). Recuperado de http://www.redalyc.org/resumen.oa?id=92502902

Roig, A. A. (2014, abril 11). EL MODELO DE EVALUACION DE LAS UNIVERSIDADES ECUATORIANAS: APUNTES CRITICOS PARA EL DEBATE. Recuperado 22 de julio de 2018, de https://lalineadefuego.info/2014/04/11/el-modelo-de-evaluacion-de-las-universidades-ecuatorianas-apuntes-criticos-para-el-debate/

Sadiq, H., Neama Abdulaziz, D., Fadl Mutaher, B.-A., & Najoua, R. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Scienc*, *9*(2), 447-459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Saito, T., & Rehmsmeier, M. (2015). *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. http://dx.doi.org/10.6084/m9.figshare.1245061

SENESCYT. (2016). *Informe costos, óptimo por carrera por estudiante establecido por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación SENESCYT* (p. 10) [Económico]. Recuperado de SENESCYT website: http://webcache.googleusercontent.com/search?q=cache:dTeTwF6AFLEJ:financiamiento.cti.espol.edu.ec/documentos2/descargarArchivoT/20/1/1/+&cd=2&hl=es-419&ct=clnk&gl=ec

Sierra Araujo, B. (2006). *Aprendizaje Automático: conceptos básicos y avanzados*. Madrid, España: Cofás, S.A.

Strandjev, I. (2017, agosto 30). machine learning - what is f1-score and what its value indicates? Recuperado 17 de octubre de 2018, de Stack Overflow website: https://stackoverflow.com/questions/45963174/what-is-f1-score-and-what-its-value-indicates

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (Segunda). Recuperado de http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf

Telégrafo, E. (2017, diciembre 27). Salario básico para 2018 será de $ 386. Recuperado 4 de enero de 2019, de El Telégrafo website: https://www.eltelegrafo.com.ec/noticias/economia/4/salario-basico-para-2018-es-de-usd-386

UNESCO UIS. (2004). Recuperado 22 de julio de 2018, de http://uis.unesco.org/

Vila, D., Cisneros, S., Granda, P., Ortega, C., Posso-Yépez, M., & García-Santillán, I. (2018). Detection of Desertion Patterns in University Students using Data Mining Techniques: A Case Study. *Communications in Computer and Information Science*, *895*, 420 a 429. https://doi.org/DOI https://doi.org/10.1007/978-3-030-05532-5_31

Villacís, B., & Carrillo, D. (2012). *País atrevido: la nueva cara sociodemográfica del Ecuador*. Recuperado de http://www.ecuadorencifras.gob.ec/wp-content/descargas/Libros/Economia/Nuevacarademograficadeecuador.pdf

Waikato. (2018). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Recuperado 17 de octubre de 2018, de https://www.cs.waikato.ac.nz/ml/weka/

ANEXOS

Anexo 1. Cálculo del Coeficiente Kappa: http://bit.ly/2RBCgBe

Anexo 2. Cálculo de Regresión logística: http://bit.ly/2Xzyooy

Anexo 3. Reglas de decisión del algoritmo RandomTree: http://bit.ly/2Op5gMo

Anexo 4. Reglas de decisión algoritmo Random Forest: http://bit.ly/2HFme8Q

Anexo 5. Probabilidades del algoritmo Naive Bayes: http://bit.ly/2uuGmSm

Anexo 6. Coeficientes del algoritmo Logistic: http://bit.ly/2FAzsBx