

UNIVERSIDAD TÉCNICA DEL NORTE



Facultad de Ingeniería en Ciencias Aplicadas
Carrera de Ingeniería en Sistemas Computacionales

**DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO
TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN
MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN LA UNIVERSIDAD
TÉCNICA DEL NORTE.**

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

Autor:

Saúl Andrés Cisneros Buitrón

Director:

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:	100354707-0		
APELLIDOS Y NOMBRES:	Cisneros Buitrón Saúl Andrés		
DIRECCIÓN:	Ezequiel Rivadeneira 8-36 y Ramón Teanga		
EMAIL:	sacisnerosb@utn.edu.ec		
TELÉFONO FIJO:	062932354	TELÉFONO MÓVIL:	0983822286

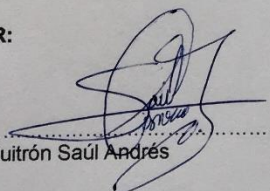
DATOS DE LA OBRA	
TÍTULO:	DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN LA UNIVERSIDAD TÉCNICA DEL NORTE.
AUTOR (ES):	Cisneros Buitrón Saúl Andrés
FECHA: DD/MM/AAAA	11/04/2019
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	<input checked="" type="checkbox"/> PREGRADO <input type="checkbox"/> POSGRADO
TÍTULO POR EL QUE OPTA:	Ingeniero en Sistemas Computacionales
ASESOR /DIRECTOR:	PhD. Iván Danilo García Santillán

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

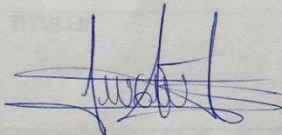
EL AUTOR:


.....
Cisneros Buitrón Saúl Andrés

CERTIFICADO TUTOR

En mi calidad de Tutor del Trabajo de Grado presentado por el egresado, **Cisneros Buitrón Saúl Andrés** para optar por el Título de Ingeniero en Sistemas Computacionales, cuyo tema es: **DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN 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.

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



PhD. Iván Danilo García Santillán

DIRECTOR TRABAJO DE GRADO



UNIVERSIDAD TÉCNICA 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

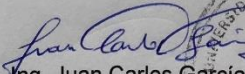
QUE: El señor SAÚL ANDRÉS CISNEROS BUITRON con cédula identidad 1003547070 estudiante de la Facultad de Ingeniería en Ciencias Aplicadas – de la Carrera de Ingeniería en Sistemas Computacionales, ha desarrollado con los datos entregados de la Dirección de Desarrollo Tecnológico e Informático, el Proyecto de Tesis **“DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN LA UNIVERSIDAD TÉCNICA DEL NORTE”**.

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

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

Ibarra, 10 de abril del 2019

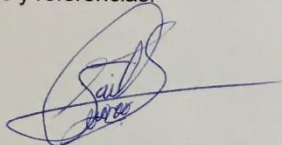
Atentamente
CIENCIA Y TÉCNICA AL SERVICIO DEL PUEBLO


Ing. Juan Carlos García
DIRECTOR



AUTORÍA

Yo, **Cisneros Buitrón Saúl Andrés**, portador de la cédula de ciudadanía número **1003547070**, declaro bajo juramento que el trabajo aquí descrito es de mi autoría, **DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN 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.



Cisneros Buitrón Saúl Andrés

1003547070

Dedicatorias

Este trabajo de titulación va dedicado a mis padres Saúl Cisneros Vásquez y Guadalupe Buitrón Rojas y a mi hermana Karla Cisneros quienes con su abnegado apoyo y constancia me han permitido culminar mis estudios.

Agradecimientos

Mis más profundo agradecimiento en primer lugar a mi familia quienes con su constante apoyo me han permitido llegar al final de mi carrera, de igual forma agradezco infinitamente a todos quienes conforman la carrera de Ingeniería en Sistemas Computacionales y a la institución, quienes a lo largo de este tiempo, me han formado tanto humana como profesionalmente, de igual forma extendo un agradecimiento a mi tutor de tesis quien me ha brindado todo sus conocimientos con la finalidad de culminar este trabajo de titulación

Tabla de contenido

Autorización de uso y publicación a favor de la Universidad Técnica del Norte	¡Error! Marcador no definido.
Certificado tutor	¡Error! Marcador no definido.
Certificado institución.....	¡Error! Marcador no definido.
Autoría.....	¡Error! Marcador no definido.
Dedicatorias.....	VI
Agradecimientos.....	VII
Resumen.....	XVI
Abstract.....	XVI
Introducción.....	XVII
Antecedentes.....	XVII
Problema.....	XVIII
Objetivos.....	XIX
Objetivo General.....	XIX
Objetivos Específicos.....	XIX
Justificación.....	XIX
Alcance.....	XX
CAPÍTULO 1.....	1
Marco teórico.....	1
1.1. Introducción a la minería de datos.....	1
1.2. Relación de la minería de datos con otras disciplinas.....	1
1.2.1. Bases de datos.....	2
1.2.2. Visualización.....	2
1.2.3. Estadísticas.....	2
1.2.4. Aprendizaje automático.....	2
1.2.5. Sistema de toma de decisiones.....	3
1.2.6. Recuperación información.....	3
1.2.7. Otras.....	3
1.3. Tipos de datos.....	3
1.3.1. Bases de datos relacionales.....	4
1.3.2. Bases de datos desnormalizadas.....	5

1.3.3.	Otros tipos de bases de datos	5
1.4.	Proceso KDD.....	6
1.4.1.	Recopilación de datos	7
1.4.2.	Selección, limpieza y transformación de datos	8
1.4.3.	Minería de datos	12
1.4.4.	Evaluación e interpretación.....	18
1.5.	ISO/IEC 25012:2008	21
CAPÍTULO 2.....		25
Desarrollo del Proceso de Descubrimiento de Conocimiento en Bases de Datos		25
2.1.	Vista General del Proyecto.....	25
2.2.	Entregables del Proyecto	26
2.3.	Organización del Proyecto	26
2.3.1.	Participantes del Proyecto	26
2.3.2.	Roles y Responsabilidades.....	27
2.4.	Gestión del Proceso	28
2.4.1.	Estimaciones	28
2.4.2.	Plan del Proyecto	29
2.5.	Recopilación de Datos.....	30
2.5.2.	Implementación de la norma ISO/IEC 25012:2008	36
2.5.3.	Construcción del data warehouse.....	37
2.6.	Fase de selección, limpieza y transformación.....	43
2.6.1.	Selección	43
2.6.2.	Transformación	45
2.6.3.	Limpieza	58
2.7.	Minería de Datos.....	59
2.7.1.	Agrupamiento	60
2.7.2.	Asociación	61
2.7.3.	Atípicos.....	61
CAPÍTULO 3.....		62
Validación de Resultados		62
3.1.	Evaluación e interpretación	62
3.1.1.	Evaluación, análisis e interpretación de tareas de asociación	62
3.1.2.	Evaluación e interpretación de la tarea de agrupamiento	67

3.2. Atípicos	80
3.3. Obtención del conocimiento	91
3.4. Análisis de impactos.....	94
3.4.1. Impacto Educativo.....	95
3.4.2. Impacto Sociocultural.....	96
3.4.3. Impacto Económico.....	96
3.4.4. Impacto General.....	97
3.5. Discusión.....	98
Conclusiones y Recomendaciones	99
Conclusiones.....	99
Recomendaciones.....	100
Glosario de términos	102
Bibliografía.....	103
Anexos	110

Índice de Figuras

Fig. 1. Disciplinas que contribuyen a la minería de datos - Fuente (Lara, 2014).....	2
Fig. 2. Base de datos relacional - Fuente: Propia.....	4
Fig. 3. Esquema de base de datos desnormalizado - Fuente: Propia	5
Fig. 4. Proceso KDD - Fuente:(Lara, 2014)	7
Fig. 5. Etapa de recopilación de datos, fuente:(Lara, 2014).....	8
Fig. 6. Etapa de selección, limpieza y transformación, fuente (Lara, 2014)	11
Fig. 7. Etapa de minería de datos, fuente (Lara, 2014)	17
Fig. 8. Etapa de evaluación e interpretación.....	21
Fig. 9. Transformación PDI para Dimensión LOCALIDADES	38
Fig. 10. Transformación PDI para Dimensión DEPENDENCIAS.....	39
Fig. 11. Transformación PDI para Dimensión ESTUDIANTE_CARRERA.....	39
Fig. 12. Primera transformación PDI para Dimensión PERSONA	40
Fig. 13. Segunda transformación PDI para Dimensión PERSONA.....	40
Fig. 14. Tercera transformación PDI para Dimensión PERSONA.....	41
Fig. 15. Transformación PDI para el Data Warehouse.....	41
Fig. 16. Atributos de la tabla ESTUDIANTE_CARRERA.....	43
Fig. 17. Atributos relevantes al estudio	44
Fig. 18. Atributos seleccionados para realizar el análisis	44
Fig. 19. Parte de la transformación que calcula la edad hasta el año 2018	51
Fig. 20. Cálculo del promedio general de cada estudiante.....	53
Fig. 21. Error ortográfico en Clase CONVIVIENTE categoría FAMILIAR	58
Fig. 22. Error de transformación en la Clase CARRERA	59
Fig. 23. Vista minable para asociación en formato *.csv	60
Fig. 24. Vista minable para agrupamiento en formato *.csv	60
Fig. 25. Resultado con Apriori	63
Fig. 26. Resultado con Apriori	64
Fig. 27. Resultados obtenidos del algoritmo EM	74
Fig. 28. Resultados obtenidos del algoritmo EM	75
Fig. 29. Diagrama de caja de atributo RANGO_EDAD	82
Fig. 30. Diagrama de caja de atributo RANGO_INGRESO_MENSUAL	83
Fig. 31. Diagrama de caja de atributo RANGO_PROMEDIO.....	83
Fig. 32. Estadísticas del atributo ETNIA.....	85

Fig. 33. Estadísticas del atributo PAÍS_NACIONALIDAD	86
Fig. 34. Estadísticas del atributo RANGO_DISCAPACIDAD	86
Fig. 35. Estadísticas del atributo ESTADO_CIVIL	87
Fig. 36. Estadísticas del atributo TIPO_SANGRE	87
Fig. 37. Estadísticas del atributo FINANCIAMIENTO	88
Fig. 38. Estadísticas del atributo GÉNERO	88
Fig. 39. Estadísticas del atributo CONVIVIENTE	89
Fig. 40. Estadísticas del atributo TIPO_VIVIENDA.....	89
Fig. 41. Estadísticas del atributo ACTIVIDAD_ESTUDIANTE.....	90
Fig. 42. Estadísticas del atributo PROVINCIA_PROCEDENCIA	90
Fig. 43. Estadísticas del atributo MOTIVO_ABANDONO.....	91

Índice de Cuadros

TABLA 1.1 ENTIDAD ESTUDIANTE.....	3
TABLA 1.2 TABLA CON ELEMENTOS SIMILARES	13
TABLA 1.3 MATRIZ DE CONFUSIÓN DE DOS VARIABLES.....	19
TABLA 1.4 CARACTERÍSTICAS DEL MODELO DE CALIDAD DE DATOS	22
TABLA 2.1 DIRECTORES DE LAS ÁREAS COMPRENDIDAS	27
TABLA 2.2 PARTICIPANTES DIRECTOS DEL PROYECTO	27
TABLA 2.3 ROLES Y RESPONSABILIDADES	27
TABLA 2.4 TALENTO HUMANO	28
TABLA 2.5 RECURSOS MATERIALES.....	28
TABLA 2.6 COSTO TOTAL DEL PROYECTO.....	29
TABLA 2.7 DISTRIBUCIÓN DE HORAS	29
TABLA 2.8 HITOS IMPORTANTES	29
TABLA 2.9 ESTRUCTURA TABLA CICLO_ACADEMICOS_102018	30
TABLA 2.10 ESTRUCTURA TABLA DEPENDENCIAS_102018.....	31
TABLA 2.11 ESTRUCTURA TABLA DETALLE_MATRICULAS_102018.....	31
TABLA 2.12 ESTRUCTURA TABLA ESTUDIANTE_CARRERA_102018.....	32
TABLA 2.13 ESTRUCTURA TABLA LOCALIDADES_102018	33
TABLA 2.14 ESTRUCTURA TABLA MATRICULAS_102018	33
TABLA 2.15 ESTRUCTURA TABLA NOTAS_102018	34
TABLA 2.16 ESTRUCTURA TABLA PERSONAS_102018.....	35
TABLA 2.17 ESTRUCTURA TABLA FICHA_112018.....	36
TABLA 2.18 EVALUACIÓN ISO/IEC:25012.....	36
TABLA 2.19 CARACTERÍSTICAS DE LOS EQUIPOS	37
TABLA 2.20 TIEMPOS DE RESPUESTA DIMENSIÓN LOCALIDADES	38
TABLA 2.21 TIEMPOS DE RESPUESTA DIMENSIÓN DEPENDENCIAS.....	39
TABLA 2.22 TIEMPOS DE RESPUESTA DIMENSIÓN ESTUDIANTE_CARRERA.....	39
TABLA 2.23 TIEMPOS DE RESPUESTA DIMENSIÓN PERSONA.....	41
TABLA 2.24 ESTRUCTURA DATA_WAREHOUSE	42
TABLA 2.25 TIEMPOS DE RESPUESTA DATA_WAREHOUSE	43
TABLA 2.26 CATEGORIZACIÓN CLASE CONVIVIENTE	45
TABLA 2.27 CATEGORIZACIÓN CLASE FINANCIAMIENTO	45

TABLA 2.28 CATEGORIZACIÓN CLASE INGRESO_MENSUAL.....	46
TABLA 2.29 CATEGORIZACIÓN CLASE CARRERA FACAE.....	47
TABLA 2.30 CATEGORIZACIÓN CLASE CARRERA FCCSS	47
TABLA 2.31 CATEGORIZACIÓN CLASE CARRERA FECYT	48
TABLA 2.32 CATEGORIZACIÓN CLASE CARRERA FICA.....	48
TABLA 2.33 CATEGORIZACIÓN CLASE CARRERA FICAYA.....	49
TABLA 2.34 CATEGORIZACIÓN CLASES GENERO Y TIPO_IDENTIFICACION.....	49
TABLA 2.35 CATEGORIZACIÓN CLASE ESTADO_CIVIL.....	50
TABLA 2.36 CATEGORIZACIÓN CLASE EDAD	51
TABLA 2.37 CATEGORIZACIÓN CLASE ESTADO_CIVIL.....	51
TABLA 2.38 CATEGORIZACIÓN CLASE PROVINCIA_PROCEDENCIA	52
TABLA 2.39 CATEGORIZACIÓN CLASE PORCENTAJE_DISCAPACIDAD	52
TABLA 2.40 CATEGORIZACIÓN CLASE PROMEDIO.....	53
TABLA 2.41 TIEMPOS DE RESPUESTA ETAPA DE TRANSFORMACIÓN	53
TABLA 3.1 COMPARATIVA ENTRE WEKA Y SPSS PARA APRIORI.....	62
TABLA 3.2 CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO APRIORI	63
TABLA 3.3 REGLAS DE ASOCIACIÓN, SOPORTE Y CONFIANZA RESULTANTE DEL ALGORITMO APRIORI	64
TABLA 3.4 10 MEJORES REGLAS DE ASOCIACIÓN	66
TABLA 3.5 COMPARATIVA ENTRE WEKA Y SPSS PARA K-MEANS.....	68
TABLA 3.6 CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO KMEANS	69
TABLA 3.7 CENTROIDES INICIALES ALEATORIOS DE CADA CLÚSTER.....	69
TABLA 3.8 CENTROIDES FINALES DE CADA CLÚSTER.....	70
TABLA 3.8 RESULTADOS CON ANOVA	70
TABLA 3.10 CLÚSTERES DESNORMALIZADOS DEL ALGORITMO K-MEANS.....	72
TABLA 3.11 COMPARATIVA ENTRE WEKA Y SPSS PARA EM.....	73
TABLA 3.12 CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO EM.....	73
TABLA 3.13 VARIABLES ESTADÍSTICAS DESCRIPTIVAS DEL ALGORITMO EM	76
TABLA 3.14 CLÚSTERES DESNORMALIZADOS DEL ALGORITMO EM	79
TABLA 3.15 CUARTILES, MEDIANA Y LÍMITES	81
TABLA 3.16 NIVELES DE IMPACTOS.....	94
TABLA 3.17 IMPACTO EDUCATIVO.....	95

TABLA 3.18 IMPACTO SOCIOCULTURAL	96
TABLA 3.19 IMPACTO ECONÓMICO	96
TABLA 3.20 IMPACTO GENERAL.....	97

Resumen

Actualmente, la deserción estudiantil es un fenómeno que afecta a las instituciones de educación superior y como consecuencia sus estándares de calidad bajan. En el presente trabajo de investigación se obtuvieron patrones de deserción estudiantil y los principales factores que influyen en esta problemática en la Universidad Técnica del Norte (Ecuador), por medio de técnicas descriptivas de minería de datos (agrupamiento, asociación y atípicos), para analizar los datos personales, académicos y socioeconómicos de los estudiantes desde del año 2017 a 2018. Para obtener la vista minable con 11200 registros se desarrolló el proceso KDD (Proceso de descubrimiento de conocimiento en bases de datos), con el objetivo de obtener el conocimiento deseado con las herramientas Weka y SPSS. Para definir el mejor algoritmo se evaluaron cuantitativamente cada uno de ellos mediante medidas estadísticas. Los principales resultados demostraron que el mejor algoritmo es EM, para obtener el conocimiento de atípicos se emplearon los diagramas de cajas.

Palabras clave: deserción estudiantil, descubrimiento de patrones, minería de datos, técnicas descriptivas

Abstract

Currently, student desertion is a phenomenon that affects higher education institutions and as a result their quality standards go down. In the present research work, student desertion patterns and the main factors that influence this problem were obtained at the Technical University of the North (Ecuador), by means of descriptive data mining techniques (clustering, association and atypical), to analyze personal, academic and socio-economic data of students from 2017 to 2018. To obtain the minable view with 11200 records, the KDD process (Knowledge discovery process in databases) was developed, with the aim of obtaining the desired knowledge with the Weka and SPSS tools. To define the best algorithm, each of them was quantitatively evaluated by means of statistical measures. The main results showed that the best algorithm is EM, to obtain the knowledge of atypical box diagrams were used.

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

Introducción

Antecedentes

Entre 1988 y 1996, en Ecuador se implementan varias iniciativas educativas orientadas a ampliar la cobertura y mejorar la calidad. Sin embargo, a partir de 1996 la situación del país en general, y la de su sistema educativo en particular, estuvo caracterizada por inestabilidad política y frecuente cambio de autoridades, así como por crisis financiera y duros ajustes macroeconómicos. En 1999, 56% de ecuatorianos vivían bajo los límites de la pobreza, y existía un crónico desfinanciamiento de la educación pública y otros servicios sociales, por este motivo los estudiantes optaban por abandonar sus estudios (Araujo & Bramwell, 2015).

En el período 2006- 2007, empezaron a darse importantes cambios en la situación de la política pública educativa en el país. El más importante fue el Plan Decenal de Educación (PDE), aprobado mediante consulta popular en noviembre de 2006 (Araujo & Bramwell, 2015). El PDE tiene ocho políticas, cuatro de las cuales se centran en el incremento de la cantidad de personas atendidas por servicios educativos.

La deserción estudiantil es un fenómeno que hace referencia al abandono de la educación, esta situación se ha venido presentando de manera significativa en la educación superior en el Ecuador, ya que, 8 de cada 10 estudiantes que ingresaron a la universidad o escuela politécnica publica en el año 2012 continuaron con sus estudios en el año 2013 y 7 de cada 10 continuaron en el 2014 (SENESCYT, 2015), contribuyendo al aumento del desempleo y pobreza del país (Gonzalez et al., 2016).

(Mishra et al, 2017) usaron tareas de agrupación en 84 datos de estudiantes de grado de psicología en el norte de España, agrupándolos en tres clústeres de acuerdo con sus calificaciones finales del curso. El clúster más pequeño fue el más comprendido, mientras que los dos más grandes resultaron complejos de interpretar.

(Amelio & Tagarelli, 2018) con información académica, socioeconómica y demográfica de 300 estudiantes de 3 colegios de la India y 24 atributos, aplicaron técnicas predictivas y descriptivas de minería de datos con la herramienta WEKA, los principales resultados mostraron que Random Forest fue el clasificador más preciso frente a J48, PART y Redes Bayesianas, mientras que el modelo descriptivo que se empleó fue la asociación con su algoritmo Apriori determinaron que se obtuvieron buenas reglas de asociación.

(García López et al. 2008) presentaron las principales metodologías de agrupación, particularmente describen las diferentes medidas de similitud y distancia que son usadas usualmente en agrupamiento, y realizan una evaluación por categoría para determinar que algoritmos son los más efectivos para cada tipo de problema a resolver y la evaluación de su efectividad.

(Espinoza & Gallegos, 2018) argumentan que a lo largo del tiempo las empresas comerciantes de productos y servicios han ido almacenando información en sus bases de datos, por ello surge la necesidad de contar con un profesional que interprete los datos para la toma de decisiones estratégicas. Realizaron una revisión sistemática de la literatura para identificar el perfil profesional del llamado Científico de los Datos, puesto que es una carrera emergente no se cuenta con perfiles definidos. Concluyeron que entre las actitudes y aptitudes fundamentales que debe tener este profesional debe ser el liderazgo, comunicación e investigación, así como conocimientos en tecnologías de la información, matemáticas y estadística.

(Palacios-Pacheco et al., 2018) Manifiestan que la deserción estudiantil a nivel universitario se ha vuelto una problemática en especial Latinoamérica, para afrontar dicho problema, algunos países han puesto en marcha diferentes proyectos con la finalidad de que los estudiantes de los primeros años no abandonen sus carreras. se enfocan en atacar este problema con ayuda de minería de datos. Para este estudio emplearon algoritmos en Weka de Árboles de Decisión, Naive Bayes, Agrupamiento y Redes Neuronales. Hasta la publicación de su trabajo la interpretación de resultados se encuentra en desarrollo.

Problema

En la Universidad Técnica del Norte existe un sinnúmero de causas de tipo académica, social, psicológica y económicas por las que los estudiantes desertan o en tal motivo repiten sus niveles (Baquerizo & López, 2014). Gran parte de las instituciones de educación superior en el Ecuador atraviesan una constante lucha por elevar los índices de permanencia y egreso de los estudiantes, por este motivo han elaborado un sinnúmero de políticas y estrategias integrales que contribuyen a elevar los niveles de eficiencia académica, sin embargo, no se ha considerado identificar las características de los estudiantes que son candidatos potenciales a abandonar su carrera universitaria, y mucho menos a detectar los casos singulares de deserción.

Objetivos

Objetivo General

Detectar patrones de deserción estudiantil utilizando técnicas descriptivas de agrupamiento, asociación y atípicos en minería de datos para la gestión académica en la Universidad Técnica del Norte.

Objetivos Específicos

Construir un marco teórico que fundamente las técnicas descriptivas de minería de datos y el proceso de descubrimiento de conocimiento en base de datos (Knowledge Discovery in Databases - KDD).

Obtener un data warehouse a partir de los datos académicos y socioeconómicos de los estudiantes utilizando el proceso KDD en la herramienta Pentaho

Aplicar técnicas descriptivas de agrupación, asociación y detección de atípicos a la vista minable utilizando el software Weka.

Adquirir patrones de deserción estudiantil y validar los resultados obtenidos mediante métodos estadísticos, métricas cuantitativas de calidad y característica de consistencia de la ISO/IEC 25012.

Justificación

Identificar de manera temprana los estudiantes que tienen mayor probabilidad de abandonar su carrera universitaria, significa un aporte importante para optimizar los estándares de calidad en la educación. Por este motivo, el uso de minería de datos en estudios de esta naturaleza resulta fundamental, ya que su análisis y aplicación ha tenido un impacto significativo en los últimos años, puesto que el uso de sus técnicas permite, entre otras cosas, prever cualquier hecho dentro del ámbito de investigación (Lara, 2014). En la presente investigación se elaborará un estudio para detectar los patrones, clústeres y atípicos que rigen la deserción estudiantil en la Universidad Técnica del Norte, utilizando los datos académicos y socio económicos, de los estudiantes a nivel de grado de la UTN de los últimos 5 años.

El ámbito esencial de la presente investigación es que, mediante técnicas descriptivas de agrupamiento y asociación en minería de datos, obtener los diferentes clústeres que generará dichos algoritmos; información que será utilizada para que las personas encargadas tomen decisiones estratégicas en beneficio de la Universidad Técnica del Norte con el fin de alcanzar

mejores estándares de calidad, ya que uno de los factores que influye en la calidad de la educación superior es la retención de los estudiantes.

Alcance

La finalidad de este estudio propuesto es obtener los patrones, clústeres y atípicos de deserción estudiantil mediante las técnicas descriptivas de agrupamiento, asociación y atípicos de minería de datos; para lo cual se llevará a cabo un proceso que inicia con la obtención de los datos académicos y socio económicos de los estudiantes de nivel de grado de las diferentes facultades de la Universidad Técnica del Norte de la base de datos Oracle, el siguiente paso a seguir es realizar el proceso ETL (Extracción Transformación y Limpieza), para lo cual se empleará la herramienta de Pentaho 7, después de este proceso se obtendrá la vista minable, la cual se usará para aplicar los diferentes algoritmos descriptivos de agrupamiento y asociación de minería de datos. Con los datos que se obtengan de los diferentes algoritmos, se evaluará cuál de estos obtuvo mejores resultados, para realizar las respectivas validaciones con métodos cuantitativos de calidad tales como: matriz de confusión, curvas ROCC, el coeficiente capa, error cuadrático, entre otros, así mismo métodos estadísticos que permitirán realizar una buena interpretación de la información estudiantil que servirán para en un futuro identificar de manera apropiada a los estudiantes que son propensos a abandonar sus estudios.

CAPÍTULO 1

Marco teórico

1.1. Introducción a la minería de datos

Actualmente, el desarrollo tecnológico ha permitido analizar y procesar grandes volúmenes de información que reposa en las bases de datos, generando información estratégica para la toma de decisiones, por lo que el término minería de datos toma un papel importante en el análisis de datos.

El origen de la palabra minería se relaciona con el arte de extraer minerales de la corteza terrestre, en este caso la extracción se realiza de las bases de datos existentes, y la palabra datos se define específicamente como el valor que toma una variable. La minería de datos nace en la década de 1990, ya que la concepción que se tenían de los datos empezó a cambiar, hasta entonces los datos que se almacenaban únicamente servían como soporte para la operativa diaria, sin embargo, esta mentalidad cambió y los datos empezaron a considerarse como una fuente importante de conocimiento útil para generar beneficios. Años atrás para referirse a minería de datos, se hablaba de “data fishing” o “data dredging”, dichos términos desaparecieron con el tiempo dando origen al término minería de datos (Lara, 2014).

La minería de datos es un área extensa de la informática, que abarca numerosas técnicas de análisis de datos y extracción de modelos, que nace específicamente por la necesidad de analizar la gran cantidad de datos que se generan cada día, debido a la automatización de la mayoría de los procesos en las instituciones, el crecimiento acelerado del internet, de igual forma la evolución de los sistemas de almacenamiento masivo y que dichos datos se encuentran subutilizados (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

1.2. Relación de la minería de datos con otras disciplinas

La minería de datos no nace como un área de la informática totalmente nueva, ya que se nutre de la investigación y avances que se produce en las áreas relacionadas a ella, En la Fig. 1 se detallan las disciplinas que más influyen en la minería de datos según Lara (2014).

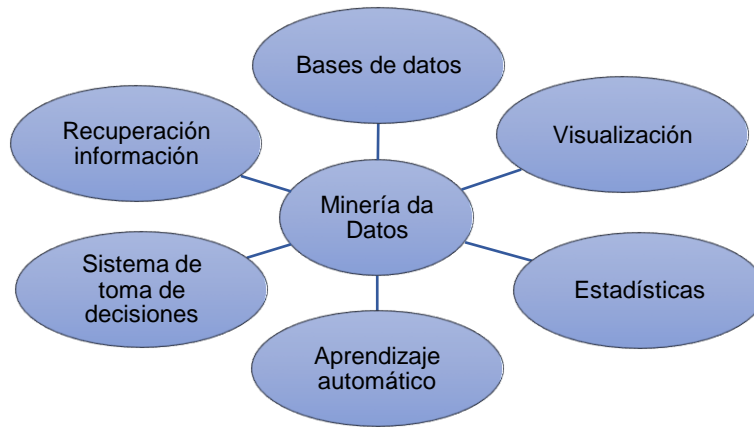


Fig. 1. Disciplinas que contribuyen a la minería de datos - Fuente (Lara, 2014)

1.2.1. Bases de datos

La mayoría de la información recopilada viene de las bases de datos transaccionales (por ejemplo, Oracle, SQL Server, PostgreSQL, MySQL, etc.). Por ello las técnicas de indexación y de acceso eficiente a los datos son muy importantes para el diseño de algoritmos de minería de datos (Hernández Orallo et al., 2004).

1.2.2. Visualización

Las técnicas de visualización se tornan indispensables, ya que permiten al usuario entender los patrones que son más difíciles de percibir por medio de descripciones matemáticas o resultados textuales, entre ellas están los diagramas de barras, graficas de dispersión, histogramas, etc. (Lara, 2014).

1.2.3. Estadísticas

La estadística es una disciplina sumamente importante para la minería de datos ya que varios de sus conceptos y técnicas se utilizan en ella, por ejemplo, la media, la varianza, las distribuciones, entre otros.

1.2.4. Aprendizaje automático

El aprendizaje automático que se encuentra estrechamente relacionado con la minería de datos ya que los principios que siguen son los mismos: la máquina aprende un modelo partiendo de ejemplos y lo usa para resolver un problema (Hernández Orallo et al., 2004)

1.2.5. Sistema de toma de decisiones

El objetivo de ambas disciplinas es brindar información necesaria para la toma de decisiones efectivas en el ámbito empresarial o en las tareas de diagnóstico.

1.2.6. Recuperación información

Habitualmente una de las tareas principales es encontrar documentos por medio de palabras claves para ello se utiliza medidas de similitud entre varios documentos que son aplicadas más generalmente en minería de datos.

1.2.7. Otras

Dependiendo del tipo de información que se pretende analizar la minería de datos emplea técnicas de otras disciplinas tales como el análisis de imágenes, procesamiento de señales, visión por computador etc.

1.3. Tipos de datos

La minería de datos se puede aplicar a cualquier tipo de información, siempre y cuando se apliquen las técnicas apropiadas para la información recopilada. El dato es una representación simbólica, numérica, alfabética, algorítmica, etc., de un atributo o variable que puede ser cualitativa o cuantitativa (Lara, 2014), por ejemplo, para la entidad estudiante se tienen varios atributos con su respectivo dato, a continuación, en la Tabla 1.1 se muestran los datos asociados a esta entidad.

TABLA 1.1
ENTIDAD ESTUDIANTE

Atributo	Dato
Nombre	Saúl Cisneros
Edad	26
Carrera	CISIC
Nivel	10

Fuente: Propia

A continuación, se explicarán las diferencias entre los datos que vienen de bases de datos relacionales de otros tipos de estructuras de datos, espaciales, temporales y multimedia, así como de datos provenientes de diferentes tipos de repositorios.

1.3.1. Bases de datos relacionales

Una base de datos relacional está formada por un conjunto de relaciones que son representadas por medio de tablas con la finalidad de dar soporte a los procesos básicos de la organización, ventas, producción, personal, etc. (Pérez López & Santín González, 2007). Las tablas están compuestas por un conjunto de filas denominadas instancias, cada instancia representa a un objeto en la tabla que tendrá una clave primaria o identificador. Cada instancia tendrá diferentes columnas que serán los atributos de cada objeto, que dependiendo del caso se relacionará con una o varias instancias de una segunda tabla (Lara, 2014). En la Fig. 2 se muestra un esquema en el cual se tiene una base de datos relacional.

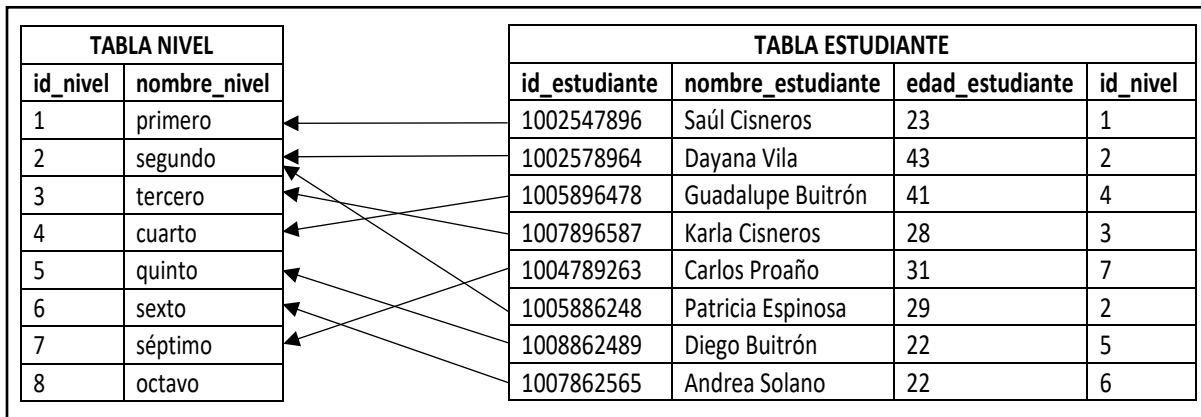


Fig. 2. Base de datos relacional - Fuente: Propia

Entre los diferentes atributos que se pueden almacenar en una base de datos están los enteros, reales, fechas, cadenas de texto, entre otros, sin embargo, en la minería de datos los tipos más utilizados son los numéricos y los categóricos.

a. Atributos numéricos

Los atributos numéricos son aquellos que contienen valores enteros o reales, por ejemplo, la edad de una persona o el número de visitas a un sitio web.

b. Atributos categóricos

Los atributos categóricos también conocidos como nominales son un conjunto finito y preestablecido de categorías por ejemplo el estado civil, el sexo, etc.

1.3.2. Bases de datos desnormalizadas

Las bases de datos desnormalizadas se encuentran diseñadas de forma diferente que las bases de datos relacionales, ya que guardan información en torno a ellos, en este caso de acuerdo si el estudiante abandonó o no su carrera. Estos hechos se caracterizan por una serie de dimensiones, como por ejemplo los estudiantes de que facultad son los que más abandonan su carrera (Lara, 2014). En la Fig. 3 se muestra un esquema de bases de datos desnormalizado.

id_estudiante	nombre_estudiante	edad_estudiante	nombre_nivel
1002547896	Saúl Cisneros	23	primero
1002578964	Dayana Vila	43	segundo
1005896478	Guadalupe Buitrón	41	cuarto
1007896587	Karla Cisneros	28	tercero
1004789263	Carlos Proaño	31	séptimo
1005886248	Patricia Espinoza	29	Segundo
1008862489	Diego Buitrón	22	quinto
1007862565	Andrea Solano	22	sexto

Fig. 3. Esquema de base de datos desnormalizado - Fuente: Propia

1.3.3. Otros tipos de bases de datos

En la actualidad existen un sin número de aplicaciones que requieren otro tipo de estructura de la organización de la información, otros tipos de bases de datos que contienen datos complejos son las siguientes:

a. Base de datos espaciales

Este tipo de base de datos contienen información específica como datos geográficos, médicos, de transporte, etc., en las cuales las relaciones especiales cumplen un papel fundamental. Al aplicar minería de datos sobre este tipo de bases, se puede encontrar información tal como las características de una zona geográfica, movilidad en una ciudad, etc.

b. Base de datos temporales

Las bases de datos temporales incluyen atributos que se relacionan con el tiempo, pueden ser intervalos temporales, generalmente lo que se obtiene después de aplicar minería de datos son características de la evolución o tendencias del cambio basados en distintas medidas o valores.

c. Base de datos documentales

Las bases de datos documentales contienen información que puede ir de palabras claves a resúmenes que describen objetos. La minería de datos en este tipo de bases permite obtener asociación entre contenidos, agrupar o clasificar los objetos textuales.

d. Base de datos multimedia

Las bases de datos multimedia soportan objetos de gran tamaño como imágenes, audio o video, para la minería de este tipo de bases de datos es indispensable integrar técnicas de almacenamiento.

1.4. Proceso KDD

En si la minería de datos es una fase de un proceso mayor llamado proceso de descubrimiento en conocimiento, o mejor conocido como proceso KDD (Knowledge Discovery in Databases). Generalmente, el proceso KDD es requerido por los encargados de velar por el giro de negocio de las de las organizaciones con la finalidad de buscar una solución a diferente problema de negocio que existan, para lo cual este trabajo está orientado a especialistas en minería de datos.

El proceso KDD consiste en emplear a una determinada base de datos operaciones de: selección, exploración, muestreo, transformación y métodos de modelado, para que posterior a este proceso se obtendrá patrones que serán evaluados y de igual forma obtener el conocimiento, el cual es el objetivo principal de dicho proceso, en la Fig. 4 se muestra gráficamente las fases del proceso KDD.

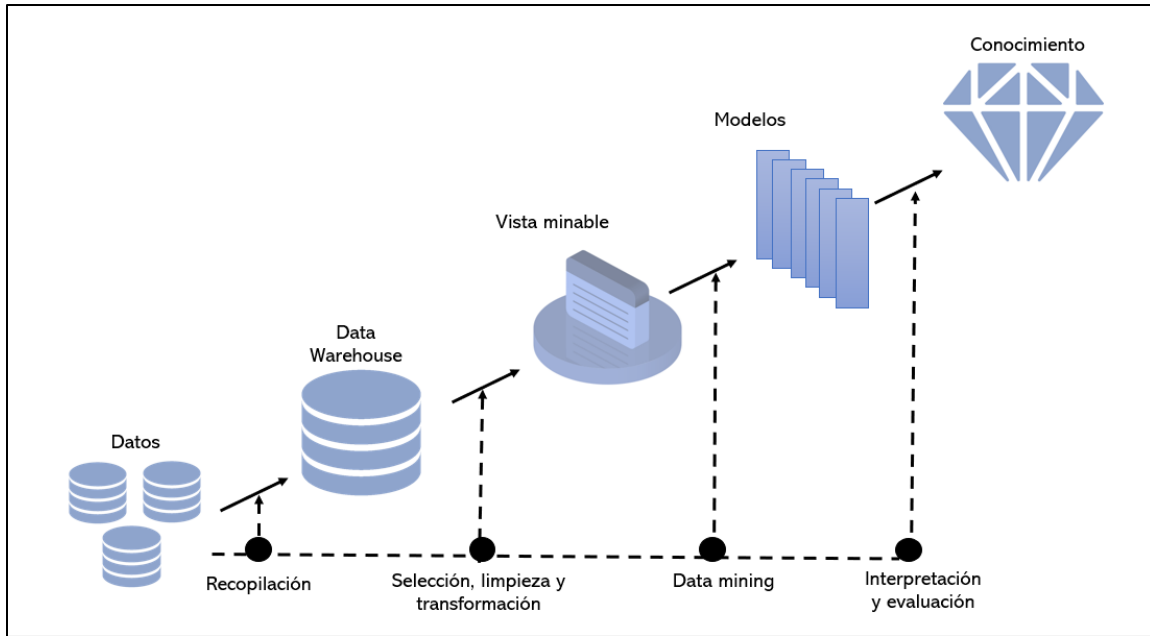


Fig. 4. Proceso KDD - Fuente:(Lara, 2014)

1.4.1. Recopilación de datos

La recopilación de datos es el paso inicial del proceso KDD, el cual es el encargado de compilar toda la información necesaria para su posterior análisis.

Todas las organizaciones tienen sus sistemas de bases de datos basados en el procesamiento tradicional en línea (OLTP), básicamente son bases de datos las cuales están preparadas para realizar todo tipo de transacciones que una organización ejecuta diariamente, sin embargo, no son aptas para el procesamiento, análisis, predicción, toma de decisiones estratégicas, etc.; ya que para realizar este tipo de proceso se necesitan técnicas especializadas para efectuar este tipo de análisis. Los datos necesarios para realizar un análisis de información completo pueden provenir de diferentes departamentos de una misma organización e incluso, es necesario la recopilación de datos que provienen de organizaciones externas. Una vez que se tiene claro que datos se usará para el posterior procesamiento, serán almacenados en un solo repositorio llamado almacén de datos (data warehousing) (Hernández Orallo et al., 2004).

Un almacén de datos se forma a partir de la integración de las bases de datos bajo un esquema unificado con el que se trabajará posteriormente como se aprecia en la Fig. 5. Los almacenes de datos están diseñados a partir de un esquema desnormalizado, los cuales se modelan de una estructura multidimensional (Lara, 2014). Esta perspectiva multidimensional de

un almacén de datos es adecuada para el procesamiento analítico en línea (OLAP), ya que permite un análisis multidimensional de los datos, ya que se puede utilizar el conocimiento previo sobre el dominio de los datos para mostrar la información a diferentes niveles de abstracción (Hernández Orallo et al., 2004).



Fig. 5. Etapa de recopilación de datos, fuente:(Lara, 2014)

Para realizar la recopilación de datos se requiere el uso de una herramienta que ‘permita integrar los datos en este caso se empleará la herramienta Pentaho, la cual es una plataforma moderna de integración de datos, orquestación y análisis de negocios; generalmente se emplea este software integral para acceder, preparar, combinar y analizar cualquier dato de cualquier fuente (Hitachi, 2019).

1.4.2. Selección, limpieza y transformación de datos

La calidad de los datos a ser procesados es indispensable en las tareas de minería de datos, por lo que es muy importante eliminar los datos que son irrelevantes, o innecesarios. La elección de los atributos con los cuales se trabajará es muy importante, ya que los datos deben ser selectos específicamente para obtener la vista minable.

Selección de Datos

Por lo general, las herramientas de minería de datos son las encargadas de elegir cual variable es o no necesaria en su análisis para obtener el conocimiento deseado, no obstante esta no es una buena práctica ya que para conseguir buenos resultados, lo más factible es elegir cuidadosamente las variables con las cuales se trabajará (Hernández Orallo et al., 2004).

El objetivo de la selección de datos, es filtrar los datos que no serán utilizados en el posterior análisis, la depuración de estos datos se puede realizar en varios niveles que se encuentren los datos (Lara, 2014):

- **Filtrado de atributos**

En ciertas ocasiones, será necesario la eliminación de atributos los cuales son de importancia para el posterior análisis, por ejemplo, en un almacén de datos de una Universidad, es necesario filtrar atributos como cédula, nombre de los estudiantes ya que no son relevantes al momento de realizar el análisis.

- **Filtrado de registros**

Al igual que el filtrado de atributos, existen registros innecesarios, en ciertas ocasiones no serán de ayuda para el posterior análisis, los cuales se deben eliminar. Por ejemplo, Al momento de ejecutar una minería de datos con estudiantes que son de primero a quinto nivel de alguna carrera en específico, y existen registros de estudiantes que están de sexto nivel en adelante, es necesario eliminar dichos registros que no serán de ayuda.

De igual forma, existen ocasiones en que es necesario realizar un filtrado de registros ya que tienen un número elevado de estos. Para lo cual se realiza la llamada muestra, que consiste en obtener un conjunto reducido de registros. Existen técnicas de muestreo que sin las siguientes (Lara, 2014):

- **Muestreo aleatorio simple:** Quiere decir que todos los registros tienen las mismas probabilidades de ser elegidos para su ser tratado posteriormente.
- **Muestreo aleatorio estratificado:** El objetivo de este método es que todos los registros que se todos los grupos o registros que se elegirán serán simbolizados de forma equilibrada.
- **Muestreo de grupos:** Este tipo de muestreo va especificado a un área en común con parámetros que hayan sido tomados en cuanta.

Limpieza de datos

Habitualmente, se encontrarán datos los cuales son incoherentes, los cuales se deberá apartarlos de los registros. Por ejemplo, existen datos los cuales no tienen sentido en comparación de los demás datos, por ejemplo, la edad de un estudiante el cual se encontrará registro de edades de 90 años lo cual se procederá a hacer la limpieza de dichos datos.

Existen dos problemas que regularmente se necesitan ser tratados los registros con una limpieza de datos lo cuales son (Lara, 2014):

- **La ausencia de valores**

Frecuentemente, los registros que se encuentran vacíos o incompletos, regularmente por errores humanos o del sistema al momento de digitar los mismos (Hernández Orallo et al., 2004). De igual forma, la inexistencia de estos es por causas que el dueño de los datos no quiso que dichos valores sean visibles; en ambos casos los datos que necesitan una limpieza, ya que no ayudaran en nada con el objetivo del análisis (Lara, 2014).

- **La existencia de valores erróneos**

En este caso los datos existen y están con alguna información, no obstante, son datos erróneos, que no tiene sentido al momento de compararlos con los datos de su mismos atributos (Hernández Orallo et al., 2004). Por ejemplo, se puede encontrar en algún registro en donde se esperaba el campo etnia algo que no tiene nada que ver con la etnia, entonces se puede decir que es un valor erróneo.

Trasformación de datos

La transformación de datos consiste en convertir los datos originales a un solo formato el cual puede ser cualitativo o cuantitativo (Lara, 2014).

De igual forma, los diferentes algoritmos que posteriormente se aplicarán para realizar las tareas de minería de datos, necesitan que los datos de entrada se encuentren en un formato acorde con cada una de los algoritmos(Pérez López & Santín González, 2006). Otra de las necesidades por la cual se debe cambiar el formato de los datos es que los atributos originales no tiene mucho poder predictivo por si solos (Hernández Orallo et al., 2004). Existen diferentes técnicas de transformación de datos:

- **Numerización**

Es el proceso por el cual se convierte un atributo de tipo cualitativo a cuantitativo, por ejemplo, al momento de reemplazar el valor aprobado o reprobado en un registro, dichos datos se cambia por 0 que equivale a reprobado y 1 a aprobado (Lara, 2014).

- **Discretización**

La discretización es el proceso mediante el cual los valores se incluyen en rangos, para que haya un número limitado de estados posibles (Minewiskan, 2018).Por ejemplo, al momento de

tener una variable de tipo edad, entonces para proceder con la discretización, se coge los valores de los diferentes rangos de edad y se agrupa en tres diferentes categorías que puede ser edad alta, edad media y edad baja.

- **Creación de características**

El proceso de creación de características, es esencialmente unir atributos existentes para formar un nuevo (Lara, 2014).

- **Normalización**

Este método tiene por objetivo la transformar los atributos en un rango moderado, generalmente se aplica la **normalización lineal uniforme**, la cual consiste en transformar todos los atributos mediante una fórmula a un rango entre [0,1] (Lara, 2014).

- **Reducción de dimensionalidad**

Esta técnica multivariante, trata de reducir el número de variables originales en un número menor de variables, denominadas componentes principales. Son una combinación lineal de las variables iniciales, que extraen la mayor parte de la información contenida en los datos originales (Sierra, 2006). Los algoritmos de dimensionalidad más conocidos son el Análisis de Componentes Principales (PCA) y el Análisis de Componentes Independientes (ICA) (Hernández, Delgado, Rivera, & Castellanos, 2006).

A continuación, en la Fig. 6 se representa la etapa de selección, limpieza y transformación

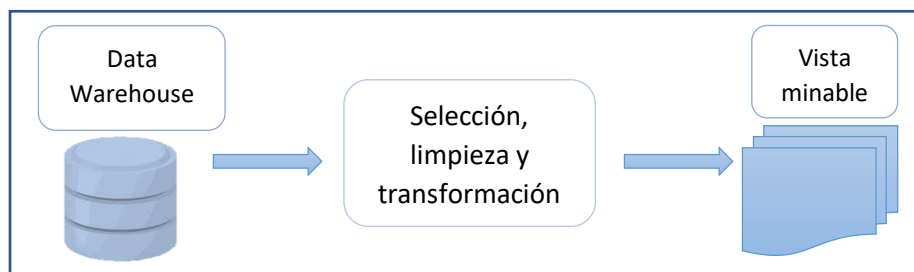


Fig. 6. Etapa de selección, limpieza y transformación, fuente (Lara, 2014)

Al igual que para la etapa de recopilación de información, para esta etapa se empleará la herramienta Pentaho; sin embargo, adicionalmente en ciertos casos puntuales se utilizará la herramienta Microsoft Excel (Microsoft, 2019).

1.4.3. Minería de datos

La etapa de minería de datos es la más significativa del proceso KDD, ya que es aquí en donde se generará el nuevo conocimiento a partir del almacén de datos. Primeramente es necesario estar claro de que tarea, y un algoritmo de minería de datos se usará en el estudio (Hernández Orallo et al., 2004).

- **Tareas de minería de datos**

En el mundo de la minería de datos, existen tareas, las cuales son encargadas de integrar toda la información necesaria para iniciar un proceso de preparación, y que estas serán resueltas por algoritmos de minería de datos específicos para cada tarea existente. Existen dos tipos de tareas, las predictivas y las descriptivas (Hernández Orallo et al., 2004).

Técnicas predictivas

El objetivo principal de las tareas predictivas es obtener modelos o patrones los cuales serán de suma importancia para en un futuro realizar aportes y dar solución a los problemas que la organización aqueja (Pérez López & Santín González, 2006).

- **Clasificación**

Este método es uno de las más usadas en cuanto a tareas predictivas, su principal objetivo es en base a los registros que se esté analizando, de igual forma a los atributos de los mismos, se encontrará un nuevo modelo (Pérez López & Santín González, 2006).

- **Regresión**

En este tipo de problemas, las variables a ser analizadas son específicamente numéricas, como por ejemplo, al momento de predecir entre un histórico de notas de un registro de estudiantes, en este caso se está utilizando valores los cuales darán a conocer una predicción (Pérez López & Santín González, 2006).

Técnicas descriptivas

Las tareas descriptivas conllevan principalmente a buscar grupos de características similares entre los datos de estudio y describirlas según las características que se busque en especial, puede ser esta descripción de forma detallada , usando los diferentes atributos que tiene los

objetos o de una forma simple (Pérez López & Santín González, 2006). Las principales técnicas descriptivas en la minería de datos son:

- **Agrupamiento**

Más conocido como tareas de “agrupamiento” o “clustering”, hace referencia al problema típico que nace al querer agrupar objetos homogéneos en el data warehouse, los datos son unidos con el objetivo de maximizar la similitud entre ellos, de igual forma se minimizará la similitud entre todos los grupos que se generaron. Este tipo de tareas se basan en aprendizaje no supervisado. En algunos casos, brinda la posibilidad de elegir el número de grupos que se pretende obtener después de aplicar el método, como también existe la posibilidad de que el algoritmo determine el número de grupos que se generarán (Hernández Orallo et al., 2004).

Por ejemplo, un sistema que obtiene datos referentes a edad y peso de los usuarios, como se muestra en la Tabla 1.2. Se busca encontrar objetos similares entre sí, para posteriormente realizar grupos de objetos o clúster.

TABLA 1.2
TABLA CON ELEMENTOS SIMILARES

Identificador	Edad	Peso (Kg)
1001	23	70
1002	22	67
1003	14	45
1004	13	49
1005	45	79

Fuente: Propia

A simple vista, después de analizar la tabla resulta fácil identificar los grupos de datos similares, de los cuales se puede destacar:

- Personas de edad y peso bajos,
- Personas de edad y peso medios
- Personas de edad y peso altos

Aparentemente resulta sencillo establecer los diferentes grupos que existen en este ejemplo, pero en otras ocasiones se encontrará datos que resulte imposible realizar este procedimiento sin la ayuda de herramientas o alguna técnica en este caso técnicas de clustering.

La mayoría de las técnicas de clustering tienen mucha relevancia en los diferentes áreas de estudio como, por ejemplo: en el marketing, compañías de seguros, planificación urbana, la World Wide Web, entre otras. Para poder aplicar las técnicas de clustering particional, es necesario analizar cuan similares son dos objetos, una forma es utilizar las medidas de distancia. Existen medidas de distancia, entre las más conocidas se encuentran Manhattan, Euclídea y Minkowski (Lara, 2014).

Cada una de estas medidas de distancia, tienen sus ventajas y desventajas según el trabajo que se lo asigne a alguna de estas, no se puede decir a priori cuál es mejor o peor.

- **Agrupamiento particional**

Esta técnica de clustering se basa específicamente en dividir un conjunto de datos en subconjuntos con una intersección vacía. Para lo cual, inicia el proceso asignando a cada uno de los objetos en análisis a un clúster según su semejanza.

La técnica de clustering particional más conocida es el algoritmo **K-medias**(K-means). Este algoritmo funciona de una forma sencilla, primeramente, para que el algoritmo inicie su proceso es necesario asignar a priori el número de clúster que se generará. Seguidamente el algoritmo elegirá un punto inicial (centroide) para representar cada uno de los clúster, a los cuales, se asignaran objetos cuyo centroide es más cercano a él (Hernández Orallo et al., 2004).

Una de las ventajas de usar k-medias es su eficiencia y facilidad para obtener particiones óptimas. Se considera como desventaja el hecho de definir el número de clústeres que se generará, aun que para otros podría no ser una desventaja este punto (Lara, 2014).

Según (Sierra, 2006) este algoritmo puede resumirse fundamentalmente en los siguientes pasos:

1. De entre los m casos elegir k que llamaremos centroide y denotaremos $c_j, j=1, \dots, k$. Cada centroide c_j , representará al clúster $C_j (j=1, \dots, k)$.
2. Asignar el caso i al clúster C_j cuando con $d(x_i, c_j) = \min_{j=1, \dots, k} d(x_i, c_j)$. Es decir, cada caso se asigna al clúster que representa el centroide que tiene más cerca.

Los pasos 1 y 2 proporcionan una partición inicial de los casos

3. Calcular la mejora que se producirá en el criterio elegido al asignar un caso a otro clúster en el que no está actualmente.
4. Hacer el cambio que mayor mejora produce en el criterio.

5. Repetir los pasos 3 y 4 hasta que ningún cambio haga mejorar el criterio.

- **Agrupamiento basado en probabilidades**

Existen varios algoritmos basados en procesos no deterministas, denominados Modelos Ocultos de Markov, que es una máquina de estados finita probabilística, es decir, un conjunto de $(N+1)$ estados conectados unos a otros por medio de transición, el objetivo principal del análisis es encontrar un conjunto de datos que pueden considerarse como una muestra de los posibles casos de las categorías ocultas. Estadísticamente se puede asumir que una categoría oculta es una distribución espacial de los datos que se puede representar usando una función de densidad de probabilidad (función de distribución); a esta categoría oculta se la denomina clúster probabilístico (Han et al., 2001).

Uno de los algoritmos más conocidos de este modelo es el algoritmo **Exactitud-Maximización (EM)**, que es una aproximación iterativa de máxima verosimilitud que utiliza para encontrar una estimación del conjunto de parámetros del modelo con el objetivo de intentar maximizar la probabilidad de generación de los datos $O = \{O_0, \dots, O_k\}$, partiendo del modelo λ , $P(O | \lambda)$, de tal forma que la probabilidad queda asociada al modelo (λ^*) sea mayor o igual a la del modelo anterior (Sierra, 2006):

$$P(O | \lambda^*) \geq P(O | \lambda)$$

- **Agrupamiento basado en densidad**

Esta técnica se basa en algoritmos que trabajan con el concepto de densidad de un punto, que mide el número de puntos que son asequibles desde dicho punto considerando un radio determinado. Estas técnicas generalmente son muy robustas frente a errores o ruidos en los datos de entrada, en particular el algoritmo **DBSCAN**, se caracteriza por funcionar exitosamente independientemente de la forma y tamaño de los clústeres (Lara, 2014).

La mayoría de los algoritmos de agrupamiento basado en densidad utilizan los siguientes parámetros:

- Radio o épsilon que representa el radio considerado para medir la densidad de cada punto.

- Número mínimo de vecinos que representa el número mínimo de objetos en la vecindad de otro.

- **Asociación**

La obtención de reglas de asociación permite localizar relaciones de asociación o correlación entre un conjunto extenso de datos. En un principio, las reglas de asociación nacieron por la necesidad de las compañías de encontrar relación entre los registros o transacciones almacenados en sus bases de datos (Hernández Orallo et al., 2004).

El algoritmo más utilizado para obtener las reglas de asociación es conocido como **Apriori**. Este algoritmo busca un conjuntos de combinaciones de valores de atributos (ítems), con determinada cobertura (número de instancias que la regla predice correctamente), para ello el algoritmo inicia formando conjuntos de un solo ítem que supera la cobertura mínima, con este conjunto de conjuntos obtenidos, se utiliza para formar el conjunto de conjuntos de dos ítems, y así sucesivamente se continua formando conjuntos hasta llegar a un punto que el tamaño de la cobertura requerida exceda el número de ítems (Hernández Orallo et al., 2004).

De manera formal, la asociación se define como (Lara, 2014):

- a. Sea $I = \{i_1, i_2, \dots, i_m\}$ un conjunto de literales, denominados ítems. Sea D un conjunto de transacciones, donde cada transacción T es un conjunto de ítems, tal que $T \subseteq I$.
- b. Se dice que una transacción T contiene a X (un conjunto de ítems de I), si $X \subseteq T$. Una regla de asociación en una implicación de la forma $X \rightarrow Y$, donde $X \subset I$, $Y \subset I$, y $X \cap Y = \emptyset$.
- c. La regla $X \rightarrow Y$ tiene en el conjunto de transacciones D , una confianza de c si el $c\%$ de las transacciones de D que contiene X , también contienen Y .
- d. La regla $X \rightarrow Y$ tiene un soporte s si el $s\%$ de las transacciones D contienen $X \cup Y$.

- **Detección de atípicos**

Es una de las tareas de data mining que se utiliza para buscar dentro de un conjunto de objetos, aquellas características significativamente diferentes al resto de objetos analizados. En ciertos casos, esta tarea es usada previamente como herramienta de filtrado antes de la utilización de otra herramienta de data mining. Según Lara (2014) existen algunas técnicas de detección de atípicos las cuales son:

- **Aproximaciones estadísticas.**

El objetivo de esta técnica es buscar un modelo estadístico que represente a una población explícita, para posteriormente, analizar cada uno de los datos con un objeto en especial de determinar si es o no el objeto atípico (Lara, 2014).

- **Aproximaciones basadas en proximidad**

Esta técnica utiliza el concepto de distancia entre objetos, generalmente una de las técnicas más empleadas son las reglas de k vecino más próximos (K-NN por sus siglas en ingles), en las cuales la clasificación exige la definición de una medida de distancia entre los elementos del espacio en representación, con esta definición se tiene que cuando mayor sea la distancia entre dos elementos se tendrá un elemento atípico con mayor certeza (Sierra, 2006).

- **Aproximaciones basadas en clustering**

Cuando se divide un conjunto de datos en grupos con características homogéneas se analizan los datos y de ellos se extraen los atípicos para que los subconjuntos sean similares entre sí, si al eliminar el elemento mejora al clúster se considera como atípico al elemento eliminado, por este motivo se considera que los elementos que se encuentran en la zonas fronterizas de los clúster son atípicos (Lara, 2014).

En la Fig. 7 se muestra gráficamente la etapa de minería de datos a partir de la vista minable.

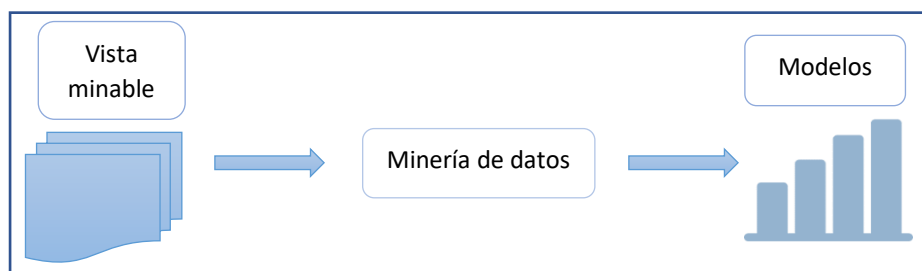


Fig. 7. Etapa de minería de datos, fuente (Lara, 2014)

Para la fase de minería de datos se empleará la herramienta Weka ya que es una colección de algoritmos de aprendizaje de máquina para tareas de minería de datos (Frank et al., 2016). Adicionalmente se puede emplear el software privativo de IBM SPSS, que ofrece una amplia biblioteca de algoritmos de aprendizaje automático (IBM, 2019).

1.4.4. Evaluación e interpretación

Al final del proceso KDD se encuentra la fase de interpretación y evaluación, su objetivo principal es medir la calidad de los patrones descubiertos, estos patrones debe ser precisos, comprensibles e interesantes (Hernández Orallo et al., 2004).

Técnicas de evaluación

- **Evaluación de modelos clustering**

Los resultados encontrados gracias a las técnicas de clustering, son segmentaciones de calidad. En comparación con las técnicas de clasificación, las técnicas de segmentación son más difíciles de evaluar, ya que evaluar los resultados de las técnicas de clasificación (supervisado) es cuestión de analizar si la clase de un objeto coincide o no con la predicción realizada por este algoritmo, mientras que determinar si una segmentación de datos encontrada por los algoritmos de segmentación (no supervisado), no es un asunto tan rápido de realizar (Lara, 2014).

Uno de los métodos más relevantes para analizar si la segmentación es de calidad es mediante el **error cuadrático** (EC) como se aprecia en la Ecuación 1.

$$EC = \sum_{i \in cluster_k} || t_i - c_k ||^2 \quad \text{Ec. 1}$$

En esta ecuación, t_i representa a cada uno de los objetos que pertenece al clúster, mientras que c_k representa el punto central de cada clúster obtenido (centroide). Mientras el error cuadrático sea menor, los puntos de cada clúster estarán más cercanos, por lo que la segmentación será de mejor calidad (Lara, 2014).

En algunos algoritmos de agrupamiento se puede aplicar también la matriz de confusión, que es una herramienta de visualización que facilita la identificación de los errores que comete el algoritmo, ya que en casos como el presente estudio la variable de clase ESTADO_CARRERA cuenta con dos categorías A (activo) e I (inactivo).

En casos como el anteriormente descrito, el error cuadrático no define si la segmentación se realizó bien o mal (Hamilton, 2018). A continuación, en la Tabla 1.3 se puede apreciar la estructura de la matriz de confusión.

TABLA 1.3
MATRIZ DE CONFUSIÓN DE DOS VARIABLES

		Asignado al Clúster (predicho)	
		Clúster 1 (+)	Clúster 2 (-)
Clúster	Clúster 1 (+)	TP	FN
	Clúster 2 (-)	FP	TN

Fuente: (Zelada, 2017)

Donde según Pina (2018):

True Positives (TP)

Los true positive o verdaderos positivos en español, son el número de agrupamiento correcto para la clase positiva.

True Negatives (TN)

Los true negatives o verdaderos negativos en español, son el número de agrupamientos correctos para la clase negativa.

False Positives (FP)

Los false positives o falsos positivos en español, son el agrupamiento positivo cuando realmente debía ser negativo. Este tipo de agrupamiento se conoce como errores de tipo I.

False Negatives (FN)

Los false negatives o falsos negativos en español, son el agrupamiento para la clase negativa cuando realmente debía agruparse en la clase positiva. Este tipo de agrupamiento se conoce como errores de tipo II y son peores que los errores tipo I.

De igual forma de la matriz de confusión se pueden extraer las siguientes medidas de calidad para determinar la bondad del algoritmo (Sierra, 2006):

Tasa de error

La validación de un algoritmo generalmente es medida basándose en la tasa de error, entendiéndose como error el agrupamiento incorrecto, a continuación, en la Ecuación 2 se muestra la fórmula para calcular la tasa de error:

$$Tasa\ de\ Error = \frac{Número\ de\ errores}{Número\ total\ de\ casos} \quad Ec. 2$$

Especificidad

La especificidad es la proporción de verdaderos negativos, en la Ecuación 3 se aprecia la fórmula para calcular la especificidad.

$$Especificidad = \frac{TN}{TN+FP} \quad Ec. 3$$

Accuracy

Esta medida representa el porcentaje de los aciertos del modelo, en la Ecuación 4, se muestra la fórmula para calcular el Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad Ec. 4$$

Recall

Es la proporción del agrupamiento de los verdaderos positivos entre todos los positivos siempre que $TP + FP$ sea mayor que 0 de lo contrario se define Recall como 1, en la Ecuación 5, se muestra la fórmula para calcular Recall.

$$Recall = \frac{TP}{TP+FN} \quad Ec. 5$$

Precisión

La precisión denota el agrupamiento verdadero entre los positivos, siempre que $TP + FP$ sea mayor que 0 caso contrario la precisión será 1, en la Ecuación 6 se muestra la fórmula de la precisión.

$$Precisión = \frac{TP}{TP+FP} \quad Ec. 6$$

- **Evaluación de modelos de asociación**

El método utilizado para evaluar las reglas de asociación, normalmente se trabaja con dos medidas, tales como: soporte y confianza de una regla, para posteriormente evaluar la calidad de esta (Hernández Orallo et al., 2004).

- **Soporte:** También conocido como cobertura, se conoce como el número de instancias en las que la regla se puede aplicar.
- **Confianza:** Es el porcentaje de veces que la regla se cumple cuando se puede aplicar

En la Ecuación 7, se aprecia la fórmula para calcular el soporte, donde x es igual al número de ítems en un conjunto de datos d:

$$\text{soporte}(x) = \frac{\|x\|}{\|d\|} \quad \text{Ec. 7}$$

En la Ecuación 8, se detalla la fórmula para calcular la confianza de la relación $x \rightarrow y$:

$$\text{confianza}(x \rightarrow y) = \frac{\text{soporte}(xy)}{\text{soporte}(x)} \quad \text{Ec. 8}$$

Para revisar un caso práctico donde se calcula el soporte y la confianza en una cesta de supermercado visitar el siguiente enlace <http://bit.ly/2GcfVI1>

En la Fig. 8, se encuentra explicada gráficamente la etapa de interpretación y evaluación.

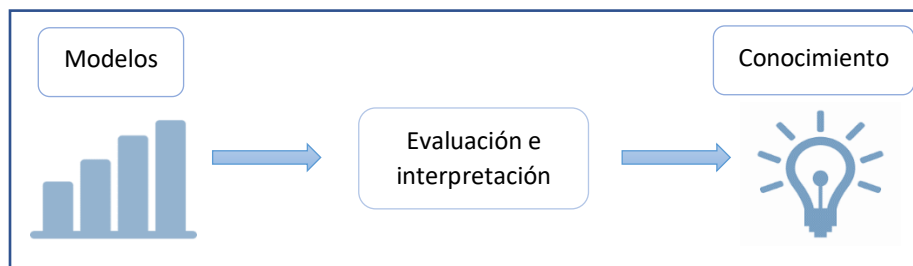


Fig. 8. Etapa de evaluación e interpretación

1.5. ISO/IEC 25012:2008

Las Normas Internacionales se delinean de acuerdo con las reglas establecidas por la Organización Internacional de Normalización (ISO) y la Comisión Electrotécnica Internacional (IEC), cuya tarea principal es preparar Normas Internacionales. La Norma ISO/IEC 25012 forma parte de un conjunto de Normas Internacionales bajo el título general Ingeniería de Software,

Requerimientos y Evaluación del producto de Software (SQueRE, por sus siglas en ingles) que se dividen en (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)

Generalmente un requisito previo para los proyectos tecnológicos es verificar la calidad de la información que es intercambiada, procesada y utilizada para los procesos de negocio. La calidad de la información se define en Norma Internacional ISO/IEC 25012:2008 con la finalidad de satisfacer estas necesidades considerando el ciclo de vida del software. El principal objetivo de esta Norma Internacional es generar un modelo general de la calidad de la información. Esta Norma Internacional define características de calidad para la información. El modelo de calidad de datos definidos en esta Norma Internacional está categorizado en diferentes atributos consideradas desde dos puntos de vista: inherentes y dependientes del sistema (ISO & IEC, 2014).

1.5.1. Calidad inherente de datos

La calidad de datos inherentes se refiere al grado en el cual las características de calidad tienen el potencial propio para satisfacer las necesidades manifestadas e implícitas cuando la información es usada bajo condiciones específicas.

1.5.2. Calidad de datos de un sistema dependiente

La calidad de datos de un sistema dependiente se refiere al grado en el que la calidad de los datos es alcanzada y preservada dentro de un sistema computacional cuando la información es usada en condiciones específicas. En la Tabla 1.4, se explican las Características del Modelo de Calidad de Datos

TABLA 1.4
CARACTERÍSTICAS DEL MODELO DE CALIDAD DE DATOS

Características	CALIDAD DE LOS DATOS	
	Inherentes	Dependientes del sistema
Exactitud	X	
Integridad	X	

Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Cumplimiento	X	X
Confidencialidad	X	X
Eficiencia	X	X
Precisión	X	X
Trazabilidad	X	X
Comprensión	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

Fuente: (ISO & IEC, 2014)

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 9, donde A es el número de registros en el campo especificado sintácticamente exacto y B es el número total de registros:

$$exactitud\ sintáctica\ del\ registro = \frac{A}{B} \quad EC. 9$$

- **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 10 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.

$$\text{integridad de datos en un archivo} = \frac{A}{B} \quad \text{EC. 10}$$

- **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 11, 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:

$$\text{credibilidad de datos} = \frac{A}{B} \quad \text{EC. 11}$$

CAPÍTULO 2

Desarrollo del Proceso de Descubrimiento de Conocimiento en Bases de Datos

El presente trabajo de minería de datos busca analizar la información personal, académica demográfica y socioeconómica de los estudiantes de la Universidad Técnica del Norte almacenada en las bases de datos transaccionales, con el objetivo de proporcionar información que permita a los directivos de la universidad tomar decisiones en base a la realidad. De igual forma fue desarrollado simultáneamente con el trabajo de titulación denominado “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” elaborado por Dayana Patricia Vila Espinosa, siendo partes de este capítulo un tronco en común de los dos trabajos.

A continuación, se detallarán las fases del Proceso de Descubrimiento del Conocimiento en Bases de Datos que se tomaron para obtener la información deseada.

En el desarrollo del proyecto, se tomará algunos puntos específicos los cuales se muestra a continuación.

2.1. Vista General del Proyecto

- **Suposiciones**

El desarrollo de la minería de datos inicia con la recopilación de información proveniente de los datos históricos de la UTN, para ser analizadas mediante técnicas descriptivas de minería de datos, para que posteriormente la información obtenida pueda ser analizada por los expertos del área y la empleen para mitigar esta problemática.

El desarrollo del proyecto depende de los roles que tomarán cada uno de los actores de este, el jefe del proyecto tiene el deber de determinar usuarios funcionales con conocimientos asociados al proyecto, usuarios técnicos que participarán aportando conocimientos relacionados con la problemática planteada y un patrocinador para el proyecto que contara con influencia directa o indirecta con las entidades asociadas (Guzmán, 2015)

El proyecto se deberá cumplir en función al cronograma establecido, considerando las etapas del proceso KDD.

- **Restricciones**

El proyecto para desarrollarse esta estimado a realizarse en un tiempo de 6 meses, para lo cual existe un cronograma de actividades definidas en el cual consta las diferentes actividades con su respectivo tiempo de cada una de ellas, por esto es dispensable cumplir con los tiempos establecidos porque cada actividad es secuencia del anterior, tomando en cuenta con la disponibilidad de los especialistas.

Para ejecutar la minería de datos, se debe tener en regla todos los permisos, asimismo, se debe adquirir un conocimiento previo para el uso de los diferentes softwares que se utilizará.

2.2. Entregables del Proyecto

Los entregables del proyecto se mencionarán a continuación, tomando en cuenta que mientras duran el proceso KDD estos pueden ir cambiando.

- Check list en base a la ISO/IEC 25012:2008
- Data Warehouse: Correspondiente a la fase de recopilación de datos
- Vista minable: Correspondiente a la fase de Selección, limpieza y transformación de datos
- Modelos descriptivos (asociación, agrupación y atípicos): Correspondientes a la fase de minería de datos.
- Conocimiento (agrupación y atípicos): Correspondiente a la fase de interpretación y evaluación

2.3. Organización del Proyecto

2.3.1. Participantes del Proyecto

En la Tabla 2.1 se especifica los directores de las áreas comprendidas:

TABLA 2.1
DIRECTORES DE LAS ÁREAS COMPRENDIDAS

Dependencia	Participante	Función
Coordinación de la carrera de Ingeniería en Sistemas Computacionales.	Ing. Pedro Granda	Asignar especialista en aprendizaje supervisado
Dirección de Desarrollo Tecnológico e Informático.	Mgs. Juan Carlos García	Asignar especialista en base de datos
Departamento de Bienestar Universitario	Dra. Eugenia Orbes	Especialista en la problemática

Fuente: Propia

En la Tabla 2.2 se especifica los participantes directos del 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 base de datos	Dirección de Desarrollo Tecnológico e Informático	Ing. Evelin Enríquez Ing. Fernanda Rivera
Analista de Sistemas	Carrera de Ingeniería en Sistemas Computacionales	Sr. Saúl Cisneros

Fuente: Propia

2.3.2. Roles y Responsabilidades

En la Tabla 2.3 se describen los roles de cada uno de los integrantes directos e indirectos del proyecto.

TABLA 2.3
ROLES Y RESPONSABILIDADES

Rol	Responsabilidad
Jefe de proyecto	Es el responsable de planificar, ejecutar y monitorizar las acciones que son parte de un proceso. Además se encuentra en colaboración directa con el patrocinador para su consecución de los objetivos, de igual forma tiene el deber de dirigir y coordinar los recursos empleados en todas las fases del proyecto (Business School, 2018).
Administrador de base de datos	Es el encargado de proporcionar los datos necesarios para su posterior análisis, dichos datos son requeridos por el analista de sistemas.
Analista de Sistemas	El analista de sistemas es el encargado de analizar la información proporcionada por el administrador de base de datos, validar e interpretar resultados.

Fuente: Propia

2.4. Gestión del Proceso

2.4.1. Estimaciones

En la Tabla 2.4, Tabla 2.5, y Tabla 2.6 se detalla el presupuesto estimado y recursos involucrados. Por tal motivo 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 primer paso para el desarrollo del proyecto es verificar la calidad de los datos, para posteriormente aplicar el proceso KDD que consta de varias fases hasta obtener el conocimiento, a continuación, 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 Recopilación de Datos	30
Implementación de la ISO/IEC 25012:2008	10
Fase de Selección Limpieza y Transformación de Datos	30
Fase de Minería de datos	50
Fase de Evaluación e Interpretación	170
Documentación	40
Análisis de Resultados	40
Análisis de Impactos	40
TOTAL	400

Fuente: Propia

A continuación, la Tabla 2.8 se detallan los hitos que determinan el término de cada fase.

TABLA 2.8
HITOS IMPORTANTES

HECHO	DESCRIPCIÓN
Revisión de la bibliografía	Obtener los conocimientos necesarios con la finalidad de entender ampliamente que pasos se pueden dar en el proceso KDD.

Recibimiento de los datos base	Es el hito más importante porque de este depende todo el proceso, ya que si se tienen datos que no aportan conocimiento, los resultados serán irrelevantes.
Finalización Check-list ISO	Se verifica la calidad de los datos y el grado en el que estos se comportan intrínsecamente.
Desarrollo del data warehouse	Es la construcción del almacén de datos con información personal, académica, social y económica de los estudiantes de la UTN.
Desarrollo de la vista minable	Se definen si los atributos serán categorizados o numerizados, con el objetivo que sean aplicables en la siguiente etapa del proceso KDD.
Desarrollo de los modelos descriptivos	Aplicación de algoritmos a la vista minable con el fin de obtener la información de acuerdo con un modelo descriptivo.
Verificación de métricas de calidad	Verificación de los niveles de calidad que cumple el modelo descriptivo y si son confiables los resultados.
Obtención del conocimiento	Interpretación del conocimiento y puesto de forma que personas que no son familiarizadas con el área los entiendan claramente.

Fuente: Propia

2.5. Recopilación de Datos

2.5.1. Tipos de datos base

Para el desarrollo de la primera etapa del proceso KDD, se emplearon los datos académicos, socioeconómicos, demográficos y personales de los estudiantes de la UTN que se almacenados en la base de datos Oracle 11g de la institución. Entre los diferentes tipos de datos con los cuales se trabajó se encuentran: enteros, reales, fechas y cadenas de texto, por ello para efectos del presente estudio interesa diferenciar entre dos tipos: numéricos (enteros y reales) y categóricos o discretos (toman valores en un conjunto finito de categorías) (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	Carácter
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

- **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	Carácter
DEPEN_CODIGO	Cadena de caracteres
DESCRIPCION	Cadena de caracteres
SIGLAS	Cadena de caracteres
OBSERVACION	Cadena de caracteres
ESTADO	Carácter
COD_SUBAREA_UNESCO	Cadena de caracteres
SECTOR	Entero
DEPEN_ANTIGUA	Cadena de caracteres

Fuente: DDTI

- **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	Carácter
NUMERO_MATRICULA	Entero
ANULACION	Carácter
FECHA_ANULACION	Fecha
PENSUM_CODIGO	Carácter
ESTADO_EVAL_DOC	Carácter

Fuente: DDTI

- **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	Carácter
MOTIVO_CODIGO	Cadena de caracteres
ESTADO	Carácter
FECHA_ULTIMA_MATRICULA	Fecha
TERMINA_CARRERA	Carácter
PIERDE_TERCERA	Carácter
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	Carácter
PENSUM_CICLO_ACAD_CODIGO	Cadena de caracteres
PENSUM_MODAL_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	Carácter

Fuente: DDTI

- **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	Carácter
LOCALIDAD_CODIGO	Cadena de caracteres
GENTILICIO	Cadena de caracteres
FUNCION	Carácter
OBSERVACION	Cadena de caracteres
ZONA_PLANIFICACION	Entero

Fuente: DDTI

- **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	Carácter
ARRASTRES	Entero
LEGALIZADO	Carácter
FECHA_LEGALIZACION	Fecha
CARNETIZADO	Carácter
CONTINGENCIA	Carácter

Fuente: DDTI

- **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	Carácter
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	Carácter
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	Carácter

Fuente: DDTI

- **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	Carácter
FECHA_NACIMIENTO	Fecha
GENERO	Carácter
ESTADO_CIVIL	Carácter
ESTADO	Carácter
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
IDENTIFICACIÓN	Cadena de caracteres

Fuente: DDTI

- **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

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:

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 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. 9 se muestra la transformación en PDI para obtener la dimensión LOCALIDADES

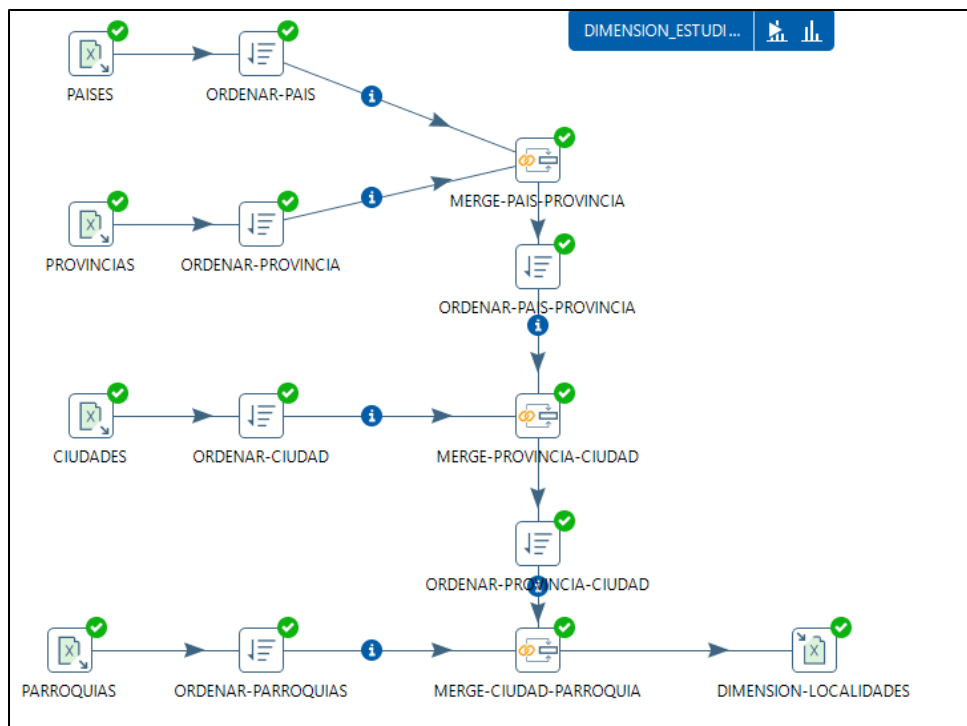


Fig. 9. Transformación PDI para Dimensión LOCALIDADES

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. 10 se muestra la transformación en PDI para obtener la dimensión DEPENDENCIAS

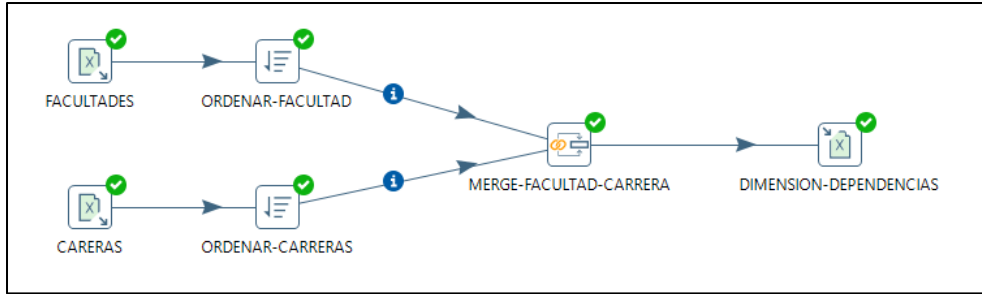


Fig. 10. Transformación PDI para Dimensión DEPENDENCIAS

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. 11 se muestra la transformación en PDI para obtener la dimensión ESTUDIANTE_CARRERA, que cuenta con 13675 filas:

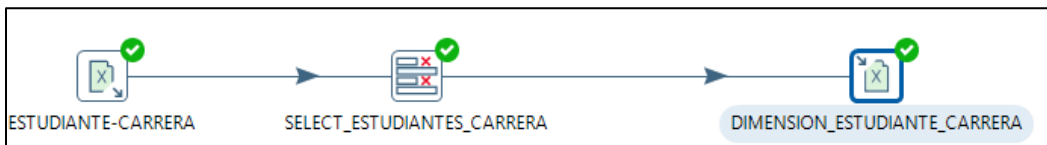


Fig. 11. Transformación PDI para Dimensión ESTUDIANTE_CARRERA

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

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. 12 se muestra la primera transformación en PDI para obtener la dimensión PERSONAS.

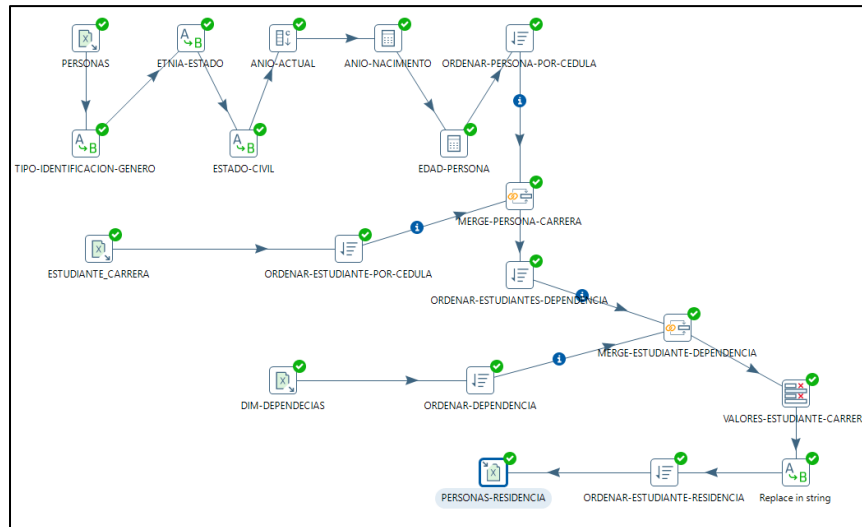


Fig. 12. Primera transformación PDI para Dimensión PERSONA

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

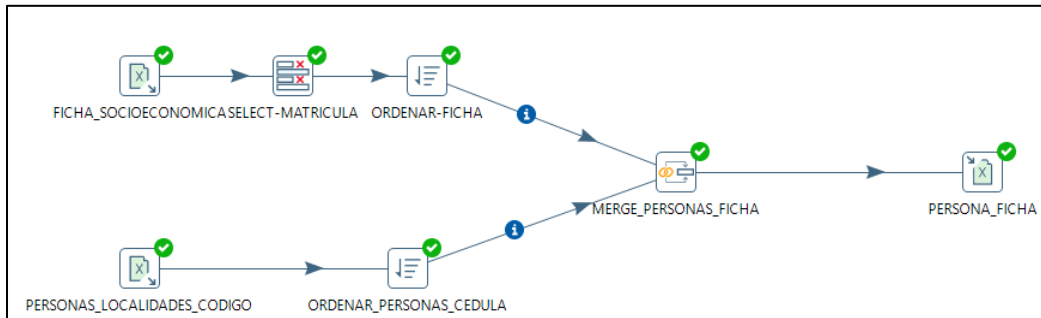


Fig. 13. Segunda transformación PDI para Dimensión PERSONA

En la Fig. 14 se muestra la tercera transformación en PDI para obtener la dimensión PERSONAS, que cuenta con 13677.

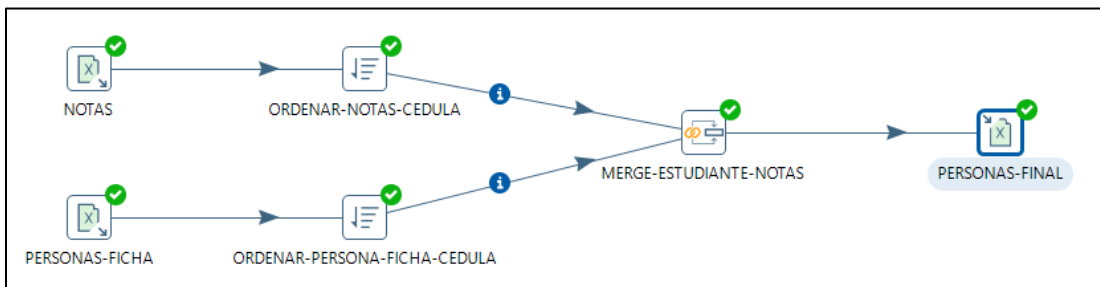


Fig. 14. Tercera transformación PDI para Dimensión PERSONA

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

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. 15 se muestra la transformación en PDI para obtener el DATA_WAREHOUSE.

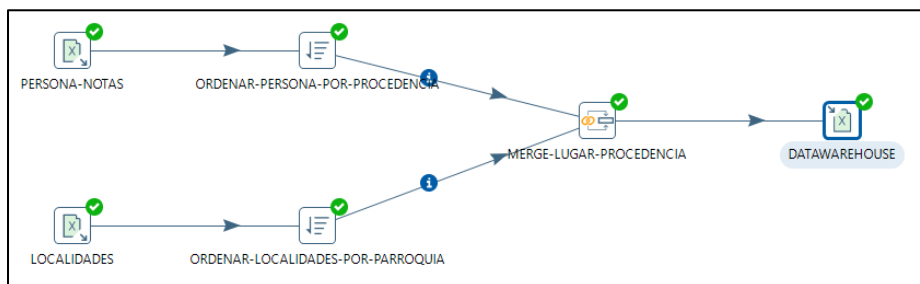


Fig. 15. Transformación PDI para el Data Warehouse

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:

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 el 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. 16 y Fig. 17.

#	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_MODAL_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. 16. Atributos de la tabla ESTUDIANTE_CARRERA

The screenshot shows a window titled 'Select / Rename values' with a step name of 'SELECT_ESTUDIANTES_CARRERA'. Below the title bar, there are tabs for 'Select & Alter', 'Remove', and 'Meta-data'. A table titled 'Fields:' contains the following data:

#	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. 17. 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. 18.

The screenshot shows a window titled 'Select / Rename values' with a step name of 'SELECT-VISTA-MINABLE'. Below the title bar, there are tabs for 'Select & Alter', 'Remove', and 'Meta-data'. A table titled 'Fields:' contains the following data:

#	Fieldname	Rename to	Length	Precision
1	CEDULA			
2	CONVIVIENTE			
3	TIPO_VIVIENDA			
4	FINANCIAMIENTO			
5	INGRESO_MENSUAL			
6	ACTIVIDAD_ESTUDIANTE			
7	FACULTAD			
8	CARRERA			
9	PAIS_NACIONALIDAD			
10	GENERO			
11	ESTADO_CIVIL			
12	TIPO_SANGRE			
13	ETNIA			
14	ESTADO_CARRERA			
15	MOTIVO_ABANDONO			
16	PROVINCIA_PROCEDENCIA			
17	RANGO_DISCAPACIDAD			
18	RANGO_EDAD			
19	RANGO_INGRESO_MENSUAL			
20	RANGO_PROMEDIO			

Fig. 18. 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 que nos servirán para la tarea de asociación, 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
	Físico Matemático
FÍSICA Y MATEMÁTICA	Pedagogía en las Ciencias Experimentales
	Contabilidad y Computación
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. 19, 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, Cisneros, Granda, Ortega, Posso-Yépez, et al., 2018).

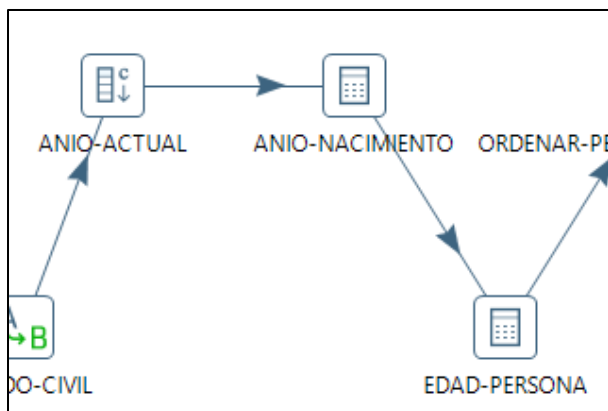


Fig. 19. Parte de la transformación que calcula la edad hasta el año 2018

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. 20.

MATRICULA_CODIGO	(Todas)
Promedio de NOTA_FINAL	
ESTUDIANTE_CEDULA	Total
0104651666	7,5
0104659115	9,285714286
0105175483	9,267
0105719074	9,166666667
0105977268	9,048571429
0106012784	7,214285714
0106121791	7,666
0106435316	8,772727273
0106861974	7,383
0107959652	8,833333333
0201796976	8,136363636
0202038766	7,305
0202040903	8,954545455
0202224259	7,916666667
0202337887	6,945
0202540183	6,347272727
0302242631	8,666666667
0302287263	6,846666667
0302287271	6,542857143

Fig. 20. 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

Para la tarea de agrupamiento se requiere trabajar con datos cuantitativos, por este motivo se normalizaron todos los datos que se van a analizar (Lara, 2014). La normalización se realizó de acuerdo con los siguientes parámetros:

- RANGO_EDAD
 - BAJA = 1
 - MEDIA = 2
 - ALTA = 3

- TIPO_SANGRE
 - A- = 1
 - A+ = 2
 - AB- = 3
 - AB+ = 4
 - B- = 5
 - B+ = 6
 - O- = 7
 - O+ = 8
 - NO ASIGNADO = 9

- PAÍS_NACIONALIDAD
 - ALEMANIA = 1
 - ARGENTINA = 2
 - COLOMBIA = 3
 - COSTA RICA = 4
 - CUBA = 5
 - ECUADOR = 6
 - ESPAÑA = 7
 - ESTADOS UNIDOS = 8
 - HOLANDA = 9
 - VENEZUELA = 10
 - NO ASIGNADO = 11

- GENERO
 - MASCULINO = 1
 - FEMENINO = 2

- ETNIA
 - AFRODESCENDIENTE = 1
 - INDÍGENA = 2
 - MESTIZO = 3
 - MONTUBIO = 4
 - NO ASIGNADO = 5

- RANGO_DISCAPACIDAD
 - NO TIENE = 1
 - LEVE = 2
 - MODERADO = 3
 - SEVERO = 4

- ESTADO_CIVIL
 - SOLTERO = 1
 - UNIÓN DE HECHO = 2
 - CASADO = 3
 - DIVORCIADO = 4
 - VIUDO = 5

- CONVIVIENTE
 - FAMILIAR = 1
 - MADRE = 2
 - OTROS = 3
 - PADRE = 4
 - PADRES = 5
 - PAREJA = 6
 - SOLO = 7

- TIPO_VIVIENDA
 - ANTICRESIS = 1
 - ARRENDADA = 2
 - CONCEDIDA POR EL TRABAJO = 3
 - HIPOTECADA = 4

- PRESTADA = 5
- PROPIA = 6
- NO ASIGNADO = 7

- FINANCIAMIENTO
 - BECA = 1
 - CRÉDITO = 2
 - FAMILIAR = 3
 - MADRE = 4
 - OTROS = 5
 - PADRE = 6
 - PADRES = 7
 - PAREJA = 8
 - SOLO = 9

- ACTIVIDAD_ESTUDIANTE
 - TRABAJA = 1
 - NO TRABAJA = 2

- RANGO_INGRESO_MENSUAL
 - MUY BAJO = 1
 - BAJO = 2
 - MEDIO = 3
 - ALTO = 4
 - MUY ALTO = 5

- FACULTAD
 - FACAE = 1
 - FCCSS = 2
 - FECYT = 3
 - FICA = 4
 - FICAYA = 5

- PROVINCIA_PROCEDENCIA

- AZUAY = 1
- BOLÍVAR = 2
- CAÑAR = 3
- CARCHI = 4
- CHIMBORAZO = 5
- COLOMBIA = 6
- COTOPAXI = 7
- CUBA = 8
- EL ORO = 9
- ESMERALDAS = 10
- ESPAÑA = 11
- GUAYAS = 12
- IMBABURA = 13
- LOJA = 14
- LOS RÍOS = 15
- MANABÍ = 16
- MORONA SANTIAGO = 17
- NAPO = 18
- ORELLANA = 19
- PASTAZA = 20
- PICHINCHA = 21
- SANTO DOMINGO DE LOS TSÁCHILAS = 22
- SUCUMBÍOS = 23
- TUNGURAHUA = 24
- VENEZUELA = 26
- NO ASIGNADO = 27
- ALEMANIA = 28
- ARGENTINA = 29
- ESTADOS UNIDOS = 30
- GALÁPAGOS = 31
- HUILA = 32
- SANTA ELENA = 33
- ZAMORA CHINCHIPE = 34

- RANGO_PROMEDIO
 - INSUFICIENTE = 1
 - SUFICIENTE = 2
 - BUENO = 3
 - EXCELENTE = 4

- MOTIVO_ABANDONO
 - NINGUNO = 1
 - PIERDE TERCERA MATRÍCULA = 2
 - CAMBIO CARRERA = 3

- ESTADO_CARRERA
 - ACTIVO = 1
 - INACTIVO = 2

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. 21 y Fig. 22.

NGO_EDAD	TIPO_SANGRE	PAIS_NACIONALIDAD	GENERO	ETNIA	RANGO_DISCAPACIDAD	ESTADO_CIVIL	CONVIVIENTE	TIPO_VIVIENDA	FINANCIAMIENTO
DIA	A+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	O+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
DIA	O+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
JA	O-	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PRESTADA	CREDITO
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	CASADO	FAMILIARs	PROPIA	MADRE
JA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	A+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	A+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
JA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
JA	O+	ECUADOR	FEMENINO	MONTEBUNO	NO TIENE	SOLTERO	FAMILIARs	PRESTADA	PADRES
JA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES
JA	A+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
JA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	MADRE
JA	O+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
JA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRES
ITA	A+	ECUADOR	MASCULINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	FAMILIAR
JA	A+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	FAMILIARs
JA	NO ASIGNADO	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	PROPIA	PADRE
DIA	O+	ECUADOR	FEMENINO	MESTIZO	NO TIENE	SOLTERO	FAMILIARs	ARRENDADA	PADRES

Fig. 21. 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	FALT
MADRE	NO TRABAJA	BAJO	FICA	MECATRÓNICA	CARCHI	FALT
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	EXCE
PADRES	NO TRABAJA	BAJO	FICA	TELECOMUNICACIONES	CARCHI	EXCE
PADRES	NO TRABAJA	ALTO	FECYT	RELACIONES PÚBLICAS	CARCHI	EXCE
SOLO	TRABAJA	MUY ALTO	FECYT	Licenciatura en ENTRENAMIENTO DEPORTIVO	TUNGURAHUA	BUE
MADRE	NO TRABAJA	MUY BAJO	FACAE	CONTABILIDAD SUPERIOR Y AUDITORIA	IMBABURA	BUE
PADRES	TRABAJA	MUY BAJO	FICA	SOFTWARE	IMBABURA	SUF
MADRE	NO TRABAJA	MUY BAJO	FACAE	ADMINISTRACIÓN DE EMPRESAS	CARCHI	FALT
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	FALT
PADRE	NO TRABAJA	MUY BAJO	FACAE	ADMINISTRACIÓN DE EMPRESAS	CARCHI	INSU

Fig. 22. 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. Minería de Datos

El objetivo principal de la presente investigación es identificar patrones de deserción estudiantil, por este motivo se optó por emplear únicamente los datos de los estudiantes que en el ítem ESTADO_CARRERA registraba el parámetro "I", que hace referencia a que el estudiante inactivo en su carrera, para de esta manera encontrar las posibles relaciones que tienen las diferentes variables en los desertores estudiantiles. Con el objetivo de tener una información de calidad se realizó una comparativa entre dos herramientas de análisis de datos, Weka y SPSS. El software Weka recibe archivos en formato *.csv (comma-separated values), por lo que el archivo resultante de la vista minable se transformó a dicho formato (Frank et al., 2016), mientras que el software SPSS recibe datos provenientes de hojas de cálculo (IBM, 2019).

A continuación, en la Fig. 23 y 24 se aprecia los datos a evaluar 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_ABA
NDONO, 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

```

Fig. 23. Vista minable para asociación 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_ABA
NDONO, ESTADO_CARRERA
1, 8, 6, 1, 1, 1, 1, 5, 2, 7, 2, 5, 4, 1, 2, 1, 1
1, 8, 6, 2, 3, 1, 1, 2, 2, 4, 2, 2, 3, 1, 4, 1, 1
1, 1, 6, 2, 3, 1, 1, 2, 2, 4, 2, 1, 1, 13, 3, 1, 1
1, 8, 6, 2, 3, 1, 1, 4, 2, 7, 2, 5, 3, 1, 3, 1, 1
2, 8, 6, 2, 3, 1, 1, 1, 2, 9, 1, 2, 3, 13, 4, 1, 1
1, 8, 6, 2, 3, 1, 3, 6, 2, 8, 2, 1, 1, 1, 4, 1, 1
2, 8, 6, 2, 3, 1, 1, 5, 6, 9, 2, 3, 1, 13, 4, 1, 2
1, 8, 6, 1, 3, 1, 1, 4, 4, 6, 2, 4, 4, 21, 2, 1, 1
1, 8, 6, 2, 3, 1, 1, 4, 6, 6, 2, 2, 3, 13, 3, 1, 1
1, 8, 6, 2, 3, 1, 1, 7, 6, 4, 2, 1, 2, 1, 3, 1, 1
1, 8, 6, 2, 3, 2, 1, 7, 2, 7, 2, 1, 3, 1, 3, 1, 1
1, 2, 6, 1, 3, 1, 1, 2, 6, 4, 2, 2, 3, 13, 3, 1, 1
2, 8, 6, 1, 3, 1, 1, 5, 2, 7, 2, 2, 4, 1, 2, 1, 1
1, 8, 6, 2, 3, 1, 1, 5, 6, 7, 2, 2, 1, 1, 3, 1, 1
3, 8, 5, 2, 3, 1, 4, 2, 6, 9, 1, 1, 3, 8, 3, 1, 2
1, 8, 6, 2, 3, 1, 1, 5, 6, 6, 2, 5, 5, 13, 3, 1, 2
1, 8, 6, 2, 3, 1, 1, 5, 2, 7, 2, 2, 5, 13, 2, 1, 1
1, 8, 6, 2, 3, 1, 1, 7, 2, 4, 2, 2, 3, 2, 3, 1, 1

```

Fig. 24. Vista minable para agrupamiento en formato *.csv

2.7.1. Agrupamiento

Algoritmo K-means

Para realizar la descripción de datos empleando k-means, se consideró que dividir un problema ayuda a priorizar esfuerzos, ya que no todos los estudiantes tienen las mismas

necesidades. El conocimiento que se obtenga después de utilizar el algoritmo k-means permitirá a los altos directivos adaptar sus propuestas de valor a cada grupo homogéneo dependiendo de sus necesidades, comportamientos, características o actitudes (Sanchez, 2017). Se empleó el algoritmo K-means ya que procede a dividir los datos en k grupos, garantizando que los objetos de un mismo grupo sean homogéneos y a su vez diferentes a los de otro grupo (Ochoa Reyes et al., 2014).

Expectativa-Maximización EM

El algoritmo de Expectativa-Maximización (EM), se usa principalmente para realizar estimaciones de máxima verosimilitud, asignando una distribución de probabilidad a cada instancia que indica la pertenencia a cada uno de los clúster, creándolos por medio de validación cruzada o como en este caso especificando a priori el número de clústeres que se van a utilizar (Huang & Chen, 2017).

2.7.2. Asociación

Algoritmo Apriori

El algoritmo Apriori es uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y uno de los más usados, debido a su antigüedad es uno de los algoritmos más robustos para realizar asociaciones de variables y es el que mayor número de versiones tiene, por este motivo para crear las reglas de asociación se utilizó únicamente este algoritmo (Han et al., 2001).

2.7.3. Atípicos

En el análisis de datos, existen problemas frecuentes en los cuales pueden aparecer observaciones inconsistentes con el resto de los datos, estas anomalías, inconsistencias o datos que a simple vista se diferencian de la masa principal de datos se denomina atípicos o en inglés outliers. Para identificar estos atípicos existen dos tipos de técnicas: (i) técnicas de acomodación, cuyo principal objetivo es emplear métodos robustos que soportan atípicos y (ii) técnicas de identificación, cuyo objetivo principal es aplicar métodos estadísticos que permiten identificar este tipo de datos en el conjunto de observaciones, como es este caso (Pérez Díez de los Ríos, 1987). Para la identificación de atípicos se utilizaron medidas de posición denominadas cuartiles que determinan la ubicación de los valores en un conjunto de observaciones de cuatro parte iguales (Lind et al., 2012).

CAPÍTULO 3

Validación de Resultados

3.1. Evaluación e interpretación

En esta última etapa de minería de datos se procederá a aplicar las métricas de calidad para cada uno de los algoritmos de las tareas descriptivas para posteriormente realizar el análisis e interpretación de la información para obtener el conocimiento.

3.1.1. Evaluación, análisis e interpretación de tareas de asociación

Apriori

Antes de ejecutar el algoritmo Apriori se evaluaron dos herramientas distintas: Weka 3.9 e IBM SPSS Statistics 22, cada una de ellas proporciona diferente información complementaria de acuerdo con la Tabla 3.1:

TABLA 3.1
COMPARATIVA ENTRE WEKA Y SPSS PARA APRIORI

CARACTERÍSTICAS	WEKA	SPSS
Permite identificar datos nominales y ordinales		X
Reglas de asociación	X	
Proporciona ítems X	X	
Proporciona ítems Y	X	
Proporciona soporte	X	
Proporciona confianza	X	
Permite establecer soporte mínimo	X	
Proporciona coeficiente de correlación		X

Fuente: Propia

De la Tabla 3.1 se obtiene que el software Weka resulta mejor que SPSS en esta tarea, ya que brinda más información para poder evaluar las reglas de asociación obtenidas. Sin embargo, SPSS permite establecer el coeficiente de correlación que permite identificar entre una relación negativa perfecta (-1) o una relación positiva perfecta (+1) (IBM, 2019). Además, con Weka se puede analizar la información de mejor manera puesto que proporciona información necesaria para identificar como está conformada la regla, mediante sus ítems X e ítems Y (Frank et al.,

2016). Por lo tanto, para el análisis de asociación se empleó Weka de acuerdo con las especificaciones que se observan en la Tabla 3.2 de acuerdo con (Lara, 2014), obteniendo los resultados que se aprecian en las Fig. 25 y 26.

TABLA 3.2
CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO APRIORI

INSTANCIAS	N° ATRIBUTOS	SOPORTE MÍNIMO	N° REGLAS
1096	16	0.4	25

Fuente: Propia

Donde las 1096 instancias analizadas corresponden a los estudiantes que se encuentran inactivos en su carrera universitaria, de igual forma los 16 atributos hacen referencia a los datos personales, académicos y socio económicos de los estudiantes antes mencionados. Mientras que el soporte mínimo de 0.4 se estableció en base a pruebas experimentales con el objetivo de identificar que soporte brinda mejores reglas de asociación, con la finalidad de verificar que cada uno de los ítems se repita por lo menos 438.4 veces para generar las 25 reglas deseadas, de las cuales se eligieron las 10 mejores reglas en base a la confianza (Frank et al., 2016).

```

=== Run information ===

Scheme:      weka.associations.Apriori -N 25 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.4 -S -1.0 -c -1
Relation:    VISTA-MINABLE-FACULTADES-ASOCIACION
Instances:   1096
Attributes:  16
             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

=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.85 (932 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 9
Size of set of large itemsets L(3): 4

```

Fig. 25. Resultado con Apriori

```

Best rules found:
1. ESTADO_CIVIL=SOLTERO 975 ==> RANGO_DISCAPACIDAD=NO TIENE 972 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.89)
2. ACTIVIDAD_ESTUDIANTE=NO TRABAJA 974 ==> RANGO_DISCAPACIDAD=NO TIENE 971 <conf:(1)> lift:(1) lev:(0) [0]
conv:(0.89)
3. PAIS_NACIONALIDAD=ECUADOR ESTADO_CIVIL=SOLTERO 969 ==> RANGO_DISCAPACIDAD=NO TIENE 966 <conf:(1)> lift:(1)
lev:(0) [0] conv:(0.88)
4. PAIS_NACIONALIDAD=ECUADOR ACTIVIDAD_ESTUDIANTE=NO TRABAJA 968 ==> RANGO_DISCAPACIDAD=NO TIENE 965 <conf:(1)>
lift:(1) lev:(0) [0] conv:(0.88)
5. PAIS_NACIONALIDAD=ECUADOR 1087 ==> RANGO_DISCAPACIDAD=NO TIENE 1083 <conf:(1)> lift:(1) lev:(-0) [0]
conv:(0.79)
6. ETNIA=MESTIZO 994 ==> RANGO_DISCAPACIDAD=NO TIENE 990 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.73)
7. PAIS_NACIONALIDAD=ECUADOR ETNIA=MESTIZO 985 ==> RANGO_DISCAPACIDAD=NO TIENE 981 <conf:(1)> lift:(1) lev:(-0)
[0] conv:(0.72)
8. MOTIVO_ABANDONO=NINGUNO 949 ==> RANGO_DISCAPACIDAD=NO TIENE 945 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.69)
9. PAIS_NACIONALIDAD=ECUADOR MOTIVO_ABANDONO=NINGUNO 940 ==> RANGO_DISCAPACIDAD=NO TIENE 936 <conf:(1)>
lift:(1) lev:(-0) [0] conv:(0.69)
10. ESTADO_CIVIL=SOLTERO 975 ==> PAIS_NACIONALIDAD=ECUADOR 969 <conf:(0.99)> lift:(1) lev:(0) [2] conv:(1.14)
11. ACTIVIDAD_ESTUDIANTE=NO TRABAJA 974 ==> PAIS_NACIONALIDAD=ECUADOR 968 <conf:(0.99)> lift:(1) lev:(0) [1]
conv:(1.14)
12. RANGO_DISCAPACIDAD=NO TIENE ESTADO_CIVIL=SOLTERO 972 ==> PAIS_NACIONALIDAD=ECUADOR 966 <conf:(0.99)>
lift:(1) lev:(0) [1] conv:(1.14)
13. RANGO_DISCAPACIDAD=NO TIENE ACTIVIDAD_ESTUDIANTE=NO TRABAJA 971 ==> PAIS_NACIONALIDAD=ECUADOR 965
<conf:(0.99)> lift:(1) lev:(0) [1] conv:(1.14)
14. RANGO_DISCAPACIDAD=NO TIENE 1092 ==> PAIS_NACIONALIDAD=ECUADOR 1083 <conf:(0.99)> lift:(1) lev:(-0) [0]
conv:(0.9)
15. ETNIA=MESTIZO 994 ==> PAIS_NACIONALIDAD=ECUADOR 985 <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.82)
16. ETNIA=MESTIZO RANGO_DISCAPACIDAD=NO TIENE 990 ==> PAIS_NACIONALIDAD=ECUADOR 981 <conf:(0.99)> lift:(1)
lev:(-0) [0] conv:(0.81)
17. ESTADO_CIVIL=SOLTERO 975 ==> PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE 966 <conf:(0.99)>
lift:(1) lev:(0) [2] conv:(1.16)
18. ACTIVIDAD_ESTUDIANTE=NO TRABAJA 974 ==> PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE 965
<conf:(0.99)> lift:(1) lev:(0) [2] conv:(1.16)
19. MOTIVO_ABANDONO=NINGUNO 949 ==> PAIS_NACIONALIDAD=ECUADOR 940 <conf:(0.99)> lift:(1) lev:(-0) [-1]
conv:(0.78)
20. RANGO_DISCAPACIDAD=NO TIENE MOTIVO_ABANDONO=NINGUNO 945 ==> PAIS_NACIONALIDAD=ECUADOR 936 <conf:(0.99)>
lift:(1) lev:(-0) [-1] conv:(0.78)
21. ETNIA=MESTIZO 994 ==> PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE 981 <conf:(0.99)> lift:(1)
lev:(-0) [-1] conv:(0.84)
22. MOTIVO_ABANDONO=NINGUNO 949 ==> PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE 936 <conf:(0.99)>
lift:(1) lev:(-0) [-1] conv:(0.8)
23. RANGO_DISCAPACIDAD=NO TIENE 1092 ==> ETNIA=MESTIZO 990 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.99)
24. PAIS_NACIONALIDAD=ECUADOR 1087 ==> ETNIA=MESTIZO 985 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.98)
25. PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE 1083 ==> ETNIA=MESTIZO 981 <conf:(0.91)> lift:(1)
lev:(-0) [-1] conv:(0.98)

```

Fig. 26. Resultado con Apriori

• Evaluación

En la TABLA 3.3 se aprecian las 25 mejores reglas de asociación resultantes del algoritmo Apriori, con el soporte y la confianza por cada regla:

TABLA 3.3
REGLAS DE ASOCIACIÓN, SOPORTE Y CONFIANZA RESULTANTE DEL ALGORITMO APRIORI

N. REGLA	ÍTEMS EN X	SOPORTE (x)	ÍTEMS EN Y	SOPORTE (x∩y)	CONFIANZA (%)
1	ESTADO_CIVIL=SOLTERO	975	RANGO_DISCAPACIDAD=NO TIENE	972	0.9969
2	ACTIVIDAD_ESTUDIANTE=NO TRABAJA	974	RANGO_DISCAPACIDAD=NO TIENE	971	0.9969
3	PAIS_NACIONALIDAD=ECUADOR ESTADO_CIVIL=SOLTERO	969	RANGO_DISCAPACIDAD=NO TIENE	966	0.9969
4	PAIS_NACIONALIDAD=ECUADOR ACTIVIDAD_ESTUDIANTE=NO TRABAJA	968	RANGO_DISCAPACIDAD=NO TIENE	965	0.9969
5	PAIS_NACIONALIDAD=ECUADOR	1087	RANGO_DISCAPACIDAD=NO TIENE	1083	0.9963
6	ETNIA=MESTIZO	994	RANGO_DISCAPACIDAD=NO TIENE	990	0.9960
7	PAIS_NACIONALIDAD=ECUADOR ETNIA=MESTIZO	985	RANGO_DISCAPACIDAD=NO TIENE	981	0.9959

8	MOTIVO_ABANDONO=NINGUNO	949	RANGO_DISCAPACIDAD=NO TIENE	945	0.9958
9	PAIS_NACIONALIDAD=ECUADOR MOTIVO_ABANDONO=NINGUNO	940	RANGO_DISCAPACIDAD=NO TIENE	936	0.9957
10	ESTADO_CIVIL=SOLTERO	975	PAIS_NACIONALIDAD=ECUADO R	969	0.9938
11	ACTIVIDAD_ESTUDIANTE=NO TRABAJA	974	PAIS_NACIONALIDAD=ECUADO R	968	0.9938
12	RANGO_DISCAPACIDAD=NO TIENE ESTADO_CIVIL=SOLTERO	972	PAIS_NACIONALIDAD=ECUADO R	966	0.9938
13	RANGO_DISCAPACIDAD=NO TIENE ACTIVIDAD_ESTUDIANTE=NO TRABAJA	971	PAIS_NACIONALIDAD=ECUADO R	965	0.9938
14	RANGO_DISCAPACIDAD=NO TIENE	1092	PAIS_NACIONALIDAD=ECUADO R	1083	0.9918
15	ETNIA=MESTIZO	994	PAIS_NACIONALIDAD=ECUADO R	985	0.9909
16	ETNIA=MESTIZO RANGO_DISCAPACIDAD=NO TIENE	990	PAIS_NACIONALIDAD=ECUADO R	981	0.9909
17	ESTADO_CIVIL=SOLTERO	975	PAIS_NACIONALIDAD=ECUADO R RANGO_DISCAPACIDAD=NO TIENE	966	0.9908
18	ACTIVIDAD_ESTUDIANTE=NO TRABAJA	974	PAIS_NACIONALIDAD=ECUADO R RANGO_DISCAPACIDAD=NO TIENE	965	0.9908
19	MOTIVO_ABANDONO=NINGUNO	949	PAIS_NACIONALIDAD=ECUADO R	940	0.9905
20	RANGO_DISCAPACIDAD=NO TIENE MOTIVO_ABANDONO=NINGUNO	945	PAIS_NACIONALIDAD=ECUADO R	936	0.9905
21	ETNIA=MESTIZO	994	PAIS_NACIONALIDAD=ECUADO R RANGO_DISCAPACIDAD=NO TIENE	981	0.9869
22	MOTIVO_ABANDONO=NINGUNO	949	PAIS_NACIONALIDAD=ECUADO R RANGO_DISCAPACIDAD=NO TIENE	936	0.9863
23	RANGO_DISCAPACIDAD=NO TIENE	1092	ETNIA=MESTIZO	990	0.9066
24	PAIS_NACIONALIDAD=ECUADOR	1087	ETNIA=MESTIZO	985	0.9062
25	PAIS_NACIONALIDAD=ECUADOR RANGO_DISCAPACIDAD=NO TIENE	1083	ETNIA=MESTIZO	981	0.9058

Fuente: Resultados de Weka

Donde:

x & y = combinación de ítems

soporte (x) = número de veces que se repite x

$$\text{soporte}(x \cap y) = \text{soporte}(x) \text{ intersección } \text{soporte}(y)$$

$$\text{confianza} = \frac{\text{soporte}(x \cap y)}{\text{soporte}(x)} \quad \text{Ec. 12}$$

Los resultados se evalúan considerando la confianza conteniendo valores entre 0 a 1, donde 1 es una confianza perfecta. Entonces se consideraron como las 9 mejores reglas de asociación aquellas cuya confianza es mayor a 0.995 (Hernández Orallo et al., 2004), tal como se especifican en la Tabla 3.4:

TABLA 3.4
10 MEJORES REGLAS DE ASOCIACIÓN

N. REGLA	ÍTEMS EN X	ÍTEMS EN Y	CONFIANZA (%)
1	ESTADO_CIVIL=SOLTERO	RANGO_DISCAPACIDAD=NO TIENE	0.9969
2	ACTIVIDAD_ESTUDIANTE=NO TRABAJA	RANGO_DISCAPACIDAD=NO TIENE	0.9969
3	PAIS_NACIONALIDAD=ECUADOR ESTADO_CIVIL=SOLTERO	RANGO_DISCAPACIDAD=NO TIENE	0.9969
4	PAIS_NACIONALIDAD=ECUADOR ACTIVIDAD_ESTUDIANTE=NO TRABAJA	RANGO_DISCAPACIDAD=NO TIENE	0.9969
5	PAIS_NACIONALIDAD=ECUADOR	RANGO_DISCAPACIDAD=NO TIENE	0.9963
6	ETNIA=MESTIZO	RANGO_DISCAPACIDAD=NO TIENE	0.9960
7	PAIS_NACIONALIDAD=ECUADOR ETNIA=MESTIZO	RANGO_DISCAPACIDAD=NO TIENE	0.9959
8	MOTIVO_ABANDONO=NINGUNO	RANGO_DISCAPACIDAD=NO TIENE	0.9958
9	PAIS_NACIONALIDAD=ECUADOR MOTIVO_ABANDONO=NINGUNO	RANGO_DISCAPACIDAD=NO TIENE	0.9957

Fuente: Weka

- **Análisis e interpretación**

Después de evaluar y establecer las mejores reglas de asociación, se observa que la mayoría de reglas poseen una confianza que se acerca al 100%, esto se debe a que los datos son no balanceados, es decir, que un ítem de dos categorías presenta una diferencia significativa en cuanto a la cantidad de datos de las ambas categorías (Saito & Rehmsmeier, 2015), en este caso en particular varios de los ítems tienen datos no balanceados, tal como el RANGO_DISCAPACIDAD cuya categoría “NO TIENE” es la que predomina, de igual forma en el ESTADO_CIVIL la categoría que tiene más registros es “SOLTERO”, entre otros casos. Por este

motivo las reglas se generan únicamente con los ítems que cumplen la condición anteriormente descrita, puesto que se ignora ítems tales como INGRESO_MENSUAL, TIPO_SANGRE, RANGO_PROMEDIO, entre otros, cuyos datos son más balanceados.

Por lo anteriormente expuesto y después de analizar las reglas de asociación generadas, los resultados obtenidos utilizando tareas de asociación no son relevantes para este estudio, ya que la información se encuentra explícita en la base de datos y según Lara (2014) donde indica que la minería de datos debe ser no trivial, ya que de nada sirve extraer conocimiento que se distingue a simple vista o que carezca de importancia, de igual forma su contenido debe ser previamente desconocido y útil.

3.1.2. Evaluación e interpretación de la tarea de agrupamiento

Para aplicar la tarea de agrupamiento se tomó en cuenta los algoritmos K-means y EM, ya que han tenido buenos resultados en trabajos investigativos previos, tales como (Yang, 2017) y (Pasin & Ankarali, 2017). A diferencia de las tareas de asociación, la vista minable que se empleó se encuentra con sus datos normalizados para poder realizar la evaluación de los algoritmos, tal como se mencionó en el capítulo anterior (sección #).

El principal problema de las tareas de agrupamiento es determinar el número de clústeres que se pretende generar; ya que no existe un criterio objetivo ni ampliamente validado que permita elegir un número óptimo de clústeres (Moya, 2016). Para este caso en particular se definió el número de clústeres después de realizar pruebas experimentales que involucraban establecer como variables de clase el ESTADO_CARRERA, RANGO_PROMEDIO y FACULTAD respectivamente, y por último se realizó un análisis sin una clase definida; obteniendo como resultado en todas las pruebas que independientemente de cómo se defina el número de clústeres, estos son similares; por este motivo se escogió $k=5$ debido al número de facultades de la UTN con el fin de buscar disminuir el error y que los grupos de datos fueran más homogéneos. Los resultados para los diferentes algoritmos de agrupamiento son los siguientes:

K-means

Para ejecutar el algoritmo K-means se evaluaron dos herramientas diferentes: Weka 3.9 e IBM SPSS Statistics 22, cada una de ellas proporciona diferente información de acuerdo con la Tabla 3.5:

TABLA 3.5
COMPARATIVA ENTRE WEKA Y SPSS PARA K-MEANS

CARACTERÍSTICAS	WEKA	SPSS
Permite identificar datos nominales y ordinales		X
Proporciona centroides iniciales	X	X
Proporciona centroides finales	X	X
Distingue a que clúster corresponde cada registro		X
Proporciona la distancia del registro al centroide del clúster al que pertenece		X
Distancias entre centros de clústeres finales		X
Número de registros en cada clúster	X	X
Medidas estadísticas		X
Proporciona el error cuadrático medio	X	
Genera gráficas de los clústeres	X	

Fuente: Propia

De la Tabla 3.5 se observa que el software SPSS es superior a Weka para esta tarea, ya que brinda más información para poder evaluar los clúster obtenidos. En SPSS se puede procesar la información de mejor manera puesto a que permite diferenciar el tipo de datos, nominales y ordinales, no muestra el error cuadrático medio (ECM), sin embargo, proporciona los instrumentos para calcularlo, adicionalmente se puede identificar a que clúster pertenece cada uno de los registros analizados. Como característica más fuerte, SPSS nos permite evaluar la calidad del agrupamiento mediante el método estadístico Analysis of Variance (ANOVA), que permitirá identificar si los clúster muestran diferencias significativas o puede suponerse que sus medias no difieren (Gorgas et al., 2011).

Por lo anteriormente expuesto, para ejecutar el algoritmo k-means se empleó el software SPSS con las especificaciones que se detallan en la Tabla 3.6, el resultado del clúster de pertenencia se adjunta en el siguiente enlace: <http://bit.ly/2UV3GVa>

TABLA 3.6
CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO K-MEANS

INSTANCIAS	N° ATRIBUTOS	N° ITERACIONES	N° CLÚSTERES	DISTANCIA
1096	16	10	5	Euclídea

Fuente: Propia

Para garantizar la exactitud del centroide se puso 10 iteraciones con la finalidad de que el algoritmo continúe trabajando hasta que se haya cumplido el criterio de convergencia, que determina cuando se detenga la iteración.

En la Tabla 3.7, se encuentran los centroides iniciales aleatorios de cada clúster proporcionado por el algoritmo:

TABLA 3.7
CENTROIDES INICIALES ALEATORIOS DE CADA CLÚSTER

	Clúster				
	1	2	3	4	5
RANGO_EDAD	2,00	1,00	2,00	1,00	1,00
TIPO_SANGRE	9,00	8,00	8,00	2,00	2,00
PAIS_NACIONALIDAD	6,00	6,00	3,00	6,00	6,00
GENERO	2,00	2,00	1,00	1,00	2,00
ETNIA	3,00	3,00	3,00	3,00	3,00
RANGO_DISCAPACIDAD	1,00	1,00	1,00	1,00	1,00
ESTADO_CIVIL	1,00	1,00	3,00	1,00	1,00
CONVIVIENTE	1,00	7,00	6,00	7,00	7,00
TIPO_VIVIENDA	6,00	2,00	2,00	6,00	2,00
FINANCIAMIENTO	3,00	6,00	9,00	7,00	9,00
ACTIVIDAD_ESTUDIANTE	2,00	2,00	2,00	2,00	2,00
RANGO_INGRESO_MENSUAL	1,00	3,00	4,00	5,00	2,00
PROVINCIA_PROCEDENCIA	4,00	34,00	13,00	21,00	4,00
MOTIVO_ABANDONO	1,00	1,00	1,00	1,00	1,00
RANGO_PROMEDIO	2,00	4,00	2,00	1,00	3,00

Fuente: SPSS

En la Tabla 3.8, se encuentran los centroides finales de cada clúster proporcionado por el algoritmo k-means:

TABLA 3.8
CENTROIDES FINALES DE CADA CLÚSTER

	Clúster				
	1	2	3	4	5
RANGO_EDAD	1,69	1,40	1,54	1,52	1,43
TIPO_SANGRE	7,94	6,00	5,75	5,87	2,03
PAIS_NACIONALIDAD	5,93	6,00	5,99	5,97	5,90
GENERO	1,52	1,50	1,55	1,51	1,62
ETNIA	3,01	2,90	2,88	2,93	2,97
RANGO_DISCAPACIDAD	1,00	1,00	1,00	1,02	1,00
ESTADO_CIVIL	1,25	1,20	1,23	1,19	1,21
CONVIVIENTE	4,55	4,40	4,11	4,38	4,48
TIPO_VIVIENDA	3,70	3,50	5,12	4,21	3,72
FINANCIAMIENTO	6,64	5,70	6,37	6,52	6,52
ACTIVIDAD_ESTUDIANTE	1,84	1,80	1,90	1,86	1,89
RANGO_INGRESO_MENSUAL	1,88	2,20	2,16	2,12	2,07
PROVINCIA_PROCEDENCIA	4,06	27,70	13,00	21,18	4,07
MOTIVO_ABANDONO	1,17	1,30	1,20	1,12	1,30
RANGO_PROMEDIO	2,66	2,00	2,57	2,31	2,67

Fuente:SPSS

- **Evaluación**

Para realizar la evaluación del algoritmo K-means se consideraron dos aspectos, el método estadístico ANOVA y el error cuadrático medio en base a los resultados obtenidos del software (Lind et al., 2012). En la Tabla 3.9 se puede observar el resultado con el método ANOVA para la siguiente hipótesis planteada:

Hipótesis: Las medias de los clústeres son iguales.

TABLA 3.9
RESULTADOS CON ANOVA

	Clúster		Error		F	sig.
	Media cuadrática	gl	Media cuadrática	gl		
RANGO_EDAD	0.797	4	0.287	1091	2.779	0.026
TIPO_SANGRE	329.853	4	7.086	1091	46.552	0.000
PAIS_NACIONALIDAD	0.176	4	0.066	1091	2.676	0.031
GENERO	0.166	4	0.248	1091	0.669	0.614

ETNIA	0.439	4	0.173	1091	2.531	0.039
RANGO_DISCAPACIDAD	0.008	4	0.004	1091	2.113	0.077
ESTADO_CIVIL	0.051	4	0.443	1091	0.114	0.977
CONVIVIENTE	6.849	4	3.285	1091	2.085	0.081
TIPO_VIVIENDA	82.300	4	2.877	1091	28.610	0.000
FINANCIAMIENTO	3.413	4	2.364	1091	1.444	0.217
ACTIVIDAD_ESTUDIANT E	0.130	4	0.099	1091	1.311	0.264
RANGO_INGRESO_MEN SUAL	0.799	4	1.101	1091	1.635	0.163
PROVINCIA_PROCEDEN CIA	5342.031	4	0.185	1091	28826.285	0.000
MOTIVO_ABANDONO	0.336	4	0.287	1091	1.172	0.321
RANGO_PROMEDIO	2.712	4	0.680	1091	3.987	0.003

Fuente: SPSS

Para el método estadístico ANOVA, se emplearon 2 factores, clúster y error, tal como se aprecia en la Tabla 3.9. Sin embargo, para este caso en específico, de datos no balanceados, las pruebas F se deben utilizar únicamente con fines descriptivos puesto que los clústeres se han elegido con el objetivo de maximizar las diferencias entre los casos de los diferente clústeres, y los niveles de significancia presentados no se encuentran corregidos para esto, por lo tanto no se pueden interpretar como pruebas de la hipótesis de que las medias de los clúster son iguales (IBM, 2019). Por ello se tomó en cuenta realizar la evaluación por medio del error cuadrático medio, considerando que no existe un rango específico para medir la calidad del agrupamiento.

- **Análisis e interpretación**

Para analizar los resultados obtenidos mediante el algoritmo K-meas en el software SPSS, se interpretaron los centroides de cada uno de los clústeres considerando que k-means busca optimizar los centroides y no las fronteras, entendiéndose que el centroide es el registro que posee las características más significativas de cada clúster (Honarkhah & Caers, 2010), tomando en cuenta que los datos se normalizaron previamente al proceso de minería se procederá a desnormalizarlos considerando los valores establecidos en el capítulo anterior en la etapa de transformación (sección 2.6.2), aproximando los datos que tienen cinco décimas (0.5) o más al inmediato superior y descartando las décimas en aquellos números que poseen cuatro décimas y nueve centésimas (0.49) o menos. Por ejemplo:

- 7.57 = 8
- 5.45 = 5

Quedando los clústeres como se aprecia en la Tabla 3.10:

TABLA 3.10
CLÚSTERES DESNORMALIZADOS DEL ALGORITMO K-MEANS

	Clúster				
	1	2	3	4	5
RANGO_EDAD	Media	Baja	Media	Media	Baja
TIPO_SANGRE	O+	B+	B+	B+	A+
PAIS_NACIONALIDAD	Ecuador	Ecuador	Ecuador	Ecuador	Ecuador
GENERO	Femenino	Femenino	Femenino	Femenino	Femenino
ETNIA	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo
RANGO_DISCAPACIDAD	No tiene	No tiene	No tiene	No tiene	No tiene
ESTADO_CIVIL	Soltero	Soltero	Soltero	Soltero	Soltero
CONVIVIENTE	Padres	Padre	Padre	Padre	Padre
TIPO_VIVIENDA	Hipotecada	Hipotecada	Prestada	Hipotecada	Hipotecada
FINANCIAMIENTO	Padres	Padre	Padre	Padres	Padres
ACTIVIDAD_ESTUDIANTE	No trabaja	No trabaja	No trabaja	No trabaja	No trabaja
RANGO_INGRESO_MENSUAL	Bajo	Bajo	Bajo	Bajo	Bajo
PROVINCIA_PROCEDENCIA	Carchi	Alemania	Imbabura	Pichincha	Carchi
MOTIVO_ABANDONO	Ninguno	Ninguno	Ninguno	Ninguno	Ninguno
RANGO_PROMEDIO	Bueno	Suficiente	Bueno	Suficiente	Bueno

Fuente: SPSS

Finalmente, después de desnormalizar, el proceso inverso al que se llevó al final de la sección 2.6.2 (normalización), cada uno de los datos nos encontramos con:

El Clúster 1 representante de la facultad

Expectation-maximization EM

Para ejecutar el algoritmo EM se evaluaron dos herramientas diferentes: Weka 3.9 e IBM SPSS Statistics 22, cada una de ellas proporciona diferente información complementaria de acuerdo con la Tabla 3.11:

TABLA 3.11
COMPARATIVA ENTRE WEKA Y SPSS PARA EM

CARACTERÍSTICAS	WEKA	SPSS
Permite identificar datos nominales y ordinales		X
Proporciona la desviación estándar	X	X
Proporciona media	X	X
Proporciona número de registros en cada clúster	X	X
Gráficas	X	
Definir previamente el número de clústeres	X	
Detalla cómo está asignada la información a cada clúster	X	

Fuente: Propia

En la Tabla 3.11 se observa que cada uno de los softwares cuenta con sus características propias, sin embargo, Weka brinda información que permitirá identificar plenamente como se realizó el agrupamiento y visualizarlo por medio de las gráficas. Uno de los puntos más importantes que permitirá escoger que software emplear es que permita definir a priori el número de clústeres que se pretenden generar, siendo Weka el software que permite escoger esta opción para esta tarea de minería de datos. Por lo anteriormente expuesto, para ejecutar el algoritmo EM se empleó el software Weka con las especificaciones que se detallan en la Tabla 3.12.

TABLA 3.12
CONDICIONES CON LAS QUE SE EJECUTÓ EL ALGORITMO EM

INSTANCIAS	N° ATRIBUTOS	N° ITERACIONES	N° CLÚSTERES	DISTANCIA
1096	16	100	5	Euclideana

Fuente: Propia

Para garantizar que el algoritmo pueda analizar los valores perdidos a cabalidad se empleó el método de validación cruzada con 10-iteraciones en donde se define que la expectativa - maximización se realice 10 veces como se establece por defecto, ya que el número de instancias es mayor que 10. Con el objetivo de obtener los mejores clústeres se estableció como número máximo de clústeres a emplear -1 y 100 iteraciones que vienen por defecto (Frank et al., 2016). Los resultados obtenidos se aprecian en las Fig. 27 y 28.

```

=== Run information ===

Scheme:          weka.clusterers.EM -I 100 -N 5 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots
1 -S 100
Relation:        VISTA-MINABLE-agrupamiento
Instances:       1096
Attributes:      16
                 RANGO_EDAD
                 TIPO_SANGRE
                 PAIS_NACIONALIDAD
                 GENERO
                 ETNIA
                 RANGO_DISCAPACIDAD
                 ESTADO_CIVIL
                 CONVIVIENTE
                 TIPO_VIVIENDA
                 FINANCIAMIENTO
                 ACTIVIDAD_ESTUDIANTE
                 RANGO_INGRESO_MENSUAL
                 PROVINCIA_PROCEDECENCIA
                 RANGO_PROMEDIO
                 MOTIVO_ABANDONO

Ignored:         FACULTAD
Test mode:       Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

Number of clusters: 5
Number of iterations performed: 1

=== Model and evaluation on training set ===

Clustered Instances

0      305 ( 28%)
1       80 (  7%)
2      533 ( 49%)
3       45 (  4%)
4      133 ( 12%)

Log likelihood: -4.9399

Class attribute: FACULTAD
Classes to Clusters:

  0   1   2   3   4 <-- assigned to cluster
69   7  87   4  23 | FACAE
75   7  87   8  56 | FECYT
34  15  74   6  10 | FICAYA
62  46 142  23  30 | FICA
65   5 143   4  14 | FCCSS

Cluster 0 <-- FACAE
Cluster 1 <-- FICA
Cluster 2 <-- FCCSS
Cluster 3 <-- FICAYA
Cluster 4 <-- FECYT

Incorrectly clustered instances :      776.0      70.8029 %

```

Fig. 27. Resultados obtenidos del algoritmo EM

```

EM
==
Attribute          Cluster
                   0      1      2      3      4
                   (0.19) (0.25) (0.32) (0.14) (0.11)
-----
RANGO_EDAD
  mean             1.3781 1.6458 1.3952 1.4889 2.0829
  std. dev.        0.503  0.4809 0.4946 0.5025 0.4783
TIPO_SANGRE
  mean             2.0109 7.7942 7.4607 1.9796 6.1
  std. dev.        0.1158 0.6318 1.0347 0.1777 2.5773
PAIS_NACIONALIDAD
  mean             6      6      6      6 5.7866
  std. dev.        0.0002 0.2572 0      0.2572 0.7601
GENERO
  mean             1.906  1.1391 1.8529 1.1737 1.4657
  std. dev.        0.2919 0.346  0.3542 0.3788 0.4988
ETNIA
  mean             2.9114 2.9567 2.8854 2.799 2.8394
  std. dev.        0.3323 0.2127 0.3904 0.5098 0.4964
RANGO_DISCAPACIDAD
  mean             1      1.0046 1      1 1.0234
  std. dev.        0.0603 0.0678 0.0603 0.0603 0.1512
ESTADO_CIVIL
  mean             1.1789 1.0143 1.2235 1.0083 2.0919
  std. dev.        0.5858 0.1446 0.6245 0.1026 1.1796
CONVIVIENTE
  mean             4.276  4.1255 4.0074 4.3057 4.7071
  std. dev.        1.8606 1.8009 1.7915 1.6561 1.8973
TIPO_VIVIENDA
  mean             4.6103 5.0164 4.6533 5.1072 4.6732
  std. dev.        1.8415 1.6606 1.8316 1.5927 1.7992
FINANCIAMIENTO
  mean             6.314  6.0644 6.0909 6.4817 8.25
  std. dev.        1.428  1.4305 1.3932 1.2634 1.3726
ACTIVIDAD_ESTUDIANTE
  mean             1.9836 1.9639 1.9872 1.9706 1.1481
  std. dev.        0.127  0.1866 0.1126 0.1688 0.3552
RANGO_INGRESO_MENSUAL
  mean             1.9451 2.1713 2.1818 2.353 1.8605
  std. dev.        0.8606 1.0784 1.0102 1.1558 1.1455
PROVINCIA_PROCEDENCIA
  mean             12.2318 12.6187 12.4627 12.582 12.251
  std. dev.        4.309  4.1533 4.1947 3.8753 4.8278
RANGO_PROMEDIO
  mean             2.835  2.1228 2.8443 2.1532 2.7313
  std. dev.        0.6781 0.792  0.7095 0.819  0.85
MOTIVO_ABANDONO
  mean             2.4837 2.4696 2.4788 2.4875 2.4557
  std. dev.        0.1142 0.2693 0.0508 0.2371 0.196

Time taken to build model (full training data) : 0.18 seconds

```

Fig. 28. Resultados obtenidos del algoritmo EM

- **Evaluación**

Para evaluar el algoritmo EM se tomaron en cuenta todas la variables descriptivas estadísticas como la media y la desviación estándar, tal como se aprecia en la Tabla 3.13

TABLA 3.13
VARIABLES ESTADÍSTICAS DESCRIPTIVAS DEL ALGORITMO EM

	Clúster									
	1		2		3		4		5	
	Media	Desviación Estándar	Media	Desviación Estándar	Media	Desviación Estándar	Media	Desviación Estándar	Media	Desviación Estándar
RANGO_EDAD	1.37	0.50	1.64	0.48	1.39	0.49	1.48	0.50	2.08	0.47
TIPO_SANGRE	2.01	0.11	7.79	0.63	7.46	1.03	1.97	0.17	6.1	2.57
PAIS_NACIONALIDAD	6	0.0002	6	0.25	6	0	6	0.25	5.78	0.76
GENERO	1.90	0.29	1.13	0.34	1.85	0.35	1.17	0.37	1.46	0.49
ETNIA	2.91	0.33	2.95	0.21	2.88	0.39	2.79	0.5	2.83	0.49
RANGO_DISCAPACIDAD	1	0.06	1.004	0.06	1	0.06	1	0.06	1.02	0.15
ESTADO_CIVIL	1.17	0.58	1.01	0.14	1.22	0.62	1.008	0.1	2.09	1.17
CONVIVIENTE	4.27	1.86	4.12	1.8	4.007	1.79	4.3	1.65	4.7	1.89
TIPO_VIVIENDA	4.61	1.84	5.01	1.66	4.65	1.83	5.107	1.59	4.67	1.79
FINANCIAMIENTO	6.31	1.42	6.06	1.43	6.09	1.39	6.48	1.26	1.25	1.37
ACTIVIDAD_ESTUDIANTE	1.98	0.12	1.96	0.18	1.98	0.11	1.97	0.16	1.14	0.35
RANGO_INGRESO_MENSUAL	1.94	0.86	2.17	1.07	2.18	1.01	2.35	1.15	1.86	1.14
PROVINCIA_PROCEDENCIA	12.23	4.309	12.61	4.15	12.46	4.19	12.58	3.87	12.25	4.82
RANGO_PROMEDIO	12.83	0.67	2.12	0.79	2.84	0.709	2.15	0.81	2.73	0.85
MOTIVO_ABANDONO	2.48	0.11	2.46	0.26	2.47	0.05	2.48	0.23	2.45	0.19

Fuente:Weka

Para evaluar este algoritmo se trabajará en base a dos medidas, la media o promedio y la desviación estándar (Frank et al., 2016), considerando que la media es la suma de los valores de la muestra, en este caso del clúster, dividido para el número total de casos, mientras que la raíz cuadrada de la varianza de la población es la desviación estándar, que en este caso mientras más se acerque a 0 quiere decir que los datos del grupo son más similares (homogéneos), de igual forma cuando la desviación estándar es alta quiere decir que el clúster tiene datos variados (Lind et al., 2012), para profundizar esta información visitar el siguiente enlace <http://bit.ly/2tFxZD5>.

- **Análisis e interpretación**

Tomando en cuenta lo anteriormente expuesto el análisis de los resultados se lo realizó por cada variable para identificar que tan similares son los atributos de un clúster con respecto al

mismo atributo de los otros clústeres, de igual forma que en la técnica anterior se aproximaron los valores al inmediato superior en el caso que fuera necesario, para posteriormente desnormalizar los datos, obteniendo los siguientes resultados de acuerdo con la Tabla 3.13:

- **RANGO_EDAD**

Este atributo en los clústeres 1, 3 y 4 tiene valores similares, en la categoría **Baja**, mientras que los clústeres 2 y 5 poseen la categoría **Media**, puesto que ambos casos la media y la desviación estándar tienen datos similares.

- **TIPO_SANGRE**

Este atributo posee datos variados, puesto que para los clústeres 1 y 4 se tiene como resultado la categoría **A+** con una desviación estándar de 0.11 y 0.17, respectivamente, que quiere decir que sus datos son similares, para el clúster 2 se tiene la categoría **O+** con una desviación estándar de 0.63 que al igual que en el caso anterior los datos guardan cierta similitud, en el clúster 3 se encuentra la categoría **O-** con una desviación de 1.03, que da a entender que sus datos tiene cierta variedad, y finalmente para el clúster 5 se tiene la categoría **B+** con una desviación de 2.57 que quiere decir que sus datos son completamente variados.

- **PAIS_NACIONALIDAD**

Como se mencionó anteriormente cuando se trabaja con datos no balanceados como es este estudio, se obtienen valores como los que se presentan para este atributo, que todos los clústeres presentan la categoría "**Ecuador**" con una desviación estándar relativamente baja en todos los casos, demostrando que todos los datos de cada clúster son similares.

- **GÉNERO**

Para los clústeres 1 y 3 se tiene la categoría **Femenino** con una desviación estándar de 0.29 y 0.35 para cada uno de ellos, mientras que para los clústeres 2, 4 y 5 se tiene la categoría **Masculino** con una desviación estándar de 0.34, 0.37 y 0.49 respectivamente; teniendo como resultado que todo los clústeres tienen datos homogéneos de acuerdo con la categoría que fueron agrupados.

- **ETNIA**

La etnia al igual que al el caso país nacionalidad posee datos no balanceados, por lo que la categoría que registran todos los clústeres es **Mestizo** con una desviación estándar que va de 0.21 a 0.5 dando a entender que los registros son similares en cada clúster.

- **RANGO_DISCAPACIDAD**

Este atributo también contiene datos no balanceados y registra la categoría **No tiene** con una desviación estándar por debajo de una décima, lo que quiere decir que al igual que en los casos anteriores los datos de los clústeres son similares.

- **ESTADO_CIVIL**

Los clústeres 1, 2, 3 y 4 poseen el registro **Soltero** con una desviación estándar muy baja que da a entender que sus registros son similares mientras que el clúster 5 tiene el atributo **Unión de hecho** con una desviación por encima de 1 dando a entender que ente clúster se registran más categorías.

- **CONVIVIENTE**

Los clústeres 1, 2, 3 y 4 en este atributo almacenan la categoría **Padre** con una desviación estándar por encima de 1.5 dando a entender que los registros de cada clúster poseen más categorías, mientras que el clúster 5 almacena la categoría **Padres**.

- **TIPO_VIVIENDA**

En este caso, todos los clústeres poseen la categoría **Prestada** y una desviación estándar de 1.5 que muestran que los registros son variados.

- **FINANCIAMIENTO**

Los clústeres 1, 2, 3 y 4 almacenan la categoría **Padre**, con una desviación estándar por encima de la unidad entendiéndose que sus datos son variados, mientras que el clúster 5 almacena la categoría **Pareja** y también posee datos variados.

- **ACTIVIDAD_ESTUDIANTE**

Todos los clústeres almacenan la categoría **No trabaja** y tienen una desviación estándar baja, por lo que los registros de cada clúster son similares

- **RANGO_INGRESO_MENSUAL**

En este caso todos los clústeres poseen la categoría **Bajo** en rango ingreso mensual, según la desviación estándar el clúster 1 muestra que sus registros son similares, mientras que en el resto de los clústeres sus registros son variados.

- **PROVINCIA_PROCEDENCIA**

En este caso los clústeres los clústeres 1, 3 y 5 almacenan la categoría **Guayas**, mientras que los clústeres 2 y 4 almacenan la categoría **Imbabura**, teniendo como desviación estándar en todos los casos valores cercanos a 4 entendiéndose que los datos de cada clúster en este atributo son muy variados.

- **RANGO_PROMEDIO**

Los clústeres 1, 3 y 5 tienen registrado en este atributo la categoría **Bueno**, mientras que los clústeres 2 y 4 tienen registrado la categoría **Suficiente**, teniendo en todos los casos una desviación estándar por debajo de 0.8, entendiéndose que los registros de cada clúster son similares respectivamente.

- **MOTIVO_ABANDONO**

En todos los clústeres se registra como motivo abandono la categoría **Pierde tercera matrícula**, teniendo una desviación estándar por debajo de 0.2 entendiéndose que sus datos son variados.

Los clústeres finales se aprecian en la Tabla 3.14

TABLA 3.14
CLÚSTERES DESNORMALIZADOS DEL ALGORITMO EM

	Clúster				
	1	2	3	4	5
RANGO_EDAD	Baja	Media	Baja	Baja	Media
TIPO_SANGRE	A+	O+	O-	A+	B+
PAIS_NACIONALIDAD	Ecuador	Ecuador	Ecuador	Ecuador	Ecuador

GENERO	Femenino	Masculino	Femenino	Masculino	Masculino
ETNIA	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo
RANGO_DISCAPACIDAD	No tiene	No tiene	No tiene	No tiene	No tiene
ESTADO_CIVIL	Soltero	Soltero	Soltero	Soltero	Unión de hecho
CONVIVIENTE	Padre	Padre	Padre	Padre	Padres
TIPO_VIVIENDA	Prestada	Prestada	Prestada	Prestada	Prestada
FINANCIAMIENTO	Padre	Padre	Padre	Padre	Pareja
ACTIVIDAD_ESTUDIANTE	No trabaja	No trabaja	No trabaja	No trabaja	Trabaja
RANGO_INGRESO_MENSUAL	Bajo	Bajo	Bajo	Bajo	Bajo
PROVINCIA_PROCEDENCIA	Guayas	Imbabura	Guayas	Imbabura	Guayas
MOTIVO_ABANDONO	Pierde tercera matrícula	Pierde tercera matrícula	Pierde tercera matrícula	Pierde tercera matrícula	Pierde tercera matrícula
RANGO_PROMEDIO	Bueno	Suficiente	Bueno	Suficiente	Bueno

Fuente: Weka

La Fig. 27 demuestra que los resultados del software asignaron un clúster a cada una de las facultades de la UTN, FACA E, FICA, FCCSS, FICAYA y FECYT respectivamente.

3.2. Atípicos

Para realizar la detección de atípicos se tomaron en cuenta las variables cuantitativas con sus datos en bruto, de igual forma se realizó el mismo proceso con las variables cualitativas que se encuentran normalizadas, tal como se explica en el capítulo anterior (sección 2.6.2), puesto que el proceso para identificar cuartiles el valor intercuartil y los atípicos exige que los datos sean ordenados de menor a mayor. Considerando que los cuartiles dividen al número total de casos en 4 partes iguales, entonces el 25% de los casos serán menores que el primer cuartil (Q_1) y el 75% de los casos serán menores que el tercer cuartil (Q_3), que representan a su definición gráfica denominada diagrama de caja (Lind et al., 2012), entonces:

$Q_1 = 25\%$ de los casos

$Q_2 = \text{mediana} = 50\%$ de los casos

$Q_3 = 75\%$ de los casos

$Q_4 = 100\%$ de los casos

Partiendo de este concepto, las fórmulas para obtener los cuartiles se aprecian en las Ecuaciones 13, 14 y 15:

$$Q_1 = (n + 1) \left(\frac{25}{100} \right) \quad \text{Ec. 13}$$

$$Q_2 = \text{mediana} = (n + 1) \left(\frac{50}{100} \right) \quad \text{Ec. 14}$$

$$Q_3 = (n + 1) \left(\frac{75}{100} \right) \quad \text{Ec. 15}$$

Donde:

n = número de casos

Para definir en qué rango se encuentran los atípicos se debe calcular el rango intercuartil (RI) que es la fórmula que se aprecia en la Ecuación 16, de igual forma en las Ecuaciones 17 y 18 se observan las fórmulas que permitirá definir los límites para obtener los valores atípicos (Lind et al., 2012):

$$\text{Rango intercuartil} = Q_3 - Q_1 \quad \text{Ec. 16}$$

$$\text{Dato atípico} > Q_3 + 1.5(\text{Rango intercuartil}) \quad \text{Ec. 17}$$

$$\text{Dato atípico} < Q_1 - 1.5(\text{Rango intercuartil}) \quad \text{Ec. 18}$$

- **Evaluación**

Para evaluar que datos son típicos y atípicos se tomaron en cuenta los siguientes rangos en cada una de las variables, que son resultados de las ecuaciones anteriormente planteadas obteniéndose los resultados que se aprecian en la Tabla 3.15:

TABLA 3.15
CUARTILES, MEDIANA Y LÍMITES

ATRIBUTO	Q ₁	MEDIANA (Q ₂)	Q ₃	RI	Q ₃ + 1.5 (RI)	Q ₁ - 1.5 (RI)
RANGO_EDAD	20.00	22.00	25.00	5.00	32.50	12.50
TIPO_SANGRE	2.00	8.00	8.00	6.00	17.00	-7.00
PAIS_NACIONALIDAD	6.00	6.00	6.00	0.00	6.00	6.00
GENERO	1.00	2.00	2.00	1.00	3.50	-0.50
ETNIA	3.00	3.00	3.00	0.00	3.00	3.00
RANGO_DISCAPACIDAD	0.00	0.00	0.00	0.00	0.00	0.00
ESTADO_CIVIL	1.00	1.00	1.00	0.00	1.00	1.00
CONVIVIENTE	2.00	5.00	5.00	3.00	9.5	-2.50
TIPO_VIVIENDA	2.00	6.00	6.00	4.00	12.00	-4.00
FINANCIAMIENTO	6.00	7.00	7.00	1.00	8.50	4.50

ACTIVIDAD_ESTUDIANTE	2.00	2.00	2.00	0.00	2.00	2.00
RANGO_INGRESO_MENSUAL	380.00	570.00	850.00	470.00	1555.00	-325.00
PROVINCIA_PROCEDENCIA	13.00	13.00	13.00	0.00	13.00	13.00
MOTIVO_ABANDONO	1.00	1.00	1.00	0.00	1.00	1.00
RANGO_PROMEDIO	7,43	8,00	8,50	1,07	10,11	5,83

Fuente: Propia

En la Tabla 3.15, se puede identificar claramente que los atributos cuyos datos se encuentran no balanceados Q_1 , mediana (Q_2) y Q_3 se encuentra el valor de la categoría que predomina, como es el caso de PAIS_NACIONALIDAD, ETNIA, RANGO_DISCAPACIDAD, ESTADO_CIVIL, ACTIVIDAD_ESTUDIANTE, PROVINCIA_PROCEDENCIA y MOTIVO_ABANDONO, y en el caso del TIPO_SANGRE, GENERO, CONVIVIENTE, TIPO_VIVIENDA y FINANCIAMIENTO coinciden la mediana y el cuartil 3, en estos casos se hace difícil definir si existen o no datos atípicos.

Considerando la Tabla 3.15, para identificar los atípicos de las variables cuantitativas se tiene el siguiente análisis con su respectivo diagrama de cajas:

RANGO_EDAD

Los datos atípicos en este atributo son aquellos que son menores a 12.5 años o mayores a 32.5, en los datos que se encuentran almacenados en las bases de datos se tiene que 574 personas tienen más de 32.5 años siendo estos los atípicos, el diagrama de cajas de este atributo que representa la distribución de sus datos típicos y atípicos se muestra en la Fig. 29.

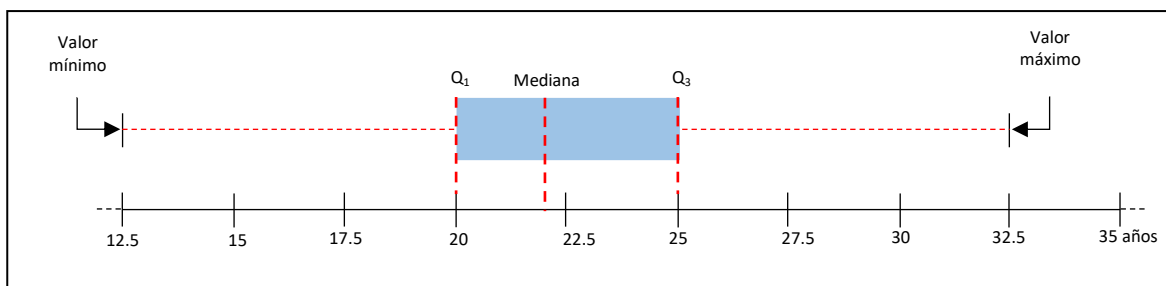


Fig. 29. Diagrama de caja de atributo RANGO_EDAD

RANGO_INGRESO_MENSUAL

En el rango ingreso mensual, se consideran como atípicos aquellos datos que se encuentre por debajo de -325 o sobre 1555, tomando en cuenta que los datos que se almacenan van de \$0 a \$130733, únicamente se consideran como atípicos aquellos datos que se encuentren sobre 1555 que son 824 registros considerados atípicos, el diagrama de cajas de este atributo se aprecia en la Fig. 30.

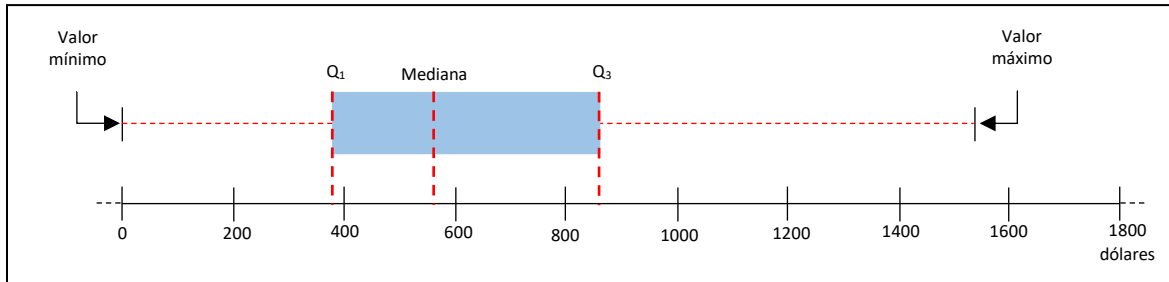


Fig. 30. Diagrama de caja de atributo RANGO_INGRESO_MENSUAL

RANGO_PROMEDIO

Para que el promedio de un estudiante sea considerado como un valor atípico, debe ser menor a 5.83 o mayor a 10.11, tomando en cuenta que el sistema de calificación en la universidad va de 1 a 10 (UTN, 2019), se tomarán en cuenta únicamente aquellos datos que son menores a 5.83, existiendo un total de 469 datos atípicos, su diagrama de cajas se encuentra especificado en la Fig. 31.

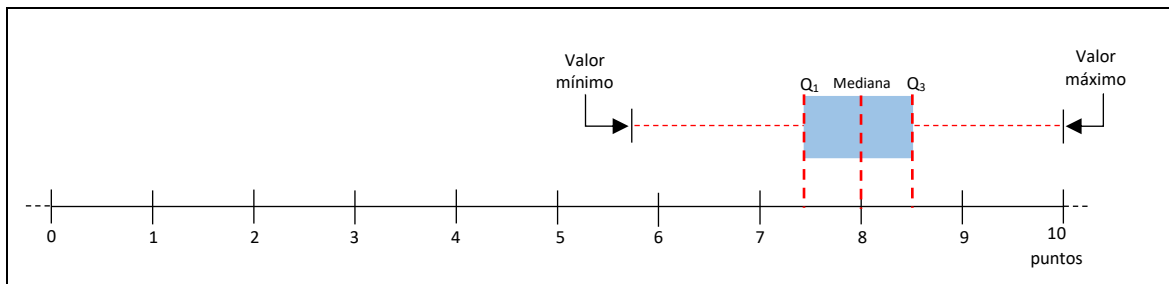


Fig. 31. Diagrama de caja de atributo RANGO_PROMEDIO

Para las variables cualitativas se empleó el mismo concepto de cuartiles, sin embargo, se obtuvieron los siguientes resultados con respecto a las variables parcialmente balanceadas:

TIPO_SANGRE

En el atributo que almacena el tipo de sangre de los estudiantes se tiene que las personas cuyo tipo de sangre después de la normalización sea menor a -7 o mayor a 17 son considerados atípicos, es decir que en este atributo no se tienen datos atípicos por que en los datos que se almacenan son del 1 al 9.

FINANCIAMIENTO

Los datos que se consideran atípicos en el financiamiento son aquellos que su valor después de ser normalizado es menor que 4.5 o mayores que 8.5, en este caso se tienen como atípicos 2475 casos menores de 4.5 y 1008 datos valores a 8.5, existiendo un total de 3483 registros considerados atípicos 11200 registros.

CONVIVIENTE

En el atributo conviviente para que un dato sea considerado atípico, su valor después de ser normalizado debe ser menor que -2.5 o mayor que 9.5, teniendo en cuenta que las categorías de este atributo van del 1 al 7, no existen datos atípicos en este atributo.

TIPO_VIVIENDA

Los datos atípicos en el atributo tipo vivienda son aquellos cuyo valor después de ser normalizado debe ser menor que -4 o mayor que 12, al igual que en el caso anterior no se tienen valores atípicos puesto que los datos que se almacena son del 1 al 7.

Cuando se trata de atributos no balanceados, no se puede identificar valores atípicos que se acerquen a la realidad, ya que se considerará como valores típicos, aquellos que se encuentren en la categoría que predomine en cada uno de los atributos no balanceados, teniendo como dato típico para cada atributo lo siguiente: (i) en el atributo **ETNIA** se entiende la categoría mestizo, (ii) en **PAIS_NACIONALIDAD** Ecuador, (iii) en el **RANGO_DISCAPACIDAD** que el estudiante no tenga discapacidad, (iv) en el **ESTADO_CIVIL** se tiene como valor típico soltero, (v) en **ACTIVIDAD_ESTUDIANTE** que el estudiante no trabaje, (vi) en **PROVINCIA_PROCEDENCIA** Imbabura y (vii) en **MOTIVO_ABANDONO** ninguno.

Puesto que los resultados anteriormente expuestos no reflejan la realidad por la naturaleza de las variables (cualitativas), se procedió a realizar el siguiente análisis estadístico en base a

los porcentajes presentes en las bases de datos institucionales con referencia a los estudiantes que abandonaron su carrera:

ETNIA

En el atributo ETNIA, se observa que la cantidad de mestizos predomina sobre todas las variables, en cuanto a las demás variables se tiene que MONTUBIO es la etnia que menos se encuentra en la UTN siendo un 0.091% del total, se tiene que 0.638% de los estudiantes no especifican su etnia, la etnia AFRODESCENDIENTE cuenta con un total de 2.372%, considerándose que las tres categorías anteriormente mencionadas son atípicas en cuanto a los estudiantes inactivos tal como se observa en la Fig. 32, teniéndose como total 34 datos atípicos según su etnia

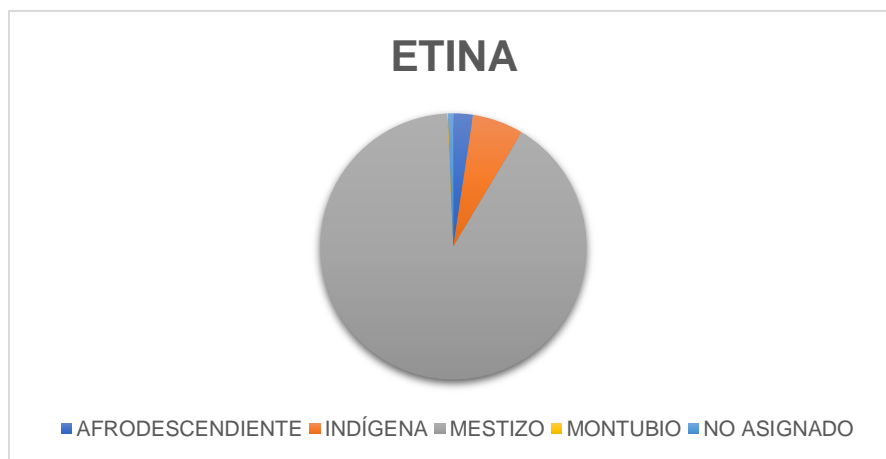


Fig. 32. Estadísticas del atributo ETNIA

PAIS_NACIONALIDAD

El país predominante que hace referencia a la nacionalidad de los estudiantes inactivos es Ecuador con un porcentaje del 99.178%, mientras que Cuba y Colombia se consideran atípicos ya que cuentan con 0.0912% y 0,729% respectivamente del total, conformando un total de 9 datos atípicos, como se aprecia en la Fig. 33.



Fig. 33. Estadísticas del atributo PAÍS_NACIONALIDAD

RANGO_DISCAPACIDAD

En el atributo RANGO_DISCAPACIDAD se considera como atípico la discapacidad leve ya que es 0.364% del total que corresponde a 4 datos atípicos, de acuerdo con la Fig. 34.



Fig. 34. Estadísticas del atributo RANGO_DISCAPACIDAD

ESTADO_CIVIL

En el atributo ESTADO_CIVIL, la variable que predomina es soltero con 88.959%, dejando como valores atípicos a viudo, divorciado y unión de hecho con 0.091%, 1.551% y 1.277%, respectivamente, porcentajes que corresponde a 32 valores atípicos, tal como se muestra en la Fig. 35.

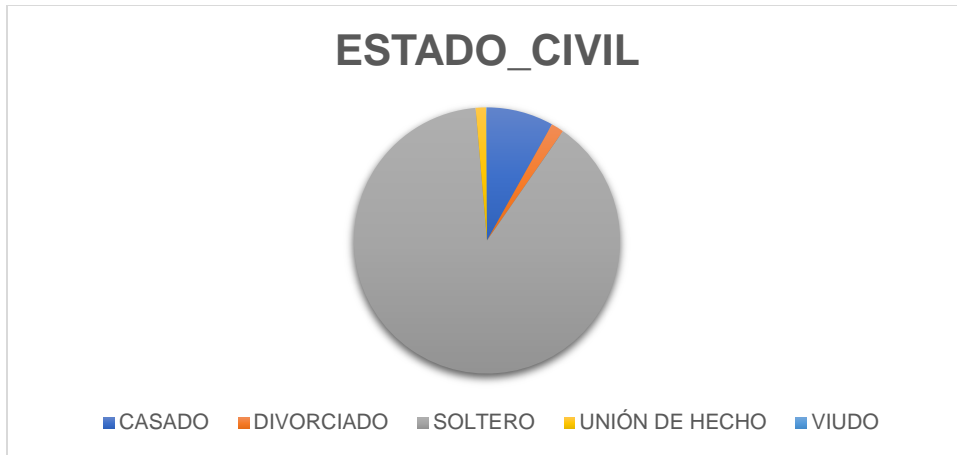


Fig. 35. Estadísticas del atributo ESTADO_CIVIL

TIPO_SANGRE

El atributo TIPO_SANGRE cuenta con algunas variables que son considerados atípicos por su bajo porcentaje, tales como: A-, AB-, AB+, B- y O- con un porcentaje de 0.364%, 0.273%, 1.003%, 0.456%, 0.729% respectivamente los cuales representan 31 datos atípicos, tal como se muestra en la Fig. 36.

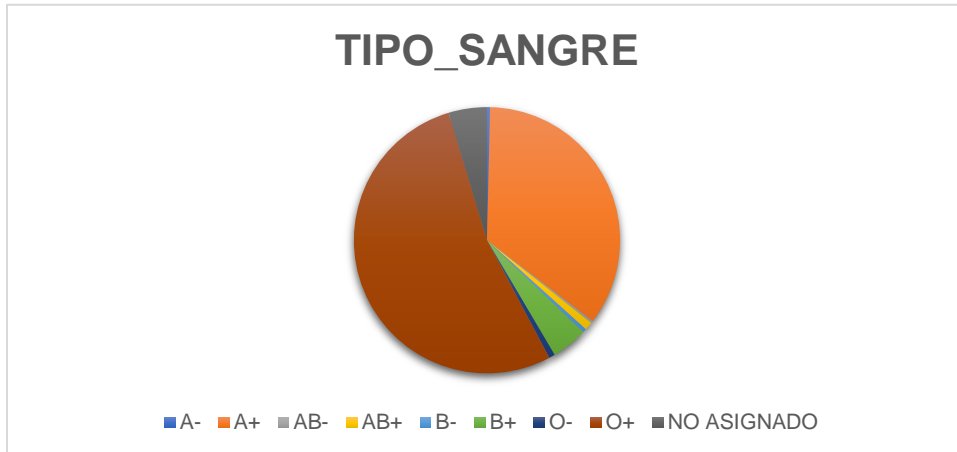


Fig. 36. Estadísticas del atributo TIPO_SANGRE

FINANCIAMIENTO

Los valores que son considerados atípicos en este atributo son: otros, pareja y familiar con un porcentaje de: 0.456%, 2.281%, 2.372% respectivamente, que corresponden a 56 valores atípicos, tal como se muestra en la Fig. 37.

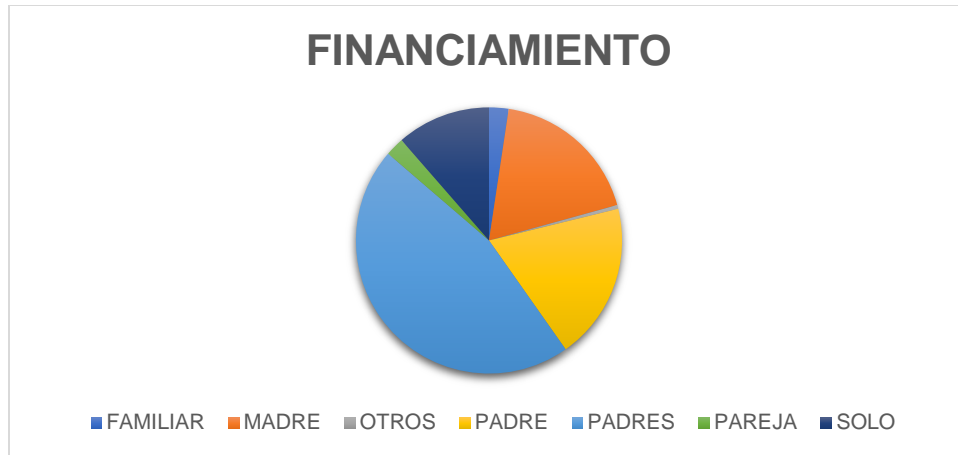


Fig. 37. Estadísticas del atributo FINANCIAMIENTO

GENERO

Es este atributo las variables: femenino y masculino con un porcentaje de: 54.744%, 45.255%, respectivamente, no se puede considerar a ninguna de las variables como atípico, ya que tienen una similitud en cuanto al porcentaje, tal y como se aprecia en la Fig. 38.

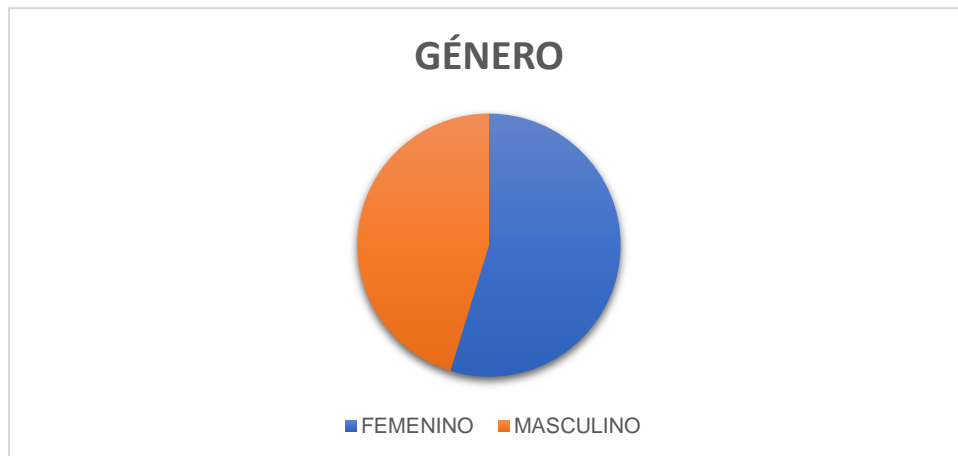


Fig. 38. Estadísticas del atributo GÉNERO

CONVIVIENTE

En este atributo, las variables que son consideradas atípicos son: otros y padre con un porcentaje de: 0.821% y 2.554% respectivamente que corresponden a 37 valores, tal y como se muestra en la Fig. 39.



Fig. 39. Estadísticas del atributo CONVIVIENTE

TIPO_VIVIENDA

La variable que predomina en este atributo es con respecto a los estudiantes que tienen casa propia, con un porcentaje de 64.324% del total, en cuanto a las variables que son consideradas atípicas son: anticresis, hipotecada y no asignada con un porcentaje de: 0.456%, 0.821%, 0.821% respectivamente que corresponden a 23 valores, tal y como se aprecia en la Fig. 40.



Fig. 40. Estadísticas del atributo TIPO_VIVIENDA

ACTIVIDAD_ESTUDIANTE

La variable predominante en este atributo hace referencia a los estudiantes que no trabajan, con un porcentaje de 88.868%, en cuanto a la variable trabaja tiene un porcentaje del 11.131%, siendo esta considerada como atípico con un total de 122 registros, tal y como se muestra en la Fig. 41.



Fig. 41. Estadísticas del atributo ACTIVIDAD_ESTUDIANTE

PROVINCIA_PROCEDENCIA

Las provincias de procedencia de los estudiantes como: Bolivar, Cuba, Colombia, Esmeraldas, Loja, Manabí, Orellana, Santo Domingo de los Tsáchilas, Sucumbíos, Zamora Chinchipe, Tungurahua son consideradas atípicas ya que tiene un porcentaje de: 0,091, 0,091, 0,365, 0,365, 0,091, 0,456, 0,091, 0,547, 0,365, 0,091, 0,182 respectivamente que representan 30 registros, tal y como se aprecia en la Fig. 41.

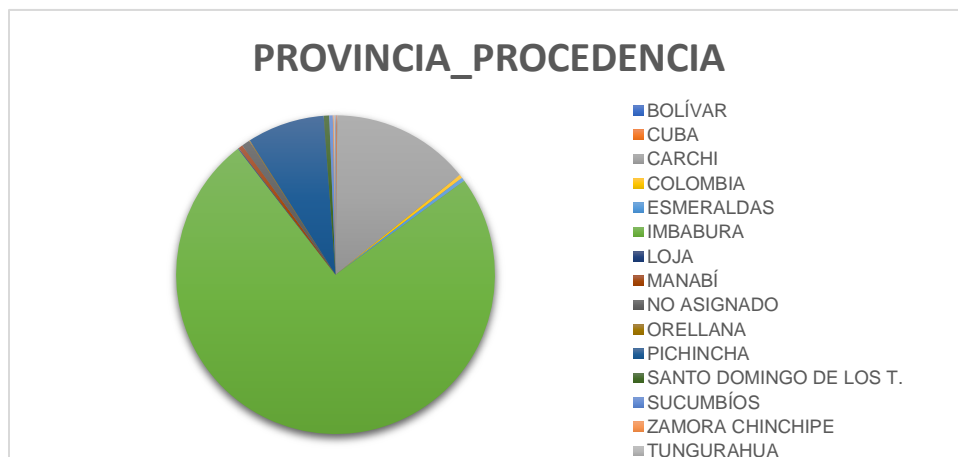


Fig. 42. Estadísticas del atributo PROVINCIA_PROCEDENCIA

MOTIVO_ABANDONO

En este atributo, las variables: pierde tercera matrícula y cambio carrera, se consideran atípicas ya que tienen un porcentaje del, 6.386% y 7.025% respectivamente que representan 147 registros, tal como se muestra el la Fig. 43.



Fig. 43. Estadísticas del atributo MOTIVO_ABANDONO

3.3. Obtención del conocimiento

- **Agrupamiento**

Para obtener el conocimiento, se tomaron en cuenta varios aspectos entre ellos si los clústeres que arrojo un algoritmo con los del otro coinciden o se complementan, obteniéndose como resultado que únicamente cuando las variables son no balanceadas coinciden en la categoría predominante de lo contrario son totalmente diferentes; esto se debe a que cada algoritmo busca que los clústeres generados sean diferentes entre sí (Lara, 2014). Tomando en cuenta que la eficacia de cada algoritmo se analizó de forma diferente, se tomó en cuenta las propiedades y las características más sobresalientes de cada algoritmo, siendo el algoritmo EM el más apropiado para la naturaleza del presente estudio ya que asigna una distribución de probabilidad de pertenencia a cada instancia a cada uno de los grupos, una de las características esenciales para elegir este algoritmo es que trata de maximizar la probabilidad del modelo a partir de datos incompletos mediante la media y la desviación estándar de cada uno de los atributos en cada clúster y se especifica que clúster describe a cada facultad (Sierra, 2006).

Por lo anteriormente mencionado, se tiene que los estudiantes que abandonaron su carrera universitaria tienen las siguientes características para cada clúster:

Clúster 1 – FACAE

Las estudiantes de género femenino, solteras, con edad baja comprendida entre 18 y 25 años, con tipo de sangre A+, que viven únicamente con su padre con un promedio entre 8.01 y 9 puntos.

Clúster 2 – FICA

Los estudiantes de género masculino, solteros, con una edad comprendida entre 26 y 39 años, con tipo de sangre O+ que viven únicamente con su padre, con un promedio de 7 a 8 puntos.

Clúster 3 – FCCSS

Las estudiantes de género femenino, solteras, con una edad entre 18 y 25 años y tipo de sangre O- que viven con su padre y tienen un promedio entre 8.01 y 9 puntos.

Clúster 4 – FICAYA

Los estudiantes de género masculino, solteros, con una edad comprendida entre 18 y 25 años, con tipo de sangre A+ que viven con su padre, con un promedio de 7 a 8 puntos.

Clúster 5 – FECYT

Los estudiantes de género masculino, cuyo estado civil es unión de hecho, que viven con sus padres y tipo de sangre B+, sus estudios son financiados por su pareja y tienen un promedio de 8 a 9 puntos.

Como se aprecia muchas de las variables no fueron tomadas en cuenta al momento de definir las características que diferencian a un clúster de los otros, esto se debe al fenómeno que se ha venido mencionando a lo largo de este capítulo que son los datos no balanceados; considerándose que las variables PAÍS_NACIONALIDAD, ETNIA, RANGO_DISCAPACIDAD, TIPO_VIVIENDA, ACTIVIDAD_ESTUDIANTE, RANGO_INGRESO_MENSUAL y MOTIVO_ABANDONO, en este caso no aportan mayor conocimiento al momento de describir a los potenciales desertores académicos de la UTN. El atributo PROVINCIA_PROCEDENCIA es uno de los datos que se encuentra no balanceados e incompletos, por esta razón el algoritmo EM llenó la información en los espacios perdidos, por lo que al momento de desnormalizar los resultados se obtuvieron los valores de Imbabura y Guayas que se mantenían como la media en

cada uno de los clústeres, por este motivo también se considera que no aporta conocimiento al estudio, teniéndose como variables más relevantes: RANGO_EDAD, TIPO_SANGRE, GÉNERO, ESTADO_CIVIL, CONVIVIENTE, FINANCIAMIENTO y RANGO_PROMEDIO.

- **Atípicos**

Para identificar qué características son poco frecuentes de los estudiantes que han abandonado su carrera, se tomaron en cuenta si las variables son cualitativas o cuantitativas, con el objetivo de aplicar las técnicas adecuadas para cada una de ellas, diagrama de cajas para las variables cuantitativas y un análisis de porcentajes para las variables cualitativas. Como resultado principal de datos atípicos en los estudiantes que abandonaron su carrera universitaria se obtuvo lo siguiente:

- Los estudiantes de grado que su edad se encuentra por encima de los 32,5 años.
- Los estudiantes cuyas entradas económicas familiar sobrepase los 1555 dólares.
- Los estudiantes que tienen un promedio académico menor a 5.83 puntos.
- Las personas que son de etnia afrodescendiente, montubio.
- Los estudiantes cuyo estado civil es viudo, divorciado o unión de hecho.
- Los estudiantes cuyo tipo de sangre es A-, AB-, AB+, O- o B- .
- Las personas que sus estudios son financiados por su pareja, familiar u otros.
- Los estudiantes que viven solo con su padre u otras personas.
- Los estudiantes cuya vivienda se encuentra hipotecada, adquirida en anticresis o que no especifican el estado de la vivienda.

Las variables que se encuentran no balanceadas como datos típicos las categorías predominantes, entendiéndose que la identificación de valores atípicos en este tipo de variables resulta ineficiente puesto que de antemano se conoce que como valores atípicos se obtendrán las categorías que cuenten con un número menor de registros, como es el caso particular del atributo RANGO_DISCAPACIDAD que en la categoría NO TIENE existen 1092 instancias mientras que en la categoría LEVE existen 4 instancias, considerándose la categoría leve como un atípico; este fenómeno se presenta en todas la variables no balanceadas. Por este motivo y considerando los resultados obtenido en el punto 3.2. que hace referencia a la obtención de valores atípicos, se han seleccionado las variables que aportan conocimiento relevante en este estudio en particular: RANGO_EDAD, RANGO_INGRESO_MENSUAL, RANGO_PROMEDIO,

ETNIA, ESTADO_CIVIL, TIPO_DE_SANGRE, FINANCIAMIENTO, CONVIVIENTE y TIPO_VIVIENDA.

3.4. Análisis de impactos

Después de aplicar las técnicas descriptivas de minería de datos (agrupamiento, asociación y atípicos) y teniendo distintas características de lo que arrojó el descubrimiento, el siguiente paso es que el conocimiento obtenido sirva para que los directivos y las personas encargadas de velar por los estudiantes y que culminen su carrera, tomen decisiones estratégicas y elaboren proyectos en base a información real para incentivar a los estudiantes, ayudando así a cumplir con el objetivo 4 de Desarrollo Sostenible (ODS) (Naciones Unidas, 2015), que hace referencia a la Educación de Calidad.

El análisis de impactos define las posibles consecuencias que se podrían presentar cuando se tomen decisiones estratégicas en base a los patrones 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 3.16 (Posso, 2013):

TABLA 3.16
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 3.17 a 3.19 y en la Tabla 3.20 se detalla el impacto general del proyecto.

3.4.1. Impacto Educativo

TABLA 3.17
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

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

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

Nivel de Impacto Educativo = Medio positivo

Fuente: Posso, 2013

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.

3.4.2. Impacto Sociocultural

TABLA 3.18
IMPACTO SOCIOCULTURAL

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

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

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

Nivel de Impacto Sociocultural = Alto positivo

Fuente: Posso, 2013

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.

3.4.3. Impacto Económico

TABLA 3.19
IMPACTO ECONÓMICO

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

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

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

Nivel de Impacto Económico = Alto positivo

Fuente: Posso, 2013

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.

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.

3.4.4. Impacto General

TABLA 3.20
IMPACTO GENERAL

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

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

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

Nivel de Impacto Tecnológico = Alto positivo

Fuente: Posso, 2013

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 y la sociedad.

3.5. Discusión

El presente estudio se relaciona con la investigación realizada por (Vila, et al., 2018), ya que desarrollaron su trabajo con datos personales y académicos de los estudiantes de la UTN para obtener patrones de deserción estudiantil mediante el uso de técnicas predictivas mediante la tareas de clasificación, mientras que el presente trabajo se realizó en base a la misma información y adicionalmente con los datos socioeconómicos de los mismos estudiantes con técnicas descriptivas, obteniéndose resultados similares en base a las variables que se relacionan directamente con el abandono escolar, tales como edad del estudiante, estado civil y promedio. Sin embargo, el lugar de procedencia del estudiante es un factor importante en esta problemática en el trabajo realizado por Vila et al., siendo por el contrario en este trabajo una de las variables poco relevantes.

Igualmente, comparte puntos en común con la investigación realizada por (Grijalva Arriaga, et al., 2018), utilizaron 12 variables cualitativas de información académica y personal del estudiante, para posteriormente procesar esta información por medio del algoritmo k-means, ya que llagaron a la conclusión que el género, promedio y edad son factores que contribuyen directamente a la deserción estudiantil mediante la técnica de clustering, teniendo como principal diferencia que los estudiantes con promedio menor a 6 tienden a abandonar sus estudios.

(Timarán, et al., 2013) utilizaron una vista minable de 11 variables compuesta por atributos cualitativos y cuantitativos para formar dos clústeres con información académica del estudiante, así como el género y su estrato económico, y determinaron que, en el grupo relacionado con las ciencias de la educación, desertan únicamente las personas de género masculino al igual que en el presente estudio, mientras que en las materias relacionadas con las ciencias económicas y administrativas se diferencian ambos trabajos puesto que Timarán en sus resultados hace referencia a que los hombres abandonan sus estudios.

Por otro lado, entre las principales restricciones de este estudio, se encuentran las siguientes: (i) los datos no balanceados, ya que limitan el estudio al no poder aplicar cualquier algoritmo por la naturaleza de los datos; (ii) la capacidad de procesamiento tanto del hardware como del software, ya que al existir una gran cantidad de datos los recursos no se abastecen, debido a la complejidad de procesamiento de ciertos algoritmos de agrupamiento (como el algoritmo jerárquico) y; (iii) falta de información psicológica de los estudiantes para realizar este análisis desde otro enfoque.

Como trabajos a futuro en esta rama de inteligencia artificial se sugiere: (i) abordar las problemáticas que afecten al desempeño del estudiante, tales como movilidad, prioridad que le dieron a la carrera que se encuentran cursando en el Sistema Nacional de Nivelación y Admisión (SNNA), afinidad con su carrera, entre otros; (ii) recolectar información del uso del aula virtual, con el fin identificar si el desempeño varía o no de los estudiantes que utilizan esta plataforma de los que no y; (iii) establecer un modelo por cada facultad con el número de clústeres de acuerdo al número de carreras que posee cada una de ellas con las variables relevantes obtenidas en esta investigación.

Conclusiones y Recomendaciones

Conclusiones

Se llevó a cabo una minería de datos con información personal, académica y socioeconómica de los estudiantes del nivel de grado de la UTN, y se determinó que mediante las técnicas descriptivas se pudieron identificar 5 clústeres principales correspondientes a cada facultad.

En el presente proyecto se pudo evidenciar que existe mucha información relevante con relación a las técnicas de agrupamiento para obtener patrones de abandono escolar; sin embargo, para las tareas de asociación y atípicos no se encontraron trabajos relacionados que sirvan como base para el presente estudio.

Se determinó que al aplicar el proceso KDD se puede realizar un análisis de datos más ordenado, para obtener la vista minable se emplearon los softwares de Pentaho Data Integration y Microsoft Excel; se procesaron datos de ciclos académicos, dependencias, localidades, matrículas y notas provenientes de los repositorios Oracle de la universidad, para llegar a conformar la vista minable conformada por 17 atributos, sin embargo no todos son aplicables al estudio.

Se aplicaron técnicas descriptivas de minería de datos tales como agrupamiento, asociación y atípicos; puesto que el software de Weka no cumplía con las expectativas planteadas al aplicar ciertos algoritmos, por ejemplo, al ejecutar el algoritmo K-means no brindó la información necesaria para realizar la evaluación, por esta razón se realizó una comparativa con el software SPSS. No se obtuvieron resultados con el algoritmo Apriori perteneciente a la tarea de asociación, ya que los datos con los que se trabajaron son no balanceados, por este motivo se obtuvieron resultados muy obvios (sección 3.1.1) que no aportan conocimiento. En la tarea de

agrupación se realizó una comparativa con el algoritmo EM y K-means (sección 3.1.2). Para la tarea de atípicos se emplearon técnicas correspondientes a la naturaleza de las variables (cualitativa y cuantitativas) (sección 3.2).

Para validar que los datos empleados sean de calidad, se aplicó la ISO/IEC 25012:2018; sin embargo los resultados obtenidos no aportaron de manera relevante al estudio puesto que el proceso KDD en si ya verifica en cada una de sus fases que la vista minable sea de calidad; en cuanto a las tareas de agrupamiento se tomó en cuenta las características principales de cada uno de los algoritmos, se tomó en cuenta el algoritmo EM puesto que es robusto a datos perdidos mientras que el algoritmo K-means es débil frente a datos no balanceados como es en este caso. En cuanto a la obtención del conocimiento de los datos atípicos, se determinó que los diagramas de cajas devuelven resultados efectivos únicamente con datos cuantitativos por lo que para los datos cualitativos se realiza un análisis estadísticos descriptivo en cuanto a los porcentajes que cada una de las variables posee frente al abandono escolar.

Recomendaciones

Para trabajo de minería de datos se recomienda seguir el proceso KDD de manera ordenada y precisa, para obtener resultados de calidad al momento de describir cierta población ya que su uso es sencillo y de rápido desarrollo para proyectos relacionados con esta temática.

Se recomienda revisar la literatura antes de determinar que técnicas se van a aplicar y si estas son apropiadas para la información seleccionada, para obtener resultados de calidad y que vayan de acuerdo con las expectativas planteadas inicialmente.

Para asegurar que los resultados obtenidos sean de calidad se recomienda que antes de la etapa de recolección de datos se preste mucha atención a la información para evitar que existan datos perdidos, nulos, incongruentes para evitar inconvenientes en la siguientes etapas del proceso KDD, ya que de esto depende la precisión de los resultados.

Se recomienda realizar una comparativa previa de las herramientas tecnológicas que son aplicables para este tipo de análisis con el objetivo de emplear algoritmos con diferentes funcionalidades para generar nuevos conocimientos.

Se recomienda a las personas encargadas usar esta información de forma estratégica en bien de la comunidad universitaria; y a su vez se puede plantear otros estudios por esta línea de investigación para que la información registrada en las bases de datos institucionales pueda ser utilizada en beneficio de la universidad, cubriendo aspectos tales como el factor psicológico de

los estudiantes, docentes y personal administrativo, la movilidad de los mismos o su nivel de satisfacción dentro de la universidad e información proveniente del aula virtual de la UTN, esto permitirá cubrir mayor parte de los indicadores institucionales.

GLOSARIO DE TÉRMINOS

Atípicos: Valores que son diferentes al conjunto de datos principal

Centroide: Punto central de un objeto.

Cuartil: Uno de los tres puntos que dividen un conjunto de datos numéricamente ordenados en cuatro partes iguales

Data dredging: Término que se utilizaba para referirse a minería de datos

Dimensionalidad: Proceso de reducción del número de variables en el cual se esté trabajando

Entidad: Representación de un objeto o concepto del mundo real que se describe en una base de datos

Media: La media es el valor promedio de un conjunto de datos numéricos, calculada como la suma del conjunto de valores dividida entre el número total de valores

Mediana: En el ámbito de la estadística, la mediana representa el valor de la variable de posición central en un conjunto de datos ordenados.

Rango intercuartil: Valor absoluto de la diferencia numérica entre el primer y tercer cuartiles

BIBLIOGRAFÍA

- Unesco. (2015). Recuperado de <http://unesdoc.unesco.org/images/0023/002324/232430s.pdf>
- Amelio, A., & Tagarelli, A. (2018). Data Mining: Clustering. *Encyclopedia of Bioinformatics and Computational Biology*, 1, 437-438. <https://doi.org/doi.org/10.1016/B978-0-12-809633-8.20489-5>
- Asamblea Nacional del Ecuador. (2016). Ley orgánica de gestión de la identidad y datos civiles. 16.
- Baquerizo, D. C. R. P., Tam, D. C. O. A., & López, D. C. J. G. (2014). La deserción y la repitencia en las instituciones de Educación Superior: algunas experiencias investigativas en el Ecuador. *Universidad y Sociedad*, 6(1). Recuperado de <https://rus.ucf.edu.cu/index.php/rus/article/view/177>
- Business School, O. (2018). Project Management. Recuperado 5 de diciembre de 2018, de OBS Business School website: <https://www.obs-edu.com/int/blog-project-management/salidas-profesionales-pm/que-es-un-jefe-de-proyecto-y-cual-es-su-perfil-profesional>
- Espinoza, M., & Gallegos, D. (2018). Data Scientist: A Systematic Review of the Literature. © Springer Nature Switzerland AG 2019, 477-487. 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). The Weka Workbench (Cuarta). Recuperado de https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf?fbclid=IwAR07ih1OByDfK4KQMR2yTua80LU3L38XbVUn6NJse5TEbghb4nhAowAdccA

- García López, A., Berlanga, R., & Danger, R. (2008). A Description Clustering Data Mining Technique for Heterogeneous Data. En J. Filipe, B. Shishkov, & M. Helfert (Eds.), *Software and Data Technologies* (pp. 361-373). Springer Berlin Heidelberg.
- Gonzalez, A. G. H., Armenta, R. A. M., Rosales, L. A. M., Barrientos, A. G., Xihuitl, J. L. T., & Algreto, 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>
- Gorgas, J., Cardiel, N., & Zamorano, J. (2011). *Estadística básica para estudiantes de ciencias* (17.^a ed.). Recuperado de http://webs.ucm.es/info/Astrof/users/jaz/ESTADISTICA/libro_GCZ2009.pdf
- Grijalva Arriaga, P., Freire Avilés, V., Real Avilés, K., & Arellano Arcentales, A. (2018). *Application of Data Mining Techniques for the Analysis of Academic Efficiency*. Vol. 3, 17.
- Guzmán, A. (2015). Repositorio Digital Universidad Técnica del Norte: Implementación de una solución de inteligencia de negocios acerca de la información de los docentes, estudiantes y personal administrativo de la Universidad Técnica del Norte para el Instituto de Altos Estudios (Universidad Técnica del Norte). Recuperado de http://repositorio.utn.edu.ec/handle/123456789/7720?fbclid=IwAR3YJlbyYcDhwxidFzLw9mYxktshkUvtWHSs7COZ_5nEbDoqfRosHk9B0ug
- Hamilton, H. (2018, septiembre 7). Matriz de confusión [Computer Science]. Recuperado 8 de enero de 2019, de Knowledge Discovery in Databases website:
<http://www.esacademic.com/dic.nsf/eswiki/788754?fbclid=IwAR0UXEbmIZiak6tsqz2Bzm5B7dC69L8VpVH1KdXb5Vy8oJgXIBwYXAvVT6U>
- Han, J., Kamber, M., & Pei, J. (2001). *Data Mining: concepts and techniques* (Tercera). Recuperado de <http://myweb.sabanciuniv.edu/rdekharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and->

- Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf?fbclid=IwAR2SvytVmOOkgt9eh4k6AsbXqJdtZDUqf_IXLgUsU8JF-qyRrBG0i60cok
- Hasperué, L. W. (2013). Extracción de Conocimiento en Grandes Bases de Datos Utilizando Estrategias Adaptativas. p. 209.
- Hernández, A., Delgado, E., Rivera, J., & Castellanos, G. (2006). Reducción de dimensiones para clasificación de datos multidimensionales usando medidas de información. *Scientia et Technica* Año XII, 32, 181-186.
- Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid, España: PEARSON EDUCACIÓN, S.A.
- Hitachi. (2019). About Pentaho. Recuperado 17 de febrero de 2019, de <https://www.hitachivantara.com/go/pentaho.html>
- Honarkhah, M., & Caers, J. (2010). Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling. *Mathematical Geosciences*, 42(5), 487-517. <https://doi.org/10.1007/s11004-010-9276-7>
- Huang, W., & Chen, Y. (2017). The multiset EM algorithm. *Statistics & Probability Letters*, 126, 41-48. <https://doi.org/10.1016/j.spl.2017.02.021>
- IBM. (2019). IBM SPSS Software | IBM. Recuperado 12 de febrero de 2019, de https://www.ibm.com/analytics/spss-statistics-software?fbclid=IwAR3zO__NZjFXtg_RP_GFP9DLtvtedmQ92I83fCsezPXugPQZmrCxKoehEY
- 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_Analíticas_nov_2018.pdf

- 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/?fbclid=IwAR2uuTFb-t3MfrTgFQSza7S322IEBli0zXEvgSA2x5iS-3tN9g3fpAnvI_U#iso:std:iso-iec:25012:ed-1:v1:en
- Lara, J. (2014). Minería de Datos (CENTRO ESTUDIOS FINANCIEROS).
- Lind, D., Marchal, W., & Wathen, S. (2012). Estadística aplicada a los negocios y la economía (15.ª ed.). Mexico.
- Microsoft. (2019). Software de hojas de cálculo: prueba gratuita de Excel, Microsoft Excel. Recuperado de <https://products.office.com/es-ww/excel>
- Minewiskan. (2018). Métodos de discretización (minería de datos). Recuperado 11 de octubre de 2018, de <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/discretization-methods-data-mining>
- Mishra, A., Bansal, R., & Singh, S. N. (2017). Educational data mining and learning analysis. 2017 7th International Conference on Cloud Computing, Data Science Engineering - Confluence, 491-494. <https://doi.org/10.1109/CONFLUENCE.2017.7943201>
- Moya, R. (2016, septiembre 12). Selección del número óptimo de Clusters. Recuperado 19 de febrero de 2019, de Jarroba website: <https://jarroba.com/seleccion-del-numero-optimo-clusters/>
- Naciones Unidas. (2015, septiembre 17). Objetivos y metas de desarrollo sostenible. Recuperado 11 de marzo de 2019, de Desarrollo Sostenible website: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>
- Ochoa Reyes, A. J., Orellana García, A., Sánchez Corales, Y., & Davila Hernández, F. (2014). Componente web para el análisis de información clínica usando la técnica de Minería de Datos por agrupamiento. Revista Cubana de Informática Médica, 6(1), 5-16.

- Palacios-Pacheco, X., Villegas-Ch, W., & Luján-Mora, S. (2018). Application of Data Mining for the Detection of Variables that Cause University Desertion. © Springer Nature Switzerland AG 2019, 510-520. https://doi.org/10.1007/978-3-030-05532-5_38
- Pasin, O., & Ankarali, H. (2017). Comparison of EM and Two-Step Cluster Method for Mixed Data: An Application. *International Journal of Medical Science and Clinical Inventions*, 2768 a 2773. <https://doi.org/10.18535/ijmsci/v4i3.08>
- Perez, 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 Enterprice Miner*. RA-MA Editorial.
- Pérez López, C., & Santín González, D. (2007). *Minería de datos Técnicas y Herramientas*. Madrid, España: Paraninfo, S.A.
- Posso, M. (2013). *Proyectos, Tesis y Marco Lógico: Planes e Informes de Investigación*. Quito, Ecuador. [presentación-rendición-de-cuentas.pdf](http://www.senescyt.gob.ec/rendicion2015/assets/presentaci%C3%B3n-rendici%C3%B3n-de-cuentas.pdf). (s. f.). Recuperado de <http://www.senescyt.gob.ec/rendicion2015/assets/presentaci%C3%B3n-rendici%C3%B3n-de-cuentas.pdf>
- 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>
- Sanchez, A. (2017, enero 11). Cómo puede el algoritmo K-means ayudarte a mejorar el conocimiento de tus clientes, y de paso, si eres fans de los Pokemons a mejorar tu juego. Recuperado 7 de enero

de 2019, de A un Clic de las TIC website: <https://aunclidelastic.blogthinkbig.com/algorithmo-k-means-clientes-pokemons/>

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, B. (2006). Aprendizaje Automático: Conceptos básicos y avanzados B. Madrid, España.

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>

Timarán, R., Calderón, A., & Jiménez, J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. Vínculos, Vol. 10. <https://doi.org/10.14483/issn.2322-939X>

UTN. (2019). Universidad Técnica del Norte / UniPortal Web UTN | Vive, sueña, construye. Recuperado 20 de febrero de 2019, de <https://www.utn.edu.ec/web/uniportal/>

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

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>
- Yang, J. (2017). Clustering analyzing of undergraduate schools based on k-means algorithm. 2017 International Conference on Advanced Mechatronic Systems (ICAMechS), 309-311. <https://doi.org/10.1109/ICAMechS.2017.8316489>
- Zelada, C. (2017, octubre 5). RPubs - Matriz de Confusión - Evaluación de modelos de predicción. Recuperado 8 de enero de 2019, de https://rpubs.com/chzelada/275494?fbclid=IwAR3r_PVAXv96kZiSgAT1Mso4hFys8398JRdv2TKbOPZFn3BzKP13yiQ63wA

ANEXOS

Anexo 1. Cálculo del soporte y confianza: <http://bit.ly/2GCfVl1>

Anexo 2. Desviación estándar: <http://bit.ly/2tFxZD5>

Anexo 3. Resultados algoritmo K-means: <http://bit.ly/2UV3GVa>