UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
FACULTAD CIENCIAS DE LA INGENIERÍA
CARRERA DE INGENIERÍA EN SISTEMAS

Proyecto de Investigación previo a la obtención del título de Ingeniero en Sistemas

Título del Proyecto de Investigación:

"ANÁLISIS INTELIGENTE DE DATOS PARA IDENTIFICAR LOS FACTORES QUE INFLUYEN EN LA DESERCIÓN DE LOS ESTUDIANTES DE LA UNIDAD DE ADMISIÓN Y NIVELACIÓN DE LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO"

Autor:

William Daniel Burbano Ferrin

Director de Proyecto de Investigación:

Lic. Amilkar Yudier Puris Cáceres PhD.

Quevedo- Los Ríos- Ecuador

2016

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, **William Daniel Burbano Ferrin,** declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

f. _____

William Daniel Burbano Ferrin

C. C. 1309743605

CERTIFICACIÓN DE CULMINACIÓN DEL PROYECTO DE INVESTIGACIÓN

El suscrito, Lic. Amilkar Yudier Puris Cáceres PhD, Docente de la Universidad Técnica Estatal de Quevedo, certifica que el estudiante William Daniel Burbano Ferrin, realizó el Proyecto de Investigación de grado titulado "**ANÁLISIS INTELIGENTE DE DATOS PARA IDENTIFICAR LOS FACTORES QUE INFLUYEN EN LA DESERCIÓN DE LOS ESTUDIANTES DE LA UNIDAD DE ADMISIÓN Y NIVELACIÓN DE LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO**", previo a la obtención del título de Ingeniero en sistemas, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

…..……..

Lic. Amilkar Yudier Puris Cáceres PhD

DIRECTOR DE PROYECTO DE INVESTIGACIÓN

CERTIFICADO DEL REPORTE DE LA HERRAMIENTA DE PREVENCIÓN DE COINCIDENCIA Y/O PLAGIO ACADÉMICO

Sra.

Ing. Marlene Medina Villacis, M.Sc.

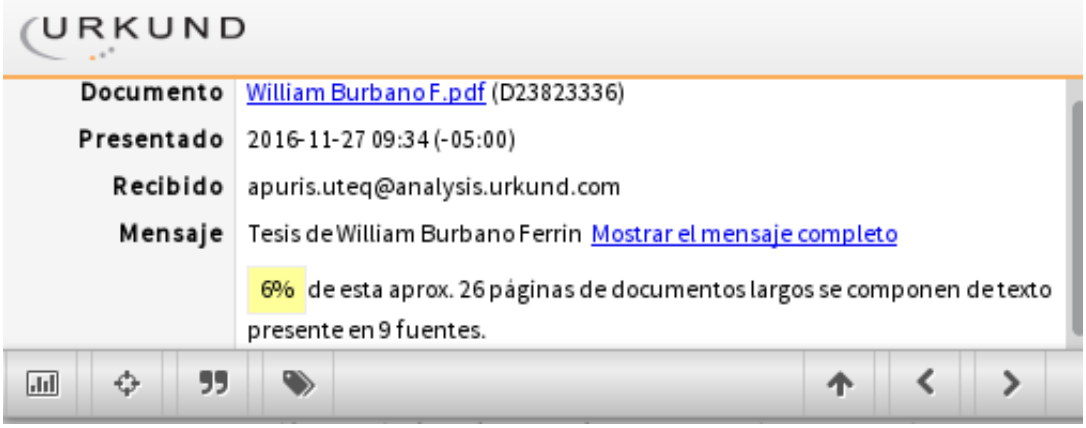
DECANA SUBROGANTE

FACULTAD DE CIENCIAS DE LA INGENIERÍA

En su despacho.

De mi consideración.-

Yo, Lic. Amilkar Yudier Puris Cáceres, PhD., en calidad de Director de la Tesis: **"**ANÁLISIS INTELIGENTE DE DATOS PARA IDENTIFICAR LOS FACTORES QUE INFLUYEN EN LA DESERCIÓN DE LOS ESTUDIANTES DE LA UNIDAD DE ADMISIÓN Y NIVELACIÓN DE LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO**"**, de la autoría **William Daniel Burbano Ferrin**, egresado de la carrera de Ingeniería en Sistema, certifico que ha cumplido con las correcciones pertinentes, y su tesis ha sido ingresada al **sistema URKUND** para determinar el porcentaje de similitud existente con otras fuentes. La evaluación realizada en el sistema Urkund determinó en su informe que existe un 6% de similitud.



Atentamente,

_____

Lic. Amilkar Yudier Puris Cáceres, Phd.

DIRECTOR DE PROYECTO DE TITULACIÓN

CERTIFICACIÓN DE REDACCIÓN TÉCNICA DEL PROYECTO DE INVESTIGACIÓN

La suscrita, Soc. **Teddy Elizabeth De la Cruz Valdiviezo** M.Sc, Docente de la Universidad Técnica Estatal de Quevedo, certifica que al estudiante **Burbano Ferrin William Daniel**, se le procedió a la respectiva revisión y a su vez las correcciones realizadas por el estudiante de su Proyecto Titulado "**ANÁLISIS INTELIGENTE DE DATOS PARA IDENTIFICAR LOS FACTORES QUE INFLUYEN EN LA DESERCIÓN DE LOS ESTUDIANTES DE LA UNIDAD DE ADMISIÓN Y NIVELACIÓN DE LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO**" previo a la obtención del título de Ingeniero en Sistemas bajo mi revisión, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

Soc. Teddy Elizabeth De la Cruz Valdiviezo M.Sc.
REDACCIÓN TÉCNICA

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
FACULTAD CIENCIAS DE LA INGENIERÍA
CARRERA DE INGENIERIA EN SISTEMAS

PROYECTO DE INVESTIGACION

Título:

"Análisis inteligente de datos para identificar los factores que influyen en la deserción de los estudiantes de la unidad de admisión y nivelación de la Universidad Técnica Estatal de Quevedo"

Presentado a la Comisión Académica como requisito previo a la obtención del título de Ingeniero en Sistemas.

Aprobado por:

PRESIDENTE DEL TRIBUNAL DE TESIS

Ing. Pavel Novoa Hernández P.hD

MIEMBRO DEL TRIBUNAL DE TESIS

Ing. Washington Chiriboga Casanova M.Sc

MIEMBRO DEL TRIBUNAL DE TESIS

Ing. Eduardo Samaniego Mena M.Sc

QUEVEDO – LOS RIOS – ECUADOR

2016

AGRADECIMIENTO

En primer lugar a agradezco a Dios quien con su eterno amor, fidelidad y misericordia me ha permitido alcanzar una meta más en mi vida.

Le agradezco a mis padres, quienes me brindaron su apoyo incondicional en cada uno de los objetivos que me he planteado y han sido ejemplo de perseverancia y amor.

Le debo mucho agradecimiento a mis hermanos y a sus respectivos conyugues, por su ayuda y su apoyo en cada momento que lo necesité.

Les doy las gracias a mis compañeros y docentes que hicieron memorables las experiencias vividas a lo largo de mi carrera universitaria

DEDICATORIA

Dedico el presente proyecto al Dios todopoderoso que me permite avanzar en mis logros día a día, llevándome de lo bueno a lo mejor y de lo mejor a lo excelente.

A mi madre Luisa

A mi padre Williams

A mis hermanos Daniel y Daniela

A mis cuñados Kenya y Roberto

A mis demás familiares

A mis memorables amigos, en especial a Angie, Karla y Gabriel.

RESUMEN EJECUTIVO Y PALABRAS CLAVES

La presente investigación se centra en la búsqueda de conocimiento de los repositorios de información de los estudiantes del curso de nivelación de la Unidad de Admisión y Registro de la Universidad Técnica Estatal de Quevedo. Dicha información se caracterizaba por ser variada y almacenaba algunos planos de la realidad de los estudiantes.

En el presente proyecto se muestra la metodología para obtener los factores que influyen en la deserción estudiantil. Esta información que será tomada a partir de un conjunto de datos presente en la Unidad de Admisión y Registro de la Universidad Técnica Estatal de Quevedo. Para lograr obtener los factores de deserción fue necesario la aplicación de la minería de datos y un proceso de análisis inteligente.

La presente investigación aplica los procesos de extracción de conocimientos mediante el uso de los árboles de decisión, los cuales brindaron una serie de modelos que pudieron ser comparados para determinar el más óptimo y los resultados a utilizar, permitiendo detectar las variables o factores más influyentes en la deserción estudiantil del curso de nivelación de la Unidad de Admisión y Registro de la Universidad Técnica Estatal de Quevedo.

Finalmente se establecen las conclusiones y las recomendaciones, que principalmente establecen la importancia de aplicar este tipo de investigaciones a nivel educativo y la posibilidad de expandir la presente investigación con el uso de otros algoritmos para modelado de datos.

Palabras clave: Análisis inteligente, KDD, árboles de deserción, deserción estudiantil.

ABSTRACT AND KEYWORDS

The present research focuses on the search of knowledge of the information repositories of the students of the leveling course of the Admission and Registration Unit of Quevedo State Technical University. This information was characterized by being varied and stored some maps of the reality of students.

The present project shows the methodology to obtain the factors that influence student dropout. This information will be taken from a data set present in the Admission and Registration Unit of Quevedo State Technical University. In order to obtain the desertion factors it was necessary the application of data mining and an intelligent analysis process.

The present research applies the processes of knowledge extraction through the use of decision trees, which provided a series of models that could be compared to determine the most optimal and the results to be used, allowing to detect the variables or factors most influential in The student desertion of the leveling course of the Admission and Registration Unit of Quevedo State Technical University.

Finally, the conclusions and recommendations are established, which mainly establish the importance of applying this type of research at the educational level and the possibility of expanding the present research with the use of other algorithms for data modeling.

Keywords: Intelligent analysis, KDD, decision trees, student desertion.

ÍNDICE DE CONTENIDO

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

CÓDIGO DUBLIN

Título	Análisis inteligente de datos para identificar los factores que influyen en la deserción de los estudiantes de la unidad de admisión y nivelación de la Universidad Técnica Estatal de Quevedo			
Autor	Burbano Ferrin, William Daniel			
Palabras clave:	Análisis inteligente	KDD	Árboles de decisión	Deserción estudiantil
Fecha de publicación				
Editorial	Quevedo: UTEQ, 2016			
Resumen	**Resumen.-** En el presente proyecto se muestra la metodología para obtener los factores que influyen en la deserción estudiantil. Esta información será tomada a partir de un conjunto de datos presente en la Unidad de Admisión y Registro de la Universidad Técnica Estatal de Quevedo. Para determinar los factores de deserción más influyentes fue necesario la aplicación de la minería de datos, un proceso de análisis inteligente y modelos de árboles de decisión. **Abstract.-** The present project shows the methodology to obtain the factors that influence the student's desertion. This information that was taken from a data set present in the Admission and Leveling Unit of Quevedo State Technical University. To determine the most influential desertion factors it was required the application of data mining, a process of intelligent analysis and decision trees models.			
Descripción	FORMATO: A4 29cm x 21cm			
URI	(En blanco hasta cuando se dispongan los repositorios)			

INTRODUCCIÓN.

En el Ecuador el Consejo de Educación Superior del Ecuador (CES) estableció el 12 de octubre de 2010 la Ley Orgánica de Educación Superior (LOES). En el artículo 8 de esta misma ley se estipulan los objetivos de las universidades. Uno de estos objetivos es formar académicos y profesionales responsables, con conciencia ética y solidaria, capaces de contribuir con la sociedad que los rodea.

La Secretaría de Educación Superior, Ciencia, Tecnología e Innovación establece en el artículo 183 literal C de la LOES el deber de diseñar, implementar, administrar y coordinar el Sistema Nacional de Nivelación y Admisión (SNNA). El SNNA tiene como objetivo principal garantizar el acceso a la educación superior gratuita basado en igualdad de oportunidades, meritocracia y transparencia. Por sus características este sistema sirve de apoyo a los bachilleres que no cuentan con las bases de conocimiento necesarias para iniciar una carrera universitaria exitosa.

La Universidad Técnica Estatal de Quevedo en cumplimiento con la Ley Orgánica de Educación Superior ha implementado los cursos del SNNA. Estos cursos son actualmente dirigidos por la Unidad de Admisión y Registro. Con ellos se brinda a los estudiantes la oportunidad de nivelar sus conocimientos para empezar favorablemente una carrera universitaria. Sin embargo se debe indicar que muchos estudiantes que inician el curso de nivelación no lo logran culminar. Este problema produce pérdida de recursos para la Universidad y para el estado ecuatoriano.

La presente investigación se centra en analizar las características que corresponden a la realidad de los estudiantes en los cursos de nivelación. Para esto se realizó la toma de datos de una encuesta previamente realizada que aborda las realidades socio-económicas, psicológicas y médicas de los estudiantes. Esta información fue utilizada para identificar los factores más influyentes en la deserción estudiantil. Para determinar estos factores se aplicaron técnicas de minería de datos y modelos basados en árboles de decisión.

En esta investigación se ha logrado la obtención de la información sobre los factores que más influyen en producir la deserción estudiantil. Se espera que sirva a los altos directivos y encargados de la Unidad de Admisión y Registro como una herramienta más para la toma de decisiones. Se desea que con la información otorgada se logre disminuir la deserción presente en los cursos de nivelación.

CAPÍTULO I

CONTEXTUALIZACIÓN DE LA INVESTIGACIÓN

1.1.Problematización.

En este apartado se establece el problema que se pretende solucionar, las características del problema, el pronóstico esperado en caso de que no se llegara a corregir, la formulación y problematización.

1.1.1. Planteamiento del problema.

La Secretaría de Educación Superior, Ciencia, Tecnología e Innovación en la Ley Orgánica de Educación Superior ha establecido un curso de nivelación de conocimientos. Dicho curso es para los estudiantes que desean empezar una carrera de pregrado en cualquiera de las universidades del país, entre ellas la Universidad Técnica Estatal de Quevedo. La UTEQ ha designado a la Unidad de Admisión y Registro para la ejecución y control del curso de nivelación.

Sin embargo, a pesar del esfuerzo presentado por los docentes y autoridades de la Unidad de Admisión y Registro, muchos de los estudiantes que empiezan el curso no logran culminarlo. Esto genera una pérdida de recursos por el uso de las instalaciones, docentes y materiales académicos. Se debe destacar que en el caso de los estudiantes desertores, estos recursos terminan siendo invertidos sin obtener resultados favorables para la universidad.

1.1.1.1. Diagnóstico.

La deserción estudiantil es un problema grave que produce pérdidas de los recursos de la UTEQ. Esto es causado por las inversiones realizadas en edificaciones, en personal administrativo, en personal académico, material educativos y tiempo de planificación. Los fondos son invertidos para tratar de brindar la mejor educación para nivelar los conocimientos de los estudiantes.

1.1.1.2. Pronóstico.

La falta de datos referentes a la realidad de los estudiantes del curso de nivelación provocará perjurios en la comisión de la Unidad de Admisión y Registro en la UTEQ. Uno de ellos es que la escasez de información ocasiona que la directiva por no pueda tomar las medidas correctivas necesarias para poder contrarrestar el problema de deserción.

1.1.2. Formulación del problema.

¿Cuáles son los factores inciden en la deserción estudiantil en el curso de nivelación de la Unidad de Admisión y Registro de la UTEQ?

1.1.3. Sistematización.

- ¿Qué características presenta la información existente de los cursos de nivelación de la Unidad de Admisión y Registro?
- ¿De qué forma se pueden preparar los repositorios de datos de tal manera que se encuentren listos para la obtención de modelos?
- ¿Qué técnicas o herramientas de árboles de decisión será la más adecuada para producir un modelo que describa la naturaleza de las variables y la relación presente entre ellas?

1.2. Objetivos.

En el presente apartado se establecen los objetivos generales y específicos que se plantean cumplir con la realización del proyecto.

1.2.1. Objetivo General.

- Desarrollar un análisis inteligente que permita identificar los factores que influyen en la deserción de los estudiantes de la Unidad de Nivelación de la Universidad Técnica Estatal de Quevedo.

1.2.2. Objetivos Específicos.

- Analizar la información proporcionada por la Unidad de Admisión y Registro y construir los conjuntos de datos necesarios en base a sus características para el proceso de obtención de conocimiento.
- Aplicar las herramientas y técnicas necesarias para realizar el pre-procesamiento de los datos para preparar los conjuntos de datos para la aplicación de modelos.
- Aplicar varios modelos basados en árboles de decisión y determinar el más adecuado para la obtención de conocimientos referentes a los factores de deserción de los estudiantes del curso de nivelación.

1.3.Justificación.

La Universidad Técnica Estatal de Quevedo, institución que en cumplimiento de la Ley Orgánica de Educación Superior implemento los cursos SNNA. Estos cursos son dirigidos por la Unidad de Admisión y Registro de la universidad. Con los cursos de nivelación se busca ayudar a los estudiantes otorgándoles las instalaciones, los docentes y los materiales académicos necesarios. Todos estos beneficios se brindan para que ellos puedan alcanzar los conocimientos mínimos que son fundamentales para empezar adecuadamente una carrera universitaria de éxito.

Los cursos de la Unidad de Admisión y Registro presentan una problemática que es la deserción estudiantil en el transcurso del primer semestre del 2016. Realidad que se busca evitar que se produzca nuevamente en el año 2017. Sin embargo la unidad no cuenta con las bases de conocimientos necesarias para una adecuada toma de decisiones que consigan contrarrestar esta realidad.

Precisamente por las características del problema anteriormente mencionado, resulta necesario obtener un repositorio de datos para aplicar técnicas de minería de datos. Con dicho repositorio en una etapa conocida como el pre-procesamiento de datos se pueden llegar a corregir o mermar las características negativas que pueden presentare. Este proceso prepara los datos para realizar un análisis más exhaustivo con los algoritmos estadísticos y de minería de datos. Son estos algoritmos los que permitirán encontrar los factores que influyen en la deserción estudiantil.

Es primordial establecer que ya se han realizado investigaciones en los sistemas educativos de otras entidades universitarias que son similares a la presente. Estos estudios fueron en su mayoría realizados en unidades educativas. En dichos estudios se demuestra la importancia de la minería de datos y del análisis inteligente dentro de los sistemas educativos.

CAPÍTULO II
FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

2.1.Marco Conceptual.

A continuación se procede a establecer los conceptos y definiciones. Esta información es necesaria para poder comprender adecuadamente los procesos y actividades realizadas en el presente proyecto de investigación.

2.1.1.Minería de datos.

La minería de datos es un proceso para la extracción de conocimientos. En [1] se establece que con ella se pueden descubrir patrones, relaciones y tendencias al examinar los datos. Esta información suele ser bastante importante en muchos estudios.

En otra descripción por el mismo autor en [1], establece que es la disponibilidad de grandes volúmenes de información y el uso común de herramientas informáticas es lo que ha transformado el análisis de datos en los últimos tiempos. Estas técnicas informáticas especializadas son aquellas que se engloban bajo el nombre de minería de datos.

Por otro lado en [2] hablan del objetivo fundamental de la minería de datos. El objetivo es aprovechar el valor de la información localizada y usar patrones preestablecidos. Con esto se pretende lograr que los directivos tengan una mejor fuente de conocimiento sobre su negocio y puedan tomar decisiones más confiables.

En [3] establecen que muchas personas llaman a la minería de datos como el proceso de extracción de conocimientos. Por otro lado otros consideran a la minería de datos como una parte del proceso de extracción de conocimientos.

2.1.2.Fases del proceso de minería de datos

El proceso para aplicar una correcta minería de datos y que posibilite la adecuada extracción de conocimientos según el autor [3] es el siguiente:

- Limpieza de datos: Es la eliminación de ruido y datos incoherentes[3].
- Integración de datos: Es el proceso para combinar múltiples fuentes de datos [3].
- Selección de datos: Es la selección de los datos más importantes para el análisis [3].
- Transformación de datos: Es la transformación de los datos en formas apropiadas para realizar la minería de datos [3].
- Minería de datos: Es el aplicar métodos inteligentes para extraer patrones de datos [3].
- Evaluación de patrones: Es identificar los patrones que representan mayores conocimientos en base a la investigación realizada [3].
- Presentación de conocimiento: Proceso donde se aplican diversas técnicas para la visualización y presentar el conocimiento extraído para el usuario [3].

Por otro lado en [4] se establece una serie de pasos diferentes para la extracción de conocimientos. A continuación se enuncia cada uno de ellos con su respectiva descripción:

- **Determinación de los objetivos:** Previo al proceso de la extracción de conocimientos propiamente dicho, se necesita establecer los objetivos que se desean cumplir por parte del usuario. Esto es fundamental debido a que permite enfocar los futuros procesos, técnicas y herramientas que se utilizaran para el cumplimiento de los objetivos. Un error en este etapa puede llegar a producir la invalidez de todo el proceso y los resultados obtenidos [4].

- **Preparación de datos:** En la mayoría de los casos, los datos originales pueden presentar ruidos, outliers, ambigüedades, formato inadecuado, etc. Por lo que es necesario una adecuada preparación de los datos para la aplicación de algoritmos. Una adecuada preparación de los datos mejorará la calidad del modelo de conocimiento obtenido. Esta fase se divide en otras tres; selección, pre-procesado y transformación [4].

- **Minería de datos:** La selección de la herramienta y el método a utilizar en el proceso de minería de datos, es de suma importancia en el proceso de extracción de conocimientos y la generación del modelo. En esta etapa se suele incluir la validación del modelo, para que en caso de ser necesario se realice una reorganización de los datos y reajustar el propio algoritmo [4].

- **Análisis:** Es la etapa que se encarga de la interpretación de los datos, la evaluación y el estudio del modelo de conocimiento obtenido por el algoritmo aplicado en la minería de datos. El uso de estas técnicas facilitan al usuario la interpretación del modelo y brinda mayor seguridad cuando estos modelos se aplican en la toma de decisiones [4].

- **Aplicación:** En esta fase se lleva el conocimiento adquirido a aplicarse al campo real, en el cual los conocimientos son usados como un soporte de seguridad en la toma de decisiones [4].

2.1.3. Weka.

En [7] se define a Weka como una serie de librerías JAVA dedicadas a la extracción de conocimientos de las bases de datos. Weka usa un formato específico de datos llamado urff. Los ficheros con dicho formato además de almacenar la información de los datos, almacena meta información de los propios datos que se van a tratar. Esta meta información incluye el nombre y tipo de cada atributo, así como una descripción del origen de los datos.

Por otro lado en [8] se define a Weka como una colección de métodos y algoritmos de estado del trabajo de la máquina de aprendizaje y los datos de las herramientas de pre-procesamiento. Weka se encuentra desarrollado para que se puedan probar los métodos existentes diversos conjuntos de datos de manera sencilla y flexible. Adicionalmente ofrece un amplio soporte para todo lo que corresponde al proceso de la minería de datos experimentales, la preparación de los datos de entrada y el resultado del aprendizaje.

Weka además proporciona diversas implementaciones de los algoritmos de aprendizaje que se pueden aplicar a los conjuntos de datos. Otra de sus características es que incluye una serie de herramientas para la transformación de datos. Weka tiene la capacidad de tratar un conjunto de datos, realizar actividades de aprendizaje, analizar los resultados de la aplicación del clasificador y su rendimiento. Es destacable mencionar que puede realizar todo lo anteriormente dicho sin necesidad de escribir código de programación [8].

2.1.4.Discretización de variables.

La discretización según [5] es uno de los principales y más importantes métodos de pre-procesamiento. Este método suele ser a necesario en mucho de los casos para transformar los datos que se encuentran en formatos incompatibles y reducir la complejidad de los datos para algunos procesos de análisis. El objetivo es reducir el número de valores posibles de cada atributo continuo, agrupándolos en una serie de intervalos.

2.1.5.Árbol de decisión.

Según [6], cuando hablamos de un árbol de decisión en los procesos de minería de datos, estamos hablando de un modelo predictivo. Este modelo se puede utilizar para la representación tanto de los clasificadores como de los modelos de regresión. Los árboles de decisión que son usados para clasificar tareas, se le suele llamar como un árbol de clasificación. Los árboles que son usados en actividades y tareas de regresión, se les llama árboles de regresión.

El autor de [6] establece que los árboles de clasificación tienden a ser usados en áreas como la ingeniería, el marketing, las finanzas y la medicina. Estos árboles son útiles al ser aplicados como técnicas de exploración. Se debe destacar que los árboles de clasificación no buscan reemplazar los métodos estadísticos tradicionales y existen muchas técnicas que suelen usarse para clasificar, como las máquinas de vectores y las redes neuronales.

2.1.6.Algoritmo SMOTE.

El SMOTE cuyas siglas significan Syntetic Minority Over-sampling Technique[1] es un algoritmo de sobre-muestreo que genera una serie de instancias artificiales. Esta técnica tiene como objetivo equilibrar la muestra de datos basados en la regla del vecino más cercano. La generación se realiza extrapolando nuevas instancias en lugar de duplicarlas [9].

[1] Técnica de sobre-muestreo minoritario sintético

A continuación se muestran algunas características del algoritmo SMOTE en base a lo descrito por el autor [9]:

- El algoritmo produce ejemplos sintéticos [9].
- Opera en el espacio de atributos [9].
- Produce un ejemplo sintético a lo largo de los segmentos de línea que unen alguno o todos los k vecinos más cercanos de la clase minoritaria [9].
- Se eligen algunos de los k vecinos más cercanos de manera aleatoria [9].

El autor [9] establece que algoritmo de SMOTE para cumplir con su objetivo y dar un sobre-muestreo estable y sólido realiza los siguientes pasos:

- Recibe como parámetro el porcentaje de ejemplos a sobre-muestrear y con ello calcula el número de ejemplos que tiene que generar [9].
- Calcula los k vecinos más cercanos de los ejemplos de la clase minoritaria[9].
- Genera los ejemplos siguiendo este proceso:
    - Para cada ejemplo de la clase minoritaria, elige aleatoriamente el vecino a utilizar para crear el nuevo ejemplo[9].
    - Para cada atributo del ejemplo a sobre-muestrear, calcula la diferencia entre el vector de atributos muestra y el vecino elegido[9].
    - Multiplica esta diferencia por un número aleatorio entre 0 y 1 [9].
    - Suma este último valor al valor original de la muestra[9].
    - Devuelve el conjunto de ejemplos sintéticos[9].

2.1.7.Algoritmo Decision Stump.

El decision stump es un árbol de decisión caracterizado por utilizar solamente un atributo para dividir los atributos discretos, esto quiere decir que el árbol consiste solo en un único nodo interior. Por otra parte si el atributo es numérico, puede ocasionar que el árbol generado se vuelva mucho más complejo[10].

2.1.8.Algoritmo Hoeffding Tree.

Los árboles Hoeffding son una técnica para el procesamiento de flujos de datos de alta velocidad. Su ingenio se debe a suficientes actualizaciones estadísticas. El crecimiento del árbol se basa en tomas de decisiones que están garantizadas para ser casi idénticas a los que se hace por métodos de aprendizaje por lotes convencionales. A pesar de esta garantía, las decisiones siguen estando sujetos a cuestiones de búsqueda hacia delante y una estabilidad limitada [11].

La inducción de un árbol de decisión de flujo se llama Hoeffding Tree. Su nombre proviene del límite Hoeffding, el cual es utilizado en la inducción del árbol. La idea principal es que le limite Hoeffding da cierto nivel de confianza en el mejor atributo para dividir el árbol, por lo tanto, se puede construir el modelo basado en casos vistos previamente [12].

Un árbol Hoeffding es un algoritmo incremental de inducción de un árbol de decisión que capaz de aprender de los flujos de datos masivos, suponiendo que los ejemplos de generación de distribución no cambian con el tiempo. Los árboles Hoeffding explotan el hecho de que una pequeña muestra a menudo puede ser suficiente para elegir un atributo de división óptimo [13].

Esta idea se apoya matemáticamente por la unidad Hoeffding, que cuantifica el número de observaciones que se necesita para estimar algunas estadísticas dentro de una precisión prescrita [13].

2.1.9.Algoritmo J48.

El algoritmo J48 es una versión del árbol de decisión C4.5 desarrollado para el software Weka. El algoritmo trabaja con el uso de una técnica codiciosa que se encarga de la inducción de los árboles de decisión utilizados para la clasificación. Adicionalmente el algoritmo utiliza reducción de errores podando aquellos nodos que son erróneos [14].

2.1.10. Algoritmo LMT.

Es importante establecer que un árbol de modelo logístico (LMT) consiste básicamente en una estructura de árbol de decisión estándar con funciones de regresión logística en las hojas. Por otro lado como árbol de modelo es un árbol de regresión con funciones de regresión en sus hojas[15].

Adicionalmente el autor establece que como los árboles de decisión ordinarios, una prueba en uno de los atributos se asocia con cada nodo interno. Para un atributo nominal con valores de k, el nodo tiene nodos hijo de k, y las instancias se ordenan hacia abajo de una de las ramas k dependiendo su valor del atributo [15].

Para atributos numéricos, el nodo tiene dos nodos del hijo y la prueba consiste en comparar el valor el atributo a un umbra. Una instancia es clasificada por rama de la izquierda si su valor para ese atributo es más pequeño que el umbral y ordenada por la rama derecha de lo contrario[15].

2.1.11. Algoritmo Random Forest.

El algoritmo Random Forests es uno de los mejores entre los algoritmos de clasificación -. Capaz de clasificar grandes cantidades de datos con exactitud al azar[16].

Los bosques son un método de aprendizaje conjunto para la clasificación y la regresión que construyen una serie de árboles de decisión en el tiempo de entrenamiento y la salida de la clase que es el modo de la salida de clases de árboles individuales [16].

Un Random Forest desarrolla un conjunto de árboles de decisión. Los bosques al azar se utilizan a menudo cuando tenemos conjuntos de datos de formación muy grandes y un gran número de variables de entrada [17].

Un modelo de bosque aleatorio se compone típicamente de decenas o cientos de árboles de decisión, cada uno construido utilizando una muestra aleatoria del conjunto de datos, y mientras se construye un árbol, se considera una muestra aleatoria de las variables en cada nodo[17].

El muestreo aleatorio de los datos y las variables asegura que incluso la construcción de 500 árboles de decisión puede ser eficiente. También ofrece una considerable robustez al ruido, valores atípicos y ajuste excesivo, en comparación con un solo clasificador de árboles[17].

2.2.Marco referencial.

A continuación se describirán una serie de investigaciones similares a la presente investigación, lo que permite confirmar además, que el análisis inteligente de datos y la minería de datos han sido utilizados en diversas ocasiones a favor de la educación.

2.2.1.Minería de datos para descubrir estilos de aprendizaje.

Este estudio fue realizado para poder identificar el estilo de aprendizaje predominante que presentaban los estudiantes de un curso secundario, partiendo desde el punto multifacético que puede tener un estudiante en su aprendizaje.

Utilizando el proceso de KDD, les permitió conocer las características del estilo de aprendizaje compartido por la mayoría de los alumnos. Teniendo en cuenta la información descubierta se sugieren estrategias de intervención didáctica [18].

2.2.2.Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos.

En esta investigación se aplican las técnicas de minería de datos para adquirir conocimiento, en este caso se buscaban conocimiento referentes a la deserción escolar y los factores que los provocaban. Para esta investigación se utilizaron algoritmos de clasificación predictivos específicamente dos tipos de árboles: el algoritmo de árboles de clasificación C4.5 y el algoritmo de los k vecinos más cercanos, obteniendo modelos resultantes para establecer sus bases de conocimiento [19].

2.2.3. Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos.

Esta investigación se centra en la reprobación estudiantil de nivel superior, estableciendo que es una crisis grave que se presenta en la sociedad, indicando esto porque teniendo en cuenta en que la eficiencia de estudiantes que culminan la carrera en la educación superior oscila entre el 53% y el 63% en México es un problema que urge solucionar [20].

En este trabajo se llevó a cabo el análisis de los datos para generar un modelo que ayude a predecir las causas que llevarán a los estudiantes a reprobar, así como las materias con mayor riesgo y las características específicas de cada estudiante [20].

2.2.4. Aplicación de técnicas de minería de datos para la evaluación de la deserción estudiantil.

Este artículo presenta los resultados de la evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas. La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008 [21].

En [21] indica que la implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

2.2.5.Descubrimiento de perfiles de deserción estudiantil.

En este artículo se presentan los resultados de un proyecto de investigación [22], cuyo objetivo fue detectar patrones de deserción estudiantil utilizando técnicas de minería de datos. Para el análisis fueron tomados como referente datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado.

Se construyó un repositorio de datos con la información de los estudiantes que ingresaron a la Universidad entre el primer semestre de 2004 y el segundo semestre de 2006, con una ventana de observación hasta el 2011. Utilizando técnicas de clasificación y clustering, se descubrieron perfiles socioeconómicos y académicos de los estudiantes [22].

CAPÍTULO III
MÉTODOLOGÍA DE LA INVESTIGACIÓN

3.1.Localización.

La Universidad Técnica Estatal de Quevedo es el lugar en el que se desarrolló la presente investigación. Se encuentra ubicada en la Avenida Quito km. 1 1/2 vía a Santo Domingo de los Tsáchilas. En la calle Transversal Central entre la Avenida Carlos J. Arosemena y la calle Patria Nueva, junto a la Unidad Educativa Quevedo y la cancha sintética y bar Wembley.

Se encuentra en la parroquia 24 de Mayo del Cantón Quevedo en la provincia de los Ríos en la República del Ecuador. Las coordenadas geográficas de su ubicación son -1º 0' 45" en latitud y a -79º 28' 10" en longitud.

3.2.Tipo de investigación.

La presente investigación es de naturaleza descriptiva, este tipo de investigaciones se centran en el análisis de múltiples variables para lograr determinar las características de una situación específica. El proceso consiste en establecer las propiedades de la situación o fenómeno a investigar. Por lo cual si contrastamos con esta investigación y la necesidad de identificar los factores que influyen en la deserción estudiantil en la Unidad de Admisión y Registro, se puede indicar que efectivamente corresponde a las investigaciones descriptivas

3.2.1.Investigación descriptiva.

Como se estableció anteriormente, una investigación descriptiva se centra principalmente en el análisis en general, sin tomar en cuenta la profundidad del análisis que se vaya a realizar. Proceso que se realiza identificando las características de las variables y sus relaciones, lo que permite describir una situación en específica y lo que la produce.

La presente investigación es de tipo descriptiva, lo cual podemos atribuir a las razones que se enunciaran a continuación:

- La investigación realizada se centra en el investigar la deserción estudiantil en la Unidad de Admisión y Registro. Esto es una situación específica y que a través de un análisis especializado se pueden determinar las variables que influyen en dicha situación.

- Otra razón se puede establecer a la interpretación de los modelos de árboles de decisión aplicados. Estos modelos permiten tener un plano detallado de las variables y las relaciones que se presentan entre ellas.

3.3.Métodos y técnicas a usar en la investigación.

En el presente proyecto se utilizaron varios métodos y técnicas de investigación. Estas fueron las Técnicas de Extracción de Conocimientos de las Bases de Datos (KDD)[2] y el método de la observación. Ambos métodos permitieron el adecuado desarrollo de la investigación y la obtención de los resultados necesarios para cumplir con los objetivos del proyecto.

3.3.1.Método observación.

La metodología de la observación permite obtener la información de una situación o medio específico en el estado natural en el que se encuentra. Uno de los requisitos al utilizar esta metodología es tener presente el objetivo que se quiere conseguir al aplicar este método. Esto permitirá enfocarse principalmente en aquello que se investiga y se quiere llegar a conocer.

La investigación en sus primeras etapas utilizo el método de observación. Aplicado para adquirir una primera impresión del repositorio de datos. Este método permitió conocer la naturaleza del repositorio, las variables que se encontraban almacenadas, los tipos de datos de cada variable, particularidades generales que se presentaban y el punto de partida para los procesos futuros que se le aplicarían a la base de datos.

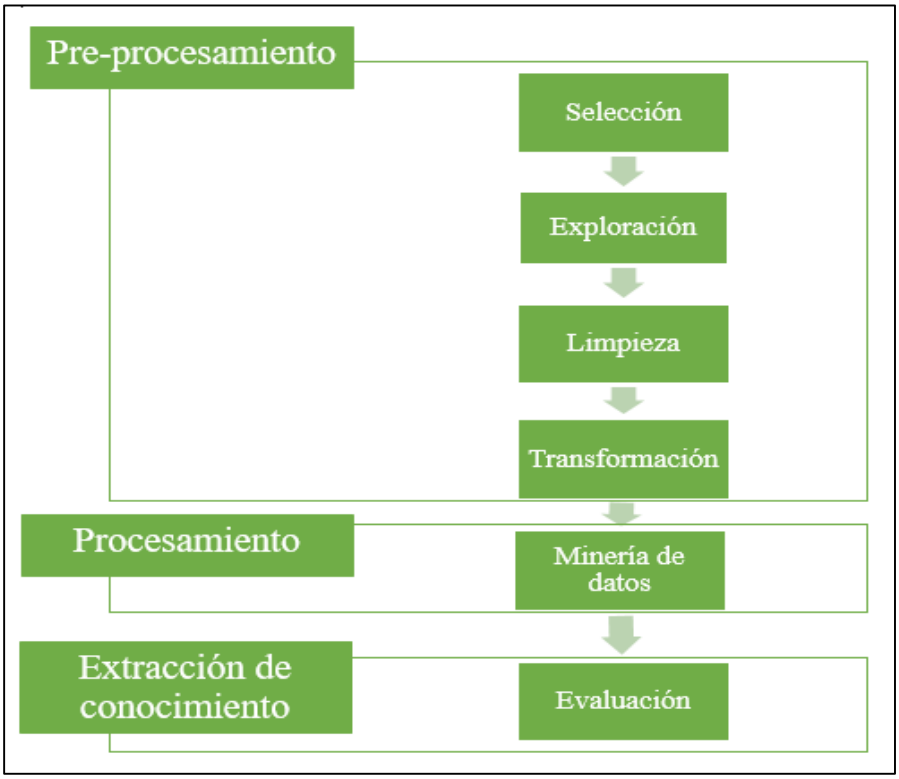
[2] Knowledge Discovery in Databases

3.3.2. Técnicas de extracción de conocimiento de bases de datos.

Cuando hablamos de la metodología de la extracción de conocimientos, el autor Pérez López Cesar en [26], indica que el proceso de extracción de conocimiento es una secuencia de 7 fases. Estas son: selección, exploración, limpieza, transformación, minería de datos, evaluación y por último difusión. Adicionalmente estable que los cuatro primeros corresponden al pre-procesamiento de datos.

Sin embargo a pesar de estar de acuerdo con la mayor parte de lo descrito por el autor previamente enunciado, es necesario indicar que no se está de acuerdo con la fase de difusión como extracción de conocimiento. La razón se debe principalmente a que el conocimiento se adquiere en la etapa correspondiente a la evaluación. Por otra parte la de difusión solo se encarga de impartir dicho conocimiento. Es por esto que en la presente investigación la fase de difusión no se tomará en cuenta, dejando así 6 fases que son las siguientes:

Gráfico 1: Proceso para la extracción de conocimientos



FUENTE: CONOCIMIENTOS PROPIOS.

ELABORADO POR: EL AUTOR.

Para una mejor comprensión de cada una de las fases de la extracción de conocimiento, se procederá a brindar una breve explicación de cada una de ellas:

3.3.2.1. Selección.

En la fase de selección se procede a comenzar estableciendo los objetivos que se quieren obtener de la extracción de conocimiento. Posteriormente se procede a recopilar los datos que permitirán cumplir con los objetivos planteados. Además se identifican y seleccionan en variables que primera instancia son relevantes en los datos.

3.3.2.2. Exploración.

La segunda fase que corresponde a la exploración, es precisamente para realizar un análisis exploratorio de datos. En esta fase también se puede identificar la distribución de los datos. Además se tiende a analizar las correlaciones existentes en la información.

3.3.2.3. Limpieza.

La limpieza de los datos es la tercera fase, esta se centra en detectar outliers[3] y tratarlos en la medida de lo posible. De la misma manera se centra en los procesos de imputación de la información faltante y la eliminación de datos erróneos e irrelevantes que pueden presentarse en los repositorios de datos.

3.3.2.4. Transformación.

Seguido tenemos la cuarta fase llamada transformación, esta corresponde al uso de técnicas para la discretización y numeración de los datos cuando sean necesarios. En el caso de que se requiera, en esta fase se puede realizar procesos de escalado, de aumento y reducción de dimensiones de los datos.

[3] Outlier = Valores atípicos que se encuentra distante de los demás datos.

3.3.2.5. Minería de datos.

La minería de datos es la quinta fase, en ella se elige si se utilizara una técnica descriptiva o predictiva. Posteriormente se escoge cuál de las técnicas se utilizaran para obtener un modelo. Es importante destacar que cada una de las técnicas y herramientas de la minería de datos presentan requerimientos y especificaciones que se deben cumplir para un correcto análisis de los datos.

3.3.2.6. Evaluación.

En la sexta y última fase, la de evaluación se procede a evaluar los patrones obtenidos por los modelos. En ella se identifica el comportamiento de las variables. Se debe destacar que en este punto se puede regresar a fases anteriores para obtener un nuevo modelo. Esto se realiza cuando se desean realizar comparaciones, en este punto ya se ha adquirido un nuevo conocimiento

3.4.Fuentes de recopilación de información.

Para obtener la información necesaria para poder cumplir con el objetivo, se planteó inicialmente dos posibilidades. La primera posibilidad era encontrar un repositorio existente de datos que cumpla con los requerimientos necesarios para la extracción de conocimiento o en su defecto realizar una encuesta para poder recopilar la información.

Favorablemente en la Unidad de Admisión y Registro de la Universidad Técnica Estatal de Quevedo constaba con un repositorio de datos bastante amplio. Dicho repositorio fue facilitado bajo la autorización del Director de la Unidad, el Lic. Harold Escobar. Es primordial destacar que se estableció un acuerdo hablado sobre la importancia de la confidencialidad, seguridad y la integridad de los datos personales de los estudiantes.

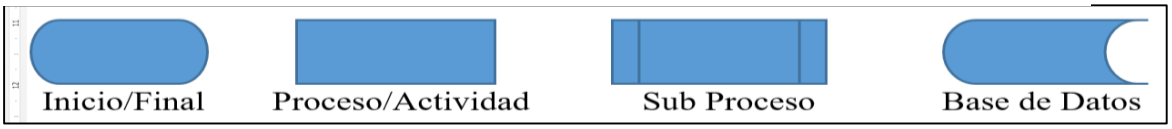
3.5.Diseño de la investigación.

El diseño de la investigación es el cuasi-experimental debido a que se encarga del análisis de entidades y variables, sin aplicar procesos de experimentación. Esto se debe a que no se manipulan las variables que inciden en ella. Sin embargo, en caso de ser necesario dicha se puede experimentar el comportamiento de la información en circunstancias específicas.

3.6.Diagrama de Flujo acerca de la investigación.

Para explicar la forma en la que se desarrollará la investigación y se obtendrán resultados se ha establecido un diagrama de flujo. En el siguiente gráfico se detallan las figuras utilizadas en el diagrama y el significado de cada una de ellas.

Gráfico 2: Figuras del diagrama de flujo utilizadas

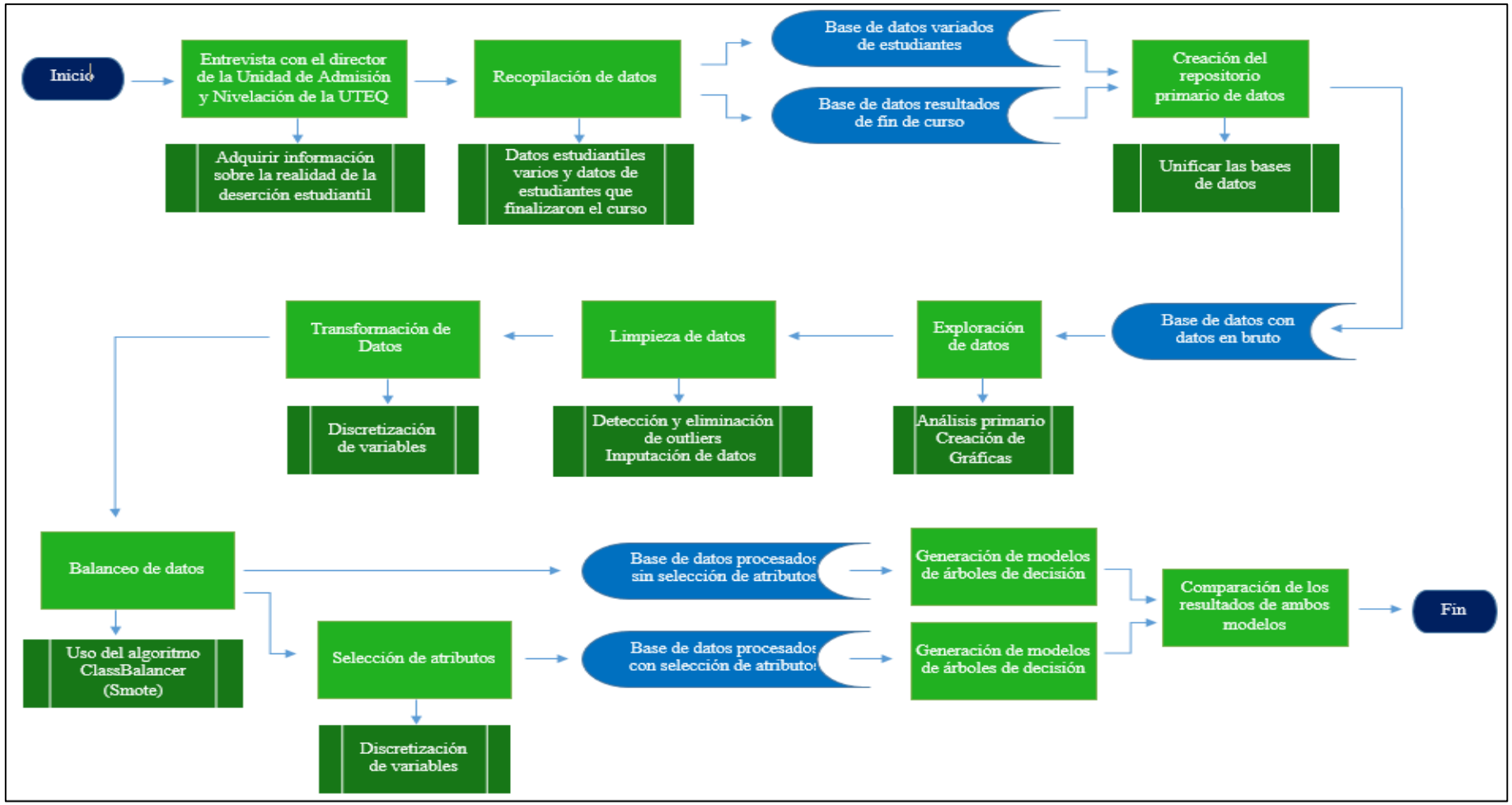


FUENTE: CONOCIMIENTOS PROPIOS.
ELABORADO POR: EL AUTOR.

El diagrama de flujo establece los procesos realizados durante el presente proyecto. En él se pueden observar las actividades realizadas desde el inicio del presente proyecto hasta su finalización. En el diagrama se observan las fases de la minería así como la de recopilación de la información y otros procesos fundamentales. Es por esto que es una herramienta fundamental para comprender lo que se ha hecho en el presente trabajo.

A continuación el siguiente gráfico corresponde al diagrama de flujo del presente proyecto:

Gráfico 3: Diagrama de flujo de la investigación



FUENTE: CONOCIMIENTOS PROPIOS.

ELABORADO POR: EL AUTOR.

3.7.Instrumentos de investigación.

El instrumento de investigación utilizado, sin tomar en cuenta aquellos que vienen sobre entendidos en las metodologías descritas es el siguiente.

3.7.1.La entrevista.

La entrevista es una comunicación interpersonal entre la persona que se encuentra bajo estudio (entrevistado) y el investigador (entrevistador). Esta técnica permite responder a las dudas y cuestionamientos que se establecen sobre un determinado problema, obteniendo respuestas de manera hablada.

La entrevista permitió establecer un primer contacto con los encargados de la Unidad de Admisión y Registro. Además se logró obtener la percepción de las autoridades con respecto a la deserción estudiantil. Por otro lado se lograron responder dudas sobre las actividades que se proponían realizar para poder contribuir a la solución de dicho problema.

3.8.Tratamiento de datos.

Para el tratamiento de datos se utilizaron principalmente 2 softwares, en primer lugar se utilizó el software de Excel para la observación inicial y la preparación inicial de datos. Por otro lado se utilizó Weka para el tratamiento y una limpieza más profunda de los datos. En esta fase se trataron los datos e incluso se realizó el modelado, siendo de vital importancia para el desarrollo del proyecto.

3.9.Recursos humanos y materiales.

Para que el presente proyecto pudiera llevarse a cabo fue necesario que se invirtieran ciertos recursos que sirvieron como base y sustento en diferentes fases de la investigación, estos recursos fueron materiales y humanos, a continuación se detallarán cada uno de ellos.

3.9.1. Equipo Humano.

El presente proyecto de investigación ha sido desarrollado por el Sr. William Daniel Burbano Ferrin, el cual ha contado con la ayuda, respaldo y dirección del PhD Amilkar Puris Cáceres. Ellos trabajaron en la recopilación de la información, tratamiento de datos, desarrollo de modelos, análisis comparativos y la documentación.

3.9.2. Equipos y Materiales.

El desarrollo del proyecto requirió el uso de una serie de equipos y dispositivos informáticos, es por ellos que se especificara por separado el uso el hardware y el software.

3.9.1.1. Hardware.

El hardware utilizado para la elaboración del presente proyecto junto con sus características se especifica en la tabla que se puede observar a continuación:

Tabla 1: Hardware utilizado para el desarrollo de la investigación

Hardware	
Equipo	**Características**
PC Portátil Asus	Procesador Intel Core i5-2450M CPU @2.5GHz 6 GB de RAM 750 Gb Disco Duro
Impresora	Multifunción Epson L555

FUENTE: CONOCIMIENTOS PROPIOS.
ELABORADO POR: EL AUTOR.

3.9.1.2. Software.

El software utilizado fue muy importante para la elaboración del proyecto, las características del software se describen en la tabla que se observa a continuación:

Tabla 2: Software utilizado para el desarrollo de la investigación

Software	
Equipo	**Características**
Sistema Operativo	Windows 10 Profesional 64bits
Microsoft Office	Professional 2013
Weka	Versión 3.8

FUENTE: CONOCIMIENTOS PROPIOS.

ELABORADO POR: EL AUTOR.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1.Entrevista.

La entrevista fue una herramienta aplicada para poder dar los primeros pasos de la investigación. Con esta herramienta se pudieron establecer 3 puntos importantes que dieron la apertura para realizar la presente investigación y son los siguientes:

4.1.1.El estado actual de la Unidad de Admisión y Registro.

Con la entrevista se consultó cuáles eran las principales dificultades y necesidades que se presentaban en la Unidad de Admisión y Registro. Esto permitió descubrir un problema relacionado la deserción que se estaba presentando en el curso de nivelación.

4.1.2.La existencia de un conjunto de datos de los estudiantes.

Durante la entrevista se cuestionó sobre la existencia de datos de los estudiantes, su naturaleza y el tiempo que había pasado desde que los datos fueron obtenidos. En respuesta las autoridades de la Unidad de Admisión y Registro establecieron que sí contaban con un repositorio de datos. Adicionalmente indicaron que el repositorio correspondía a una significativa cantidad de datos de los estudiantes del curso de nivelación del primer semestre del año 2016.

4.1.3.Acuerdo verbal de confidencialidad de la información.

Como último punto de la entrevista, el Lic. Harold Escobar director de la Unidad de Admisión y Registro procedió a informar la sensibilidad que tenía en tratar con este tipo de datos. El director enfatizó sobre todo con el cuidado que se debe tener al trabajar con información personal de los estudiantes. Debido a esto se realizó un acuerdo de manera verbal en el cual se adquirió el compromiso de mantener en la medida de lo posible la confidencialidad de la información. Posterior a esto se recibieron los repositorios de datos necesarios para continuar con la investigación.

4.2. Herramienta para el tratamiento de datos y algoritmos de modelamiento seleccionados.

La herramienta para el tratamiento de datos que se ha seleccionado para la presente investigación es Weka. A continuación se detalla una descripción.

4.2.1. Software WEKA

Weka es un software especializado en la extracción de conocimientos de las bases de datos. Sus siglas significan Entorno de Waikato para el Análisis del Conocimiento[4]. Fue desarrollado en la Universidad de Waikato en Nueva Zelanda y se define como un conjunto de librerías JAVA. Dichas librerías se encuentran especializadas en los procesos de extracción de conocimientos provenientes de las bases de datos.

Fue desarrollada con licencia libre, esto permite que muchos usuarios desarrollen sus propios procesos y funciones adicionales. Weka trabaja con un formato propio llamado arff cuyas siglas corresponden a Formato de Atributos de Archivos y Relaciones[5]. La razón del uso de esta herramienta en la investigación es el almacenamiento y uso de los metadatos en los procesos de KDD almacenados en los archivos arff.

4.2.2. Árboles de decisión

Los árboles de decisión fueron seleccionados debido a que es necesario siempre seleccionar la mejor alternativa en la toma de decisiones para la solución de problemas. Precisamente por esto se busca aplicar un diagnóstico preciso y adecuado. Los árboles de decisión muestran una serie de ramificaciones que permiten observar el origen de los procesos y resultados. Esto permite establecer algunas alternativas ante la problemática que se pretende solucionar.

[4] Waikato Enviroment for Knowledge Analysis.
[5] Atribute Relation File Format.

4.3.Descripción y análisis preliminar de los datos.

Los datos obtenidos contienen información muy diversa de los estudiantes y conforman un conjunto de 106 variables (ver anexo 1). Estas variables almacenan información que refleja la realidad de los estudiantes. Algunas son el tiempo que le toma movilizarse hasta la universidad, su carrera, sus opiniones, su relación con familiares y padres de familia, etc. A continuación se puede observar en el siguiente gráfico una pequeña visualización de los datos almacenados.

Gráfico 4: Visualización de los datos obtenidos

	Domicilio	Teléfono Con	Teléfono Celu	Email	Si vives lejo d	¿Vives con al	Viajas de reg	¿Tienes ingre	¿Pagas alqui	¿De q
	AVENIDA JO	52771616	994089324	edandres07@	Cupo del SNI	Con ambos p	Vivo dentro d	No tengo ingr	No Vivo con f	Papá
	barrio Cemen	2657670	983133926	fer-nandop@h	Cupo del SNI	Con mis hern	Cada 15 dias	No tengo ingr	Si Pago el Al	Mamá
	SAN CRISTO	62709401	980540193	isaacdanita2(	Cupo del SNI	Con mi mama	Vivo dentro d	No tengo ingr	No Vivo con f	Papá
	Guayacan los	2780487	981745227	fatimacamacl	Quiero ser pr	Abuelos	30 minutos a	Entre 177 a 3	No Vivo con f	Mamá
	Parroquia San	2771886	967033721	heiibs1@hotr	Me interesa l	Con mi mama	Vivo dentro d	No tengo ingr	No Vivo con f	Herma
e	Ciudadela Tu	52785382	979732312	litardogenesis	Me interes la	Con ambos p	60 minutos a	No tengo ingr	No Vivo con f	Papá
	SAN LUIS KM	2900022	959923029	LADYMAHOL	Me interesa l	Con padre, m	30 minutos a	No tengo ingr	No Vivo con f	Papá
	Av. Arcos Pe	248862	980636495	minoska98@	No hay otra u	Con ambos p	30 minutos a	No tengo ingr	No Vivo con f	Papá
	Parroquia "Ni	0	1	luisomarm98(	Me interes la	Con mi mama	Vivo cerca de	No tengo ingr	No Vivo con f	Papá
in	av los colonos	3767602	985166860	jordanruales3	No hay otra u	Vivo solo	2 horas aprox	No tengo ingr	Si Pago el Al	Papá
	Isla de bejuca	52885854	988119644	quintoandrein	Quiero ser pr	Cónyuge o pa	2 horas aprox	No tengo ingr	No Vivo con f	Pareja
	Recinto San I	52900069	985595513	nigerchiqui@l	Otra	Tios	30 minutos a	No tengo ingr	No Vivo con f	Tios
	Esmeraldas	62765137	981469320	ivongonzalez(	Cupo del SNI	Sólo (a) en vi	Hasta termina	No tengo ingr	Si Pago el Al	Papá
	la venus	52750372	994166486	karlaroxanam	Cupo del SNI	Con mi mama	30 minutos a	No tengo ingr	No Vivo con f	Mamá
	Avenida walte	2784045	969257799	joelk98carran	Otra	Con ambos p	15 minutos	No tengo ingr	No Vivo con f	Papá
	Guayacan Vil	52785232	995550199	andreinitap_0	Quiero ser pr	Con mi mama	30 minutos a	No tengo ingr	No Vivo con f	Mamá
	calle dimas fr	52952124	981367160	anthonystiver	Me interesa l	Con ambos p	30 minutos a	No tengo ingr	No Vivo con f	Papá
	La Concordia	2726845	981347856	byronmauricio	No hay otra u	Vivo solo	Hasta termina	No tengo ingr	Si Pago el Al	Papá
	San Camilo A	2770424	968044856	mishel_mend	Me interes la	Con padre, m	Vivo dentro d	No tengo ingr	No Vivo con f	Papá
e	El Empalme	3889041	990091472	betsaidagarci	No hay otra u	Con mis hern	60 minutos a	No tengo ingr	Si Pago el Al	Pareja
	Parroquia El (	52786651	979228561	stevenalex_1!	Me interes la	Con padre, m	30 minutos a	No tengo ingr	No Vivo con f	Papá
e	EL EMPALM	3884091	939608088	yelinagarcia@	No hay otra u	Con mis hern	60 minutos a	No tengo ingr	Comparto el p	Papá
	Km 2 1/2 Via	780923	994272146	maidithap93@	Cupo del SNI	Con ambos p	30 minutos a	No tengo ingr	No Vivo con f	Papá
	Ciudadela Pro	2764070	985321090	hugo.h96@ho	Quiero ser pr	Con padre, m	30 minutos a	No tengo ingr	No Vivo con f	Papá
	Barrio Lindo	5234562	996423370	marco08fer@	No hay otra u	Abuelos	2 horas aprox	No tengo ingr	No Vivo con f	Mamá

FUENTE: CONJUNTO INICIAL DE DATOS.

ELABORADO POR: EL AUTOR.

El análisis preliminar realizado en los datos obtenidos se determinó que existe un total de 36 variables que por diversos motivos no contribuyen como fuentes de conocimiento a la investigación que se está realizando, por lo que se establece que estas variables no serán tomadas en cuenta a partir de este punto en ninguno de los procesos de análisis que se procedieran a realizar, para mayor información véase anexo 2.

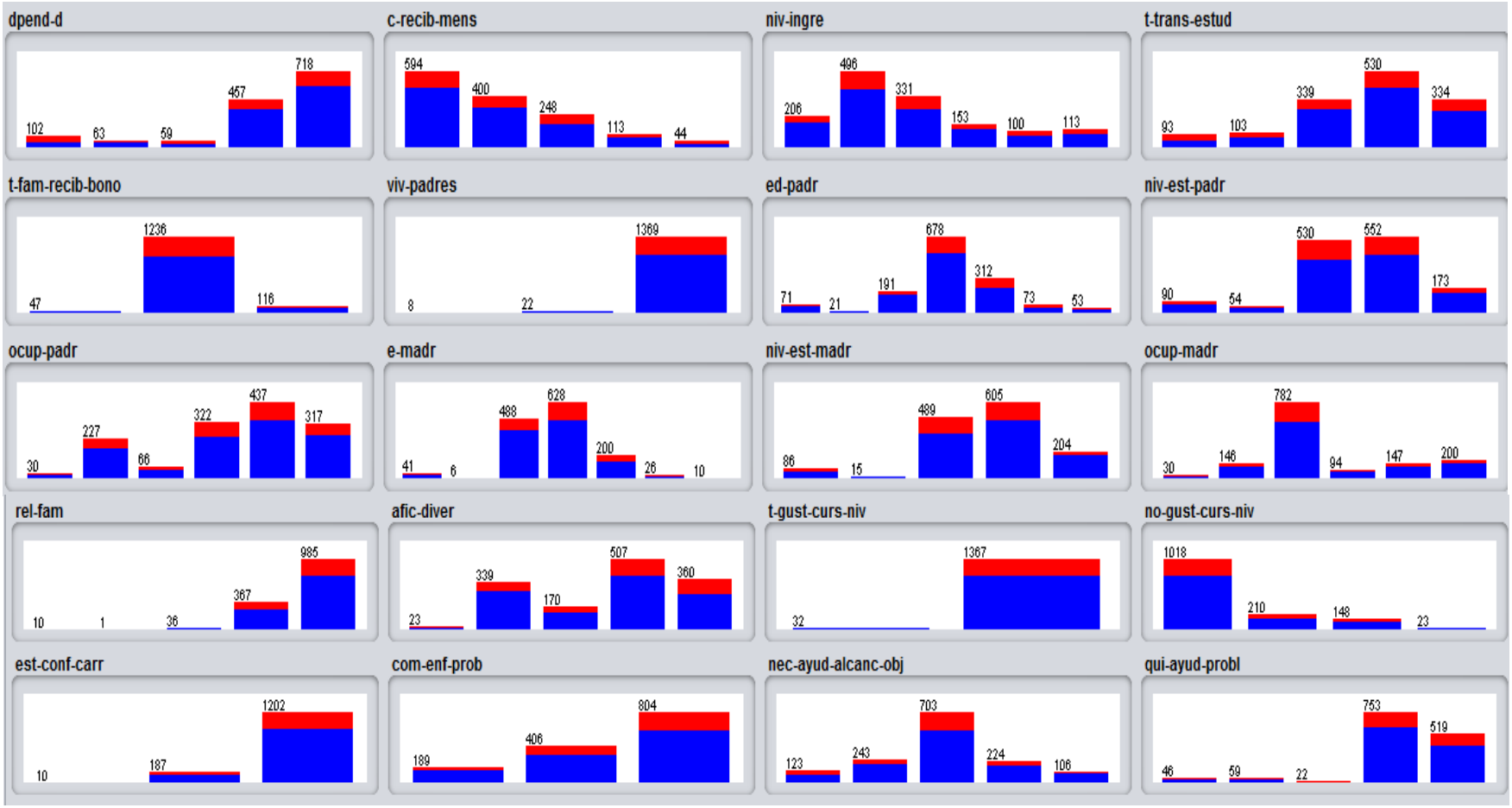
Por otra parte en el análisis a las variables restantes se puede observar y detallar que en base a su naturaleza, su contexto y sus características se requería separar el repositorio original en 3 conjuntos de datos diferentes, uno en base a los datos socio-económicos, otro para la información de índole psicológica y por último uno de datos médicos, a continuación se procede a dar una breve descripción de cada uno de ellos y de sus variables.

4.3.1.Datos socio-económicos BC1.

En estos datos se almacena todo lo referente a la realidad social y económica de los estudiantes, se presentan datos que indican cuantas personas viven en su casa, si viven con sus padres, sus ingresos económicos, su relación con sus compañeros, etc. Todo este conjunto de variables corresponden a un total de 36 y que se consideran de naturaleza socio-económica se le llamará a partir de ahora BC1, las variables se pueden ver en el anexo 3.

Desde el software Weka, que es una herramienta para tratamiento de datos, se puede observar una visualización preliminar que muestra la distribución de los datos en cada una de las variables, esto se puede observar en el siguiente gráfico:

Gráfico 5: Distribución de datos de algunas de las variables de BC1



FUENTE: WEKA WORKBENCH.

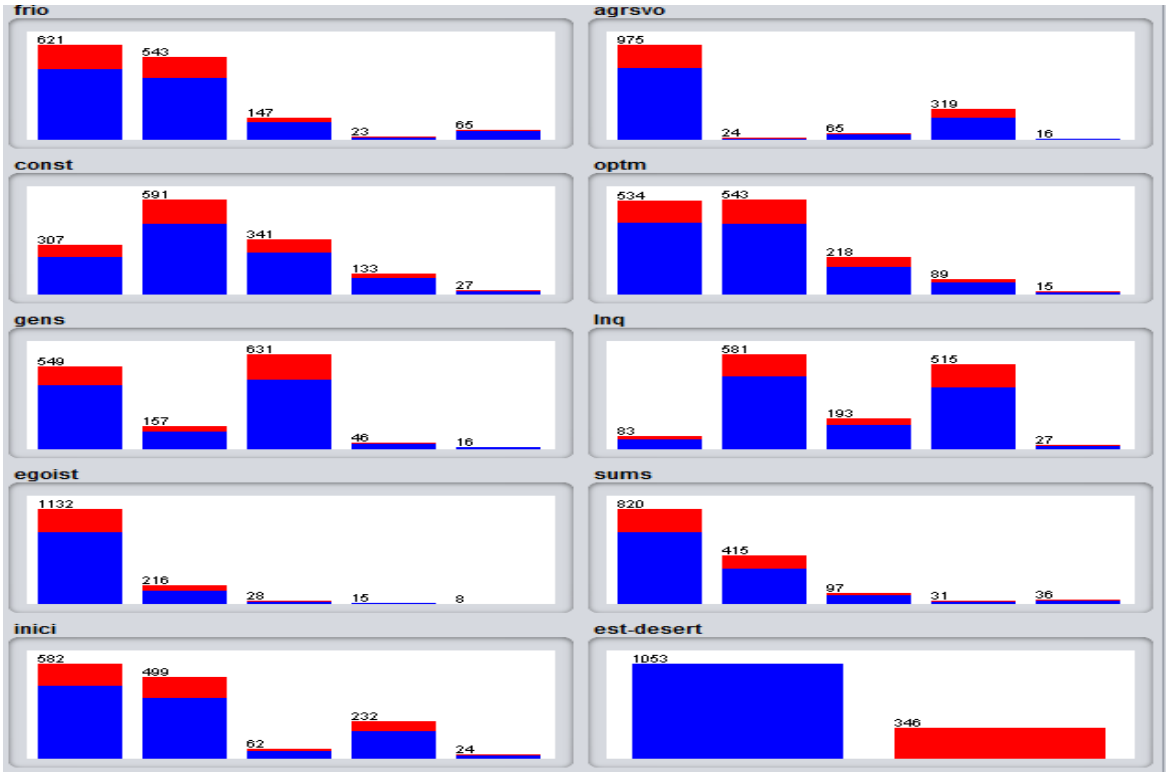
ELABORADO POR: EL AUTOR.

En el gráfico 5 se puede observar la distribución de algunas variables, una de ellas es t-trans-estud. Esta refleja información que establece que la mayoría de los estudiantes les toma más de una hora transportarse hasta la universidad. Otra variable que se puede visualizar es la distribución de tipo escalonada es c-recob-mens. Dicha variable muestra información sobre la carrera asignada de cada estudiante.

4.3.2.Datos psicológicos BC2.

El siguiente conjunto de datos almacena las variables referentes a los factores psicológicos de los estudiantes, estas almacenan información de la autoevaluación de temperamento de los estudiantes, por ejemplo si se consideraban tímidos, extrovertidos, con iniciativa, etc. A este conjunto de datos de índole psicológico se llamará a partir de ahora BC2, las variables seleccionadas para esta base de conocimiento se pueden observar en el anexo 4. A continuación se muestra un gráfico de la distribución de las variables.

Gráfico 6: Distribución de datos de algunas de las variables de BC2



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

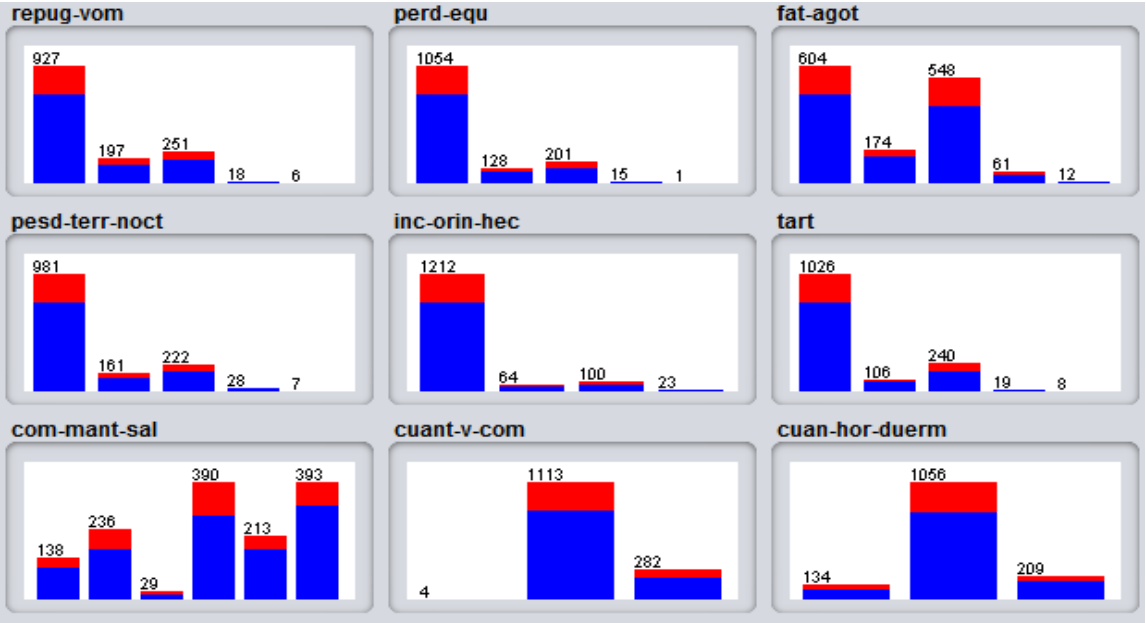
En el gráfico anterior se puede observar la realidad de algunas variables de BC2, por ejemplo en las variables const e inq. La variable const establece el autoconcepto que presenta el estudiante sobre si se considera una persona constante. La mayoría de los estudiantes no se consideran constantes. Por otra parte la variable inq establece si el estudiante se consideran inquietos se encuentra algo más distribuida.

4.3.3.Datos médicos BC3.

Se entiende como datos médicos al conjunto de datos almacenan toda la información de carácter médico, de salud y patologías que presentan los estudiantes y sus costumbres para mantenerse en forma y con buena salud, para otorgar mayor simplicidad a los futuros procesos a los datos de índole médico se les llamará BC3 las variables correspondientes se pueden observar en el anexo 5.

De la misma manera que se realizó con los conjuntos de datos vistos previamente, se mostrará a continuación un gráfico de la distribución de las variables de BC3

Gráfico 7: Distribución de los datos de algunas de las variables de BC3



FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

En el gráfico anterior se pueden observar variables com-mant-sal y repug-vom. La primera almacena las costumbres para mantener la salud por parte de los estudiantes que presenta respuestas bastante distribuidas. En la segunda almacena información del estudiante sobre si sufre repugnancia o vómitos. En ella se observa que las respuestas apuntan que la mayoría de los estudiantes dice sufrir de estos males

4.4.Pre-procesamiento de datos.

A continuación se detallará el proceso de pre-procesamiento de datos. En esta etapa se busca brindar una mayor consistencia a los conjuntos de datos y obtener mejores modelos en la etapa final de la investigación.

4.4.1.Transformación de los datos.

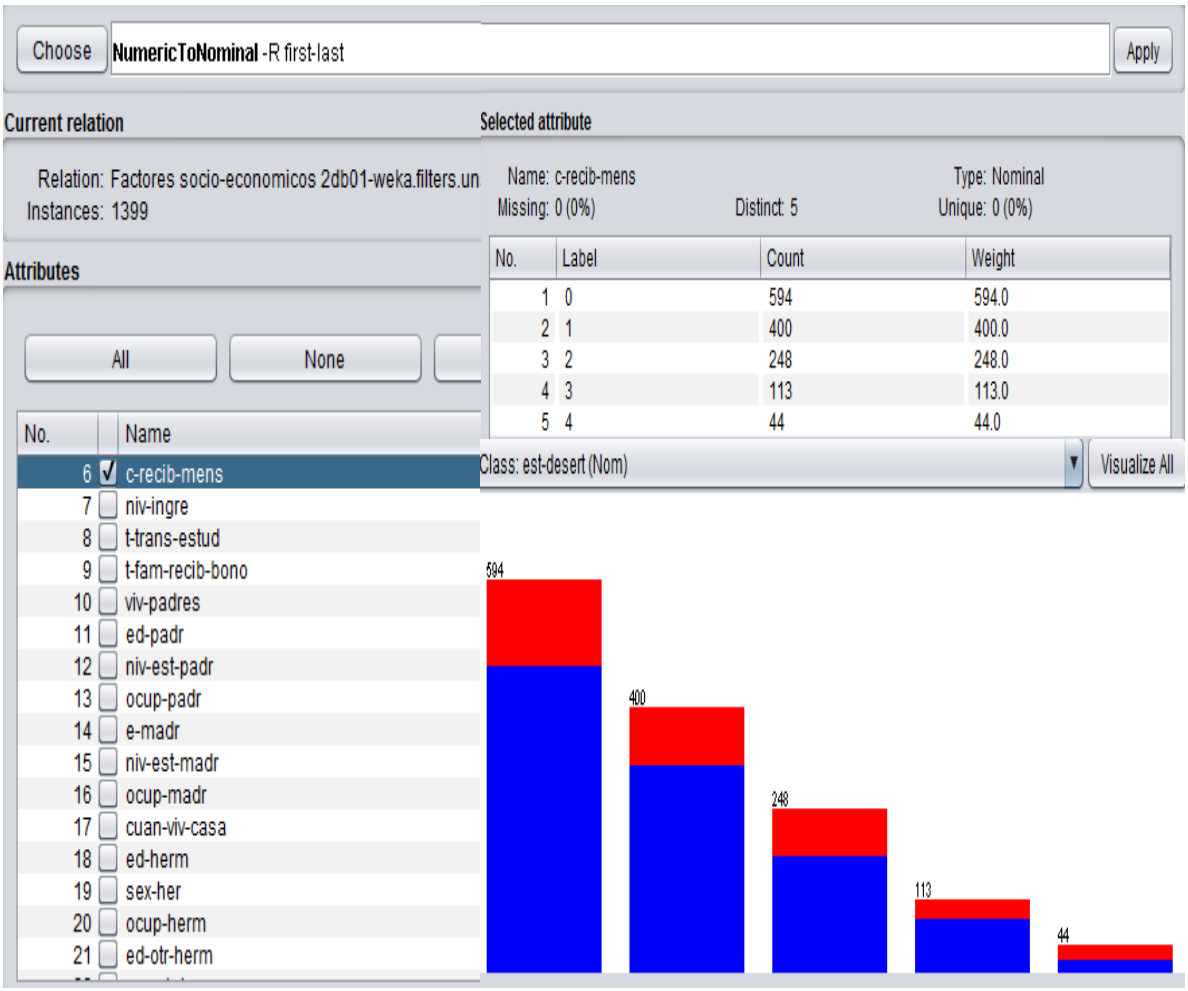
La transformación de los datos consistió en una serie de procesos y actividades aplicadas en los conjuntos de datos para poder obtener unos repositorios que sean aptos para la extracción de conocimientos y su posterior interpretación.

4.4.1.1. Limpieza de los datos

Para el proceso de detección y eliminación de outliers se aplicaron 2 herramientas de Weka llamada InterquartileRange la cual consiste en un filtro que se salta la clase de los atributos y se usa para detectar outliers y valores extremos basados en rangos y la RemoveWithValues que elimina aquellos valores atípicos que detectaba el primer filtro.

Al aplicar las técnicas y filtros para la limpieza de datos en la base de conocimiento BC1 se lograron eliminar los datos atípicos, mejorando la distribución de datos en las variables y se obtuvieron los resultados que se pueden observar en el siguiente gráfico.

Gráfico 8: Datos de BC1 posterior a la limpieza de datos



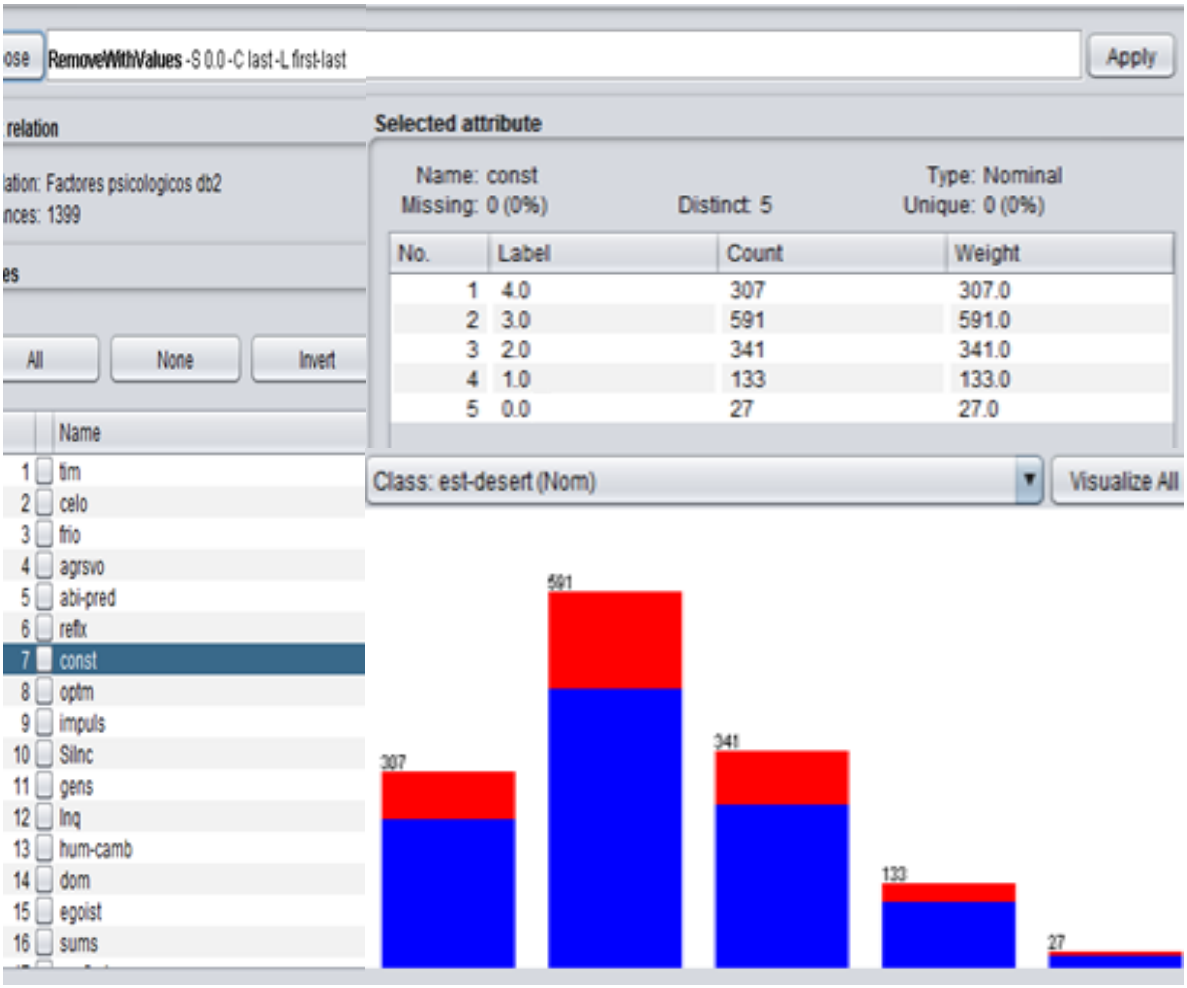
FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

En el gráfico anterior se muestra la variable cu-carr-asig posterior al proceso de limpieza de datos. Se puede observar unos resultados bastante distribuidos y ya en este punto libre de datos atípicos que se pudieron presentar.

De la misma manera se aplicaron las herramientas y técnicas para la limpieza de datos en BC2 y continuar con el proceso de tratamiento de datos, esto permitió obtener los resultados que se pueden observar en el siguiente gráfico de distribución de datos.

Gráfico 9: Datos de BC2 posterior a la limpieza de datos

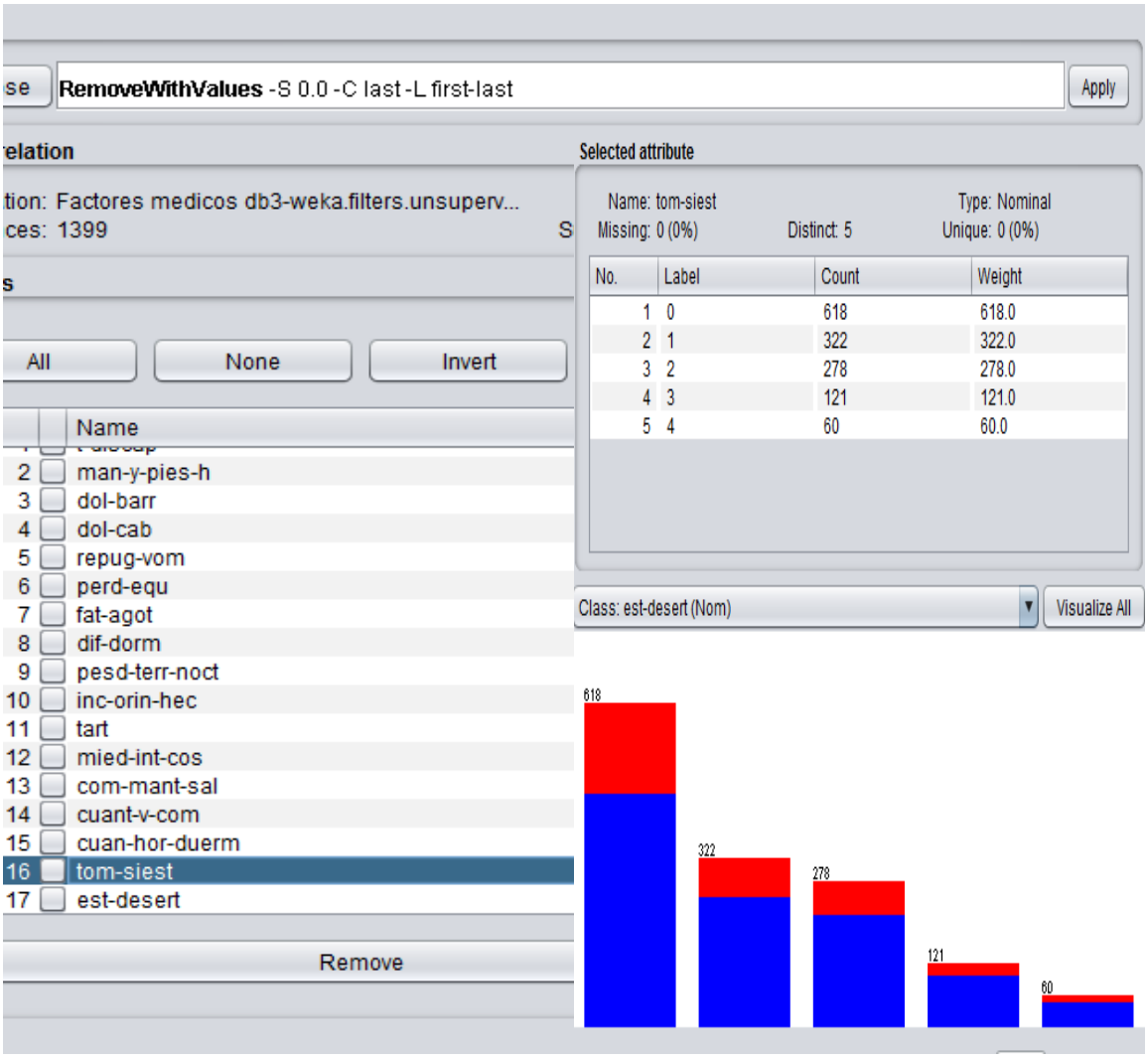


FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

El gráfico anterior se observa la variable const posterior a la limpieza de datos. Se eliminaron aquellos valores outliers que se presentaban. La distribución muestra que muchos estudiantes se consideran constantes en sus actividades.

Por último, se aplicó en BC3 las técnicas y herramientas para la limpieza de datos, se eliminaron los outliers, lo que también mejoró la distribución de las variables y permite poder observar los resultados del gráfico a continuación.

Gráfico 10: Datos de BC3 posterior a la limpieza de datos



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

El gráfico 10 mostrado anteriormente permite ver la distribución de la variable tom-siest después de la eliminación de datos. Esta variable proporciona las costumbres para tomar siestas de los estudiantes. El gráfico permite observar que la mayoría de los estudiantes suelen tomar siestas que superan las 2 horas.

Una vez finalizado aquello procedemos a continuar con el siguiente proceso de transformación de datos que es la discretización de datos. Dicho proceso que se explicará a continuación.

4.4.1.2. Discretización de los datos.

La discretización es básicamente agrupar una serie de datos de un contexto y asignar un valor que puede ser utilizado en proceso futuros del tratamiento de datos. Este proceso se pudo realizar en la mayoría de las variables debido a que poseían datos o conjuntos de datos que podían ser agrupados en rangos y asignarle un valor representativo.

A continuación se puede visualizar un extracto de los datos de BC2 antes y después de la discretización, para mayor información véase el anexo 6.

Gráfico 11: Visualización de los datos previo y post discretización

Antes

[Abierto]	[Reflexivo]	[Constante]
Mucho	Mucho	Mucho
Poco	Poco	Frecuente
Poco	Frecuente	Frecuente
Frecuente	Frecuente	Frecuente
Mucho	No	Mucho
Mucho	Frecuente	Frecuente
Frecuente	Frecuente	Poco
Frecuente	Frecuente	Frecuente
Frecuente	Frecuente	Frecuente
Poco	Mucho	Mucho
Frecuente	Frecuente	Frecuente
Mucho	Frecuente	Mucho
Frecuente	Poco	Poco
Poco	Mucho	Frecuente
Poco	Frecuente	No
Poco	Poco	Poco

Después

abi-pred	reflx	const
4.0	4.0	4.0
2.0	2.0	3.0
3.0	3.0	4.0
2.0	3.0	3.0
3.0	3.0	3.0
4.0	1.0	4.0
2.0	2.0	2.0
4.0	3.0	3.0
3.0	3.0	2.0
3.0	3.0	3.0
2.0	2.0	2.0
3.0	3.0	3.0
2.0	4.0	4.0
3.0	3.0	3.0
4.0	3.0	4.0
3.0	2.0	2.0

FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

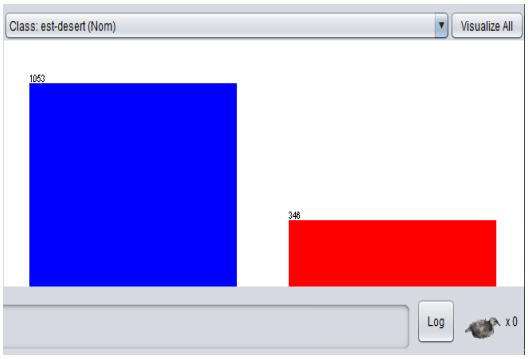
Como se puede observar en el gráfico anterior, los valores fueron agrupados en rangos establecidos. El proceso de discretización fue realizado de manera manual, entre otros motivos por la complejidad de configuración de los algoritmos para discretizar valores nominales en función de cada variable. Esto a su vez permitió realizar una valoración personalizada de cada rango que se presentaba, lo cual se puede observar en el anexo 6.

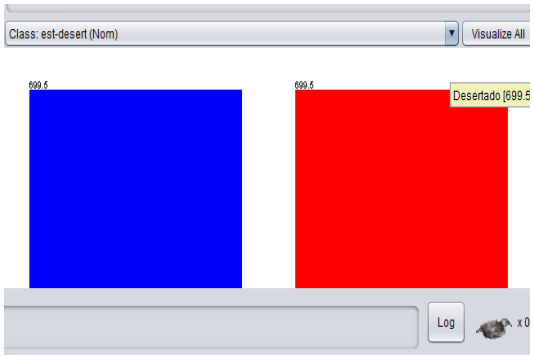
4.4.1.3. Balanceo de datos.

El balanceo de datos es una técnica que consiste en dar el mismo peso a los datos que ingresan al proceso. Esto ayuda a conseguir un análisis más equilibrado, principalmente cuando se tiene como objetivo obtener modelos aplicando algoritmos de predicción.

A continuación se muestra un gráfico del antes y después del balancear de la variable est-desert que almacena la información que detalla si el estudiante es desertor o no del curso de nivelación.

Gráfico 12: Comparación de variable est-desert previo y posterior al balance de datos





FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

Como se puede observar la distribución presento un cambio, en lugar de estar inclinado al estado “no desertó”, pasó a estar equilibrado entre “desertó” y “no desertó”. Este resultado les brinda más equilibrio a los modelos que se aplicarán posteriormente.

4.4.1.4. Selección de variables.

En esta fase se aplicará la herramienta CorrelationAttributeEval. Ella se encarga de evaluar el valor de un atributo midiendo la correlación de Pearson entre él y la clase. Además esta herramienta usa un algoritmo de búsqueda y toma de atributos y se llama Ranker. Ranker se caracteriza por asignar un rango a cada atributo, obteniendo diversos resultados según el repositorio de datos.

Con la información obtenida del proceso de selección de variables se toman aquellas que tengan un mayor rango, de esta manera se reduce la complejidad de los modelos de árboles obtenidos y permite una mejor interpretación de resultados.

Los resultados obtenidos con la herramienta CorrelationAttributeEval aplicadas en BC1 son los siguientes:

Tabla 3: Variables seleccionadas por su rango para BC1

#	Variable	Rango	Descripción
1	rang-ed	0.25157	El rango de la edad del estudiante
2	t-ingr	0.23272	¿Tienes ingresos propios?
3	dpend-d	0.21004	¿De quién dependes económicamente?
4	tien-farm	0.15858	Tienes familiares con vida
5	t-trans-estud	0.1268	¿Cuánto tiempo te toma llegar de tu casa a la UTEQ?
6	niv-est-madr	0.12115	Nivel de estudio de la madre
7	c-recib-mens	0.10406	¿Cuánto te dan mensualmente?
8	gusta-carr	0.09353	Le gusta la carrera que asignó el SNNA
9	e-madr	0.08496	Edad de la madre
10	ed-herm	0.08426	Edad del hermano

FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

La tabla anterior establece las 10 variables de BC1 más influyentes en el estado de "desertado" o "no desertado" de un estudiante. En ella se establece que la más importante es el rango de la edad del estudiante, seguido por aquella que almacena los datos de los ingresos del estudiante.

La tabla con las variables seleccionadas en base a su rango fueron obtenidos de la aplicación de la herramienta CorrelationAttributeEval en BC2, de las cuales las más importantes se muestran en la tabla a continuación:

Tabla 4: Variables seleccionadas por su rango para BC2

#	Variable	Rango	Descripción
1	gens	0.03488	¿Qué tan generoso eres?
2	tim	0.02976	¿Qué tan tímido eres?
3	celo	0.02826	¿Qué tan celoso eres?
4	inq	0.02754	¿Qué tan inquieto eres?
5	egoist	0.02644	¿Qué tan egoísta eres?
6	confiado	0.02574	¿Qué tan confiado eres?
7	agrsvo	0.02446	¿Qué tan agresivo eres?

FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

La tabla 4 mostrada previamente, establece 7 variables seleccionadas como las mas importantes en la deserción del estudiante. En esta tabla se observa que la variable que destaca en BC2 es gens, que establece cuan generoso se considera el estudiante/

Con la herramienta CorrelationAttributeEval aplicada en BC3 se obtuvieron una serie de variables ordenadas según el rango (valor obtenido con la aplicación de la herramienta), las variables más destacables son las siguientes:

Tabla 5: Variables seleccionadas por su rango para BC3

#	Variable	Rango	Descripción
1	man-y-pies-h	0.10213	Manos y/o pies hinchados
2	com-mant-sal	0.07037	¿Cómo te mantienes en buenas condiciones físicas?
3	cuan-hor-duerm	0.05906	¿Cuántas horas duermes al día?
4	tom-siest	0.05845	¿Tomas siestas en el día?
5	t-discapac	0.05366	¿Tienes alguna discapacidad?
6	cuant-v-com	0.0485	¿Cuantas comidas consumes al día?

FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

De esta manera se establecieron los 3 repositorios con variables seleccionadas. Estos repositorios ya en este punto se encuentran listos para el modelado, la interpretación y la extracción de conocimientos.

4.5.Extracción del conocimiento.

Una vez completado todo el proceso de pre-procesamiento de los datos, se tienen los repositorios listos para la extracción de conocimientos. Se debe proceder a la aplicación de los algoritmos para la generación de modelos y el análisis de cada uno de ellos.

4.5.1.Comparación de modelos.

Para un análisis de los resultados obtenidos por cada uno de los modelos se realizó un estudio comparativo. En este estudio se tomaron cada uno de los 3 repositorios antes de la selección de variables y los repositorios posterior a la selección de variables. Los resultados de la aplicación de los árboles de decisión, junto con cada uno de los conjuntos de datos a los que fueron aplicados se muestran a continuación:

Tabla 6: Comparación de la precisión de los modelos de árboles aplicados

Árboles de Decisión / Bases de Conocimiento	Decision Stump	Hoeffding Tree	J 48	LMT	Random Forest	Random Tree	REP Tree
Precisión de los modelos en las bases de conocimiento sin selección de atributos							
BC1	59.71%	63.2%	89.75%	70.28%	99.95%	99.95%	73.93%
BC2	51.55%	55.77%	87.29%	52.2%	99.05%	99.05%	52.79%
BC3	53.73%	50.14%	83.13%	56.36%	96.39%	96.39%	63.1%
Precisión de los modelos en las bases de conocimiento con selección de atributos							
BC1	59.71%	62.18%	85.82%	64.51%	98.76%	98.77%	71.23%
BC2	51.55%	54.51%	81.7%	62.16%	96.38%	96.42%	62.67%
BC3	53.73%	50.14%	68.78%	59.03%	81.29%	81.51%	58.71%

FUENTE: WEKA WORKBENCH.

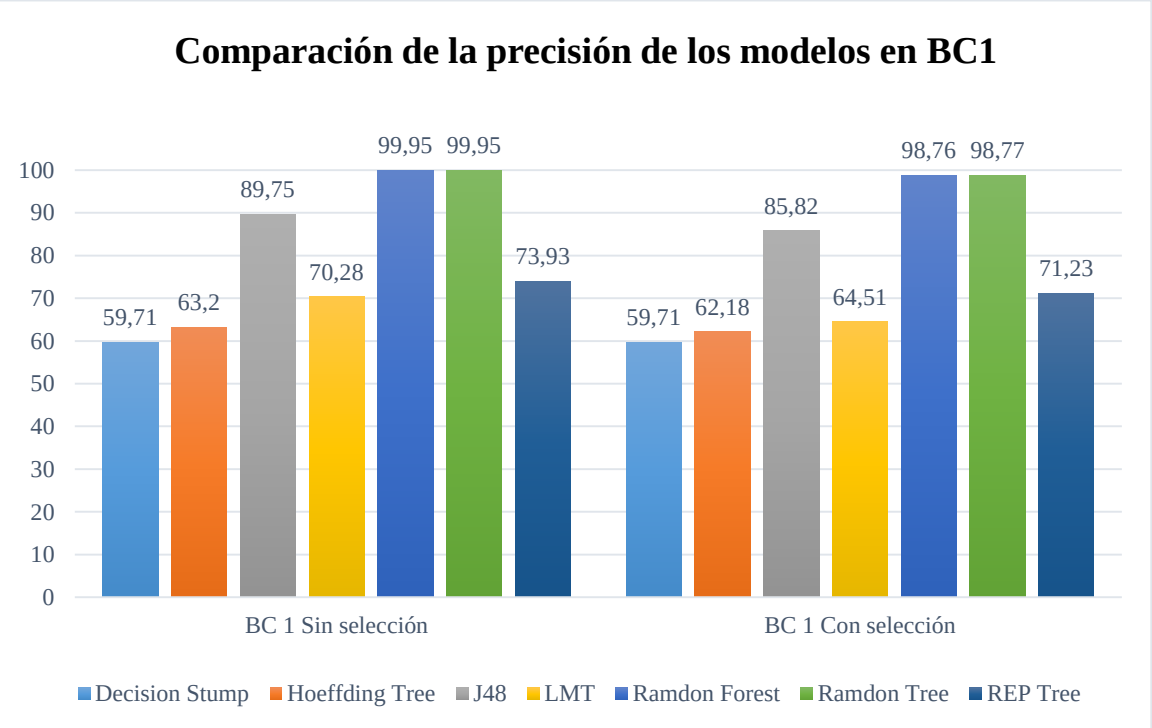
ELABORADO POR: EL AUTOR.

Como se puede observar en la matriz de resultados mostrada la mayoría de los algoritmos tienen un mayor porcentaje con las bases de conocimiento originales. Se debe resaltar que esto es un resultado esperado, pues al reducir el número de variables se disminuye la precisión. Sin embargo la selección de variables disminuye la complejidad del árbol generado por lo que era necesario pasar por este proceso.

En base a los resultados se establece al Random Tree como el mejor algoritmo de modelado para los 3 casos. Se estableció aquello debido a que son los que muestran una mayor precisión y menor complejidad.

Para una mejor apreciación se realizaron una serie de gráficos comparativos que se mostrarán a continuación se comenzará mostrando el gráfico correspondiente a BC1:

Gráfico 13: Comparación de los modelos aplicados en BC1.

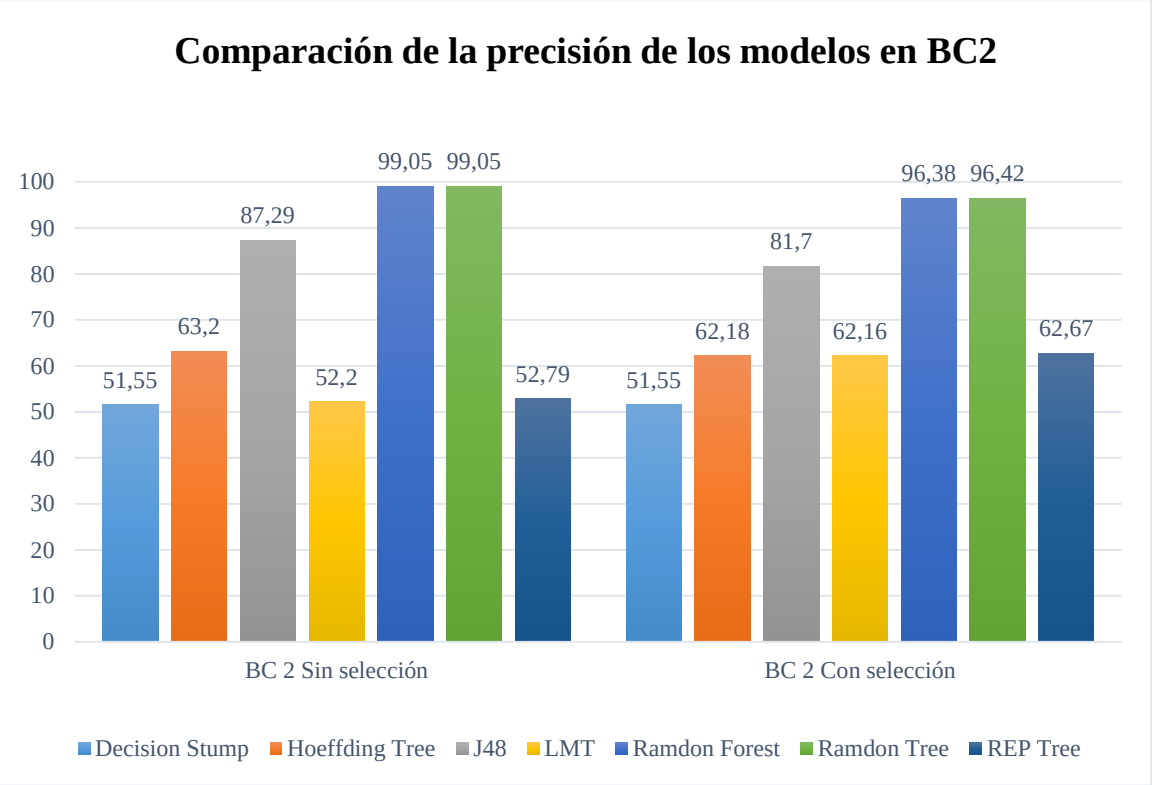


FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

En el gráfico anterior se muestran los porcentajes de precisión de cada uno de los modelos antes y después de la selección de variables. Esto permite apreciar que la perdida de precisión es mínima en todos los casos de BC1. Debido a esto se usará el modelo obtenido de BC1 posterior a la selección de variables. A continuación se muestra el gráfico correspondiente al conjunto de datos BC2.

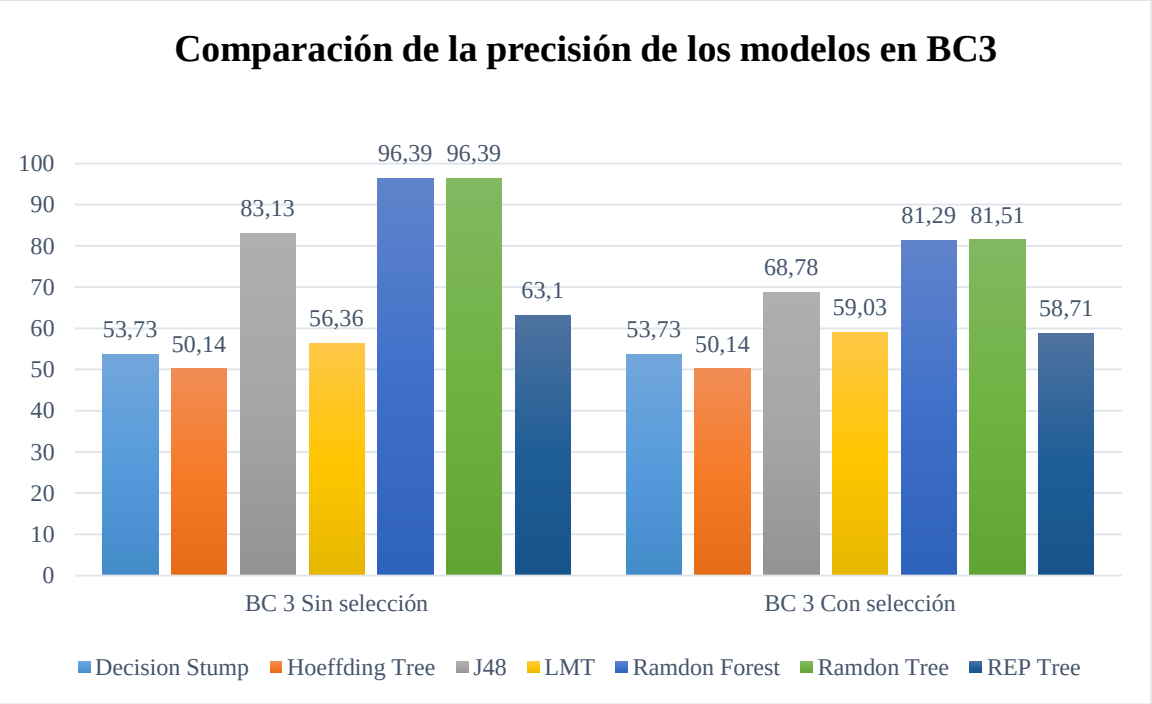
Gráfico 14: Comparación de los modelos aplicados en BC2



FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

En el gráfico mostrado previamente se puede apreciar la precisión de los modelos aplicados para BC2. Este gráfico permite observar que la precisión que se pierde después de la selección de atributos es mínima. Es por esto que se tomará el modelo posterior a la selección de variables. A continuación se muestra el gráfico para BC3.

Gráfico 15: Comparación de los modelos aplicados en BC3



FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

Se puede observar en el gráfico anterior los porcentajes de precisión de los modelos que se aplicaron en BC3. En él se puede apreciar que a diferencia de los casos anteriores, en BC3 si existe una reducción de precisión significativa en los modelos aplicados. Por lo que se establece que se utilizará el modelo obtenido con el repositorio previo a la selección de variables.

En base a los resultados mostrados previamente, se ha decidido escoger las bases de conocimiento BC1 y BC2 en su estado posterior a la selección de variables para la generación de modelos. Esto se debe a que la simplicidad del modelo generado si justifica la perdida de precisión. Sin embargo aquello no sucede con la base de conocimiento BC3 en la cual la pérdida de precisión es algo más significativa y por ello se tomará su versión previa a la selección de atributos.

4.5.2. Análisis de los resultados del modelo seleccionado.

Continuando con el proceso de obtención de resultados, es necesario detallar una síntesis de los modelos obtenidos por las herramientas utilizadas previamente. Este proceso se realizará dividiendo los resultados según cada base de conocimiento.

4.5.2.1. Datos socio-económicos – BC1.

Los estudiantes que tienden a desertar son los que dependen económicamente de sí mismos u otros familiares[6]. Otro factor importante es de aquellos que han superado los 25 años de edad y que no les toma más de 1 hora transportarse hasta la universidad. También se tomó en cuenta que el total de ingresos de estos estudiantes no suelen superar el sueldo básico, esto se puede visualizar en mayor medida en el anexo 8.

Por lo tanto se establece que aquellos factores socio-económicos que inciden en la deserción estudiantil son: la edad del estudiante, de quien depende económicamente, si tiene ingresos propios, los familiares con los que convive, el tiempo de transporte desde su hogar hasta la UTEQ y el nivel de ingresos de la familia.

4.5.2.2. Datos psicológicos – BC2.

El modelo permite establecer que la deserción en el curso de nivelación se presenta principalmente en aquellos estudiantes que frecuentemente son confiados. Otro factor importante es que los estudiantes que son algo tímidos también tienden a desertar. Un caso en particular se da cuando el estudiante suelen ser muy generosos e inquietos, pues estas características en conjunto suelen provocar deserción. Todo lo establecido anteriormente se puede observar con mayor detalle en el anexo 10, que permite visualizar el modelo de Random Tree obtenido.

[6] Se refiere a familiares distintos a los padres

El análisis del modelo permite determinar que los principales factores psicológicos que inciden en la deserción estudiantil son: la timidez, la generosidad, lo confiados que son los estudiantes y que tan inquieto se considera el estudiante del curso de nivelación.

4.5.2.3. Datos médicos – BC3.

Los datos de BC3 establecen que los estudiantes que tienden a la deserción son los que suelen tener costumbres sedentarias, realidad que es acentuada si sufren otro mal de salud. Por otro lado la falencia de salud más perjudicial para el estudiante es sufrir de manos y pies hinchados. Además influye la falta de sueño pues los estudiantes que suelen dormir menos de 8 al día tienden a ser desertores. Sin embargo una realidad bastante grave que se pudo observar es que el sufrir una discapacidad casi siempre garantiza que el estudiante se vuelva desertor del curso de nivelación. Esto se puede observar en el Random Tree del anexo 11.

Es decir se puede definir que de los factores médicos almacenados en BC3 los que inciden en la deserción estudiantil son: las afecciones de manos y pies hinchados, las costumbres para mantener salud, las horas que duermen al día y si el estudiante presenta alguna discapacidad.

4.5.3.Factores de deserción.

Los resultados obtenidos permiten establecer la siguiente lista de los factores de BC1, BC2 y BC3 que más influyen en la deserción de los estudiantes del curso de nivelación:

Tabla 7: Factores más incidentes en la deserción estudiantil del curso de nivelación

Factores de deserción estudiantil de los estudiantes del curso de nivelación		
Nº	**Factor**	**Descripción**
1	rang-ed	Corresponde a la edad del estudiante
2	t-ingr	Corresponde al monto de ingresos que genera el estudiante
3	depend-d	Establece la dependencia económica del estudiante
4	tien-fam	Establece que familiares tiene con vida el estudiante
5	t-trans-estud	Corresponde al tiempo que toma ir a la universidad

6	gens	Establece el nivel de generosidad del estudiante
7	tim	Establece el nivel de timidez del estudiante
8	confiado	Establece el nivel de confianza que otorga el estudiante a otros
9	inq	Establece el nivel de inquietud del estudiante
10	man-y-pies-h	Establece el nivel de afección de manos y pies hinchados
11	com-mant-sal	Establece costumbres para mantener la salud
12	cuan-hor-duerm	Establece la cantidad de horas que duerme el estudiante
13	t-discapac	Establece si el estudiante presenta discapacidad

FUENTE: WEKA WORKBENCH.
ELABORADO POR: EL AUTOR.

La tabla mostrada previamente permite observar los factores que inciden en la deserción estudiantil, esta establece un total de 13 factores que fueron tomados de 106 posibles y que fueron hallados y seleccionados a través de procesos de extracción de conocimientos y evaluación de modelos de árboles de decisión.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1.Conclusiones.

La presente investigación fue realizada con la finalidad de poder identificar los factores que influyen en la deserción de los estudiantes de la Unidad de Admisión y Registro, estos factores fueron hallados entre 106 variables que fueron analizadas para obtener información relevante de los estudiantes, concluyendo lo siguiente:

- Se analizó la información proporcionada por la Unidad de Admisión y Registro, permitiendo encontrar características en común entre la información que almacenaban las variables. Se la ordenó y clasificó según la naturaleza de las variables, resultando la siguiente estructuración: variables de índole socio-económica se almacenarán en el repositorio BC1, las de índole psicológica en el repositorio BC2 y por último las de índole médica en el repositorio BC3.

- Se emplearon técnicas de pre-procesamiento para preparar los datos y volverlos aptos para la generación de modelos y la extracción de conocimientos. Estas técnicas consistieron en la eliminación de datos atípicos con las herramientas de InterquartileRange y RemoveWithValues. Posteriormente se discretizados las variables de forma manual tomando en cuenta sus características. Seguido se realizó el balance de datos con SMOTE, para finalizar con la selección de variables usando CorrelationAttributeEval. Todo estos procesos prepararon los repositorios BC1, BC2 y BC3 para la generación de modelos

- Mediante los modelos de árboles de decisión aplicados como Stump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree y REP Tree, se pudo establecer que el modelo más preciso para los 3 repositorios es el modelo Random Tree tanto para el modelo con o sin selección de atributos. Se eligieron los repositorios BC1 y BC2, ambos con selección de variables. Por el contrario para BC3 se escogió el repositorio sin el tratamiento de selección de variables, debido q que ocasionaba que el modelo presentara una pérdida de precisión considerable. Los porcentajes reflejados en cada uno de los repositorios son los siguientes BC1 98.76% BC2 96.38% y BC3 96.39%.

- Los análisis desarrollados dieron como resultado que existen 13 factores que presentaban preponderancia de entre las 106 posibles causas de deserción estudiantil en los cursos de la Unidad Admisión y Registro de la Universidad Técnica Estatal de Quevedo. De estos factores, 5 corresponden a la realidad socio-económica, 4 a la psicológica y 4 a la realidad médica.

5.2.Recomendaciones.

El presente proyecto de investigación es de un amplio interés para el sector educativo, principalmente para las universidades que aplican los cursos del SNNA. Además estas instituciones se esfuerzan por disminuir la deserción estudiantil que se presenta en cada curso. Por lo que estableciendo este documento como una herramienta para futuros procesos de investigación a nivel educativo se realizan las siguientes recomendaciones:

- Obtener nuevos repositorios de datos, preferiblemente aquellos de futuros cursos de nivelación. esto permitirá establecer un historial de comportamientos y poder realizar análisis comparativos entre los repositorios de datos. Lo cual aumentará la escala de esta investigación y apoyando en mayor medida a los cursos de nivelación de la Unidad de Admisión y Registro de la UTEQ.

- Aplicar nuevas herramientas de minería de datos y extracción de conocimientos. Esto con el fin de realizar a futuro un análisis comparativo entre los modelos generados por ambos procesos. Lo que se busca obtener es una serie de resultados que establezcan una serie de comportamientos específicos en los datos.

- Utilizar los resultados obtenidos por parte de los directivos de la Unidad de Admisión y Registro como una herramienta adicional para la toma de decisiones. Esto resulta de vital importancia al momento de establecer estrategias para mejorar el proceso educativo en miras de disminuir el nivel de deserción del curso de nivelación.

CAPÍTULO VI
BIBLIOGRAFÍA

6.1.Bibliografía citada.

[1] P. L. César and Daniel Santín Gonzalez, *Minería de datos: técnicas y herramientas*. Paraninfo Cengage Learning, 2007.

[2] María Isabel Ángeles Larrieta and Angélica María Santillán Gómez, “minería de datos: Concepto, características, estructura y aplicaciones,” 1998.

[3] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier Science, 2011.

[4] J. Tuya, I. R. Román, and J. J. D. Cos’\in, *Técnicas cuantitativas para la gestión en la ingeniería del software*. Netbiblo, 2007.

[5] C. Krzysztof, P. Witold, R. Swiniarski, and L. Kurgan, *A Knowledge Discovery Approach*. 2007.

[6] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Company, 2014.

[7] *Avances en Informática y Sistema Computacionales Tomo I (CONAIS 2006)*. Univ. J. Autónoma de Tabasco.

[8] D. L. González-Bañales, M. G. L. Alan’\is, and J. A. G. Reyes, *Las fuerzas competitivas de mercado y su influencia en la incorporaci{ó}n de las TIC en las PyME. Un estudio exploratorio*. Lulu.com.

[9] J. Moreno, D. Rodríguez, M. A. Sicilia, J. C. Riquelme, and R. Ruiz, “SMOTE-I : mejora del algoritmo SMOTE para balanceo de clases minoritarias,” vol. 3, no. 1, pp. 73–80, 2009.

[10] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Springer US, 2011.

[11] B. Pfahringer, G. Holmes, and R. Kirkby, “New Options for Hoeffding Trees,” in *AI 2007: Advances in Artificial Intelligence: 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2-6, 2007. Proceedings*, M. A. Orgun and J. Thornton, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 90–99.

[12] B. Bifet, A., Frank, E., Holmes, G., & Pfahringer, “Ensembles of restricted hoeffding trees,” *Proc. 14th Int. Conf. Artif. Intell. Stat.*, vol. 15, no. 212, pp. 434–442, 2012.

[13] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank, “Fast perceptron decision tree learning from evolving data streams,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), 2010, vol. 6119 LNAI, no. PART 2, pp. 299–310.

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[15] B. Chimieski and R. Fagundes, "Association and Classification Data Mining Algorithms Comparison over Medical Datasets," *J. Heal. Informatics*, vol. 5, no. 2, pp. 44–51, 2013.

[16] Michael Walker, "Random Forests Algorithm - Data Science Central," 2013. .

[17] G. J. Williams, "Rattle : A Data Mining GUI for R," *R J.*, vol. 1, no. December, pp. 45–55, 2009.

[18] E. Durán and R. Costaguta, "Minería de datos para descubrir estilos de aprendizaje," *Rev. Iberoam. Educ.*, vol. 4, 2007.

[19] S. Valero, A. Salvador, and M. García, "Minería de datos: predicción de la deserción escolar," *Recur. Digit. para la Educ. y la Cult.*, vol. KAAMBAL, p. 235, 2010.

[20] E. Rodallegas, A. Torres, G. Beatriz, E. Gastelloú, R. Lezama, and S. Valero, "Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos," *Recur. Digit. para la Educ. y la Cult.*, vol. KAAMBAL, p. 235, 2010.

[21] O. SPOSITTO and M. ETCHEVERRY, "Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil Deserción," *… en Sist. Cibernética e …*, 2010.

[22] R. Timar and J. Jim, "Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos," *Rev. Vínculos*, vol. 10, no. 1, pp. 373–383, 2013.

CAPÍTULO VII

ANEXOS

Anexo 1: Tabla de variables del repositorio inicial

Nº	Variable	Descripción
1	p-acad	Periodo Académico
2	parl	Paralelo
3	id	Cédula
4	estud-aped	Ingrese su Apellidos
5	estud-nomb	Ingrese sus Nombres
6	f-nac	Fecha de Nacimiento
7	p-nac	País de Nacimiento
8	prov-nac	Provincia Nacimiento
9	ciud-nac	Cantón o Ciudad Nacimiento
10	Dmcl	Domicilio
11	t-conv	Teléfono Convencional
12	t-mov	Teléfono Celular
13	E-mail	Email
14	r-estud	¿Cuál es tu situación para estudiar aquí?
15	tien-fam	¿Vives con algún familiar?
16	v-regr-casa	Viajas de regreso a tu casa?
17	t-ingr	¿Tienes ingresos propios?
18	pag-alqui	¿Pagas alquiler?
19	dpend-d	¿De quién dependes económicamente?
20	alc-recur-econ	¿Los recursos económicos con que cuentas son suficientes?
21	c-recib-mens	¿Cuánto te dan mensualmente?
22	niv-ingre	Nivel de ingresos aproximado de su familia
23	t-trans-estud	¿Cuánto tiempo te toma llegar de tu casa a la UTEQ?
24	t-discap	¿Tienes alguna discapacidad y Especifique?
25	t-fam-recib-bono	¿Su familia recibe el bono solidario?
26	viv-padres	¿Viven sus padres?
27	nomb-d-padr	Nombres y Apellidos del Padre
28	ed-padr	Edad del Padre
29	niv-est-padr	Nivel de estudio del padre
30	ocup-padr	Ocupación del Padre
31	nomb-madr	Nombres y Apellidos de la Madre

32	e-madr	Edad del Madre
33	niv-est-madr	Nivel de estudio del Madre
34	ocup-madr	Ocupación del Madre
35	cuan-viv-casa	¿Cuántos viven en tu casa?
36	nomb-her	Nombre del Hermano
37	ed-herm	Edad del Hermano
38	sex-her	Sexo
39	ocup-herm	Estudios y/o trabajo
40	nomb-otr-herm	Nombre de otro Hermano
41	ed-otr-herm	Edad de otro Hermano
42	sex-otr-herm	Sexo de otro Hermano
43	est-otr-herm	Estudios y/o trabajo del otro hermano
44	man-y-pies-h	[Manos y/o pies hinchados]
45	dol-barr	[Dolores de barriga]
46	dol-cab	[Dolores de cabeza]
47	repug-vom	[Repugnancia y/o vómitos]
48	perd-equ	[Pérdida de equilibrio]
49	fat-agot	[Fatiga y agotamiento]
50	dif-dorm	[Dificultades para dormir]
51	pesd-terr-noct	[Pesadillas o terrores nocturnos]
52	inc-orin-hec	[Incontinencia (orina o heces)]
53	tart	[Tartamudeo al explicarse]
54	mied-int-cos	[Miedo intenso ante cosas]
55	otr-enf	¿Tienes otras enfermedades?
56	com-mant-sal	¿Qué hace usted para mantenerse en buenas condiciones físicas?
57	cuant-v-com	¿Cuantas comidas consumes al día?
58	cuan-hor-duerm	¿Cuántas horas duermes al día?
59	tom-siest	¿Tomas siestas en el día?
60	rel-comp	¿Cómo es tu relación con tus compañeros?
61	rel-fam	¿Cómo te relacionas con sus hermanos, padres, etc?
62	afic-diver	¿Cuáles son tus aficiones y diversiones favoritas en tiempo libre?
63	t-gust-curs-niv	¿Estas contento de venir al curso de nivelación?
64	no-gust-curs-niv	¿Qué no te agrada del curso de nivelación?

65	doc-invit-pas	¿Algún docente te invita a ir a otros lugares fuera de la Universidad?
66	t-pid-din-doc	¿Te piden dinero algún docente?
67	dond-rev-tar	¿Dónde te revisan las tareas?
68	tim	[Tímido]
69	celo	[Celoso]
70	frio	[Frío]
71	agrsvo	[Agresivo]
72	abi-pred	[Abierto]
73	reflx	[Reflexivo]
74	const	[Constante]
75	optm	[Optimista]
76	impuls	[Impulsivo]
77	Silnc	[Silencioso]
78	gens	[Generoso]
79	Inq	[Inquieto]
80	hum-camb	[Cambia tu humor fácilmente]
81	dom	[Dominante]
82	egoist	[Egoísta]
83	sums	[Sumiso]
84	confiado	[Confiado en sí mismo]
85	imag	[Imaginativo]
86	inici	[Con iniciativa]
87	esc-grad	¿En qué escuela terminaste la primaria?
88	grad-curs-prim	Grado cursado de primaria
89	otr-esc	Otra escuela que has estudiado
90	grd-crs-otr-esc	Grado cursado de otra primaria
91	col-estud	¿En qué colegio has estudiado?
92	grad-curs-col	Grados cursado del colegio
93	otr-col-est	Otro colegio que has estudiado
94	otr-col-grad-curs	Grados cursado de otro colegio
95	tiens-dif-estud	Problema o dificultad con el estudio
96	est-conf-carr	¿Estas conforme con la carrera que estás estudiando en este momento?

97	prob-dif-enfr	¿Qué problemas y dificultades tiene en su vida personal e íntima?
98	com-enf-prob	¿Qué ha hecho o está haciendo usted para resolver sus problemas?
99	nec-ayud-alcanc-obj	¿Piensa usted que necesita ayuda u orientación para resolver sus problemas
100	qui-ayud-probl	¿A quién recurre en busca de ayuda de orientación en sus problemas?
101	sug-denunc	Sugerencia o denuncias
102	otr-pers-ayud-prob	¿A qué otra persona que recurre por ayuda?
103	rang-ed	Seleccione el rango que esta su Edad
104	cu-carr-asig	¿Cuál es la carrera que asigno cupo en SNNA?
105	gusta-carr	¿Le gusta la carrera que asigno en el SNNA?
106	est-desert	¿El estudiante es desertor?

FUENTE: REPOSITORIO DE DATOS DE LA UNIDAD DE ADMISION Y REGISTRO.
ELABORADO POR: EL AUTOR.

Anexo 2: Tabla de variables descartadas

1	Periodo Académico
2	Paralelo
3	Apellidos
4	Nombres
5	País de Nacimiento
6	Provincia de Nacimiento
7	Cantón de Nacimiento
8	Domicilio
9	Teléfono Convencional
10	Teléfono Celular
11	E-mail
12	¿Viajas de regreso a tu casa?
13	Los recursos económicos con que cuentas son suficientes para
14	Nombres y Apellidos del Padre
15	Nombres y Apellidos de la Madre
16	Nombre del Hermano
17	Nombre de otro Hermano
18	Tienes otras enfermedades
19	Algún docente te invita a ir a otros lugares fuera de la Universidad?
20	¿Te piden dinero algún docente?
21	¿En qué lugar le revisan las tareas?
22	Escuela en la que estudiaste

23	Grados cursados en tu escuela principal
24	Otra primaria en la que has estudiado (colegio secundario)
25	Grados cursados en tu otra escuela (si es que tuviste)
26	Colegio en el que estudiaste
27	Grados cursados en tu colegio principal
28	Otro colegio en el que has estudiado (colegio secundario)
29	Grados cursados en colegio secundario (si es que tuviste)
30	En la actualidad ¿tienes algún problema o dificultad con el estudio?
31	Tienes problemas y dificultades en su vida personal e íntima, trato social, relaciones familiares
32	Sugerencias o denuncias
33	¿A qué otra persona que recurre en busca de ayuda?
34	Cédula
35	Fecha de nacimiento
36	Otras enfermedades

FUENTE: REPOSITORIO DE DATOS DE LA UNIDAD DE ADMISION Y REGISTRO.
ELABORADO POR: EL AUTOR.

Anexo 3: Variables que corresponden al conjunto de datos BC3

#	Variable	Descripción
1	r-estud	¿Cuál es tu situación para estudiar aquí?
2	tien-fam	¿Viven sus padres?
3	t-ingr	¿Tienes ingresos propios?
4	pag-alqui	¿Pagas alquiler?
5	dpend-d	¿De quién dependes económicamente?
6	c-recib-mens	¿Cuánto te dan mensualmente?
7	niv-ingre	Nivel de ingresos aproximado de su familia
8	t-trans-estud	¿Cuánto tiempo te toma llegar de tu casa a la UTEQ?
9	t-fam-recib-bono	¿Su familia recibe el bono solidario?
10	viv-padres	¿Vives con algún familiar?
11	ed-padr	Edad del Padre
12	niv-est-padr	Nivel de estudio del padre
13	ocup-padr	Ocupación del Padre
14	e-madr	Edad de la Madre
15	niv-est-madr	Nivel de estudio del Madre
16	ocup-madr	Ocupación del Madre
17	cuan-viv-casa	¿Cuántos viven en tu casa?
18	ed-herm	Edad del Hermano 1
19	sex-her	Sexo Hermano 1
20	ocup-herm	Estudios y/o trabajo hermano 1
21	ed-otr-herm	Edad del otro Hermano
22	sex-otr-herm	Sexo de otro Hermano
23	est-otr-herm	Estudios y/o trabajo hermano 2

24	rel-comp	¿Cómo es tu relación con tus compañeros?
25	rel-fam	¿Cómo te relacionas con sus hermanos, padres, etc?
26	afic-diver	¿Cuáles son tus aficiones y diversiones favoritas en tiempo libre?
27	t-gust-curs-niv	¿Estas contento de venir al curso de nivelación?
28	no-gust-curs-niv	qué no le gusta el curso de nivelación
29	est-conf-carr	¿Estas conforme con la carrera que estudias en este momento?
30	com-enf-prob	¿Qué ha hecho o está haciendo usted para resolver sus problemas?
31	nec-ayud-alcanc-obj	¿Necesita ayuda para alcanzar sus objetivos?
32	qui-ayud-probl	A quién recurre en busca de ayuda con sus problemas
33	rang-ed	Seleccione el rango que esta su Edad
34	cu-carr-asig	Cuál es la carrera que asigno cupo en SNNA
35	gusta-carr	Le gusta la carrera que asigno en el SNNA
36	est-desert	Informa si el estudiante ha desertado

FUENTE: REPOSITORIO DE DATOS DE BC1.
ELABORADO POR: EL AUTOR.

Anexo 4: Variables que corresponden al conjunto de datos BC2

#	Variable	Descripción
1	tim	¿Sueles ser tímido?
2	celo	¿Sueles ser celoso?
3	frio	¿Sueles ser frío (que no suele mostrar sus emociones)?
4	agrsvo	¿Sueles ser agresivo?
5	abi-pred	¿Sueles ser abierto (que estás dispuesto a participar en todo)?
6	reflx	¿Sueles ser reflexivo?
7	const	¿Sueles ser constante?
8	optm	¿Sueles ser optimista?
9	impuls	¿Sueles ser impulsivo?
10	Silnc	¿Sueles ser generoso?
11	gens	¿Sueles ser inquieto?
12	Inq	¿Sueles cambiar de humor repentinamente?
13	hum-camb	¿Sueles ser dominante?
14	dom	¿Sueles ser egoísta?
15	egoist	¿Sueles ser sumiso?
16	sums	¿Sueles ser confiado en ti mismo?
17	confiado	¿Sueles ser egoísta?
18	imag	¿Sueles ser imaginativo?
19	inici	¿Sueles tener iniciativa?
20	est-desert	Informa si el estudiante ha desertado

FUENTE: REPOSITORIO DE DATOS DE BC2.
ELABORADO POR: EL AUTOR.

Anexo 5: Variables que corresponden al conjunto de datos BC3

#	Variable	Descripción
1	t-discap	Tienes alguna discapacidad
2	man-y-pies-h	¿Sufres de manos y pies hinchados?
3	dol-barr	¿Sufres de dolores de barriga?
4	dol-cab	¿Sufres de dolores de cabeza?
5	repug-vom	¿Sufres de Repugnancia y/o vómitos?
6	perd-equ	¿Sufres de pérdida de equilibrio?
7	fat-agot	¿Sufres fatiga y agotamiento?
8	dif-dorm	¿Sufres de dificultades para dormir?
9	pesd-terr-noct	¿Sufres de pesadillas y/o terrores nocturnos?
10	inc-orin-hec	¿Sufres de incontinencia (orina o heces)?
11	tart	¿Sufres de tartamudeo?
12	mied-int-cos	¿Sufres de miedo intenso a algunas cosas?
13	com-mant-sal	¿Qué hace usted para mantenerse en buenas condiciones físicas?
14	cuant-v-com	¿Cuantas comidas consumes al día?
15	cuan-hor-duerm	¿Cuántas horas duermes al día?
16	tom-siest	Tomas siestas en el día
17	est-desert	Informa si el estudiante ha desertado

FUENTE: REPOSITORIO DE DATOS DE BC3.
ELABORADO POR: EL AUTOR.

Anexo 6: Informe del proceso de discretización

Informe de variables discretizadas y de sus valores asignados

Se revisaron un total de 106 preguntas, de las cuales se descartaron 36 debido a que no contribuían con la presente investigación, de las 70 preguntas restantes, 34 preguntas corresponden a factores socioeconómicos, 18 preguntas corresponden a factores psicológicos y 15 preguntas corresponden a factores médicos de los estudiantes, en todos los conjuntos de datos se le agregó si el estudiante ha desertado.

A continuación se procederá a mencionar las preguntas discretizadas

Datos socio-económicos - BC1

1) ¿Cuál es tu razón para estudiar en esta universidad?

0 = Otros motivos
1 = Estudia parcialmente obligado por el sistema
2 = Porque quería venir a la universidad

2) ¿Vives con algún familiar?

0 = Otros no familiares
1 = Formó su propia familia
2 = Familia (no padres)
3 = Un solo padre
4 = Familia completa

3) Tienes ingresos

0 = No tengo ingresos
1 = Menos de 177 dólares
2 = Entre 177 y 353 dólares
3 = Salario básico (354 dólares)
4 = Más de 354 dólares

4) ¿Pagas alquiler?

0 = Si (Alquilo habitación solo)
1 = Comparto habitación y el pago con compañeros
2 = No (Vivo con familiares)

5) De quien dependes económicamente

0 = Independiente
1 = Otros familiares
2 = Conyugue
3 = Sólo un padre
4 = Ambos padres

6) ¿Cuánto te dan mensualmente?

0 = 0 a 25
1 = 25 a 50
2 = 50 a 100
3 = 100 a 200
4 = Más de 200

7) Nivel de ingresos aproximado de su familia

0 = No sabe
1 = Menos de 300 dólares
2 = 300 a 399 dólares
3 = 400 a 499 dólares
4 = 500 a 599 dólares
5 = Más de 600 dólares

8) ¿Cuánto tiempo te toma llegar de tu casa a la UTEQ?

0 = Menos de 15 minutos
1 = De 15 minutos a 30 minutos
2 = De 31 minutos a 1 hora
3 = De 1 hora a 1.15 horas
4 = De 1.15 horas a 2 horas
5 = Más de 2 horas

9) ¿Su familia recibe el bono solidario?

0 = No
1 = Si

10) ¿Viven sus padres??

0 = No vive ningún familiar
1 = No, pero tengo mis hermanos
2 = Si viven

11) Edad del padre

0 = Ha fallecido o no tiene
1 = No sabe
2 = 40 o menos
3 = 41 a 45 y 46 a 50
4 = 51 a 55 y 56 a 60
5 = 61 a 65
6 = Mayor de 65

12) Nivel de estudio del padre

0 = No tiene o fallecido
1 = No sabe
2 = Primaria
3 = Secundaria
4 = Universitario
5 = Postgrado

13) Ocupación del Padre

0 = No tiene o fallecido
1 = 1= No sabe
2 = Desempleo
3 = Empleo inestable o temporal
4 = Empleo Informal
5 = Labor Autónoma
6 = Empleo Estable

14) Edad del madre

0 = Ha fallecido o no tiene
1 = No sabe
2 = 40 o MENOS
3 = 41 a 45 y 46 a 50
4 = 51 a 55 y 56 a 60
5 = 61 a 65
6 = Mayor de 65

15) Nivel de estudio del madre

0 = No tiene o fallecido
1 = No sabe
2 = Primaria
3 = Secundaria
4 = Universitario
5 = Postgrado

16) Ocupación del madre

0 = No tiene o fallecido
1 = 1= No sabe
2 = Desempleo
3 = Empleo inestable o temporal
4 = Empleo Informal
5 = Labor Autónoma
6 = Empleo Estable

17) ¿Cuántos viven en tu casa?

0 = De 1 a 2 personas
1 = De 3 a 4 personas
2 = De 5 a 6 personas
3 = 7 personas o más

18) Edad del Hermano

0 = Ha fallecido o no tiene
1 = Menor de 5
2 = 6 hasta 12
3 = 13 hasta 18
4 = 19 hasta 25
5 = 26 hasta 35
6 = Mayor de 35

19) Sexo del hermano

0 = Masculino
1 = Femenino

20) Estudios y/o trabajo del hermano

0 = No tiene o fallecido
1 = 1= No sabe
2 = Desempleo o estudiante de segundo nivel o menos
3 = No siempre tiene trabajo, estudiante universitario o bachiller
4 = Empleo Informal
5 = Labor Autónoma
6 = Empleo Estable

21) Edad del hermano 2

0 = Ha fallecido o no tiene
1 = Menor de 5
2 = 6 hasta 12
3 = 13 hasta 18
4 = 19 hasta 25
5 = 26 hasta 35
6 = Mayor de 35

22) Sexo del hermano 2

0 = Masculino
1 = Femenino

23) Estudios y/o trabajo del hermano 2

0 = No tiene o fallecido
1 = 1= No sabe
2 = Desempleo o estudiante de segundo nivel o menos
3 = No siempre tiene trabajo, estudiante universitario o bachiller
4 = Empleo Informal
5 = Labor Autónoma
6 = Empleo Estable

24) ¿Cómo es tu relación con tus compañeros?

0 = Muy mala
1 = Mala
2 = Regular
3 = Buena
4 = Muy buena

25) ¿Qué tan buena o mala es tu relación con tu familia?

0 = No tiene
1 = Mala
2 = Regular
3 = Buena
4 = Muy buena

26) ¿Qué tan buena o mala es tu relación con tu familia?

0 = Ninguna
1 = Actividades tendentes al sedentarismo
2 = Actividades varias
3 = Actividades deportivas y activas
4 = Actividades educativas y formativas

27) ¿Te agrada el curso de nivelación?

0 = Sin opinión
1 = No me gusta el curso de nivelación (explique porque)
2 = Si

28) ¿Que no te agrada del curso de nivelación?

0 = Sin opinión
1 = Tiempo
2 = Geográfico económico
3 = Factores académicos

29) ¿Estas conforme con la carrera que estás estudiando en este momento?

0 = Inconforme
1 = Poco conforme
2 = Conforme

30) ¿Qué haces cuando tienes problemas?

0 = Nada
1 = Aprender de ellos
2 = Afrontarlos

31) ¿Necesita ayuda para resolver sus problemas?

0 = No contesta
1 = Nunca
2 = Pocas veces
3 = Casi siempre
4 = Siempre

32) ¿A quién recurre en busca de ayuda de orientación en sus problemas?

0 = Otros especifique
1 = Amigos o Profesores
2 = Conyugue
3 = Padres
4 = Entidad Superior

33) ¿Edad?

0 = Menor de 17 años
1 = Entre 18 y 21 años
2 = Entre 22 y 25 años
3 = Entre 26 y 29 años
4 = Entre 30 y 34 años
5 = Entre 35 y 38 años
6 = Mayor de 38 años

34) ¿Le gusta la carrera que asigno en el SNNA?

0 = Nada
1 = Poco
2 = Mucho
3 = Totalmente de acuerdo

35) ¿El estudiante ha desertado?

0 = No desertado
1 = Desertado

Datos psicológicos - BC2

1) ¿Sueles ser tímido?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

2) ¿Sueles ser celoso?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

3) ¿Sueles ser frío (que no suele mostrar sus emociones)?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

4) ¿Sueles ser agresivo?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

5) ¿Sueles ser abierto (que estás dispuesto a participar en todo)?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

6) ¿Sueles ser reflexivo?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

7) ¿Sueles ser constante?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

8) ¿Sueles ser optimista?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

9) ¿Sueles ser impulsivo?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

10) ¿Sueles ser generoso?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

11) ¿Sueles ser inquieto?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

12) ¿Sueles cambiar de humor repentinamente?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

13) ¿Sueles ser dominante?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

14) ¿Sueles ser egoísta?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

15) ¿Sueles ser sumiso?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

16) ¿Sueles ser confiado en ti mismo?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

17) ¿Sueles ser egoísta?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

18) ¿Sueles ser imaginativo?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

19) ¿Sueles tener iniciativa?

0 = No responde
1 = No
2 = Poco
3 = Frecuente
4 = Mucho

20) Estudiante desertado

0 = No desertado
1 = Desertado

Datos médicos - BC3

1) Tienes alguna discapacidad

0 = No
1 = Si

2) ¿Sufres de manos y pies hinchados?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

3) ¿Sufres de dolores de barriga?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

4) ¿Sufres de dolores de cabeza?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

5) ¿Sufres de Repugnancia y/o vómitos?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

6) ¿Sufres de pérdida de equilibrio?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

7) ¿Sufres fatiga y agotamiento?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

8) ¿Sufres de dificultades para dormir?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

9) ¿Sufres de pesadillas y/o terrores nocturnos?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

10) ¿Sufres de incontinencia (orina o heces)?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

11) ¿Sufres de tartamudeo?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

12) ¿Sufres de miedo intenso a algunas cosas?

0 = Nunca
1 = Antes
2 = A veces
3 = Frecuente
4 = Muy frecuente

13) ¿Qué hace usted para mantenerse en buenas condiciones físicas?

0 = Ninguna Actividad
1 = Cuidar la alimentación
2 = Controlar el estrés
3 = Salir a Caminar
4 = Voy al Gimnasio
5 = Se ejercita en casa o realiza deporte

14) ¿Cuantas comidas consumes al día?

0 = Tres veces al día solo comida chatarra
1 = Menos de 3 veces al día
2 = Tres veces al día
3 = Más de tres comidas al día

15) ¿Cuántas horas duermes al día?

0 = Menos de 6 horas
1 = Entre 6 y 8 horas
2 = Más de 8 horas

16) Tomas siestas en el día

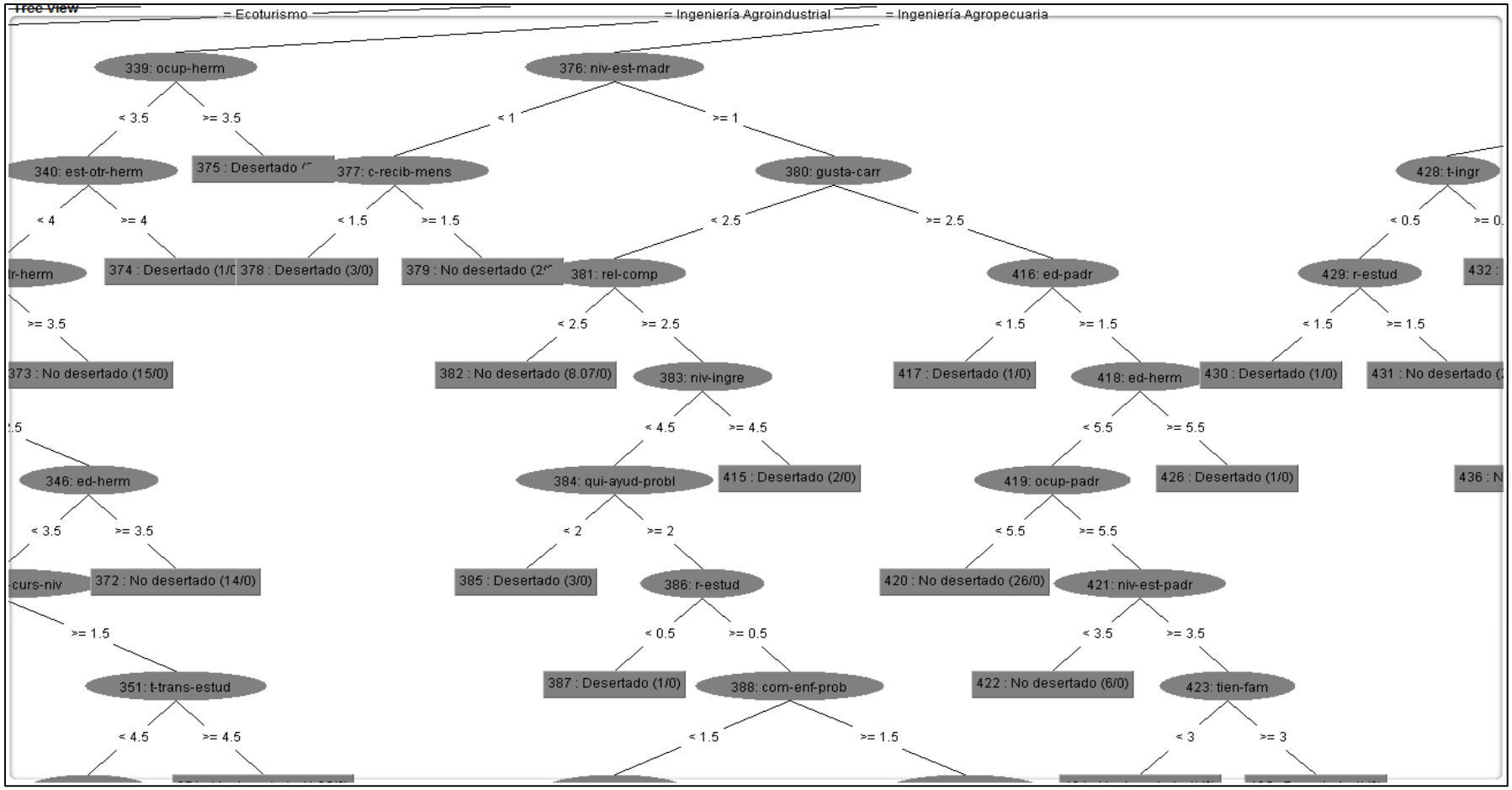
0 = No tomo siestas
1 = Por menos de 20 minutos
2 = Entre 20 y 40 minutos
3 = Más de 40 minutos

17) Estudiante desertado

0 = No desertado
1 = Desertado

FUENTE: INFORME DE DISCRETIZACIÓN.
ELABORADO POR: EL AUTOR.

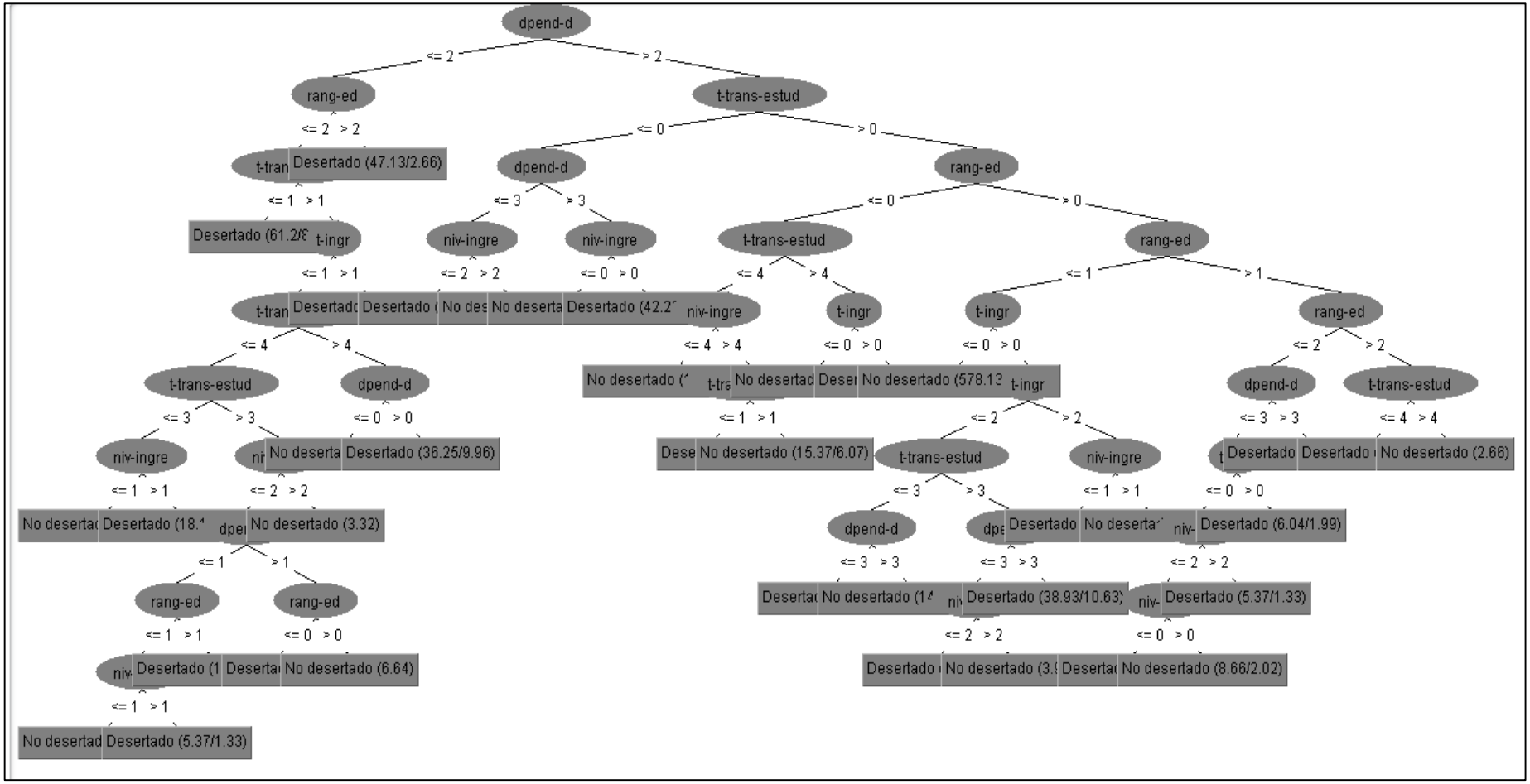
Anexo 7: Extracto de modelo Random Tree de BC1 (factores socio-económicos)



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

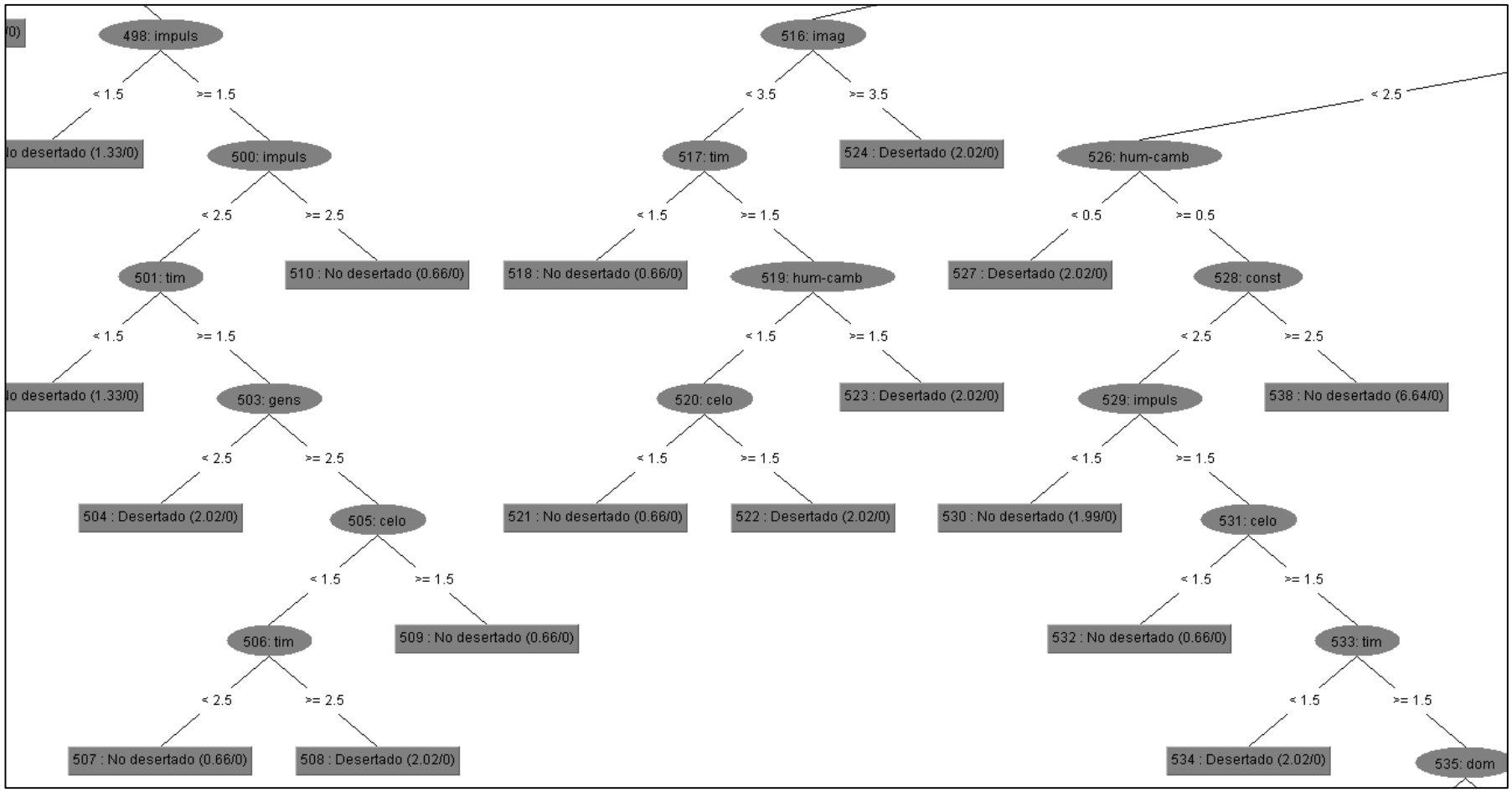
Anexo 8: Modelo de Random Tree obtenido de BC1 (factores socio económicos) posterior a selección de variables



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

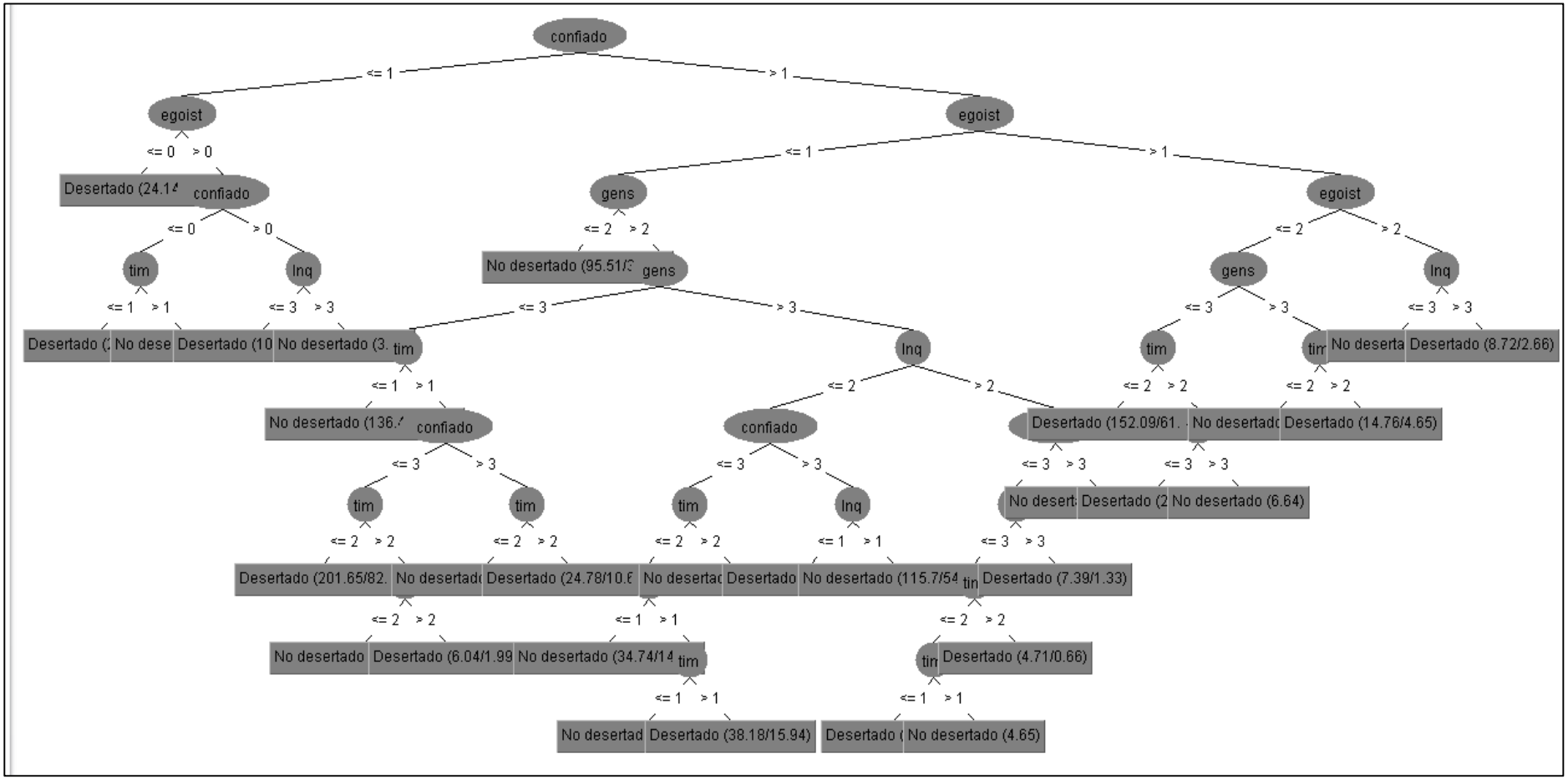
Anexo 9: Extracto del modelo Random Tree obtenido de BC2 (factores psicológicos)



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

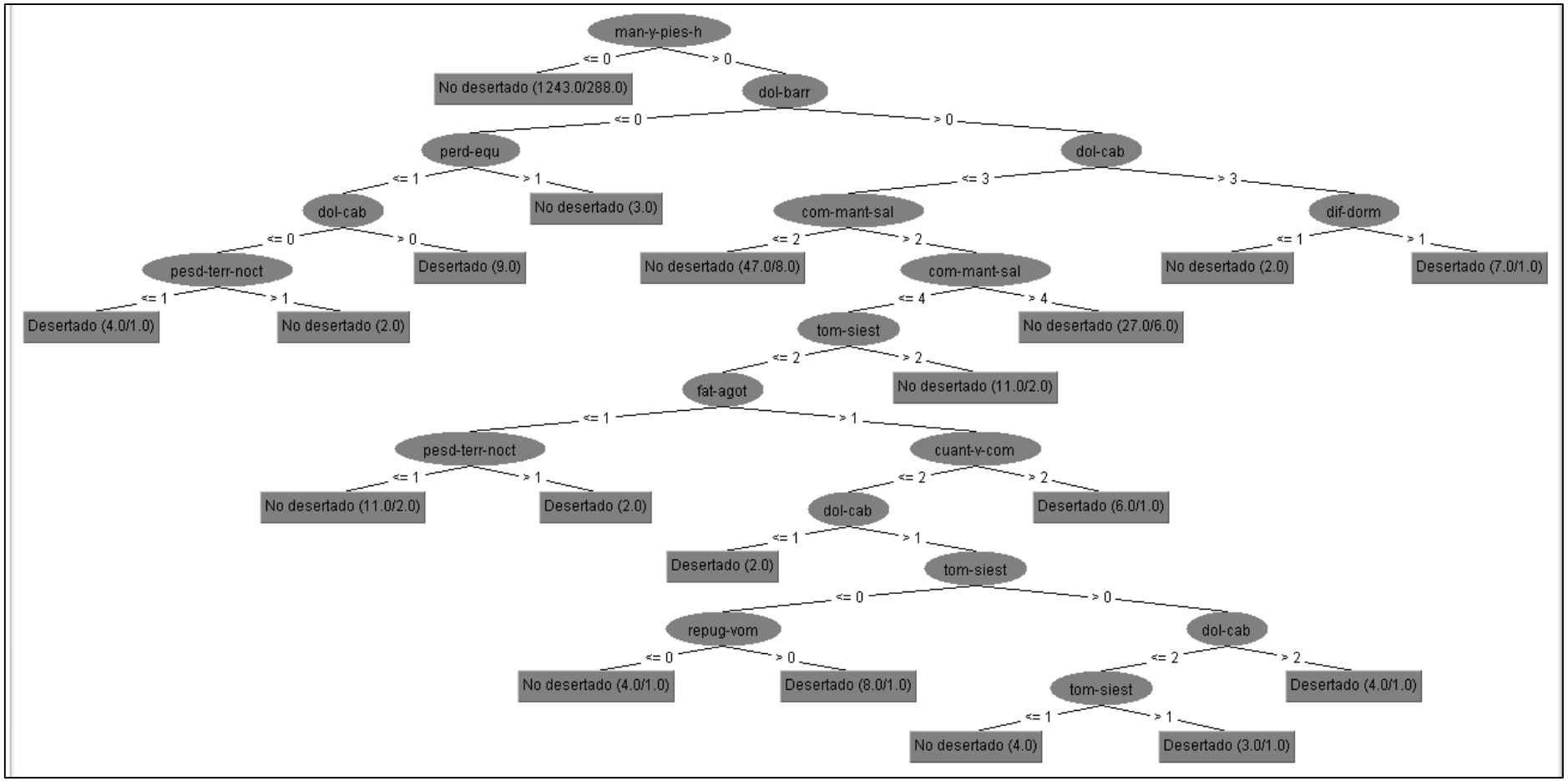
Anexo 10: Modelo de Random Tree obtenido de BC2 (factores psicológicos) posterior a selección de variables



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

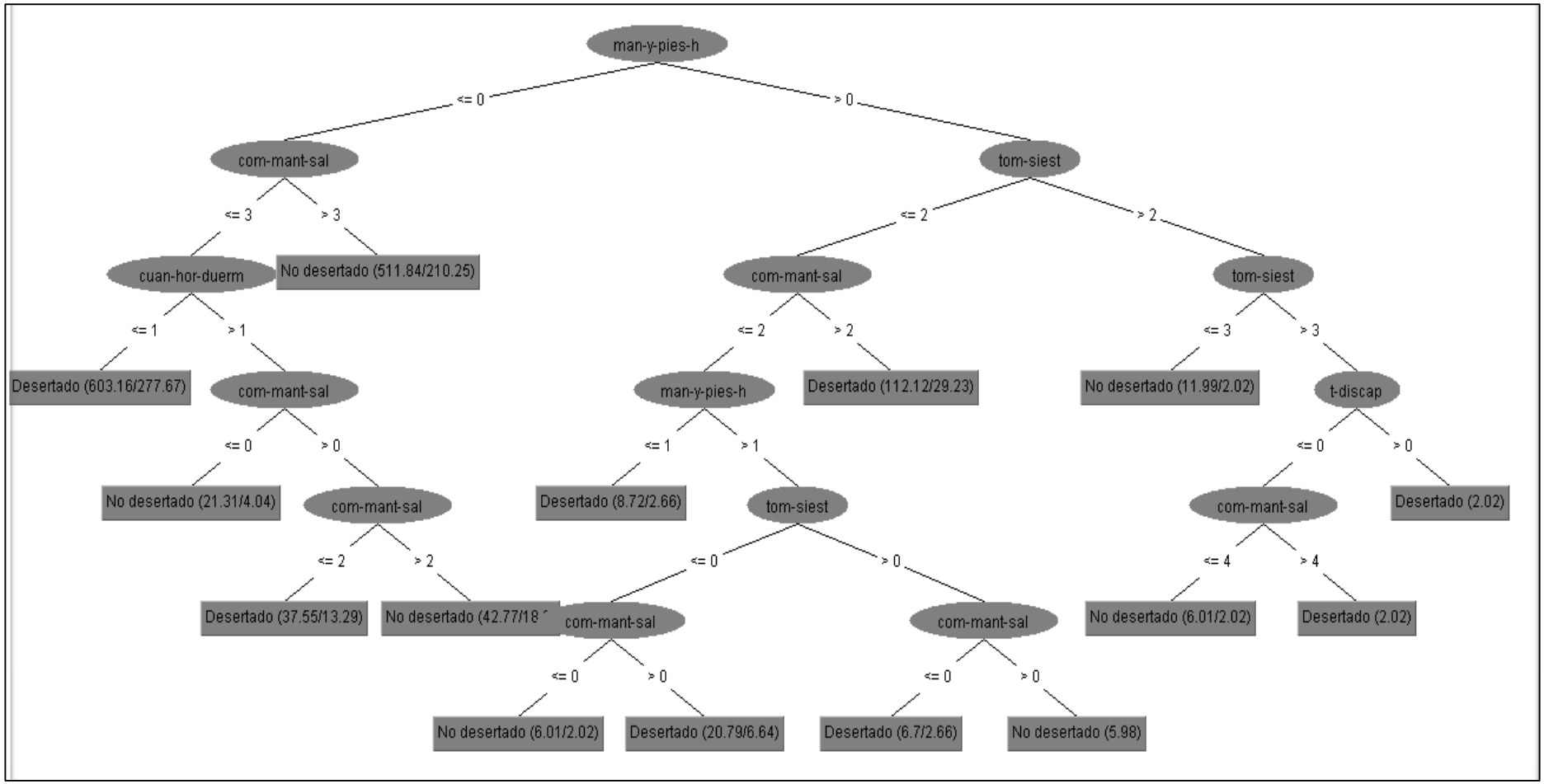
Anexo 11: Modelo de Random Tree obtenido de BC3 (factores médicos)



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.

Anexo 12:Modelo de Random Tree obtenido de BC3 (factores médicos) posterior a selección de variables



FUENTE: WEKA WORKBENCH.

ELABORADO POR: EL AUTOR.