Universidad de las Ciencias Informáticas

Facultad 6

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Aplicación de la técnica de Minería de Datos agrupamiento sobre el área de Gestión Académica de la Universidad de las Ciencias Informáticas.

Autora:

Yenei Martínez Garcia

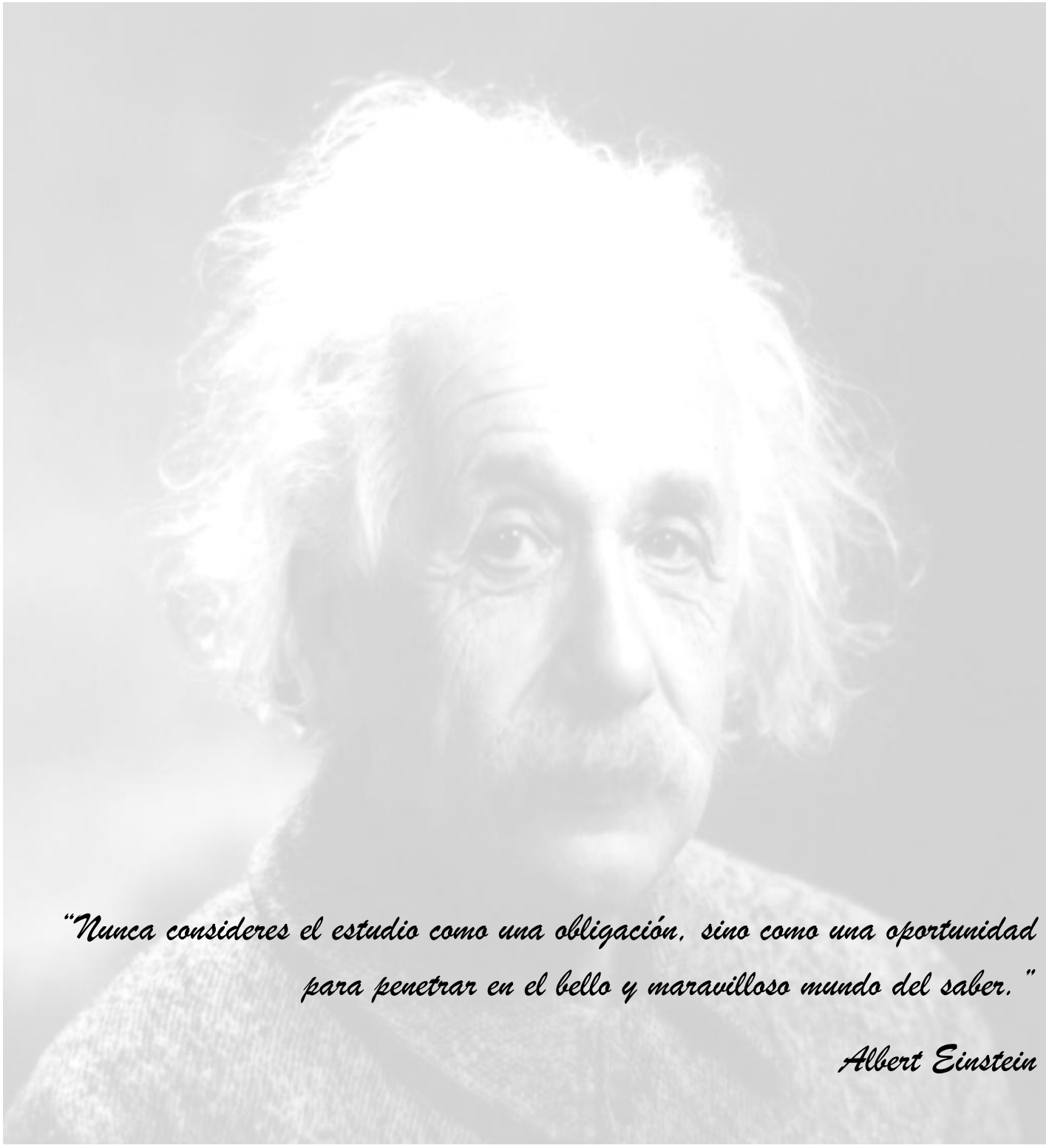
Tutor(es):

Ing. Yurima Ibañez Alfonso

Ing. Lazaro José Estupiñan Cutiño

La Habana, junio del 2012

¨Año 54 de la Revolución¨



"Nunca consideres el estudio como una obligación, sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber."

Albert Einstein

DECLARACIÓN DE AUTORÍA

Declaro ser autora de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Yenei Martínez Garcia

_____

Firma del Autor

Yurima Ibañez Alfonso

_____

Firma del Tutor

Lazaro José Estupiñan Cutiño

_____

Firma del Tutor

DATOS DE CONTACTO

Tutor:

Ing. Yurima Ibañez Alfonso.

Correo Electrónico: yibanez@uci.cu

Ingeniero Informático, Universidad de Ciencias Informáticas, 2007.

Categoría Docente: Instructor.

Tutor:

Ing. Lazaro José Estupiñan Cutiño.

Correo Electrónico: ljestupinan@uci.cu

Ingeniero Informático, Universidad de Ciencias Informáticas, 2011.

Categoría Docente: Instructor.

Después de muchos años de estudio y esfuerzo, hoy me gradúo de Ingeniera en Ciencias informáticas. Para hacer realidad este sueño muchas personas me ayudaron, apoyaron y depositaron su confianza en mí. Hoy tengo la oportunidad de hacerles saber cuan agradecida estoy.

A mis padres por todo su esfuerzo y sacrificio, por su cariño, su apoyo, gracias por estar siempre cuando los necesite. Quiero que sepan que los quiero y los admiro mucho y este logro es de ustedes también, nunca le podré estar lo suficientemente agradecida por todo lo que ha hecho por mí.

A mis abuelos por su apoyo incondicional, el siempre estar al tanto de cómo iban las cosas con los estudios, por cuidarme, quererme mucho y sentirse orgullosos de mi.

A mi hermana, mi tío Ernesto y mi primo Pocholo por estar conmigo siempre, escucharme y apoyarme en todo momento.

A mi madrina pues he sido para ella la hija que nunca tuvo, gracias por toda la ayuda, el apoyo y la confianza que me has brindado, te estoy muy agradecida.

A mi tito, mi ne por estar siempre conmigo, por tu amor, tu comprensión, por darme tu apoyo incondicional, siendo más que mi novio mi amigo. Gracias por estar ahí cuando lo necesité y no dejarme sola en ningún momento, aconsejándome y dándome toda tu ayuda en especial en estos últimos meses que han sido muy duros para mí.

Agradecer a dos personas que son muy especiales para mí a Dailyn y a Yesi, amigas en las buenas y en las malas, mis confidentes. Gracias por su apoyo y haber compartido alegrías y tristezas durante los años de estancia en la escuela.

A mis suegros por su preocupación, sus consejos, por apoyarme siempre y quererme como una hija más.

Quiero agradecer de manera especial a mis tutores, por su esfuerzo y por ser mis guías en la realización de mi trabajo de diploma.

Agradecerle también a las amistades que a lo largo de estos 5 años han estado ahí conmigo: Lisdany, Maylen, Gladys, Pompa, Yadiris, Yuri, Yaime, Karla, Yoandy, Ruso, Julio, en fin a todos muchas gracias….

A mis padres, mis abuelos y a todas las personas que me quieren y siempre confiaron en mí brindándome su apoyo para que viera realizado mi sueño, quiero regalarles este momento y honrarlos por tanto amor y dedicación. Los quiero mucho.

RESUMEN

La Minería de Datos se ha convertido en una herramienta muy poderosa que es utilizada para la extracción de conocimiento en grandes bases de datos. Sus aplicaciones son múltiples: en la medicina, la biología, en el sector bancario, en el marketing docente.

La presente Investigación describe todo el proceso realizado para la obtención de un modelo que permita efectuar el diagnóstico y detección de las causas de bajas docentes en la Universidad de las Ciencias Informáticas, haciendo uso de las técnicas y algoritmos de la MD. Para la extracción del conocimiento se utiliza la base de datos: Gestión Académica de la Universidad de las Ciencias Informáticas y la investigación se guía por la metodología CRISP-DM 1.0 que se apoya en la herramienta de libre distribución Weka 3.6.0.

Palabras claves: CRISP-DM, KDD, Minería de Datos, Weka.

TABLA DE CONTENIDO.

Índice de Tablas.

Índice de Imágenes.

Introducción.

La Revolución cubana, desde sus inicios, tuvo como objetivo fundamental solucionar los problemas que existían antes de su triunfo en el sector de la educación; tomando medidas inmediatas para erradicar el analfabetismo, garantizar la extensión de los servicios educacionales y la reorganización del Ministerio de Educación.

Los logros en la educación, han despertado el interés y la admiración en todo el mundo por el compromiso que existe con la educación primaria, la municipalización de la enseñanza superior, el establecimiento de la educación especial, así como otros muchos programas para mejorar la calidad de la educación del pueblo.

La enseñanza superior es vital, pues tiene el compromiso de crear profesionales que tengan una formación integral y de calidad, que sean los impulsores para fomentar el desarrollo del país.

Desde hace algunos años, el gobierno cubano ha realizado un gran esfuerzo, para promover el uso de las nuevas tecnologías en los centros educacionales del país, con el objetivo de profundizar en la formación integral de los jóvenes, aprovechando los avances de la Tecnología, la Información y las Comunicaciones con fines educacionales.

La Universidad de las Ciencias Informáticas (UCI), es un centro de estudios universitarios, nacido como un proyecto de la revolución, con el objetivo de formar jóvenes ingenieros en Ciencias Informáticas capaces de informatizar todo el país y contribuir al desarrollo de software. La UCI, desde sus inicios, se ha destacado por tener altos valores de promoción y retención académica, elevada calidad en la formación cultural, deportiva así como de valores humanitarios y revolucionarios de sus estudiantes. En esta institución se implementa un modelo de formación, en el cual se integra, la formación, producción e investigación, por lo que los estudiantes desde el tercer año de la carrera, son vinculados a los proyectos productivos que desarrolla la universidad con otras entidades dentro y fuera del país. Sin embargo, en los últimos años, en la institución han aumentado en gran medida las bajas de los estudiantes, por varias causas, como son: bajo rendimiento académico, por decisión propia y en un buen por ciento abandonan el centro por sanción al incurrir en indisciplinas graves, constituyendo estos elementos una fuerte derrota para la universidad.

Actualmente no existe un estudio que contribuya a determinar las principales causas por las cuales los estudiantes son dados de baja en la Universidad. Tampoco se cuenta con un componente inteligente que arroje patrones de comportamiento mediante los cuales se puedan tomar decisiones importantes y estratégicas para el diagnóstico y prevención de las bajas de los estudiantes, aprovechando la

información almacenada de cada uno de ellos, por lo que se dificulta el análisis rápido y efectivo por parte de la dirección de la universidad para prevenir este suceso.

Por la situación anteriormente descrita se define como **problema de la investigación** ¿Cómo contribuir al proceso de toma de decisiones sobre el área de gestión académica de la UCI a partir de los datos de los estudiantes? Teniendo como **objeto de estudio** la minería de datos y como **campo de acción**, la técnica agrupamiento de minería de datos sobre el área de gestión académica de la UCI.

Para dar solución al problema de la investigación se define como **objetivo general:** Definir un modelo de conocimiento empleando la técnica de minería de datos agrupamiento, el cual tribute al diagnóstico y detección de las causas de bajas docentes en la Universidad de las Ciencias Informáticas apoyando el proceso de toma de decisiones.

En correspondencia con el objetivo general se plantean los siguientes **objetivos específicos**:

- ✓ Fundamentar los conceptos relacionados con la minería de datos, así como sus técnicas, metodologías y herramientas.
- ✓ Aplicar la técnica de minería de datos, agrupamiento, al conjunto de datos seleccionados para el estudio.
- ✓ Interpretar los resultados del modelo de conocimiento obtenido con la aplicación de la técnica de minería de datos agrupamiento.

Para dar cumplimiento a los objetivos específicos se trazan las siguientes **tareas de la investigación:**

- ✓ Elaboración del diseño teórico metodológico de la investigación.
- ✓ Determinación del objetivo a cumplir con la aplicación de la técnica de minería de datos.
- ✓ Selección de los datos que van a ser utilizados para el análisis.
- ✓ Limpieza de los datos seleccionados.
- ✓ Integración de los datos seleccionados.
- ✓ Selección de la técnica de minería de datos a utilizar para la obtención del modelo de conocimiento.
- ✓ Construcción del modelo de conocimiento.
- ✓ Interpretación del modelo de conocimiento.
- ✓ Evaluación de los resultados obtenidos

- ✓ Informe definitivo del producto.

Al concluir las tareas de la investigación se obtiene como resultado un modelo de conocimiento, del cual se puede obtener patrones de comportamiento, observados después de aplicada la técnica de minería de datos entre las principales variables del problema.

Estrategia de Investigación.

Para la realización de esta investigación, se empleó una estrategia descriptiva con el objetivo fundamental de profundizar en la teoría del planteamiento investigativo, describiendo el problema, mostrando lo más notable y significativo del mismo para llegar a los resultados esperados. A continuación se describen los métodos investigativos aplicados en la investigación para conseguir una solución sintetizada de la misma:

Métodos Teóricos.

- ✓ **Análisis Histórico-Lógico:** Durante la investigación se pone de manifiesto, en la realización de estudios de las causas que originaron el problema, así como para el análisis de las técnicas y algoritmos existentes en la actualidad para la minería de datos.
- ✓ **Analítico-Sintético:** A través de este método se realiza el análisis de las distintas fuentes bibliográficas, para el estudio de las diferentes técnicas y herramientas aplicadas en la minería de datos para la obtención de patrones y plantear la solución propuesta.

Métodos Empíricos.

- ✓ **Entrevista:** Este método fue aplicado para validar los datos necesarios para la solución del problema de la investigación y para la obtención de un conocimiento manejando términos, principales causas y razones por la que el estudiante es dado de baja.

El trabajo de diploma se ha estructurado en tres capítulos:

Capítulo 1: Fundamentos Teóricos: En este capítulo se abordan definiciones y conceptos importantes de la minería de datos especificándose características, ventajas y desventajas de las técnicas, herramientas y metodologías que son utilizadas para la realización de un proyecto de minería de datos con el objetivo de seleccionar las más factibles y dar solución al problema expuesto en la presente investigación.

Capítulo 2: Propuesta de Solución: En este capítulo se define la arquitectura del proyecto de minería de datos, así como los objetivos del negocio, se analizan los datos necesarios para solucionar el

problema planteado y se describen las transformaciones realizadas sobre los datos seleccionados, para obtener la vista minable que permite generar el modelo de conocimiento.

Capítulo 3: Descripción de los resultados: Se describe la aplicación de la técnica y el algoritmo seleccionado para la generación del modelo, el cual es interpretado, obteniendo un grupo de patrones que describen el comportamiento de las principales causas de baja de la Universidad de las Ciencias Informáticas.

Capítulo 1: Fundamentos Teóricos.

Resumen.

En este capítulo se abordan definiciones y conceptos importantes de la minería de datos especificándose características, ventajas y desventajas de las técnicas, herramientas y metodologías que son utilizadas para la realización de un proyecto de minería de datos con el objetivo de seleccionar las más factibles y dar solución al problema expuesto en la presente investigación.

Introducción.

Con el desarrollo tecnológico alcanzado en los últimos años el volumen y variedad de la información que se encuentra almacenada en diferentes fuentes de datos también ha tenido un aumento impresionante, por lo que resulta imprescindible convertir los mismos en conocimiento, especialmente en las grandes organizaciones y proyectos científicos, pues genera un incremento en la gestión de la información así como el descubrimiento de patrones de comportamiento a partir de los datos almacenados.

Generalmente el estudio de los datos, se hacía mediante un proceso manual o semiautomático: una o más personas que tenían conocimiento de los mismos y con la ayuda de técnicas estadísticas generaban informes. Desde hace unos años se comienza a utilizar un nuevo sistema para apoyar la toma de decisiones: los Almacenes de Datos que son *"*un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones, permiten la transformación de los datos y brinda información desde diferentes perspectivas*"* (1), pero tiene como inconveniente que no generan patrones hipotéticos, sino que solo comprueba la existencia de los mismos. Para dar solución a este problema se emplea la Minería de Datos (MD) sobre los datos contenidos en los almacenes.

La MD facilita la extracción de información, encontrar relaciones o patrones, permitiendo la creación de modelos, es decir, representaciones de la realidad que apoyen la toma de decisiones a partir de los resultados alcanzados.

1.1 Proceso de extracción del conocimiento en bases de datos.

Después de haber realizado un estudio en diferentes fuentes bibliográficas, todas coinciden que el descubrimiento de conocimientos en base de datos o KDD por sus siglas en inglés (Knowledge Discovery in Databases) se define como: "proceso no trivial de identificación en los datos, de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles" (2). En fin se trata de obtener

conocimiento valioso, desconocido, eficiente, a partir de la aplicación de un grupo de algoritmos de aprendizaje automático.

1.1.1 Pasos para realizar KDD.

Para realizar el proceso de descubrimiento del conocimiento sobre un conjunto de datos es necesario realizar un grupo de pasos, a continuación se encuentran estructurados por diferentes etapas.

- ✓ Comprensión de los objetivos del usuario final.
- ✓ Creación del conjunto de datos: consiste en la selección de las variables candidatas de las cuales se va a obtener conocimiento.
- ✓ Limpieza y pre procesamiento de los datos: Se realiza un grupo de transformaciones sobre los datos, para manejar problemas como valores nulos o la falta de homogeneidad de los mismos.
- ✓ Reducción de los datos y proyección: Seleccionar las variables más significativas en dependencia del objetivo del proceso.
- ✓ Elegir la tarea de Minería de Datos a aplicar en el proceso.
- ✓ Elección del algoritmo(s) de Minería de Datos.
- ✓ Interpretación de los patrones encontrados. Si los resultados no son los esperados se vuelven a efectuar algunos de los pasos anteriores.
- ✓ Consolidación del conocimiento descubierto: Incorporar el conocimiento adquirido al funcionamiento del sistema, de lo contrario se almacena esta información.

1.1.2 Metas de KDD.

Durante el desarrollo del proceso de extracción del conocimiento, es de vital importancia realizar un procesamiento automático de grandes cantidades de datos en crudo y de diferentes fuentes. Con la finalidad de obtener patrones que sean significativos, relevantes y que tengan un impacto en el resultado esperado y de este modo presentarlos como un conocimiento adecuado, para satisfacer las necesidades del usuario, al culminar el proceso de KDD.

1.2 Minería de Datos.

La idea de MD no es nueva, desde los años sesenta se manejaban diferentes términos como: datos de la pesca, minería de datos o datos arqueológicos. En la actualidad, esta tecnología tiene un gran desarrollo y ha sido motivo de debate entre personas que pertenecen al ámbito académico y al de los negocios. Existen varias definiciones acerca de este término, a continuación se citan algunas de ellas:

"La minería de datos es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma autorizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" (3).

"La Minería de datos es el conjunto de metodologías y herramientas que permiten extraer el conocimiento útil (patrones de comportamiento, modo de operación, información útil para descubrir fallos, tendencias, etc.) para la ayuda en la toma de decisiones, partiendo de grandes cantidades de datos" (4).

"Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (5).

Luego de haber estudiado los diferentes conceptos citados, se concluye que la MD es el proceso de extraer conocimiento útil, previamente desconocido, a partir de grandes volúmenes de datos.

1.1.3 Características y objetivos de la Minería de Datos.

La MD es una tecnología que es de gran utilidad en el proceso de extracción de conocimiento, la aplicación de los patrones descubiertos apoyan el proceso de toma de decisiones. Por tanto, la minería de datos tiene dos desafíos fundamentales: el trabajo con grandes volúmenes de datos que pueden tener errores y la obtención de conocimiento novedoso y útil. A continuación se muestra un grupo de características que la describen:

- ✓ Es una fase del proceso de extracción de conocimiento oculto en grandes bases de datos.
- ✓ Presenta una arquitectura cliente-servidor.
- ✓ Las herramientas son de gran ayuda en la extracción del conocimiento oculto en grandes fuentes de datos.
- ✓ Se obtienen resultados valiosos que pueden ser útiles e inesperados acaparando todo el interés por parte del minero.
- ✓ Según la técnica de la MD seleccionada, se pueden obtener diferentes modelos de conocimiento, que pueden ser: Asociación, Secuencia, Clasificación, Agrupamiento y Pronóstico.

1.1.4 Alcance de la Minería de Datos.

La minería de datos, se deriva de buscar valiosa información de negocios en grandes bases de datos. Este proceso requiere examinar una inmensa cantidad de material e investigar inteligentemente, hasta encontrar exactamente donde residen los valores. Debido al tamaño que han adquirido en la actualidad las bases de datos y su calidad, la minería de datos es utilizada para generar conocimiento, que puede ser aplicado en diferentes campos, al proporcionar un grupo de facilidades como son:

- ✓ La predicción automatizada de tendencias y comportamientos: la minería de datos automatiza el proceso de buscar información predecible en bases de datos. Preguntas que tradicionalmente requerían un análisis manual exhaustivo, pueden ser contestadas rápidamente a partir del análisis de los datos.
- ✓ Descubrimiento de modelos de conocimiento: A partir del análisis rápido sobre los datos, se logra identificar modelos de conocimiento previamente desconocidos en los grandes volúmenes de información.

1.1.5 Aplicaciones de la Minería de Datos.

Por la potencialidad que ofrece la minería de datos, así como el auge que ha estado teniendo en los últimos tiempos es aplicada en grandes investigaciones científicas y en el mundo de los negocios. A continuación se mencionan algunas de las áreas a las que es aplicada esta tecnología:

Marketing:

- ✓ Para identificar patrones de compra de los clientes, entre los que se encuentran: el grado de interés por un producto, las características y el momento en que son adquiridos los mismos.
- ✓ Segmentación de clientes: con el objetivo de agrupar a los clientes según la similitud de sus características. Constituye una herramienta importante para el diseño de estrategias de marketing, para la realización de ofertas en dependencia del comportamiento de los compradores.
- ✓ Análisis de cestas de la compra: se trata de descubrir la relación que existen entre los diferentes productos, con el fin de determinar cuales se adquieren juntos y de esta manera realizar una distribución adecuada de los mismos.

Sector Bancario: Se emplea la minería de datos para detectar patrones que describan el uso fraudulento de las tarjetas de créditos, precisar el gasto de las mismas en un grupo de clientes

determinado y la identificación de reglas de mercado de valores, a partir de los datos que se encuentran almacenados en las fuentes.

Telecomunicaciones: A partir de la información almacenada de las llamadas realizadas por los clientes, se aplica la técnica agrupamiento, para detectar patrones de comportamiento, que permitan descubrir conductas fraudulentas.

Medicina: En el sector de la salud, existe un gran volumen de información almacenada sobre los pacientes, tales como: padecimientos, diagnósticos, tratamiento de enfermedades y pruebas médicas. Las técnicas de minería de datos, son aplicadas sobre estas fuentes de datos, para identificar terapias que han tenido una aplicación satisfactoria en diferentes enfermedades, poder relacionar síntomas, con el objetivo de encontrar tratamientos médicos, para ser aplicados a pacientes enfermos.

Marketing Educativo: Se realizan estudios para detallar las principales características de los estudiantes universitarios que causan baja de los centros educacionales, teniendo en cuenta las principales razones por la que son separados de las instituciones. Estos aspectos se relacionan de tal manera que la unión de ellos permite obtener una caracterización apropiada de los estudiantes. Las variables de análisis de algunos estudios son:

- ✓ Datos demográficos.
- ✓ Entorno familiar.
- ✓ Características socio-económica.
- ✓ Notas de las asignaturas.
- ✓ Las relaciones con sus compañeros o pareja.
- ✓ Empleo del tiempo libre.

1.1.6 Técnicas de la Minería de Datos.

Una técnica constituye el enfoque conceptual para extraer la información de los datos y en general es implementada por varios algoritmos. Cada algoritmo representa, en la práctica, la manera de desarrollar una determinada técnica paso a paso, de forma tal que es preciso un entendimiento de alto nivel de los algoritmos para saber cuál es la técnica más apropiada para cada problema. Asimismo es preciso entender los parámetros y las características de los algoritmos para preparar los datos a analizar. (6)

Las técnicas de la MD provienen de la Inteligencia Artificial y de la Estadística, se aplican sobre un conjunto de datos y posibilitan la obtención de patrones (Imagen 1).

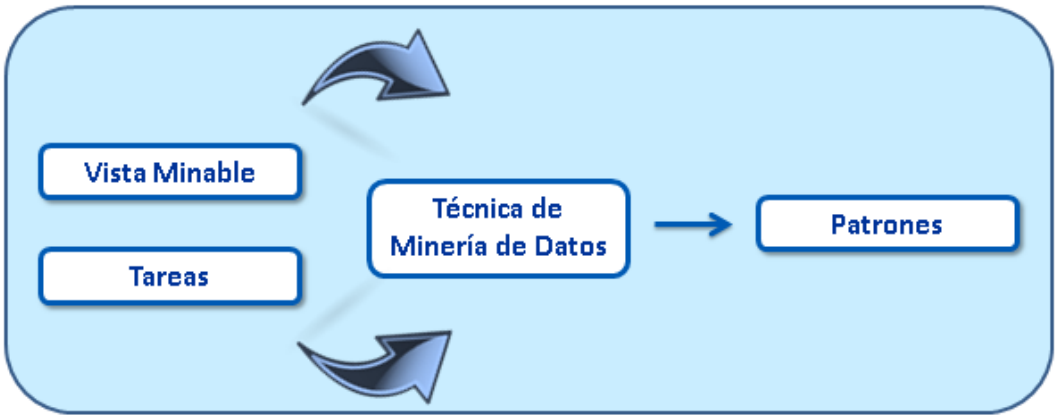


Imagen 1: Proceso de Minería de Datos.

Las técnicas que utiliza la MD se clasifican en:

- ✓ Supervisadas o predictivas: a partir de un conjunto de ejemplos, denominados de entrenamiento de un cierto dominio, se pueden construir criterios para determinar el valor del atributo clase en un ejemplo cualquiera del dominio. Esos criterios están basados en los valores de uno o varios de los otros pares (atributo; valor) que intervienen en la definición de los ejemplos. Es sencillo transmitir esa idea al caso en el que el atributo que juega el papel de la clase sea uno cualquiera o con más de dos valores (7).
- ✓ No supervisadas o del descubrimiento del conocimiento: se aborda el aprendizaje sin supervisión, trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial clase. Este es el proceder de los sistemas que realizan agrupamiento conceptual y de los que se dice también que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es la de sintetizar conocimiento cualitativo o cuantitativo (7).

Entre las principales técnicas Supervisadas o predictivas se encuentran:

Redes Neuronales: Constituyen una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro. Posteriormente se comprobó que tales modelos no eran del todo adecuados para describir el aprendizaje humano. Las redes de neuronas constituyen una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender complejos patrones y características dentro de los datos. Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y del pasado, y aplicando tal conocimiento a la resolución de problemas nuevos. Este aprendizaje se obtiene como

resultado del adiestramiento y éste permite la sencillez y la potencia de adaptación y evolución ante una realidad cambiante y muy dinámica. Una vez adiestradas las redes de neuronas pueden hacer previsiones, clasificaciones y segmentación (8).

Árboles de decisión: Un árbol de decisión puede interpretarse esencialmente como una serie de reglas compactadas para su representación en forma de árbol. Dado un conjunto de ejemplos, estructurados como vectores de pares ordenados atributo-valor, de acuerdo con el formato general en el aprendizaje inductivo a partir de ejemplos, el concepto que estos sistemas adquieren durante el proceso de aprendizaje consiste en un árbol.

El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado. La representación que se utiliza para las descripciones del concepto adquirido es el árbol de decisión, que consiste en una representación del conocimiento relativamente simple y que es una de las causas por la que los procedimientos utilizados en su aprendizaje son más sencillos que los de sistemas que utilizan lenguajes de representación más potentes, como redes semánticas, representaciones en lógica de primer orden, etc. No obstante, la potencia expresiva de los árboles de decisión es también menor que la de esos otros sistemas. El aprendizaje de árboles de decisión suele ser más robusto frente al ruido y conceptualmente sencillo, aunque los sistemas que han resultado del perfeccionamiento y de la evolución de los más antiguos se complican con los procesos que incorporan para ganar fiabilidad. La mayoría de los sistemas de aprendizaje de árboles suelen ser no incrementales, pero existe alguna excepción (1).

Clasificación: es la técnica más utilizada. En ella, cada instancia (o registro de la base de datos) pertenece a una clase, la cual se indica mediante el valor de un atributo que llamamos la clase de la instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase (1).

Algoritmos Genéticos (AG): Son otra técnica que tiene su inspiración en la Biología como las Redes de Neuronas. Estos algoritmos representan el modelado matemático de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más óptima en aras de la supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios de los genes o cromosomas, los AG hacen uso de las técnicas biológicas de reproducción (mutación y cruce) para ser utilizadas en todo tipo de problemas de búsqueda y optimización. (9)

En cuanto a las técnicas no supervisadas o del descubrimiento del conocimiento que más se utilizan tenemos:

Asociación: Esta técnica se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y concurrencias de eventos. Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos, destacándose el campo comercial, ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes. (10)

Agrupamiento o Clustering: es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes (11) de tal manera que se maximice la similitud entre los vectores de un mismo grupo y se minimice la similitud entre los grupos, además esta técnica puede ser combinada fácilmente con cualquier otra. El objetivo fundamental del agrupamiento es determinar el comportamiento de un nuevo vector, a partir de las características del mismo, se define a qué grupo pertenecerá y que acción podrá realizar.

Esta técnica es la seleccionada en la investigación para generar el modelo, ya que es la que más se ajusta para la solución del objetivo general, pues mediante su aplicación se pueden obtener un grupo de patrones que describen el comportamiento de los estudiantes que causan baja del centro. El algoritmo propuesto para esta técnica es el K-Medidas (Simple K-Means en inglés).

Descripción del algoritmo seleccionado.

Simple K-means: Es el algoritmo de agrupamiento más conocido y se encuentra clasificado como un algoritmo particional. Está basado en la minimización de la distancia interna entre los elementos de los K grupos definidos por el minero y esto constituye una de sus debilidades porque en ocasiones la cantidad de particiones no es las más óptima. Es importante destacar las características, pasos y ventajas de este algoritmo que fundamentan su selección para la investigación.

Características del Simple K-means.

- ✓ Escalabilidad: normalmente corre con pocos datos.
- ✓ Manejo de ruido: sensible a datos erróneos.
- ✓ Grupos de formas arbitrarias: basado en distancias numéricas.

- ✓ Requerimientos mínimos para especificar parámetros, como el número de grupos.

Pasos para aplicar el algoritmo Simple K-means.

1. Especificar la cantidad de grupos (K) que se van a crear y se selecciona por cada uno elementos de manera aleatoria que serán los centros de los grupos.
2. Cada una de las instancias, es asignada al grupo con características similares más cercano.
3. Se calcula el valor del centroide de todas sus clases y se toma como el centro de sus respectivos grupos.
4. Se repite el paso anterior, hasta que el valor de los centroides no varíe más, después de cada iteración.

Las ventajas que presenta este algoritmo es que permite el trabajo con atributos: numéricos, binarios, nominales, ordinales, funciona eficientemente con una gran cantidad de datos, los grupos pueden ser interpretados y convertir esta información en conocimiento.

1.3 Metodologías para guiar el proceso de Minería de Datos.

Las metodologías son un conjunto de procedimientos o métodos que indican qué hacer y cómo actuar cuando se quiere obtener una gama de objetivos que rigen una investigación científica. Desde el punto de vista de la informática se utiliza para estructurar, planear y controlar el proceso de desarrollo de un proyecto. Para guiar un proyecto de MD existe un grupo de metodologías entre las que se encuentra:

1.3.1 SEMMA.

SAS Institute desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a los cinco pasos básicos del proceso: Sample (Muestra), Explorer (Explorar), Modify (Modificar), Model (Modelo) y Assess (Evaluar) (12).

Pasos básicos.

Muestreo

Extracción de la población sobre la cual se va a aplicar el análisis. En ocasiones se trata de una muestra aleatoria pero puede ser también un subconjunto de datos del almacén de datos que cumplan unas condiciones determinadas. El objeto de trabajar con una muestra de la población en lugar de con toda ella, es simplificar el estudio y la disminución de la carga del proceso (12).

Exploración

Una vez determinada la población que sirve para la obtención del modelo se deberá determinar cuáles son las variables explicativas que van a servir como entradas al modelo. Para ello es importante hacer una exploración de la información disponible de la población que permita eliminar variables que no influyen y agrupar aquellas que presentan efectos similares. El objetivo es simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo (12).

Manipulación

Tratamiento realizado sobre los datos de forma previa al modelado, en base a la exploración realizada, de forma que se definan claramente las entradas del modelo a realizar (selección de variables explicativas, agrupación de variables similares) (12).

Modelado

Permite establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibilitan inferir el valor de las mismas con un nivel de confianza determinado (12).

Valoración del modelo

Después de todo el trabajo realizado, se comparan los modelos a partir de los cuales se obtienen patrones de negocio.

1.3.2 CRITIKAL

Desarrollada en el marco de un proyecto ESPRIT 22700 se caracteriza por su fuerte integración con el desarrollo del almacén de datos y no es de completa distribución libre. Los pasos que plantea son similares a los de la metodología CRISP-DM (12).

1.3.3 CRISP-DM

Procedimiento Industrial Estándar para realizar Minería de Datos (CRISP-DM por sus siglas en inglés, Cross-Industry Standard Process for Data Mining) fue concebido a fines de 1996 por tres compañías DaimlerChrysler (liderando aplicaciones de minería a negocios), SPSS (servicios de MD) y NCR (Compañía que se dedica al desarrollo de los almacenes de datos). Es una metodología de libre distribución que está basada en la experiencia práctica, de cómo las personas efectúan proyectos de MD. La misma, estructura un proyecto de minería de datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

Características de la Metodología CRISP-DM.

- ✓ La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): *fase, tarea genérica, tarea especializada e instancia de procesos* (13).
- ✓ En el nivel superior, el proceso de MD es organizado en seis fases; cada fase consta de varias tareas genéricas de segundo nivel, las cuales deben cubrir el proceso entero de minería de datos y todas las aplicaciones posibles; el modelo debe ser válido para acontecimientos normales y aún para desarrollos imprevistos como técnicas de modelado nuevo (13).
- ✓ El nivel de tarea especializada, es el lugar para describir cómo las acciones en las tareas genéricas deben ser realizadas en ciertas situaciones específicas. Además describe como una tarea se diferencia en distintas situaciones, cómo la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo (13).

Fases de la Metodología CRISP-DM.

- ✓ Comprensión del negocio: Se enfoca en la comprensión de los objetivos de proyecto desde una perspectiva del negocio y a partir de este conocimiento se define el problema de la MD de la investigación.
- ✓ Comprensión de los datos: Comienza con la recopilación de los datos iniciales, se realiza un análisis para detectar problemas con la calidad y veracidad de los mismos y finalmente definir las posibles variables seleccionadas para formar hipótesis en cuanto a la información oculta y desconocida.
- ✓ Preparación de datos: Se selecciona el conjunto de datos finales para la construcción del modelo partir de los datos iniciales. Las tareas de la fase muchas veces tienen que ser repetidas y en cualquier orden, estas pueden ser: la selección de tablas y atributos, así como la transformación y la limpieza de datos para las herramientas de modelado.
- ✓ Modelado: Se selecciona la técnica de modelado, se genera el modelo, se interpreta y se obtienen un grupo de patrones. Algunas técnicas tienen limitantes con el tipo de datos de los atributos y a veces es necesario volver a la fase de preparación de los datos.
- ✓ Evaluación: Después de generado uno o varios modelos y antes de proceder al despliegue final del proyecto, es importante evaluar y revisar cada uno de los pasos que

se ejecutaron para su creación, para demostrar que el modelo responde a los objetivos del negocio de la investigación. Al final de esta fase, se debe tomar una decisión con el uso de los resultados obtenidos después de la aplicación de la MD.

- ✓ Desarrollo: La creación del modelo no es generalmente el final del proyecto, el conocimiento obtenido debe ser organizado y presentado en el modo en el que el cliente pueda usarlo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa.

Se selecciona CRISP-DM como metodología de desarrollo a utilizar en el proceso de MD de esta investigación, esta elección está respalda por las siguientes ventajas:

- ✓ Por ser de libre distribución, puede trabajar con cualquier herramienta de modelado.
- ✓ Proporciona facilidades en la planificación, documentación y comunicación en los proyectos de minería de datos.
- ✓ Facilita y organiza el trabajo del minero a partir de especificar un grupo de tareas en cada una de las fases.

1.4 Herramienta para la transformación de los datos.

Pentaho Data Integration 4.2.0: es una de las herramientas utilizadas en el proceso de Extracción, Transformación y Carga (ETL), contiene un número de componentes que facilitan el modelado y la ejecución de transformaciones sobre flujos de datos. Es una de las soluciones ETL de código abierto más antiguas, desarrolladas y mejor valoradas del mercado. Dicha aplicación funciona sobre disímiles plataformas a través de un sistema que soporte Java 1.4 o alguna versión superior, cuenta con un lenguaje de programación Java para lograr conectarse a las bases de datos. Permite operar con los campos en el flujo de datos, renombrando valores, obteniendo nuevos campos en función de otros y realizando búsquedas auxiliares en bases de datos.

Se realiza el proceso de ETL utilizando la herramienta Pentaho Data Integration 4.2.0, porque aporta los componentes necesarios, para obtener los atributos que van a ser utilizados, desde la fuente de datos. A los cuales se les realizan las transformaciones siguientes: renombrar, limpiar, validar y finalmente integrar en una única tabla denominada Vista Minable, donde se encuentran las variables principales a partir de las cuales se va a obtener el modelo de conocimiento.

1.5 Herramientas para el modelado de los datos.

Para realizar la extracción de patrones a partir de un conjunto de datos iniciales, existe un grupo de herramientas, donde las más utilizadas son SPSS Clementine, Oracle Data Miner, SAS Enterprise Miner y Weka.

Clementine de SPSS: es un sistema integrado de MD que permite encontrar patrones en la información, para facilitar la toma de decisiones a los usuarios. Se centra en la integración de la MD con otros procesos y sistemas de negocio que ayuden a entregar conocimiento en un tiempo eficiente durante las operaciones de negocio diarias. La funcionalidad abierta de la MD en bases de datos que posee esta herramienta, permite que muchos de los procesos de la minería se realicen en entornos que mejoran tanto el rendimiento como el despliegue de los resultados de la misma. En su última versión, extiende las funcionalidades de la MD, al incluir un conjunto de reglas de evaluación automática, modelos de árboles de decisión y carga de resultados que se obtienen durante la aplicación de la minería en la base de datos.

Ventajas de Clementine

- ✓ Facilita el acceso, la preparación e integración de variables: numéricas, nominales y datos provenientes de páginas web.
- ✓ Utilizando las técnicas de MD disponibles en la herramienta, se construyen y validan modelos predictivos, para apoyar el proceso de toma de decisiones.
- ✓ Esta herramienta permite seleccionar campos, mostrar propiedades de los datos, encontrar relaciones e integrarlos para realizar el proceso de modelado.
- ✓ Tiene implementado múltiples Algoritmos de MD y herramientas de visualización.

SAS Enterprise Miner: pertenece a la compañía SAS: Análisis de Sistemas Estadísticos o por sus siglas en inglés (statistical analysis systems), es una solución de minería de datos que permite incorporar patrones inteligentes a los procesos de marketing, tanto operativos como estratégicos. El software de SAS, es un sistema de entrega de información que provee acceso transparente a cualquier fuente de datos, incluyendo archivos planos, archivos jerárquicos, y los más importantes manejadores de bases de datos relacionales. También incluye su propia base de datos de información para almacenar y manejar los datos, es decir, un "Almacén de Datos", soporta los principales protocolos de comunicación, cubre los cinco modelos de procesamiento cliente/servidor de acuerdo a Gartner Group y cumple con las doce reglas de OLAP. El sistema permite un grupo de aplicaciones, destacándose el análisis estadístico,

análisis de los datos para mejorar su calidad, administración de proyectos, se obtienen reportes, gráficas, despliegue de imágenes, etc.

Oracle Data Mining: (ODM) es una opción de Oracle Database 11g Enterprise Edition que permite producir información predictiva útil y crear aplicaciones con inteligencia de negocio integrada. La funcionalidad de extracción inteligente de datos incorporada en Oracle Database 11g, permite a los clientes buscar patrones y conocimientos ocultos en sus datos. Los desarrolladores de aplicaciones pueden automatizar enseguida el descubrimiento y la distribución de nueva inteligencia de negocio, predicciones, patrones y hallazgos en toda la organización (14).

Este conjunto de herramientas de modelado, tiene como desventaja que son privativas, por lo que el uso de las mismas se dificulta teniendo en cuenta el costo, el soporte exclusivo del propietario y la privación o restricción de derechos y libertades de forma general.

Weka: Es una extensa colección de algoritmos de máquinas de conocimiento, desarrollados por la universidad de Waikato (Nueva Zelanda) e implementados en Java. Esta herramienta permite realizar un grupo de transformaciones necesarias sobre los datos, trae implementada un grupo de tareas de la MD: clasificación, regresión, agrupamiento, asociación y visualización. Esta herramienta admite añadir nuevas funcionalidades y brinda la posibilidad de modificar su código, además, puesto que está programado en Java, es independiente de la arquitectura y la herramienta puede ser utilizada en cualquier ordenador que tenga instalada la máquina virtual de java.

Características de Weka:

- ✓ Contiene un grupo de herramientas que permiten el análisis de los datos.
- ✓ Tiene implementados algoritmos de clasificación, agrupamiento y reglas de asociación.
- ✓ Al cargar los datos, Weka realiza un análisis de los mismos y obtiene las mejores variables para obtener el modelo de conocimiento.
- ✓ La licencia de Weka es GPL, lo que significa que este programa es de libre distribución y difusión.

La herramienta seleccionada como entorno para la aplicación de la tarea de MD es Weka en su versión 3.6.0 puesto que proporciona una serie de ventajas para la generación del modelo:

- ✓ Tiene implementado diferentes algoritmos de la tarea agrupamiento entre ellos el Simple K-means.
- ✓ Es de libre distribución.

- ✓ Es multiplataforma.

1.6 Conclusiones

En este capítulo se definen los conceptos referentes a: descubrimiento de conocimientos en bases de datos y de minería de datos, fundamentos teóricos necesarios para la solución del problema de investigación propuesto. Se selecciona como metodología para guiar el proyecto de minería de datos, la metodología CRISP-DM, así como la técnica de minería de datos agrupamiento y el algoritmo Simple K-means, para extraer conocimiento de la fuente de datos del área de gestión académica de la UCI. Se propone la herramienta Pentaho Data Integration en su versión 4.2.0, para realizar el proceso de Extracción, Transformación y Carga sobre los datos y Weka en su versión 3.6.0 como entorno para la aplicación de la técnica de minería de datos: agrupamiento.

Capítulo 2: Propuesta de Solución.

Introducción.

En este capítulo se define la arquitectura del proyecto de minería de datos, así como los objetivos del negocio, se analizan los datos necesarios para solucionar el problema planteado y se describen las transformaciones realizadas sobre los datos seleccionados, para obtener la vista minable que permite generar el modelo de conocimiento.

Es importante señalar que los datos seleccionados para la aplicación de la MD corresponden a la Base de Datos del área de gestión académica de la UCI.

2.1 Arquitectura de un proyecto de Minería de Datos.

La arquitectura empleada consta de tres subsistemas y tres niveles. Cada subsistema es ubicado en un nivel. A continuación se detalla cada subsistema presente en la arquitectura propuesta:

El subsistema de integración se abastece de la fuente de datos del área de gestión académica de la UCI y se encarga de integrar, estandarizar y limpiar los datos que serán extraídos, utilizando la herramienta Pentaho Data Integration, los cuales serán empleados para la generación del modelo de conocimiento. Solo accederán a este subsistema los clientes que administrarán el proceso de integración mediante el protocolo de conexión TCP/IP.

El subsistema de almacenamiento: almacena los datos que son tratados por el subsistema de integración. Solo accederán a este subsistema los clientes que administrarán el proceso de integración mediante el protocolo de conexión TCP/IP.

El subsistema de visualización: consulta los datos que están en el subsistema de almacenamiento mediante la herramienta Weka. A dicho subsistema acceden los distintos clientes para obtener los modelos deseados.

A continuación se presenta una imagen que muestra la arquitectura antes descrita.

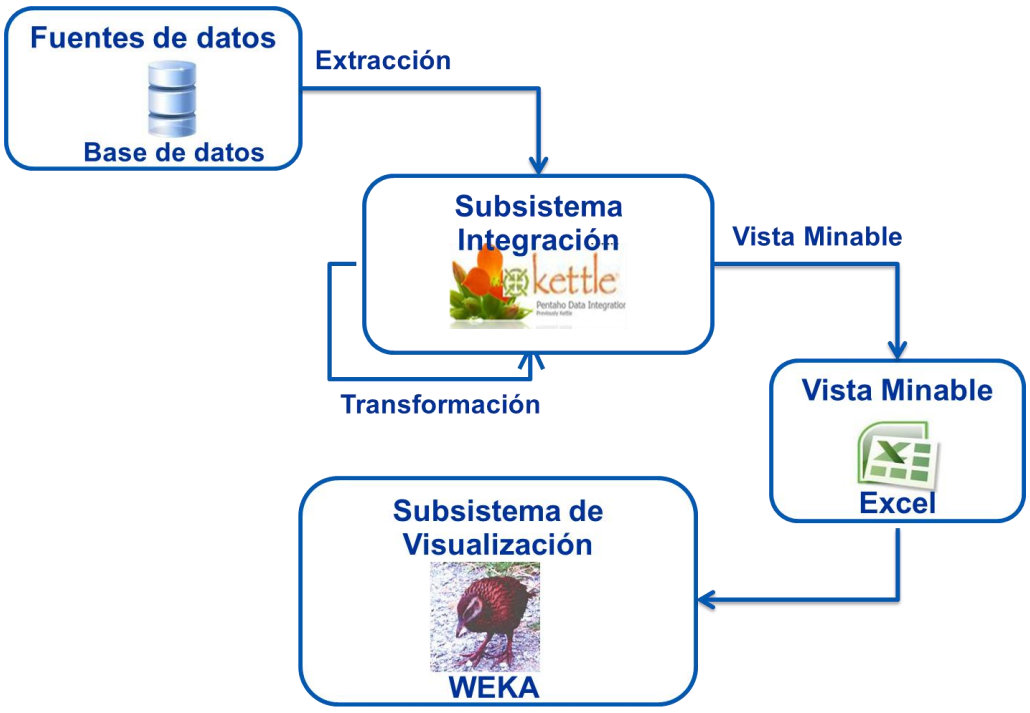


Imagen 2: Arquitectura del proyecto de minería de datos.

2.2 Fases de la metodología CRISP-DM.

2.2.1 Comprensión del negocio.

- **Determinación de los objetivos de negocio.**

En varias entrevistas realizadas al cliente, el mismo mostró interés en todo momento por el análisis de las causas de las bajas de los estudiantes del centro, para poder trazar alguna estrategia y evitar este suceso, por lo que se define como objetivo del negocio: analizar las causas por las que los estudiantes, son dados de baja de la Universidad de las Ciencias Informáticas.

- **Criterios de éxito de negocio.**

Los criterios para lograr el éxito de la investigación son los siguientes:

- ✓ Obtener un modelo de conocimiento y comprobar que sea el correcto.
- ✓ Hacer uso de la herramienta Weka para la obtención del modelo.
- ✓ Aplicar la MD realizando los pasos que se indican en la metodología.

- **Evaluación de la situación.**

Teniendo en cuenta el desarrollo de la presente investigación, se decide comenzar evaluando la situación existente en el entorno, que puede influir en el éxito de la misma; por tanto se definen un conjunto de medidas de seguridad:

- ✓ Restringir el acceso a la base de datos del personal no autorizado.
- ✓ Proteger los recursos computacionales destinados a la realización de la investigación.
- ✓ Documentar cada una de las fases de la metodología.
- ✓ La investigación será entregada en formato digital.

- **Determinación de los objetivos de la Minería de Datos.**

Teniendo en cuenta el objetivo general propuesto en el negocio y el problema a resolver con el desarrollo de este trabajo de diploma se define como objetivo o fin de la aplicación de la minería de datos sobre el Sistema de Gestión Académica de la UCI:

Obtener un modelo, empleando la técnica de minería de datos, agrupamiento, mediante el cual se genere un conjunto de patrones de comportamiento, que describan las causas de bajas docentes en la Universidad de las Ciencias Informáticas.

2.2.2 Comprensión de datos.

- **Recolección de datos iniciales.**

Los datos provienen de una base de datos que está montada sobre el gestor PostgreSQL, es un sistema de bases de datos objeto-relacional, que contiene los datos del área de gestión académica de la UCI. Los mismos fueron seleccionados tras un análisis de las variables más significativas que aporten conocimiento durante el proceso de modelado.

Para recolectar los datos necesarios que sean capaces de aportar al proceso de diagnóstico y detección de las causas de bajas docentes, se analizaron las tablas de la base de datos “Baja” (Imagen 3) y “Estudiante” (Imagen 4) y, en busca de los atributos que dieran respuesta a un grupo de preguntas realizadas al cliente mediante las entrevistas.

Las preguntas aplicadas fueron:

- ✓ ¿Cuáles son las causas de baja que se manejan en la educación superior?
- ✓ ¿Cuáles son los centros de procedencia de los estudiantes que ingresan a la UCI?
- ✓ ¿Qué sexo y raza predomina entre los estudiantes del centro?

- ✓ ¿Cuáles son las provincias que más estudiantes ingresan en la universidad de las Ciencias Informáticas?

	id_baja [PK] serial	id_estudiante integer	id_curso integer	causas text
1	221	2934	9	perdida de
2	222	473	4	academica
3	223	1414	10	inasistenci
4	224	1332	8	perdida de
5	225	2264	9	sancion
6	226	2735	8	perdida de
7	227	972	8	perdida de
8	228	1810	3	inasistenci
9	229	657	9	sancion
10	230	2967	2	inasistenci
11	231	1044	9	definitiva
12	232	1966	3	voluntaria
13	233	432	5	desercion
14	234	2131	3	perdida de
15	235	241	2	perdida de

Imagen 3: Tabla Baja.

	id_estudiante [PK] serial	nombre_estu character var	apellido1_est character var	apellido2_est character var	solapin_estu character var	sexo character var	provincia character var	municipio character var	procedencia character var	via_ingreso character var	raza character var
1	1	iwpalynp	goybq	avgvxeofnkr	57173	Masculino	Ciego de Av	Baraguá	Deporte	Orden 18 ex	Mestiza
2	2	lfwkk	cnhmashfddq	shri	76522	Mas	Mayabeque	Nueva Paz	Pre Extranj	Orden 18 ex	Blanca
3	3	fpizlorkjqv	vygpeqcabmc	giyornhewez	38402	Femenino	Ciego de Av	Majagua	IPVCE	Orden 18	M
4	4	nqpq	fdlyfjqub	ciorcpgiivh	88485	Fem	Matanzas	Unión de R	Deporte	Concurso	M
5	5	mtxvnhttwsv	plwrraczjyu	hnuzwlwiqgt	32173	M	Camaguey	Sibanicú	EMCC	MININT	B
6	6	sv	pirshcnsr	wfndmcvztoy	10523	Masculino	Artemisa	Artemisa	IPUEC	MININT	B
7	7	hbuudvmnmlg	hv	ne	95465	Mas	Cienfuegos	Palmira	IVP-MININT	Concurso	B
8	8	zavufsmzcyd	fzyzshvniq	kk	38304	Masculino	Sancti Spir	Sancti Spí	IPU	MINFAR	Mestiza
9	9	hmapv	yuoolbnyqhr	ckphjnakcp	46666	Mas	Santiago de	Contramaes	EMCC	MINFAR	Mestiza
10	10	ibcirtvpaik	vcy	yvtwknxntvh	68971	Masculino	Villa Clara	Placetas	IVP-MININT	Preuniversi	Negra
11	11	gofrczlxrea	unyttxrvggt	zjcxnku	32831	F	Villa Clara	Sagua La Gr	IPVCP	MININT	N
12	12	qiecstmfpok	npqoalafnfn	vtvmwbgmcpg	44431	M	Artemisa	Artemisa	IPVCE	Concurso	Negra

Imagen 4: Tabla Estudiante.

Tabla 1: Atributos seleccionados.

Variables	Tabla	Descripción
Id_ estudiante	Estudiante	Identificador del estudiante
Sexo	Estudiante	Sexo del estudiante
Provincia	Estudiante	Provincia del estudiante
Raza	Estudiante	Color de piel del estudiante
Causa_Baja	Baja	Tipo de causa por la que el estudiante es dado de baja
Centro_Procedencia	Estudiante	Centro educacional del que procede

La tabla anterior muestra la primera selección de los atributos a utilizar en la construcción de la vista minable.

- **Describir los datos.**

A continuación se realiza una descripción de las variables más importantes seleccionadas en las tablas de la base de datos anteriormente descritas, que son de mayor utilidad en la generación del modelo, especificando los tipos de datos de cada uno para una posible transformación según el algoritmo a implementar

Tabla 2: Descripción de los atributos seleccionados.

Variables	Tabla	Tipo de dato
Id_ estudiante	Estudiante	Numérico
Sexo	Estudiante	Nominal
Provincia	Estudiante	Nominal
Raza	Estudiante	Nominal

Causa_Baja	Baja	Nominal
Centro_Procedencia	Estudiante	Nominal

- **Explorar los datos**

La exploración absoluta de los datos es lo primero que se realiza antes de cualquier análisis que tiene como fin obtener conocimiento a partir de ellos. La misma abarca diferentes aspectos como son disímiles tipos de gráficos y índices que caracterizan una distribución de frecuencias.

En el desarrollo de la investigación no fue necesario realizar esta tarea de la metodología CRISP-DM, ya que no fue preciso almacenar los datos, porque fueron tomados directamente del área de gestión académica de la UCI.

- **Verificar la calidad de los datos**

Esta tarea es importante dentro de la fase de Comprensión de los datos porque en ella se examina la calidad de los datos, por tanto es primordial realizar un análisis exhaustivo de los mismos.

Dentro de esta tarea fueron consultados los datos contenidos en la base de datos del área de gestión académica de la UCI y se les realizó un riguroso análisis teniendo en cuenta los elementos siguientes:

- ✓ Representación de la realidad.
- ✓ Campos sobrados.
- ✓ Existencia de campos vacíos.

Error encontrado en la Base de Datos

Los datos no constituyen una representación de la realidad**.**

2.2.3 Preparación de datos

Al realizar la recolección preliminar de los datos, se comienza su preparación para adecuarlos a la técnica de MD que será utilizada posteriormente. La preparación de datos contiene las tareas de selección de datos, limpieza de datos e integración de diferentes orígenes de datos.

- **Selección de datos**

El objetivo de esta tarea, es detallar los atributos que serán incluidos o excluidos del proceso de MD. La selección de los datos se realizó de la siguiente manera:

En la base de datos del área de gestión académica de la UCI, se relaciona las tablas “Baja” (Imagen 3) y “Estudiante” (Imagen 4) a partir del atributo id _estudiante, seleccionando los atributos: Sexo, Provincia, Raza, Centro Procedencia y el tipo de causa por la que son dados de baja; estos atributos son los datos iniciales para la generación del modelo.

Tabla 3: Atributos seleccionados para realizar el proceso de modelado.

Variables	Tabla	Tipo de dato
Sexo	Estudiante	Numérico
Provincia	Estudiante	Nominal
Raza	Estudiante	Nominal
Centro_Procedencia	Estudiante	Nominal
Causa_Baja	Baja	Nominal

Es importante destacar que el atributo id _estudiante fue excluido, pues es irrelevante para la generación del modelo.

- **Limpieza de datos**

Después de realizar un análisis sobre los datos seleccionados, se detectó un problema con respecto a la variable id_estudiante, la cual no es importante en la generación del modelo, pero si en el proceso de construcción de la vista minable.

Problema detectado

Se encontraban repetidos en la tabla Baja, valores de la variable id_estudiante.

Solución al problema detectado

En la tabla “Baja” los id _estudiante que tenían más de una causa de baja, se procedió a dejar solamente uno de ellos.

- **Construir datos**

En esta tarea se comienzan a actualizar valores de columnas, crear nuevas columnas, en caso que la minería lo necesite.

Transformaciones realizadas

- ✓ Desaparece la columna Causa_Baja y aparecen siete nuevas columnas con los valores de la variable Causa_Baja: *Académica*, *Definitiva*, *Deserción*, *Inasistencia*, *Pérdida de requisitos*, *Sanción* y *Voluntaria*. Los valores que van a tomar estas columnas serán de (0 ó 1); toma valor 0 cuando el estudiante no tiene ese tipo de baja y 1 en caso contrario. Se decidió trabajar con atributos numéricos pues facilita la construcción del modelo aplicando la técnica de agrupamiento.
- ✓ Todas las variables a partir de la transformación realizada, toman valores numéricos, facilitando el trabajo con algoritmo seleccionado para generar el modelo de conocimiento.

Transformaciones realizadas a las variables

Listado de Variables

Sexo

Sexo =1 entonces Femenino.

Sexo =2 entonces Masculino.

Provincia

Provincia =1 entonces Las Tunas.

Provincia = 2 entonces Cienfuegos.

Provincia = 3 entonces Isla de la Juventud.

Provincia = 4 entonces Mayabeque.

Provincia = 5 entonces Villa Clara.

Provincia = 6 entonces Camagüey.

Provincia = 7 entonces Granma.

Provincia = 8 entonces Ciudad Habana.

Provincia = 9 entonces Ciego de Ávila.

Provincia = 10 entonces Santi Spíritus.

Provincia = 11 entonces Santiago de Cuba.

Provincia = 12 entonces Matanzas.

Provincia = 13 entonces Artemisa.

Provincia = 14 entonces Holguín.

Provincia = 15 entonces Guantánamo.

Procedencia

Procedencia =1 entonces Facultad Obrera Campesina: **Facultad**.

Procedencia =2 entonces Instituto Preuniversitario Urbano: **IPU**.

Procedencia =3 entonces Orden 18: **IPUM**.

Procedencia =4 entonces Pre en otros países: **Pre Extranjero**.

Procedencia =5 entonces Instituto Preuniversitario Vocacional de Ciencias Exactas: **IPVCE**.

Procedencia =6 entonces Escuela Militar Camilo Cienfuegos: **EMCC**.

Procedencia =7 entonces Escuelas de Deportes: **Deporte**.

Procedencia =8 entonces Instituto Preuniversitario Escuela al Campo: **IPUEC**.

Procedencia =9 entonces Instituto Preuniversitario Vocacional de Ciencias Pedagógicas: **IPVCP**.

Procedencia =10 entonces Instituto Vocacional Preuniversitario del Minint: **IVP-MININT**.

Raza

Raza =1 entonces Blanca.

Raza =2 entonces Mestiza.

Raza =3 entonces Negra.

Descripción de las causas de baja.

Baja: Se entiende por baja la suspensión temporal o definitiva de la condición de estudiante universitario. Es válida para estudiantes matriculados en cualquier tipo de curso. A los efectos de la promoción académica las bajas se consideran como año cursado y desaprobado.

A continuación se describe cada uno de los tipos de causa de baja, según el Reglamento de Organización Docente de la Educación Superior:

Baja Académica

ARTÍCULO 66: Se considera que un estudiante matriculado en el curso diurno o en curso para trabajadores causa baja por insuficiencia docente cuando:

- ✓ Desaprueba más de dos asignaturas en el año académico matriculado y ya ha agotado todas sus posibilidades de repitencias.
- ✓ Desaprueba más de dos asignatura en el año académico matriculado y no es autorizado a repetir el año.
- ✓ Desaprueba el año que repite.

Para los estudiantes matriculados en la educación a distancia, esta baja se producirá cuando hayan agotado las posibilidades de matricular una asignatura, según se establece en el artículo 17 del presente Reglamento, sin aprobarla.

Baja por Sanción

ARTÍCULO 67: Se considera que un estudiante, matriculado en cualquier tipo de curso, causa baja por sanción disciplinaria, cuando incurre en faltas establecidas en el Reglamento Disciplinario vigente y que implican la separación indefinida o temporal de la educación superior.

Baja Voluntaria

ARTÍCULO 68: Se considera que un estudiante, matriculado en cualquier tipo de curso, causa baja voluntaria cuando es solicitada por este. La solicitud la dirigirá por escrito al decano de la Facultad o al directivo designado en los municipios, según corresponda, especificando las causas que la fundamentan.

Baja por Deserción

ARTÍCULO 69: Se considera que un estudiante causa baja por deserción cuando:

- ✓ Matriculado en cualquier tipo de curso, no ratifique su matrícula en cada curso académico y en el período que se establezca por la dirección del Centro de Educación Superior, según se establece en el artículo 15 del presente Reglamento.
- ✓ Matriculado en los cursos diurnos y para trabajadores, no renueve la licencia de matrícula en cada curso académico y en el período que se determine por la dirección del Centro de Educación Superior, según se establece en el artículo 22 del presente Reglamento.
- ✓ Matriculado en cualquier tipo de curso abandone los estudios sin justificar la causa.

Baja por Pérdida de Requisito

ARTÍCULO 70: Se considera que un estudiante causa baja por pérdida de requisitos cuando:

- ✓ Matriculado en el curso para trabajadores, abandone su trabajo injustificadamente antes de haber finalizado el compromiso de tiempo asumido en el programa en que está insertado; o deje de mantener sin causa justificada la condición por la cual ingresó al mismo.
- ✓ Matriculado en cualquier tipo de curso, pierda aptitudes físicas o mentales de tal envergadura que no le permitan continuar sus estudios universitarios. Esta situación debe estar avalada por una institución de salud.
- ✓ Matriculado en cualquier tipo de curso, muestre una conducta social inconsecuente con los principios éticos y morales que propugna nuestra sociedad. Esta decisión debe aprobarse en el consejo de dirección de la Facultad en que esté matriculado el estudiante.
- ✓ Matriculado en el curso para trabajadores pierda el vínculo laboral estable
- ✓ Matriculado en el curso para trabajadores, su centro de trabajo no ratifica el aval que se exige en el acto de matrícula.

Baja por Inasistencia

ARTÍCULO 71: Se considera que un estudiante, matriculado en los cursos diurnos o para trabajadores, causa baja por inasistencia cuando haya desaprobado alguna asignatura según lo establecido en el artículo 41 del presente Reglamento, y además, haya consumido todas las posibilidades de repitencias y el número de asignaturas desaprobadas en el año que cursa exceda la cantidad de arrastres permisibles en el tipo de curso en que esté matriculado.

Baja Definitiva

ARTÍCULO 72: Se considera que un estudiante causa baja definitiva cuando:

- ✓ Matriculado en cualquier tipo de curso cause baja nuevamente por cualquiera de los tipos de baja que pueden presentarse, y además haya agotado todas las posibilidades previstas en este Reglamento, incluyendo la posibilidad de reingreso.
- ✓ Matriculado en cualquier tipo de curso, incurra en faltas disciplinarias que impliquen la expulsión de la educación superior, según se establece en el Reglamento Disciplinario vigente.
- ✓ Fallezca.

El jefe del organismo de la administración central del estado con Centros de Educación Superior adscritos podrá autorizar, de manera excepcional, el reingreso a los estudios superiores a aquellos estudiantes que hayan causado baja definitiva, cuando existan argumentos suficientes que lo ameriten.

Integrar datos

Se analizan los datos que son necesarios para generar el modelo y son combinados de múltiples tablas o registros para crear la vista minable. La misma se crea a partir de la transformación que se le realiza a las tablas de la base de datos “Estudiante” y “Baja”, utilizando la herramienta Pentaho Data Integration y así se puede obtener la Vista_Minable_Final (imagen 6) que será utilizada para la creación del modelo, mediante la aplicación de la técnica de minería de datos: agrupamiento.

Descripción de la transformación realizada a las tablas “Estudiante” y “Baja”

1. Se extraen de la base de datos del área de gestión académica de la UCI, los atributos de las tablas “Estudiante” y “Baja”.
2. Se realiza una unión por clave (id_estudiante) y se seleccionan los valores de las tablas “Estudiante” y “Baja” que serán utilizados para la construcción de la Vista_Minable_Final.
3. Se realizan dos mapeos de valores para lograr homogeneidad entre los valores de los atributos columnas: Sexo y Raza.
4. Se utiliza el componente Java Script para crear siete nuevas columnas con los tipos de causa por las cuales los estudiantes son dados de baja (Académica, Definitiva, Deserción, Inasistencia, Pérdida de requisitos, Sanción y Voluntaria), las cuales toman valores de (0/1), 0 para cuando el estudiante no tiene ese tipo de baja y 1 en caso contrario.
5. Se utiliza el componente Filas Únicas para eliminar los valores de id_estudiante que se encuentran repetidos en la tabla “Baja”.
6. Se elimina la columna id_estudiante y Causa_Baja por resultar irrelevante para la obtención del modelo de conocimiento.
7. Se realizan varios mapeos de valores para convertir los valores de los atributos a tipo de dato numérico y tratar los valores que tengan campos vacíos.

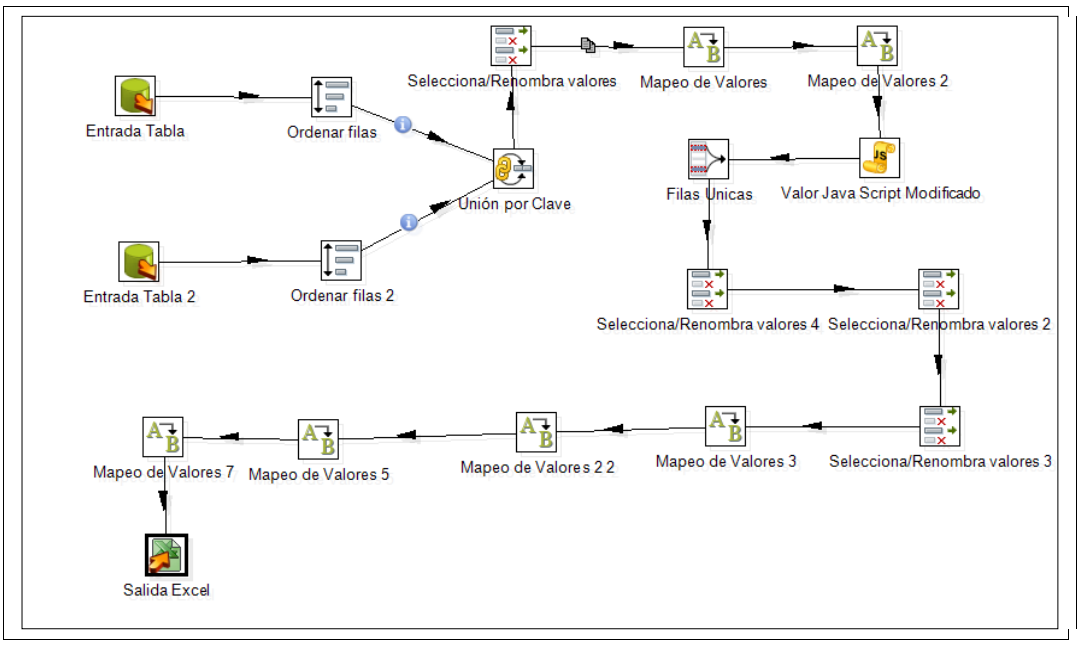


Imagen 5: Transformación de los datos.

Después de las transformaciones antes mencionada, se obtiene la Vista_Minable_Final, la cual contiene las variables principales para la generará el modelo de conocimiento; siendo estas el Sexo, Provincia, Centro de Procedencia, Raza, así como las 7 causas de baja descritas anteriormente en el desarrollo del capítulo.

Sexo	Provincia	Procedencia	Raza	Académica	Definitiva	Deserción	Inasistencia	Pérdida_requisitos	Sanción	Voluntaria
2	1	1	1	0	0	0	0	0	1	0
1	2	2	1	0	1	0	0	0	0	0
1	3	3	3	0	0	0	0	0	1	0
1	4	4	2	0	1	0	0	0	0	0
1	1	5	3	0	0	0	1	0	0	0
2	4	6	1	0	1	0	0	0	0	0
2	5	5	3	0	0	0	0	0	1	0
1	6	7	1	0	0	1	0	0	0	0
1	4	3	1	0	0	1	0	0	0	0
2	7	6	1	0	0	0	1	0	0	0
2	8	4	2	0	1	0	0	0	0	0
2	9	8	2	0	0	1	0	0	0	0
2	10	5	1	0	0	0	1	0	0	0
1	1	4	2	0	0	0	0	1	0	0
2	11	3	3	0	0	0	0	1	0	0
1	6	9	1	0	0	0	0	0	0	1
1	3	2	1	0	1	0	0	0	0	0
2	8	1	3	0	0	1	0	0	0	0
2	8	3	3	0	0	0	0	0	1	0
2	9	4	2	0	0	0	0	0	0	1
2	9	7	3	0	0	0	0	1	0	0
2	2	1	2	0	0	0	0	1	0	0
1	9	10	2	0	0	0	1	0	0	0
1	5	8	3	0	0	1	0	0	0	0
1	5	2	1	0	1	0	0	0	0	0
1	5	7	2	1	0	0	0	0	0	0
2	8	6	2	0	0	0	0	1	0	0
1	10	10	3	0	0	0	1	0	0	0

Imagen 6: Vista Minable.

2.3 Conclusiones

En el este capítulo se define la arquitectura para el proyecto de minería de datos, la cual se encuentra dividida en tres subsistemas (Integración, Almacenamiento y Visualización). Se selecciona los atributos de interés para la construcción del modelo, de las tablas “Estudiante” y “Baja” de la base de datos área de gestión académica de la UCI. Se realiza una transformación sobre los datos que fueron seleccionados, permitiendo obtener la Vista_Minable_Final con los atributos principales para generar el modelo de conocimiento.

Capítulo 3: Descripción de los resultados.

Introducción.

Se describe la aplicación de la técnica y el algoritmo seleccionado para la generación del modelo, el cual es interpretado, obteniendo un grupo de patrones que describen el comportamiento de las principales causas de baja de la Universidad de las Ciencias Informáticas.

3.1 Modelado.

- **Selección de la técnica de modelado.**

La técnica de minería de datos seleccionada fue el agrupamiento y el algoritmo Simple K-means.

Simple K-Means en WEKA

El algoritmo Simple K-medias se encuentra implementado en la clase *weka.clusterers.SimpleKMeans.java*. A continuación se muestra en la siguiente tabla las opciones de configuración.

Tabla 4: Opciones de configuración del algoritmo.

Opción	***Configuración***
numClusters (n)	Número de clúster o grupos que se forman.
seed (n)	Semilla a partir de la cual se genera el número aleatorio para inicializar los centros de los clusters.

Tipos de datos que admite el algoritmo y las propiedades de la implementación.

- ✓ Admite atributos simbólicos y numéricos.
- ✓ Para obtener los centroides iniciales se emplea un número aleatorio obtenido a partir de la semilla empleada. Los k ejemplos correspondientes a los k números enteros siguientes al número aleatorio obtenido serán los que conformen dichos centroides (15).

- **Construcción del modelo**

En esta tarea se ejecuta la herramienta de modelado Weka, sobre el conjunto de datos seleccionados.

Construcción del modelo aplicando Simple K-Means.

A continuación se presenta el modelo que se generó a partir de la aplicación del algoritmo Simple K-means sobre los datos almacenados en la vista minable. Para poder ejecutar el algoritmo se le debe especificar el parámetro *K* que será el número de clusters o grupos que se van a formar y además se debe seleccionar un número *n* que será denominado semilla, para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comience las iteraciones siguientes. Para la selección de este número se realizaron 10 intentos consecutivos probando distintas semillas y se seleccionó aquella que minimizaba la suma del error cuadrático. En la siguiente tabla se muestran los resultados de los 10 intentos.

Tabla 5: Resultado de K-Means para 2 clúster con varias semillas.

Semilla	***Error Cuadrático***
1	709.0
2	694.0
3	733.0
4	762.0
5	691.0
6	696.0
7	699.0
8	700.0
9	758.0
10	699.0

Entonces los valores seleccionados son:

- ✓ K (número de grupos)=2.
- ✓ Semilla (valor inicial del centroide)=5.

El modelo obtenido con Weka tras la ejecución de Simple K-Means con 2 clúster y un valor de la semilla = 5, es el siguiente:

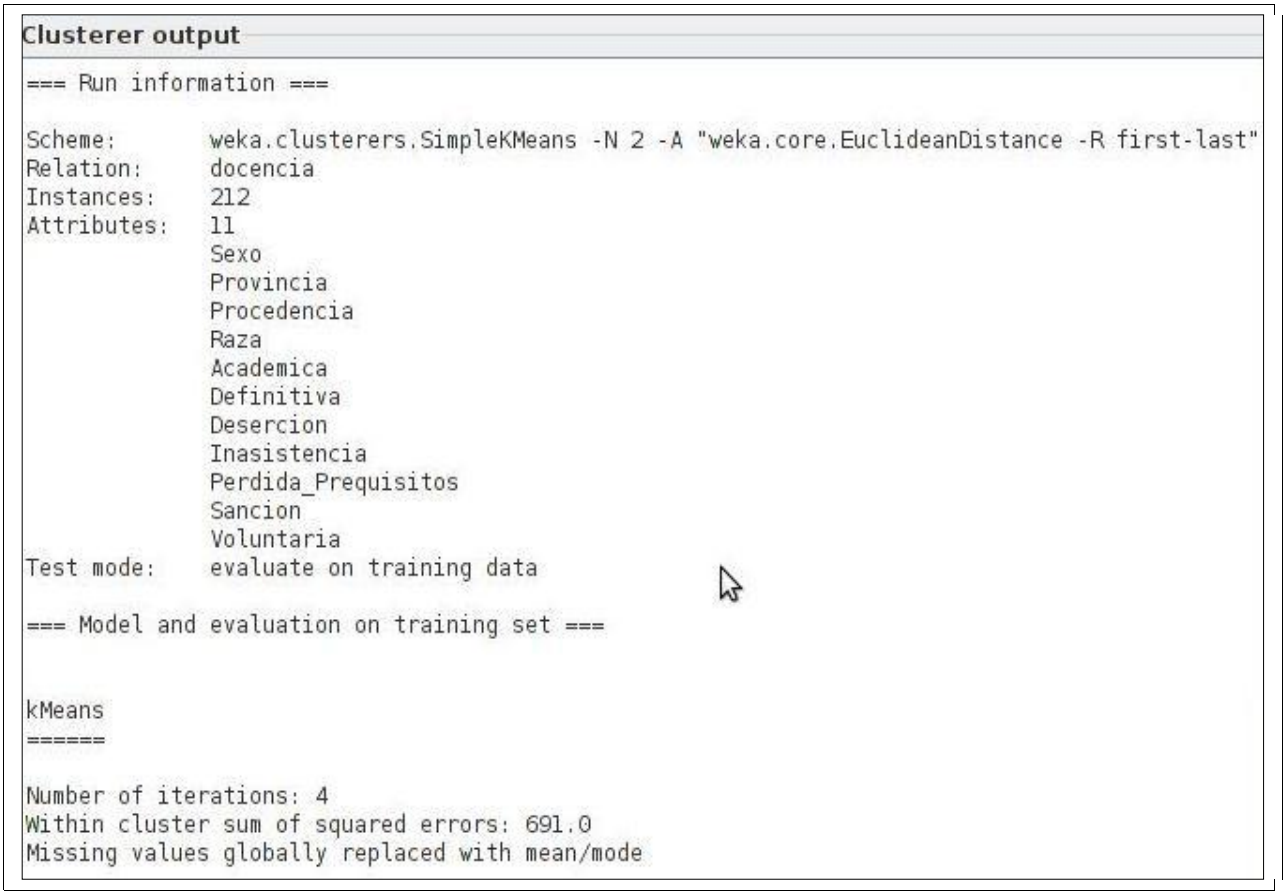
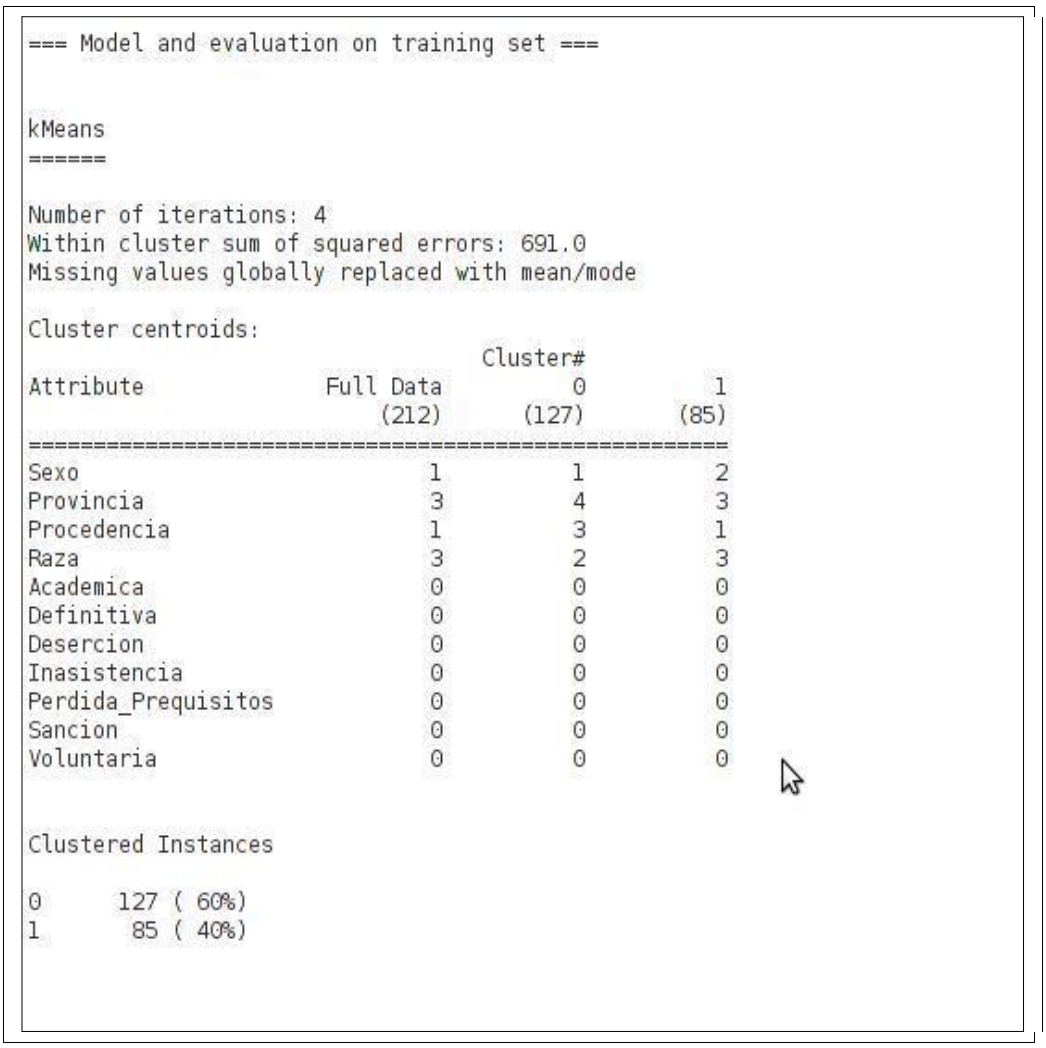


Imagen 7: Primera parte del modelo.



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


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 691.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data       0         1
                          (212)     (127)      (85)
=========================================================
Sexo                           1         1         2
Provincia                      3         4         3
Procedencia                    1         3         1
Raza                           3         2         3
Academica                      0         0         0
Definitiva                     0         0         0
Desercion                      0         0         0
Inasistencia                   0         0         0
Perdida_Prequisitos            0         0         0
Sancion                        0         0         0
Voluntaria                     0         0         0


Clustered Instances

0      127 ( 60%)
1       85 ( 40%)
```

Imagen 8: Segunda parte del modelo.

Este modelo muestra la asignación de los estudiantes que causan baja por grupos y los valores significativos de las variables en cada grupo creado.

- **Evaluación del modelo**

En esta tarea el minero interpreta los modelos según su conocimiento de dominio y los criterios de éxitos de minería de datos.

Evaluación del modelo generado por Simple K-Means.

La herramienta Weka proporciona 3 modos de prueba para realizar las opciones de test en la técnica de agrupamiento:

- ✓ Use training set.

- ✓ Supplied test set.
- ✓ Percentage Split.

Es utilizado el modo de prueba Use training set, con esta opción Weka entrena el método con todos los datos disponibles y luego lo aplica otra vez sobre los mismos. Pero es válido destacar que los modelos descriptivos en general, son difíciles de evaluar pues inicialmente el modelo va a describir un tipo de comportamiento y además no cuentan con una clase determinada con la que se pueda medir el grado de acierto del mismo. La mejor manera de evaluar este modelo es ver si tiene un comportamiento útil en el área que se vaya a aplicar (16).

Resultados obtenidos a partir del modelo generado:

Tabla Representativa

- ✓ El sexo más representativo en las bajas es el femenino.
- ✓ La provincia más representativa en las bajas es el municipio especial Isla de la Juventud.
- ✓ El centro de procedencia más representativo en las bajas es la facultad.
- ✓ La raza más representativa en las bajas es la negra.

Características de los grupos

- ✓ **Grupo 0 (60%):** El grupo 0 en su mayoría está integrado por estudiantes el sexo femenino, predominan los estudiantes de color de piel mestiza que provienen de IPUM de la provincia Mayabeque.
- ✓ **Grupo 1 (40%):** El grupo 1 en su mayoría está integrado por estudiantes el sexo masculino, predominan los estudiantes de color de piel negra que proviene de facultad del municipio especial Isla de la Juventud.

A partir de las gráficas de dispersión (GD) que se muestra en las imágenes (9 y 10) se conjugaron un conjunto de variables que permitieron la descripción de un grupo de comportamientos:

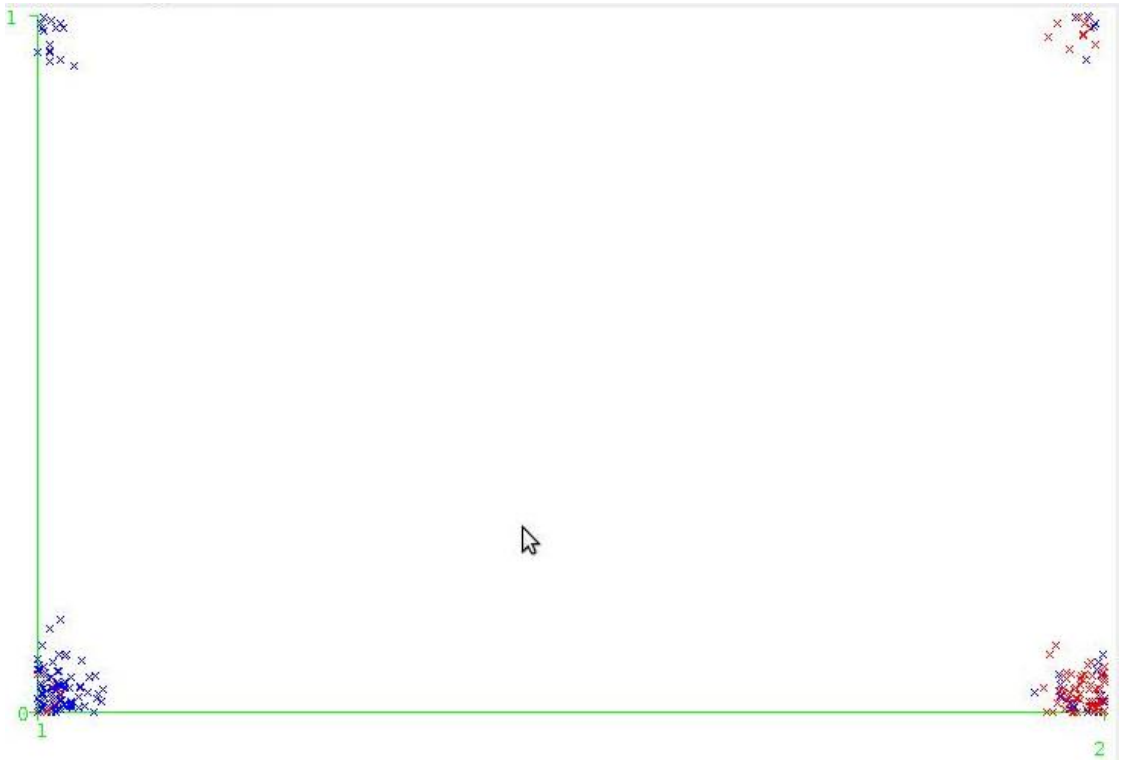


Imagen 9: Ejemplo de gráfica de dispersión, sexo-académica.

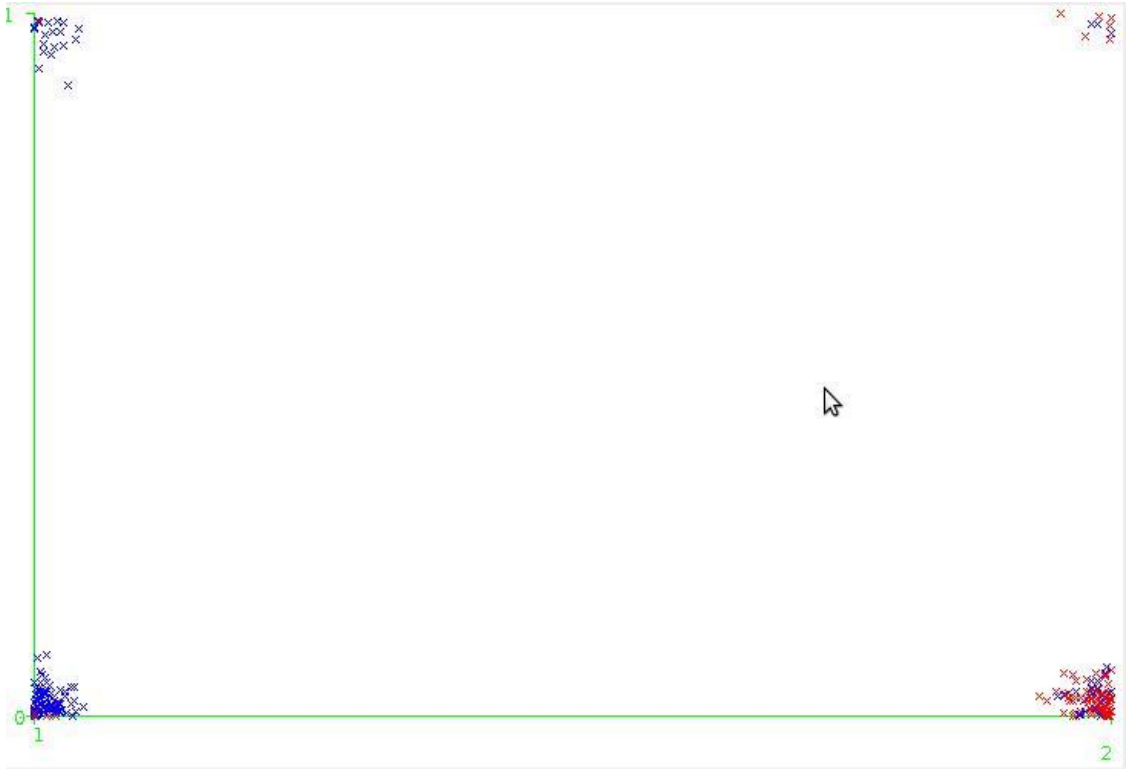


Imagen 10: Ejemplo de gráfica de dispersión, sexo-definitiva.

En cuanto a la procedencia según el tipo de baja:

- ✓ **Procedencia-Académica**

La mayoría de los estudiantes que causan baja académica proceden de IPUM y IVP-MININT.

- ✓ **Procedencia-Definitiva**

La mayoría de los estudiantes que causan baja definitiva proceden de IPU y Pre Extranjero.

- ✓ **Procedencia- Deserción**

La mayoría de los estudiantes que causan baja por deserción proceden de Facultad.

- ✓ **Procedencia-Inasistencia**

La mayoría de los estudiantes que causan baja por inasistencia proceden de Facultad e IVP- MININT.

- ✓ **Procedencia- Pérdida de requisitos**

La mayoría de los estudiantes que causan baja por pérdida de requisitos proceden de IPVCP y IVP-MININT.

- ✓ **Procedencia-Sanción**

La mayoría de los estudiantes que causan baja por sanción proceden de IPUM.

- ✓ **Procedencia-Voluntaria**

La mayoría de los estudiantes que causan baja voluntaria proceden de facultad.

En cuanto al sexo según el tipo de baja:

- ✓ **Sexo-Académica**

La mayoría de los estudiantes que causan baja académica son hombres.

- ✓ **Sexo-Definitiva**

La mayoría de los estudiantes que causan baja definitiva son mujeres.

- ✓ **Sexo-Deserción**

La misma cantidad de hombres y mujeres causan baja por deserción.

- ✓ **Sexo-Inasistencia**

La mayoría de los estudiantes que causan baja por inasistencia son mujeres.

- ✓ **Sexo- Pérdida de requisitos**

La mayoría de los estudiantes que causan baja por pérdida de requisitos son hombres.

- ✓ **Sexo- Sanción**

La mayoría de los estudiantes que causan baja por sanción son mujeres.

- ✓ **Sexo-Voluntaria**

La mayoría de los estudiantes que causan baja voluntaria son hombres.

En cuanto a la provincia según el tipo de baja:

- ✓ **Provincia-Académica**

Los estudiantes que causan baja académica provienen en su mayoría de la provincia de Ciego de Ávila.

- ✓ **Provincia-Definitiva**

Los estudiantes que causan baja definitiva provienen en su mayoría de la provincia de Mayabeque.

- ✓ **Provincia-Deserción**

La provincia de Matanzas es en la que más estudiantes causan baja por deserción.

- ✓ **Provincia-Inasistencia**

Las provincias que más estudiantes causan baja por inasistencia son: Cienfuegos, La Isla de la Juventud, Villa Clara, Granma y Santi Spíritus.

- ✓ **Provincia-Pérdida de requisitos**

La provincia que más estudiantes causan baja por pérdida de requisitos es Las Tunas.

- ✓ **Provincia-Sanción**

La provincia que más estudiantes causan baja por sanción es La Isla de la Juventud.

- ✓ **Provincia-Voluntaria**

La provincia que más estudiantes causan baja voluntaria es La Isla de la Juventud.

Una vez realizado un análisis profundo sobre el modelo y las gráficas de dispersión se obtuvieron los patrones siguientes:

- ✓ El 60% de los estudiantes que causan baja son mujeres, de color de piel mestiza que provienen de IPUM, de la provincia Mayabeque.

- ✓ El 40% de los estudiantes que causan baja son hombres, predominan los estudiantes de color de piel negra que provienen de facultad y son del municipio especial Isla de la Juventud.
- ✓ Los estudiantes que causan baja académica provienen de la provincia de Ciego de Ávila, procedentes de IPUM y de IVP-MININT y son la misma cantidad de hombres que de mujeres.
- ✓ Los estudiantes que causan baja definitiva provienen de la provincia de Mayabeque son mujeres y proceden de IPU y de Pre Extranjero.
- ✓ Los estudiantes que causan baja por deserción provienen de la provincia de Matanzas procedentes de Facultad y son mujeres.
- ✓ Los estudiantes que causan baja por inasistencia provienen de las provincias de Cienfuegos, la Isla de la Juventud, Villa Clara, Granma y Santi Spíritus procedentes de Facultad y IVP-MININT y son mujeres.
- ✓ Los estudiantes que causan baja por pérdida de requisitos provienen de la provincia de las Tunas procedentes de IPVCP y son hombres.
- ✓ Los estudiantes que causan baja por sanción provienen de la Isla de la Juventud procedentes de IPUM y son mujeres.
- ✓ Los estudiantes que causan baja voluntaria provienen de la Isla de la Juventud procedentes de Facultad y son hombres.

3.2 Evaluación

Evaluación de los resultados

Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente.

Para la evaluación de los resultados finales, se realiza una comparación entre los patrones de comportamiento y los datos que se encuentran almacenados en el área de gestión académica de la UCI, haciendo uso de gráficos de pastel. A continuación se muestran dos ejemplos en las imágenes (11 y 12), donde se comprueba que el sexo que predomina en los tipos de baja (académica y definitiva), son los obtenidos con la interpretación realizada, a partir del modelo y las GD. Aclarar que de igual manera se comprobó la veracidad de las demás interpretaciones efectuadas, obteniendo un resultado satisfactorio.

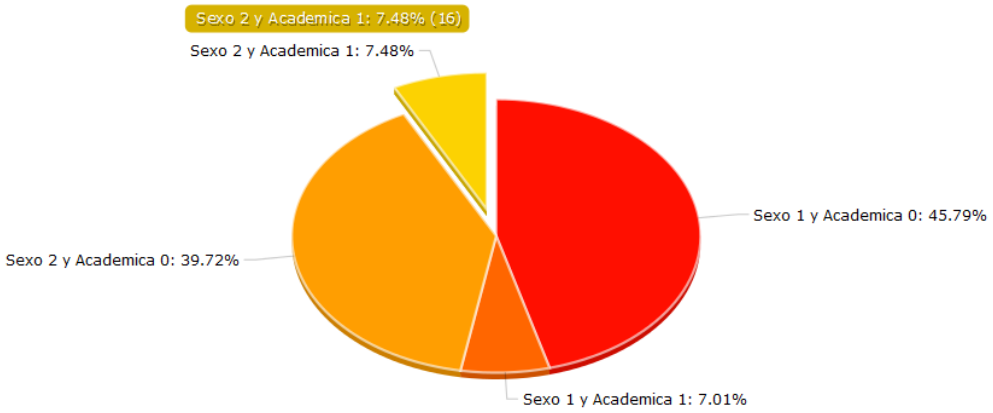


Imagen 11: Gráfico de pastel, sexo-académica.

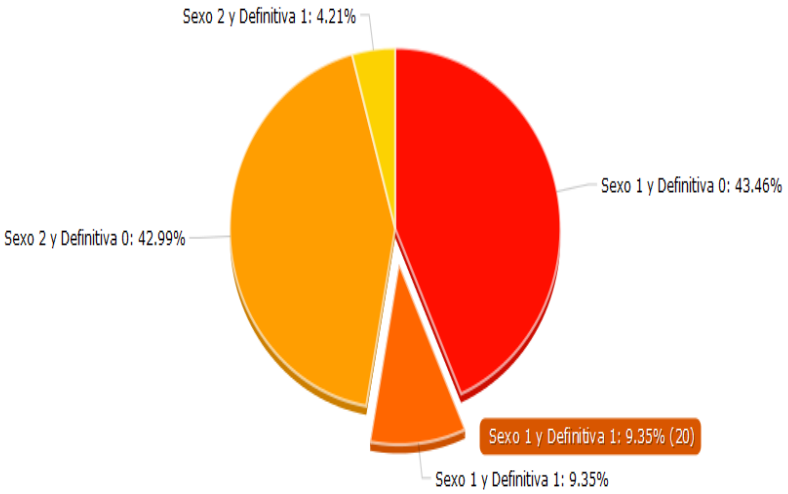


Imagen 12: Gráfico de pastel, sexo-definitiva.

Propuesta de mejora: Realizar los modelos de MD utilizando el algoritmo Simple K-Means, donde exista una mayor cantidad de datos de los estudiantes que causan baja y que sean reales.

- **Proceso de revisión**

En este punto, los modelos resultantes pasan a ser satisfactorios cubriendo las necesidades de negocio. Fue apropiado hacer una revisión más cuidadosa de los compromisos de la MD para determinar si hay cualquier factor importante o tarea que de algún modo ha sido pasada por alto.

Después de la revisión de todo el procedimiento el modelo generado es el óptimo, pues responde al objetivo del negocio de la MD, por lo que no será necesario volver a realizar alguno de los pasos anteriores.

3.3 Desarrollo

- **Plan de desarrollo**

El informe final del proyecto se basa en un resumen de los pasos que se han implementado a lo largo del desarrollo de la investigación, así como los inconvenientes encontrados. El presente documento se considera como el Informe Final, ya que describe todos los pasos de la metodología seleccionada para guiar el proyecto de minería de datos, con el fin de describir los patrones de comportamiento de las causas de baja de los estudiantes de la UCI.

3.4 Conclusiones

Con la aplicación del algoritmo Simple K-means de la técnica de minería de datos agrupamiento, se obtuvo un modelo de conocimiento, a partir de la interpretación del mismo y de las gráficas de dispersión se obtuvieron trece patrones de comportamiento, resultados que fueron validados y reflejaron ser los esperados, respondiendo al objetivo del negocio planteado.

Conclusiones Generales.

A partir de los resultados obtenidos en la investigación se llegaron a las siguientes conclusiones:

- ✓ Para realizar el proceso de minería de datos se utiliza la metodología CRISP-DM y la herramienta Weka en su versión 3.6.0, mediante las cuales el minero, de manera organizada y planificada obtiene el modelo de conocimiento.
- ✓ A partir de la interpretación del modelo de conocimiento obtenido en el desarrollo de la investigación, se identificaron trece patrones de comportamiento, que describen las principales causas de bajas docentes en la Universidad de las Ciencias Informáticas.
- ✓ La aplicación de la técnica de agrupamiento de Minería de Datos sobre el área de gestión académica de la Universidad de las Ciencias Informáticas constituye una herramienta de análisis y descubrimiento de conocimiento, oculto en las fuentes de información, que tributan al diagnóstico y detección de bajas docentes de esta institución, apoyando el proceso de toma de decisiones.

Recomendaciones

- ✓ Aplicar otras técnicas de minería de datos, sobre el mismo conjunto de datos y realizar una comparación entre los modelos de conocimiento obtenidos.
- ✓ Aplicar el modelo obtenido a los datos reales del área de gestión académica de la universidad.

Trabajos citados

1. **Hernádez Orallo, José, Ramírez Quintana, José y Ferri Ramírez, César.** *Introducción a la Mineríade Datos.* 2004.

2. **Molina López, José Manuel y Garcia Herrero, Jesús.** *Técnicas de análisis de datos utilizando microsof texcel y weka.* 2004. pág. 6.

3. **Rodríguez Suárez, Yuniet y Díaz Amador, Anolandy.** *Herramientas de Minería de Datos.* 2009. pág. 3.

4. **Corréa Ramírez, Isidro Manuel y Shelton Nada, Ronald.** *Estrategia de Trabajo para el Desarrollo del Módulo de Minería de Datos de un Call Center,aplicando la metodología CRISP-DM.* 2004.

5. **Vallejos, Sofía J.** *Minería de Datos.* 2006. pág. 17.

6. **Molina López, José Manuel y Garcia Herrero, Jesús.** *Técnicas de análisis de datos utilizando microsoft excel y weka.* 2004. pág. 41.

7. **León Rodríguez, Kirenia Helen y Davila Hernández, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011. pág. 21.

8. **Molina López, José Manuel y Garcia Herrero, Jesús.** *Técnicas de análisis de datos utilizando microsoft excel y weka.* 2004. pág. 102.

9. —. *Técnicas de análisis de datos utilizando microsoft excel y weka.* 2004. pág. 113.

10. —. *Técnicas de análisis de dato su tilizando microsoft excel y weka.* 2004. pág. 58.

11. **León Rodríguez, Kirenia Helen y Davila Hernández, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011. pág. 20.

12. **Corría Ramírez, Isidro Manuel y Shelton Nadal, Ronald.** *Estrategia de Trabajo para el Desarrollo del Modulo de Minería de Datos de un Call Center,Aplicando la metodología CRISP-DM.* 2004. pág. 34.

13. **Pete Chapman, Julian Clinton, Randy Kerber,Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, y otros.** *Guía paso a paso de Minería de Datos.* 2000.

14. oracle.com. [En línea]

15. **León Rodríguez, Kirenia Helen y Davila Hernáandez, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011. pág. 52.

Bibliografía

1. **Hernádez Orallo, José, Ramírez Quintana, José y Ferri Ramírez, César.** *Introducción a la Mineríade Datos.* 2004.

2. **Molina López, José Manuel y Garcia Herrero, Jesús.** *Técnicas de análisis de datos utilizando microsof texcel y weka.* 2004. pág. 6.

3. **Rodríguez Suárez, Yuniet y Díaz Amador, Anolandy.** *Herramientas de Minería de Datos.* 2009. pág. 3.

4. **Vallejos, Sofía J.** *Minería de Datos.* 2006. pág. 17.

5. **León Rodríguez, Kirenia Helen y Davila Hernández, Frank.** *Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial.* 2011. pág. 21.

6. **Corría Ramírez, Isidro Manuel y Shelton Nadal, Ronald.** *Estrategia de Trabajo para el Desarrollo del Modulo de Minería de Datos de un Call Center,Aplicando la metodología CRISP-DM.* 2004. pág. 34.

7. **Pete Chapman, Julian Clinton, Randy Kerber,Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth, y otros.** *Guía paso a paso de Minería de Datos.* 2000.

8. oracle.com. [En línea]

10. **ETL, Grupo de.** *Guía de Componentes de Pentaho Data Integration.*

11. **Barreno, Julio Escribano.** *Análisis de los datos mediante WEKA.*

12. **Garrido, LLuís.** *Introducción al DataMining.*

13. **García-Martínez, R.** *Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina.*

14. **Durán, Elena y Costaguta, Rosanna.** *Minería de datos para descubrir estilos de aprendizaje.*

15. **Gómez, Jose Ignacio González.** *Generalidades de la Minería de Datos.* 2007.

16. **Cadena, Lisa Leonor Pinzón.** *Aplicando minería de datos al marketing educativo.* 2011.

17. **Morate, Diego García.** *Manual de Weka.*

18. **Aler, Ricardo.** *Tutorial de Weka.* 2009.

19. **Bermejo, Sergi.** *Curso de Redes Neuronales Artificiales (2000-2001).*

20. **Lozada, Carlos Alberto Cobos.** *Agrupación Jerárquica, Particional y en VLDB.*

Glosario de Términos

Atributos Nominales: Atributos con tipo de datos String, cadenas de caracteres.

Atributos Numéricos: Atributos con tipo de datos enteros, dobles, flotantes y reales.

Bases de Datos: Se define como una serie de datos organizados y relacionados entre sí, los cuales son recolectados y explotados por los sistemas de información de una empresa o negocio en particular.

Base de Dato con ruido: Existen datos almacenados que contiene errores.

Cluster: Conjunto de objetos, que tienen características comunes.

Fase: Se le denomina fase al asunto o paso dentro del proceso. CRISP-DM consta de 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación, evaluación y explotación.

Gráfica de Dispersión: Es el grado de correlación que existe entre dos variables.

Instancias de proceso: Las instancias de proceso son las acciones y resultados de las actividades realizadas dentro de cada fase del proyecto.

KDD (Knowledge Discovery in Databases): Descubrimiento de conocimientos en bases de datos.

Minería de Datos (MD): Extracción de conocimientos ocultos en grandes bases de datos.

Weka: Herramienta que se utiliza para el modelado de los datos.

Tarea genérica: Cada fase está formada por tareas genéricas, o sea, la tarea genérica es la descripción de las actividades que se realizan dentro de cada fase. Por ejemplo, la tarea Limpiar los datos es una tarea genérica.

Tarea especializada: La tarea especializada describe cómo se pueden llevar a cabo las tareas genéricas en situaciones específicas. Por ejemplo, la tarea Limpiar los datos tiene tareas especializadas, como limpiar valores numéricos, y limpiar valores categóricos.