Universidad de las Ciencias Informáticas

Facultad 10

Propuesta para la utilización de la Minería de Datos en la recuperación de información en la sección Vigilancia Tecnológica de D´TIC

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

Autores: Karina Collazo Oliva.

Yordailis Morales Díaz.

Tutor: Ing. Miguel Jaeger Rodríguez Lazo.

Ciudad de La Habana, Cuba.

Junio 2010

Año 52 del Triunfo de la Revolución Cubana

Declaración de Autoría

Declaramos ser autores 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 2010.

Karina Collazo Oliva

_____

Firma del Autor

Yordailis Morales Díaz

_____

Firma del Autor

Ing. Miguel Jaeger Rodríguez Lazo

_____

Firma del Tutor

Agradecimientos

Mis primeros agradecimientos van para mis padres a los que admiro y respeto. Mami, papi: gracias por su dedicación día a día, por su educación, por guiarme por el mejor camino, por estar siempre a mi lado y sobre todo por confiar en mí. Los amo.

A mi mami: por saberse engrandecer en todo frente a los problemas y llevarnos a mi hermano y a mí por el camino correcto…gracias y mil veces gracias mamita. Te amo.

A mi papi: por el apoyo incondicional que me ha dado siempre, por su ternura, comprensión, su infinito amor y sobre todas las cosas por la confianza que ha depositado a mí, gracias por cuidar de esta niña que te ama con todo su corazón…vas a extrañar ahora que me vas a tener cerquita los chao pescao.

A mi hermano: Por ser mi mejor amigo y por apoyarme en los momentos que más lo he necesitado, por saber enfrentar las cosas que la vida nos ha puesto enfrente con decisión y valentía….ahora no se en que te vas a entretener, porque no te puedes poner en mala con la cocinera.

A mi chikitintin: por estar apoyándome en los momentos más difíciles, por ser amigo, compañero y novio, por cuidarme y dejarme saber día a día cuanto me ama, por apoyarme en el transcurso más difícil de esta etapa en la universidad, por su dedicación, por estar a mi lado siempre que lo he necesitado, por su ternura, te amo con todo mi corazón.

A mi abuela Hildoina: por su cariño en todo momento, por estar siempre orgullosa de su nieta.

A mi tía Maida, Mariloli y Ohilda: Porque son las otras madres que siempre me aconsejan, me ayudan, me regañan. Por ayudarme y quererme tanto y cuidarme como si fuera su niña, por darme esos consejos y esas conversaciones preparándome para la vida. Las adoro.

A mi tía Martica: Por escucharme y por aguantarme todas las malcriadeces, por saber ser amiga y madre, por estar disponible siempre para mí, te quiero.

A mis tías Katty y Lidia: por quererme mucho y confiar en mí.

A mis tíos Angelito, Carlo, Oscar y Fidel: por apoyarme y por compartir tantos momentos importantes en mi vida.

A Maina y Paino: Por estar en todo momento de mi vida, por ser como mis abuelos, por acordarse de cada momento significativo para mí, por su cariño, amor y ternura. A ustedes les regalo mis triunfos.

A Carmen y mi tía Kari: por ayudarme, apoyarme y aconsejarme, muchísimas gracias.

A mis primas: Saylin, Yanet, Marian y Lilian Liz: porque son el mayor tesoro que he tenido, más que primas hermanas, las quiero mucho.

A mis primos Leonid, Alejandro, Dariel, Arley y Alain: por ser parte de mi vida.

Agradecimientos.

A mi compañera de tesis Yordailis: más que compañera, amiga y más que amiga hermana. Sin tu entrega constante y tu paciencia para conmigo no hubiera podido realizar este trabajo.

A mi tutor querido Miguel (Cupi): sin ti mi amor, verdaderamente no hubiéramos salido adelante, gracias por ser el apoyo más grande que tuve durante el transcurso de este trabajo….que te vas a hacer sin una tesista como yo….me vas a extrañar.

Al tribunal de tesis por todo el apoyo que nos han brindado.

A mis amigas de siempre Neli y Dalinka, las quiero mucho.

A mi sobrino Jeison te adoro mi angelito.

A mis amigas: Elieyis, Yusmila, Anaivys, Ilearsi, Arlenis, gracias por compartir conmigo momentos tan lindos e importantes. Besitos.

A mi mejor amigo: Frank gracias por estar en todo momento importante de mi vida, por brindarme tu apoyo incondicional y por sobre todo por hacerme saber que cuento contigo siempre.

A mis amigos: Ernesto, Joel, Pepe, Yosbel, Yadir, Misael, gracias por compartir momentos lindos conmigo por hacerme saber que cuento con ustedes.

A mis compañeros de grupo Yelena, Laura, Daneidis, Dayana, Yudeisi, Heydi, Yudelmis, Yadir, Oreste, Eduardo Failde, Alberto, Eduardo Valdés, Mateo, El chino, Pavón, Javier Domínguez, Yamel, Adonis, Luis Carlos gracias por todos estos años juntos, los quiero.

A Yanet que a pesar de ser poquito el tiempo que nos llevamos muy bien le he sabido dar un lugar bien lindo en mi corazón, eres una niña muy dulce.

A mis compañeras de apartamento muchísimas gracias por todo el apoyo y por estar atentas a los adelantos de la tesis.

A los amigos de mi novio: Ernesto, Victor, Hansel, Ever, Yoanni, Mariño, Fuoman, José Alejandro besitos.

A Marta, Dayessy, Mimo, Blanca, Vicente que de una manera han tenido que ver en mi vida durante mi preparación como profesional, muchísimas gracias.

A la Uci y a mis profesores Hilda, Ariel, Juan José, Yanko, Rill, Alexander, Dunia, en fin a todos por contribuir en mí preparación profesional.

A todos lo que de una manera u otra han tenido que ver con mi preparación en la vida y por creer en mi superación.

A nuestro invencible Comandante en Jefe por ser el principal precursor de la Universidad de las Ciencias Informáticas y por permitirnos formar parte de este proyecto futuro.

Karina Collazo Oliva

Agradecimientos.

Antes que nada agradecerles a las personas que gracias a ellos hoy estoy aquí, a mis padres, que siempre me han brindado tanto amor y cariño y han confiado en mí, por apoyarme en todo, por su amor, dedicación, educación, preocupación y por la entrega que día a día me han brindado, hoy y por el resto de mi vida les doy las gracias. Los amo mucho.

A mi mamá Odalis: mami gracias por todo, por ser mi madre, mi amiga, mi hermana y por ser una de las cosas más bellas que tengo en la vida, por ser tan buena conmigo, por ser dura en los momentos en que si hubieras sido blandita yo no hubiera podido seguir adelante, gracias por todos tus consejos en los cuales siempre al final tenias y tienes la razón, gracias por todo, te quiero mami.

A mi papá Leonardo: papi para que decirte que eres mi vida si tú lo sabes, gracias por todo, por siempre y a cada instante confiar en mí y decirme que yo si podía, gracias por estar a mi lado cuando siempre te he necesitado, por tu amor, cariño, entrega y dedicación, por estar orgulloso de mí, gracias papito, te amo.

A mi Bebé Yadiel que durante estos años ha sabido malcriarme, darme todo su amor, darme las fuerzas para seguir adelante y decirme que yo si puedo, por estar conmigo en todo momento, por eso y por mucho más gracias mi amor, te amo.

A mi hermano Leonardo: gracias Didi antes que todo por soportarme, por celarme y por brindarme siempre tu amor infinito, por ayudarme tanto en mis primeros años de carrera, aunque en los otros siempre estuviste presente, gracias por quererme tanto, tu sabes cuánto te adoro, eres el mejor hermano del mundo, gracias por todo, mi amor.

A mis abuelas Angélica y Eloina por todo su amor y cariño infinito, y por estar orgullosas por mí, las quiero abus.

A mis tías queridas Yamir, Juanita, Daisy, Estrella y Cena por la confianza que tuvieron siempre en mí, gracias a todas

A mis tíos Manuel, Misael, Segundito, Pablo y Jesús por quererme tanto y confiar en mí, los quiero mucho.

A mi prima Midiala por todo, por ser mi amiga, por confiar en mí, por siempre estar a mi lado, tanto en los malos y buenos momentos, por brindarme tanto amor y cariño desde que tengo uso de razón, te quiero Midia.

A mis más que hermanas Saray y Sairy por tanto y tanto amor que desde chiquita me han profanado, por confiar siempre en mí, por sentirse orgullosas de su hermanita más chiquita y por siempre ayudarme en todo lo que he necesitado, gracias, las quiero mucho.

A mi cuñi Lisbety: gracias mi vida por todo, por siempre estar de mi lado, por aconsejarme, por ayudarme y por siempre confiar en mí y quererme tanto, gracias.

A tío Victor por todo, por saber que para él soy su hija más chiquita y por quererme tanto, gracias.

A mis amistades del barrio Pepe, Clara, Yunior, Pedrito, Herenia, Panchita, Nancy y a todos los que siempre le preguntaban a mis padres que como estoy, de cómo salía en las pruebas y por siempre brindarme tanto amor y cariño.

A mi otra familia, a mis suegros José Luis y Tomasita, a mi cuñado Raiko y mi concuña Aniska, a las mimas Lula y Olvido y a mis otras tías Mariana y Leonor, al Boby, Merci y toda su familia, a Edel, Niuris y a Juan Alberto gracias

por todo el amor y cariño que siempre me han brindado desde el primer día, por confiar en mí y siempre preocuparse por todo lo mío gracias a todos, los adoro.

A mis amigas de SS Danelys, Maynelis, Yaima, las Kettys, Lourdes, Yeni Cabeza, Yailenis, Yanet, Yailien y especialmente a la flaca Yure por soportarme desde hace tanto tiempo, y quererme tanto, las quiero a todas.

A mi compañera de tesis Karina: por tanto cariño que siempre me ha brindado, por confiar en mí y querer hacer la tesis conmigo, por ser más que amiga ser mi hermana, siempre tendrás un lugar especial en mi corazón.

A mi tutor y amigo Miguel (Cupi): por tanta entrega, por siempre apoyarnos en todo lo que hicimos, por siempre estar a nuestro lado y confiar siempre en nosotras, por ser mi amigo, gracias por todo.

A Yanet, Yiri y al Flaco por todo, por estar conmigo en las buenas y malas, por aguantar mis malcriadeces en estos largos años y por confiar en mí.

A Tata (Lore) y a Luis: gracias por todo, por su preocupación, entrega y cariño hacia mí.

A mis compañeros de aula Yudelmis, Yelena, Laura, Daneidys, Alberto, Mateo, El Chino, Ernesto, Pepe, Javier Domínguez, Eduardo Valdez, Adonys, Yamel, Pavon, Yadir, Luis Carlos, Ariel por todo, por compartir conmigo todos estos años de universidad, y especialmente a Dayana, Yudeisy, Elieyis, Yusmila, Anayvis, Arlenis, Ilearsi, Eduardo y Orestes por ser tan especiales conmigo y compartir momentos de alegría y tristeza conmigo, los quiero mucho.

A mis niñas Made y Maye: por siempre preocuparse por mí, por tanto amor y cariño, gracias niñas.

A Idalmis, Lisandra, Xiomara, Marisleidys, Petri, Yailín, Yaima, Roxi, Pedrito, Adalberto, el chino, Idel, Daniel, Yosbel, Frank, Sandy, Angel, Julio Cesar, Jackson, Joel a todos, gracias por tanto apoyo y cariño, los quiero mucho.

A Noraimis aunque ya no esté aquí en la universidad, siempre fue una buena amiga.

A mis compañeras de equipo de kiki Yisel, Dalvis, Yaneisy, Ivett, Myrel, Chely, Diana, Zulema en fin a todas, gracias por pasar con ustedes unos de los momentos más bonitos de mi estancia aquí en la uci, y a las que se quedan, a ganar el año que viene, arriba pingüinas.

A Maruja, Mariuska y Lucia por siempre estar preocupadas por mí, por siempre tener confianza en mí.

A Lemay y a Yuseli gracias por todo, por siempre estar atentos a todo lo de mi tesis, gracias.

A mis amigas de impresión Albis, Clarita, Isa, y Merci, gracias por todo, por su apoyo y cariño.

A mis profesores por contribuir en mi preparación profesional especialmente a Bárbaro, Hilda, Lisset, Rill, Zenaida, Juan José, Yanko, Ariel, Eluirkis, Dunia y a todos que de una manera u otra contribuyeron a mi preparación como profesional.

A todos los que de una manera u otra confiaron en mí gracias y a los que no, aquí les va mi triunfo.

A la Revolución Cubana y a Fidel.

Yordailis Morales Díaz

Dedicatoria

Este trabajo que es parte de todo lo que me va a preparar el futuro, va dedicado a mi familia, a todos ellos que de una forma u otra me han dado un pedacito de su corazón y con él toda la confianza del mundo para llegar hasta aquí. A ustedes van dedicados todos mis logros, no solo estos, todos los que a partir de este momento sea capaz de alcanzar porque gracias a ustedes hoy he llegado hasta donde estoy. Formamos la mejor familia del mundo, a veces podía decir que damos envidia de todo el amor que desprendemos, de la amistad que llevamos, del amor que hemos sido capaz de formar, aquí, donde los primos somos hermanos, los tíos son padres, mi hermano y mis padres lo son todo, los adoro y los amo por estar siempre a mi lado.

También dedico este trabajo a mi chikitintin por su amor y apoyo incondicional.

A mis amigos (as), y todos los que de una manera u otra han formado parte de mi superación todos estos años en la universidad.

Karina Collazo Oliva

Le dedico este trabajo a mis ídolos, a mis guías y a las personas que más quiero en el mundo, a mis padres por ser lo mejor que me ha dado la vida, por estar conmigo en los momentos buenos y malos, por mimarme, por sentirse orgullosos de mí, por quererme tanto día a día. Por estar orgullosos de esta personita que los lleva en lo más profundo de su corazón, gracias por confiar en mí. Y decirles que todo lo que hago es pensando primeramente en ellos. A mi novio por todo el amor que siempre me ha brindado, por estar conmigo en todo momento. A mi hermano por todos estos años de cariño, amor, entrega, de malcriarme tanto y por ser mí mejor amigo. A mis abuelas por todo el amor infinito que siempre me han dado. A toda mi familia que siempre han estado queriéndome, ayudándome y confiando en mí. Y a todas las personas que me quieren y que siempre han estado a mi lado, a mis amigos, gracias a todos.

Yordailis Morales Díaz

Resumen

Nos hallamos en una nueva era donde la información crece a ritmo vertiginoso. Gracias a las nuevas tecnologías disponemos de más canales para su transmisión, y los nuevos soportes nos facilitan su registro, su almacenamiento y su recuperación.

D´TIC es un proyecto que está destinado a poner en manos de los usuarios del Ministerio de la Informática y las Comunicaciones (MIC) gran cantidad de información para que estos se mantengan al tanto de lo último en tecnología, el proyecto cuenta con varias secciones entre ellas la de Vigilancia Tecnológica (VT) en la cual existe gran flujo de información constante.

El presente trabajo de diploma se centra en la necesidad de proveer una mejor gestión y recuperación de la información en la sección Vigilancia Tecnológica haciendo usos de la técnica de Minería de Datos. Para darle cumplimiento a la siguiente investigación, se realizó un estudio profundo de las técnicas de Minería de Datos con el fin de poder dar una propuesta de implementación para un futuro desarrollo de la misma en la sección Vigilancia Tecnológica en D´TIC.

Después de realizar la propuesta se validó la misma obteniendo como resultado un mejoramiento en la gestión de la información que se maneja en el portal.

Palabras claves: Recuperación de la información, Gestión de la información, Minería de Datos.

Índice

Índice de figuras.

Introducción

La información es considerada un recurso estratégico de gran valor para el buen desempeño de las organizaciones. Es un elemento del cual se puede extraer conocimiento y satisfacer las necesidades de personas e instituciones, razón por la cual adquiere una importancia significativa para el desarrollo, equilibrio y adaptabilidad en cualquier sector del mundo. Gran parte de esta existe en forma de texto: libros, periódicos, informes técnicos, etc. La calidad de todo este conocimiento depende de nuestra habilidad de hacer ciertas operaciones, por ejemplo:

- Buscar información.
- Comparar fuentes de información diferentes y obtener conclusiones.
- Procesar los textos (por ejemplo, traducirlos, editarlos.)

La información que se encuentra almacenada en las grandes bases de datos en muchas ocasiones se dificulta su recuperación, siempre y cuando la institución no cuente con un algoritmo para la recupera la misma.

Mediante un sistema de recuperación de información se podrá acceder a los datos que ya se encuentren almacenados, mediante herramientas informáticas que permitan establecer ecuaciones de búsquedas específicas. Dicha información ha debido ser estructurada previamente a su almacenamiento.

La gran cantidad de datos que se almacenan en las organizaciones hace imposible la utilización de métodos manuales para su análisis. Por ello son necesarias técnicas y herramientas informáticas capaces de ayudar al hombre de una forma inteligente en el análisis de grandes volúmenes (1).

La necesidad del análisis de los datos y la extracción de conocimiento no implícito en los mismos de forma automática derivó en el nacimiento de una nueva disciplina denominada Knowledge Discovery in Data Base (KDD). Con el nacimiento de esta disciplina los datos pasan de ser el producto generado por los diferentes procesos inherentes a la actividad desarrollada a ser la materia prima, de forma que a partir de ellos se extrae conocimiento útil que ayuda a la toma de decisiones en los ámbitos de donde fueron extraídos (2).

Este proceso comprende varias etapas, que van desde la obtención de los datos hasta la aplicación del conocimiento adquirido en la toma de decisiones. Entre esas etapas, se encuentran la que puede considerarse como el núcleo del proceso KDD y que se denomina Minería de Datos o Data Mining (MD) (2).

La Minería de Datos ha sido usada como sinónimo de descubrimiento de conocimiento en bases de datos (del inglés Knowledge Discovery in Data Base, KDD), sin embargo, corresponde a una de las fases de todo el proceso de descubrimiento, encargada de hacer uso de técnicas de aprendizaje automático para desarrollar algoritmos capaces de aprender y extraer conocimiento de los datos (2).

La Minería de Datos consiste en la extracción no trivial de información que se encuentra implícita en los datos. Esta información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras la Minería de Datos se basa en preparar, sondear y explorar los datos para extraer la información oculta en los mismos. Para el responsable del sistema, normalmente no son los datos en sí lo más importante, sino la información que se encierra en sus relaciones, fluctuaciones y dependencia.

El Ministerio de la Informática y las Comunicaciones (MIC), fue creado en febrero del 2000 cuya misión principal es impulsar, facilitar y ordenar el uso masivo de servicios y productos de las Tecnologías de la Información, las Comunicaciones, la Electrónica y la Automatización de todos los procesos para satisfacer las expectativas de todas las esferas de la sociedad.

A mediados del 2005 la Consultoría Delfos, como coordinadora de la Red de Información, presentó a la dirección del MIC el proyecto: "*Portal de Recursos de Información*". El cual ha compartido y puesto a disposición de los usuarios varias secciones de interés, entre las que se encuentran Biblioteca, Ofertas Formativas, Eventos, Utilidades y Vigilancia Tecnológica (VT). Esta última está destinada a poner a disposición de los usuarios del sector de las TICs en nuestro país gran variedad de información relacionada con las tendencias de las tecnologías y sus avances en la actualidad.

La Vigilancia Tecnológica es un proceso informacional y constituye una gran herramienta de apoyo y desempeño de las instituciones. La puesta en práctica de esta le proporciona a las organizaciones disponer de la información precisa para apoyar a las decisiones de manera oportuna, además de poder anticiparse a los cambios con el menor riesgo posible.

Debido al gran flujo de información en esta última sección dentro del portal D´TIC, la recuperación de

esta a pesar de contar con un buscador, que facilita este proceso al usuario, se ha vuelto crucial, por lo que se ha decidido utilizar la Minería de Datos para hacer las búsquedas más factibles y organizadas para los usuarios.

Del análisis de la problemática expuesta anteriormente surge el siguiente **problema científico**: ¿Cómo mejorar el proceso de recuperación de información en la sección Vigilancia Tecnológica de D'TIC?

Se define como **idea a defender** la propuesta para la utilización de la Minería de Datos en la recuperación de información en la sección Vigilancia Tecnológica de D´TIC permitirá organizar mejor los resultados de las búsquedas para el usuario final.

Y para lograr este propósito se identificó como **objeto de estudio**, la utilización de la Minería de Datos en la recuperación de información. Enmarcando el **campo de acción** la utilización de la Minería de Datos en la recuperación de información en la sección Vigilancia Tecnológica de D´TIC.

Para dar respuesta a este problema se asume como **objetivo general** elaborar una propuesta para la utilización de la Minería de Datos en la recuperación de información en la sección Vigilancia Tecnológica de D'TIC, derivándose a la vez los siguientes **objetivos específicos**:

- Sistematizar los aspectos teóricos que sustentan el proceso de Minería de Datos, abarcando las principales técnicas, herramientas y procesos que se utilizan a nivel mundial.
- Desarrollar una propuesta que defina las técnicas, procesos y herramientas a utilizar en la solución a desarrollar en la sección Vigilancia Tecnológica de D'TIC.

Para darle cumplimiento a estos objetivos se han establecido las siguientes **tareas investigativas**:

- Profundización teórica del proceso de Minería de Datos, incluyendo sus principales características. Tomando como referencia la estructura y los procesos de la Minería de Datos en el mundo.
- Definir los procesos para la utilización de la Minería de Datos para la recuperación de información en la sección Vigilancia Tecnológica de D'TIC.
- Realizar propuesta de una metodología, tarea, técnica y herramienta de Minería de Datos para la recuperación de información para una futura implementación en la sección Vigilancia Tecnológica de D´TIC.

Para dar cumplimiento a las tareas de investigación propuesta anteriormente, se utilizarán los métodos

científicos de investigación teóricos y empíricos.

Los **métodos teóricos** posibilitan estudiar las características del objeto de investigación que no son observables directamente, facilitan la construcción de modelos e hipótesis de investigación. De los cuáles se emplearán:

- El método **histórico – lógico**, con el objetivo de analizar el proceso de Minería de Datos, viendo las principales etapas de su desarrollo para entender la lógica interna de su evolución y alcanzar un conocimiento más profundo de su esencia.
- El método **analítico – sintético**, mediante el cual conoceremos las teorías y documentos existentes sobre el proceso de la Minería de Datos.

Por su parte los **métodos empíricos** describen y explican las características fenomenológicas del objeto, representa un nivel de la investigación cuyo contenido procede de la experiencia y es sometido a cierta elaboración racional. Dentro de estos se emplearán:

- El método de **observación**, que como instrumento universal del científico, se realiza para apreciar cómo avanza el proceso de Minería de Datos.
- El método de **encuesta**, para realizar la validación de la propuesta que daremos.

El presente trabajo consta de una introducción, tres capítulos, conclusiones generales, recomendaciones, referencias bibliográficas utilizadas durante el desarrollo del mismo, glosario de términos y por último, los anexos que completan el cuerpo del trabajo.

Capítulo 1. Fundamentación Teórica sobre Minería de Datos.

En este capítulo se ven presente los diferentes conceptos, teorías y demás aspectos que deben tenerse en cuenta para el desarrollo de la investigación; así como el análisis e investigación de las diferentes organizaciones e instituciones a nivel mundial, nacional y en la universidad que utilizan las técnicas de Minería de Datos.

Capítulo 2. Propuesta para la utilización de Minería de Datos.

En este capítulo después de un estudio sobre la Minería de Datos se realiza una propuesta para una futura implementación en la sección de Vigilancia Tecnológica en el proyecto D´TIC, quedando bien claro la metodología, la tarea, el método, el algoritmo y la herramienta a utilizar.

Capítulo 3. Validación de la propuesta.

En este capítulo se va a realizar la validación del problema planteado, en el cual de acuerdo con los resultados alcanzados se va a proponer la utilización de la técnica de Minería de Datos en la sección Vigilancia Tecnológica en el proyecto D´TIC para que alcance un nivel de calidad elevado para

satisfacer las necesidades de los usuarios del MIC.

Capítulo 1. Fundamentación Teórica sobre Minería de Datos.

En este capítulo se abordan conceptos y definiciones referentes al tema de la Minería de Datos, con el fin de ofrecer una visión más detallada y profunda sobre el contenido a tratar en el presente Trabajo de Diploma. Además se recogen los aspectos más significativos relacionados con la temática, abordados tanto en distintas fuentes bibliográficas como en criterios emitidos por diferentes especialistas. Se hará un estudio del estado del arte del tema referenciado, donde se realizará un análisis del enfoque de la Minería de Datos en la recuperación de información a nivel mundial, y luego se hará un análisis de las distintas metodologías, tareas, técnicas, algoritmos y herramientas de la Minería de Datos para posteriormente pasar a una comparación antes de dar una propuesta para una futura implementación.

1.1. Minería de Datos enmarcado en el ámbito internacional.

Vivimos hoy un proceso cada vez más acelerado de renovación tecnológica en Internet. Esta rapidez tiene como consecuencia que, cada día la interacción con esta sea más dinámica, lo cual ha generado grandes volúmenes de datos que al analizarlos correctamente le aportará información de gran utilidad a las organizaciones. Debido a la necesidad de la extracción de información de manera organizada se hizo necesaria la utilización de la Minería de Datos, que permiten de manera rápida y eficiente el descubrimiento de información relevante dentro de las grandes bases de datos.

Son muchas las empresas que usan Minería de Datos. En todo lo que tiene que ver con ventas, se emplea para identificar clientes potenciales. Los bancos también recurren a ella para detectar fraudes y verificar datos de tarjetas de crédito. En áreas más específicas como seguridad computacional, se utilizan estas técnicas para detectar intrusos. En las distintas empresas para la extracción de información relevantes y previamente desconocida de las bases de datos (3).

Por esta razón, adaptar la práctica internacional de la Minería de Datos a la realidad cubana constituye un reto a la creatividad.

1.2. Minería de Datos enmarcado en el ámbito nacional.

En nuestro país la informática en los últimos años ha alcanzado un gran avance, tanto es así que no se ha quedado atrás con el progreso en la aplicación de la Minería de Datos para un mejoramiento en la gestión de información en las distintas empresas u organizaciones.

El Sistema Cubano de Farmacovigilancia se vio en la necesidad de utilizar herramientas de análisis, por lo que se trazó el objetivo de definir, diseñar y desarrollar los sistemas de tratamiento de la información y administrar la base de datos nacional "VigiBaseCuba". Aplicando una serie de transformaciones, validaciones y la adecuación de la metodología CRISP-DM para la elaboración de proyectos de Minería de Datos, se conformó la base de datos nacional, en un sistema de gestión de bases de datos relacional con los registros de las notificaciones de sospechas de reacciones adversas a los medicamentos y un proceso de descubrimiento de conocimiento que permite gestionar eficazmente la seguridad de los medicamentos, así como desarrollar aplicaciones para la visualización de las señales de reacciones adversas y su evolución.

También existe el Centro de Estudios de Reconocimiento de Patrones y Minería de Datos el cual fue creado el 5 de abril del 2005 a partir de un grupo de profesores del Departamento de Computación perteneciente a la Facultad de Matemática y Computación de la Universidad de Oriente. Está orientado a la investigación básica y aplicada en el área del Reconocimiento de Patrones y su aplicación a la Minería de Datos y Textos.

Las investigaciones actuales incluyen el desarrollo de algoritmos para el procesamiento y análisis de grandes volúmenes de información estructurada o textual. Además, el Centro se dedica a la impartición de cursos de postgrado y una especialización en Reconocimiento de Patrones dentro de la maestría en Ciencia de la Computación de la Universidad de Oriente.

En la Universidad de las Ciencias Informáticas (UCI) existe un área específica donde se utiliza la Minería de Datos de manera constante, esta es el departamento de seguridad informática de la universidad.

En esta área es muy utilizada esta técnica para realizar registros de los logs constantemente. Se han desarrollado varias herramientas encaminadas a dar solución a objetivos específicos, entre las que se encuentran, una herramienta desarrollada para realizar escaneos a la red para verificar si las computadoras conectadas a la misma se encuentran en el dominio UCI, si está infectada por algún virus en particular y si existen conexiones desde otras computadoras a una en específico.

Se utilizan también herramientas reconocidas a nivel mundial, como son:

Lan guardian: Es una herramienta inteligente, la cual trabaja directamente con el proxy y posee una base de conocimientos muy amplia de manera tal que cuando un usuario visite un sitio que sea ocio

aun si no está bloqueado, esta basándose en la información contenida en su base de conocimiento es capaz a partir de los patrones establecidos bloquear dicha página.

Sawmill Enterprise: es un mecanismo muy potente que estable el procesamiento a través de patrones de reconocimiento arrojando como resultado los datos especificados sobre un conjunto de logs determinados.

En la facultad 10 en el proyecto Aires Web se está realizando una investigación sobre Minería Web, la cual no es más que una extensión de la Minería de Datos encargada de aplicar sus técnicas a documentos y servicios de la Web, esta investigación podrá resultar de gran utilidad para una futura implementación en el proyecto.

También se está realizando una tesis de maestría por el Lic. Darian Horacio Grass Boada, la cual se centra en aplicar esta técnica a los registros de navegación en la dirección de redes y seguridad informática de la Universidad de las Ciencias Informáticas, debido a la incapacidad de extraer modelos o patrones de navegación por los sistemas informáticos actuales.

1.3. ¿Qué es la Minería de Datos?

La Minería de Datos consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la Minería de Datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bajo el nombre de Minería de Datos se engloba un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de grandes procesos ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos.

1.3.1. Otras definiciones.

La Minería de Datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto (4).

La Minería de Datos se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos. La Minería de Datos se define también como el análisis y descubrimiento de conocimiento a partir de datos (5).

La Minería de Datos prepara, sondea y explora los datos para sacar la información oculta en ellos (6).

Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos. (Fayyad y otros, 1996). La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo[1] hacia la toma de decisión (7). (Molina y otros, 2001) (8).

1.4. ¿Para qué usar Minería de datos?

La Minería de Datos es una herramienta fundamental para la toma de decisiones. El proceso de aprendizaje de los datos juega un papel muy importante en muchas áreas de la ciencia, las finanzas y la industria, donde las entidades o las empresas han de minimizar los riesgos en la toma de decisiones estratégicas.

1.5. Descubrimiento de conocimiento en las bases de datos (KDD).

El término KKD (iniciales de Knowledge Discovery in Data Base), acuñado en 1989 se refiere a todo el proceso de extracción de conocimiento a partir de una base de datos y marca un cambio de paradigma en lo que lo importante que es el conocimiento útil que seamos capaces de descubrir a partir de los datos.

Otros conceptos dignos de destacar de KDD serían:

- KDD: Aplicación de métodos y herramientas provenientes de campos interdisciplinarios, sobre una cantidad grande de datos para lograr la extracción o generación de nuevo conocimiento útil y apoyar la toma de decisiones dentro de un sistema organizacional.

[1] oblicuidad, inclinación, través, bies, desviación cariz, curso, rumbo, dirección, sentido, orientación, tendencia, giro, marcha

- KDD: es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y por último comprensibles en los datos.

El descubrimiento de conocimiento en base de datos (KDD) hace referencia a un método que consta de una serie de fases.

KDD, consta básicamente de las siguientes fases: Preparación de los datos, Minería de Datos, Evaluación, Difusión y Uso de Modelos, las cuales se explicarán brevemente.

1.5.1. Fase de Preparación de los Datos.

Se va a dividir en diferentes componentes para facilitar la comprensión de este proceso.

Integración y recopilación:

Lo primero que se debe hacer para poder realizar un análisis de datos, es contar con ellos. Por lo que se hace necesaria la integración y recopilación desde diferentes fuentes, sean estas internas o externas. Cuando la integración de datos está presente, es porque se sobrentiende que éstos provienen de diferentes fuentes, que comúnmente están en diferentes formatos y que se necesita integrarlos a un mismo sistema de almacenamiento que cumpla con las características propias de una estructura ya definida.

En tanto la recopilación permite que estos datos sean encontrados, después de analizar cuáles son los necesarios para poder decir definitivamente que estos representan las actividades y procesos del negocio, representadas también en las correspondientes variables que caracterizan hechos en particular.

Normalmente para llevar a cabo la recopilación e integración de los datos se utiliza la tecnología de Almacenes de datos (Data Warehouse), y diferentes metodologías que permiten llevar a cabo su uso de manera apropiada según las características de los datos, y que posibiliten la extracción de conocimiento. Aunque no en todos los casos es necesaria la existencia de un Almacén de datos, estos si son la tecnología más común en las empresas. En otras ocasiones se realizan trabajos de Minería sobre bases de datos relacionales, archivos de textos planos y hojas de cálculo.

En el Data Warehousing, los datos se almacenan con una estructura de bases de datos multidimensional, donde cada dimensión corresponde a un atributo, y un conjunto de atributos

corresponde a unos hechos que almacenan el valor de alguna medida, bien sea en mayor o menor detalle. Estos pueden utilizarse de muy diferentes maneras, y pueden agilizar muchos procesos diferentes de análisis.

Limpieza y transformación:

Después de que los datos hayan sido integrados y recopilados, el siguiente paso (si se hace necesario) consiste en verificar su calidad para que existan las condiciones apropiadas para su análisis, lo cual devuelve al final (una vez aplicada también la fase de exploración y selección) lo que se conoce como vista minable, la cual no es más que una estructura de datos con propiedades necesarias para ser trabajada por algoritmos de Minería.

Gran parte de este proceso de limpieza y transformación es realizado durante las mismas etapas de integración y recopilación mencionada anteriormente, pero no se ha considerado completamente la importancia de la limpieza y transformación de los datos.

Durante la limpieza de los datos, se utilizan técnicas como los histogramas, detección de valores anómalos, y otros tipos de visualización, para detectar y solucionar problemas de los datos no resueltos durante la integración.

Las operaciones de transformación de atributos son utilizadas para presentar los datos de una manera idónea para las herramientas de Minería de Datos, haciendo uso de técnicas de discretización, numerización, de reducción y aumento de dimensionalidad, entre otras.

Exploración y selección:

Se debe realizar un reconocimiento o análisis exploratorio de los datos que se hallen en los Almacenes de Datos, con el objetivo de conocerlos mejor y poder definir cuáles datos seleccionar y determinar las tareas a realizar sobre los mismos.

Para poder iniciar la exploración de los datos, es necesario en primer lugar, que quien realice la exploración tenga un buen conocimiento del dominio de la organización, de los usuarios y respondiendo a preguntas tales como: ¿Qué aspectos son cruciales en el negocio?, ¿Qué modelos de decisión se están utilizando? Luego de tener buenas respuestas a estas preguntas y considerar que se tiene un buen dominio sobre el reconocimiento de la organización, se procede a explorar y reconocer

los datos aplicando las técnicas mencionadas anteriormente.

Luego de haber explorado a profundidad los datos, es preciso decidir por medio de la selección qué atributos o variables se van a necesitar y lo mismo, que cantidad de instancias se debe seleccionar (ejemplos), es decir, se debe seleccionar cuales columnas y cuantas filas se van a necesitar. Así, se logra reducir el tamaño para no desbordar la capacidad de análisis de algunos algoritmos de Minería de Datos y obtener resultados más rápidos.

1.5.2. Fase de Minería de Datos.

Para seguir de forma secuencial cada una de las fases que se ejecutan en el proceso de KDD, continua la fase Minería de Datos, pues de la anterior ya se obtuvo como resultado una vista minable lista para que sean aplicados los algoritmos correspondientes y luego de esto obtener como resultado un modelo que será evaluado en la siguiente fase.

En esta fase se aplican procesos a los datos para extraer patrones y obtener posteriormente conocimiento útil y apoyar la toma de decisiones en las empresas, todo esto por medio de la aplicación de diferentes técnicas. Más adelante se trabajará de una forma más extensa la Minería de Datos, argumentando las tareas, técnicas, algoritmos y herramientas más utilizadas en el mundo.

1.5.3. Fase de Evaluación e Interpretación.

Cuando se trata de evaluación, se refiere a la validación y verificación de los modelos que se van a plantear en la fase anterior, y así poder reconocer en qué nivel de madurez o de refinamiento se encuentra el modelo y si es el más adecuado para resolver el problema planteado. La cuestión en este punto es medir la calidad de los patrones encontrados luego de haber aplicado los algoritmos de Minería de Datos, con el objetivo de estimar su validez y confrontarlos con otros patrones o modelos hallados en iteraciones anteriores.

Existen dos aproximaciones que permiten evaluar los modelos, como los son la evaluación de hipótesis basada en precisión (partición de los datos en los conjuntos de entrenamiento y el de prueba), y la evaluación basada en costes.

Evaluación basada en precisión:

Consiste en separar el conjunto de datos que conforman la vista minable, en dos subconjuntos

disjuntos: el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Esto con el fin de probar la precisión del modelo como una medida independiente.

Evaluación basada en coste:

En este tipo de evaluación, se evalúa el coste de los errores cometidos por un modelo, es decir, el mejor modelo es el que comete menos errores con el menor coste asociado, y no el modelo que cometa menor número de errores.

1.5.4. Fase de Difusión, Uso y Monitorización.

En esta fase, una vez construido y validado el modelo puede usarse principalmente con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien en aplicar el modelo a diferentes conjuntos de datos. También puede incorporarse a otras aplicaciones.

Tanto en el caso de una aplicación manual o automática del modelo, es necesario su difusión, es decir que se distribuya y se comunique a los posibles usuarios, por los medios establecidos. También es importante medir lo bien que el modelo evoluciona. Aun cuando el modelo funcione bien debemos continuamente comprobar las prestaciones del mismo ya que los patrones pueden cambiar.

1.6. Inicios de la Minería de Datos.

La Minería de Datos, como ya se ha mencionado antes, hace parte importante del proceso de descubrimiento de conocimiento en bases de datos, se dice que es parte fundamental, puesto que sin esta sería imposible hablar del término “descubrimiento” en el proceso de KDD. Siendo ella la fase más importante dentro del proceso global de KDD, esta merece gran atención y estudio por parte de aquellos quienes tienen la oportunidad de aplicar conceptos y llevar a cabo esta fase dentro de la organización.

Desde un punto de vista histórico ha sido el resultado de un largo proceso de investigación que comenzó en los años 60 con el desarrollo de los primeros sistemas de almacenamiento y recuperación de datos. Fue ya a mediados de los años 80 que se dan los primeros pasos hacia la Minería de Datos como la conocemos en la actualidad cuando las bases de datos no solo se limitaron a almacenar información sino que se le adicionaron técnicas de procesado y modelado de datos. La necesidad de almacenar la información ha motivado el desarrollo de sistemas cada vez más eficientes y con una mayor capacidad de almacenamiento.

1.7. Ventajas de la Minería de Datos.

La Minería de Datos posee muchas ventajas entre las que se encuentran:

- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- Contribuye a la toma de decisiones tácticas y estratégicas.
- Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma.
- Genera modelos descriptivos: permite a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.
- Genera Modelos predictivos: permite que relaciones no descubiertas través del proceso de la Minería de Datos sean expresadas como reglas de negocio (10).
- Auxilia a los usuarios empresariales en el procesamiento de reservas de datos para descubrir relaciones de las que, en algunos casos, anteriormente ni siquiera se sospechaba.
- La información obtenida a través de la minería de datos ayuda a los usuarios a elegir cursos de acción y a definir estrategias competitivas, porque conocen información que sólo ellos pueden emplear.
- Los seres humanos tienen la capacidad para percibir excepciones y anormalidades rápidamente pero no tienen la habilidad para inferir relaciones que se encuentran en grandes volúmenes de datos, por lo que la Minería de Datos, mediante modelos avanzados y reglas de inducción, puede examinar grandes cantidades de datos y encontrar patrones difíciles de identificar a simple vista.
- Pueden trabajar siguiendo los mismos criterios con grandes cantidades de información histórica.
- El proceso de búsqueda puede ser realizado por herramientas que automáticamente buscan patrones porque así están programadas y despliegan los tópicos más importantes (11).

1.8. Extensiones de la Minería de Datos.

La Minería de Datos tiene varias extensiones como son:

1.8.1. Minería Web (Web Mining).

Una de las extensiones de la Minería de Datos consiste en aplicar sus técnicas a documentos y servicios de la Web, lo que se llama Web Mining (Minería Web) (Kosala y otros, 2000). Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador) que los servidores automáticamente almacenan en una bitácora de accesos (log). Las herramientas de Web Mining

analizan y procesan estos logs para producir información significativa, por ejemplo, cómo es la navegación de un cliente antes de hacer una compra en línea. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, investigaciones recientes usan el término multimedia Data Mining (Minería de Datos multimedia) como una instancia del Web Mining (Zaiane y otros, 1998) para tratar ese tipo de datos. Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del Web Mining.

La Minería Web, cada día es más popular y extendida, la cual es utilizada para analizar el tráfico de acceso a un determinado servidor web, previamente registrado de una manera apropiada, para ayudar, por una parte, a entender el comportamiento y hábitos de los clientes/usuarios del servidor y, por otra parte, a diseñar adecuadamente la estructura de la web o mejorar el diseño de esta inmensa colección de recursos.

Normalmente, el Web Mining puede clasificarse en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

- Web content Mining (minería de contenido web). Es el proceso que consiste en la extracción de conocimiento del contenido de documentos o de sus descripciones. La localización de patrones en el texto de los documentos, el descubrimiento del recurso basado en conceptos de indexación o la tecnología basada en agentes también pueden formar parte de esta categoría.
- Web structure Mining (minería de estructura web). Es el proceso de inferir conocimiento de la organización del WWW y la estructura de sus ligas.
- Web usage Mining (minería de uso web). Es el proceso de extracción de modelos interesantes usando los logs de los accesos al web.

Algunos de los resultados que pueden obtenerse tras la aplicación de los diferentes métodos de web Mining son:

- El ochenta y cinco por ciento de los clientes que acceden a /productos/home.html y a /productos/noticias.html acceden también a /productos/historias_suceso.html. Esto podría indicar que existe alguna noticia interesante de la empresa que hace que los clientes se dirijan a historias de suceso. Igualmente, este resultado permitiría detectar la noticia sobresaliente y colocarla quizá en la página principal de la empresa.
- Los clientes que hacen una compra en línea cada semana en /compra/producto1.html tienden a ser de sectores del gobierno. Esto podría resultar, en proponer diversas ofertas a este sector

para potenciar más sus compras.

- El sesenta por ciento de los clientes que hicieron una compra en línea en /compra/producto1.html también compraron en /compra/producto4.html después de un mes. Esto indica que se podría recomendar en la página del producto 1 comprar el producto 4 y ahorrarse el costo de envío de este producto.

Los anteriores ejemplos nos ayudan a formarnos una pequeña idea de lo que podemos obtener. Sin embargo, en la realidad existen herramientas de mercado muy poderosas con métodos variados y visualizaciones gráficas excelentes (12).

1.8.2. Minería de Texto (Text Mining).

La Minería de Texto se enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo es descubrir cosas tales como tendencias, desviaciones y asociaciones entre la gran cantidad de información textual. La Minería de Texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos.

Estudios recientes indican que el ochenta por ciento de la información de una compañía está almacenada en forma de documentos. Sin duda, este campo de estudio es muy vasto, por lo que técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al Text Mining (Minería de Texto). En ocasiones se confunde el Text Mining con la recuperación de la información (Information Retrieval o IR) (Hearst, 1999). Ésta última consiste en la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, categorización, etc. Generalmente se utilizan palabras claves para encontrar una página relevante. En cambio, el Text Mining se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo (Nasukawa y otros, 2001).

Una aplicación muy popular del Text Mining es relatada en Hearst (1999). Don Swanson intenta extraer información derivada de colecciones de texto. Teniendo en cuenta que los expertos sólo pueden leer una pequeña parte de lo que se publica en su campo, por lo general no se dan cuenta de los nuevos desarrollos que se suceden en otros campos. Así, Swanson ha demostrado cómo cadenas de

implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica. Algunas de esas claves fueron:

- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria.
- El magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que Swanson encontró mediante esas ligas. De acuerdo con Swanson (Swanson y otros, 1994), estudios posteriores han probado experimentalmente esta hipótesis obtenida por text Mining con buenos resultados.

Objetivos de la Minería de Texto.

- Búsqueda de conocimiento útil en grandes cantidades de información no estructurada.
- Utilización de este conocimiento para mejorar la organización de información.
- Ampliar los horizontes de aplicaciones y resultados que la minería de texto ofrece.
- Proporcionar una visión más amplia y selectiva de la información.
- Estructurar esta información para transformarla en conocimiento útil.

Beneficios de la Minería de Texto.

- Reduce el tiempo empleado para tomar decisiones.
- Mejora el desempeño, ahorrando dinero y horas de trabajo.
- Logra una visión más exacta de la documentación interna.
- Reconoce y anticipa oportunidades de negocio.

Campos de aplicación de la Minería de Texto.

- Estudios e investigaciones: Detecta tendencias en los estudios de ciencia y tecnología.
- Marketing y relaciones públicas: Análisis de grupos focales, entrevistas abiertas y quejas del

cliente.

- Medicina y salud pública: Detectar tendencias en historias clínicas y diagnósticos.
- Medios informativos: Análisis de noticias, encuestas de opiniones y archivos de prensa (13).

1.8. Ciclo de la Minería de Datos.

El proceso de la Minería de Datos es un ciclo, debido a que los resultados obtenidos pueden alimentar nuevamente dicho proceso; intervienen, principalmente, cuatro pasos que se describen a continuación:

- Los usuarios de la información deberán identificar los problemas del negocio y las áreas en donde los datos pueden dar valor agregado a la empresa, esto es: a raíz de un problema surge la necesidad de analizar a detalles los datos de la empresa para poder encontrar posibles soluciones al mismo, o informaciones que hagan que las decisiones tomadas sean lo más certeras posibles. Asimismo, es importante identificar las áreas en donde la información es muy cambiante, pero primordial para la competitividad de la empresa. Para esto pueden manejarse diferentes criterios, no se puede decir específicamente cuáles son los correctos debido a que esto depende de las características de la empresa, pero el objetivo a perseguir es determinar los criterios, ideas, normas y cuestionamientos que servirán como entrada para el proceso de Minería de Datos.
- El usuario para analizar la información histórica seleccionará el algoritmo o algoritmos adecuados de minería. Posteriormente, estos algoritmos son traducidos a programas mineros que realizarán las búsquedas con los criterios previamente definidos
- Existen varias dificultades que pueden interferir con el resultado que se obtenga del análisis y esto es porque los datos se pueden encontrar en diferentes formas, formatos y en múltiples sistemas, aún dado a que pueden provenir de fuentes internas o externas; para resolver este problema actualmente se ha hecho uso del data warehouse, que pretende reunir los datos más importantes de la empresa en una especie de base de datos corporativa, la cual requiere de una gran cantidad de gigabytes, no siempre disponibles en las organizaciones, sin embargo, es posible hacer Minería de Datos sin necesidad de tener el data warehouse, pero es muy importante tener claro que la información deberá estar lo más uniforme y congruente posible, ya que mucho depende de esto la certidumbre de los resultados que arroje.
- Incorporar la información obtenida a través del proceso de Minería de Datos al proceso de toma de decisiones; así como presentar los hallazgos encontrados a los responsables de las operaciones de forma que la información obtenida pueda integrarse en los procesos de la empresa y pueda aplicarse en la solución de los problemas.

- Medir los resultados: Medir el valor de los hallazgos encontrados, que se le proporcionan al tomador de decisiones con relación a la solución de los problemas identificados y a los criterios definidos en el primer punto (11).

1.9. Aplicación de la Minería de Datos.

- Ayuda a la navegación.
- Detección de ataques y fraudes en comercio electrónico
- Mejoras del diseño de los sitios.
- Caracterización de visitantes.
- Personalización de páginas.

Actualmente se aplica en áreas tales como:

- Aspectos Climatológicos: predicción de tormentas.
- Medicina: encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.
- Mercadotecnia: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos, etc.
- Inversión en casas de bolsa y banca: análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- Análisis de canastas de mercado para mejorar la organización de tiendas, segmentación de mercado (clustering).
- Determinación de niveles de audiencia de programas televisivos.
- Industria y manufactura: diagnóstico de fallas (10).

1.10. Fases del proceso de Minería de Datos.

Los pasos a seguir para la realización de un proyecto de Minería de Datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

El proceso de Minería de Datos se compone de las siguientes fases:

1.10.1. Selección y pre procesado de datos

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los Datos "en bruto".

Mediante el pre procesado se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos... según las necesidades y el algoritmo que va a usarse), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reduce el número de valores posibles (mediante redondeo, clustering) (14).

1.10.2. Selección de variables

Aún después de haber sido pre procesado, en la mayoría de los casos se tiene una cantidad ingente[2] de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la selección de los mejores atributos del problema.
- Los que buscan variables independientes mediante pruebas de sensibilidad, algoritmos de distancia o heurísticos (14).

1.10.3. Extracción de conocimiento

Mediante una técnica de Minería de Datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesado diferente de los datos (14).

1.10.4. Interpretación y evaluación

Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste

[2] Enorme, inmenso, gigantesco.

mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos (14).

1.11. Metodologías para la utilización de la Minería de Datos.

Las metodologías hacen referencia a un conjunto de procedimientos basados en principios lógicos, utilizados para alcanzar una gama de objetivos determinados que rigen una investigación científica o tecnológica. Estas son las encargadas de estudiar los métodos que se utilizan para alcanzar el objetivo de un proyecto.

La Minería de Datos utiliza varias metodologías entre las cuales se encuentran: CRoss-Industry Standard Process for Data Mining. (CRIPS – DM) y Sample, Explore, Modify, Model, Assess (SEMMA).

1.11.1. CRIPS-DM

La metodología CRISP-DM consiste en un conjunto de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada, e instancia de proceso, organizados en/de forma jerárquica en tareas que van desde el nivel más general hasta los caso más específicos.

Fase: Se le denomina fase al asunto o paso dentro del proceso. CRISP-DM consta de seis fases: comprensión del negocio o problema, comprensión de los datos, preparación de los datos, modelado, evaluación, implementación.

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.

Tarea especializada: La tarea especializada describe como se puede llevar a cabo las tareas genéricas en situaciones específicas.

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

Las fases del proyecto de la Minería de Datos de acuerdo a lo establecido por la metodología CRISP-DM es que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. La secuencia de las fases no siempre es ordenada, en ocasiones, si se determina que al realizar la evaluación de los objetivos del negocio estos no se cumplieron, se debe regresar y buscar las causas del problema para redefinirlo.

Comprensión del negocio.

Esta fase inicial se enfoca en la comprensión de los objetivos del proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de Minería de Datos y en un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos.

La fase de entendimiento de datos comienza con la colección de los datos iniciales y continúa con las actividades que le permiten familiarizar primero los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Preparación de los datos.

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (que serán provistos en las herramientas de modelado) de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescrito. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Modelado.

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores ópticos. Típicamente hay varias técnicas para el mismo tipo de problema de Minería de Datos. Algunas técnicas tienen requerimientos específicos sobre la toma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

Evaluación.

En esta etapa del proyecto, ya se ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.

Antes del proceder al despliegue final del modelo, es importante evaluar a fondo la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos del negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de Minería de Datos debería ser obtenida.

Implantación.

La creación del modelo es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización.

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. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

1.11.2. SEMMA.

SA Systems es el Instituto 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 las cinco fases básicas del proceso.

La metodología SEMMA se caracteriza principalmente por priorizar sus fases desde un punto de vista técnico, es decir, dando prioridad a las prácticas usadas para su implementación y obtención de resultados.

SEMMA procede de los cinco pasos de la fase de análisis dentro de un proyecto de Minería de Datos. Estos cinco pasos son:

- Muestreo.
- Exploración.
- Modificación.
- Modelización.
- Estimación.

El ciclo de análisis SEMMA.

Muestreo.

El proceso se inicia al crear una o más tablas utilizando muestras de los datos contenidos en el Data Warehouse. El objetivo de esta fase consiste en seleccionar muestras representativas, estas muestras deberían ser lo suficientemente grandes como para contener información significativa, aunque lo suficientemente pequeñas como para poder procesarse con rapidez. Este enfoque permite obtener resultados «coste-eficientes». Al explorar una muestra representativa en lugar del volumen completo de información se reduce de forma drástica el tiempo de procesamiento requerido. La representatividad de la muestra es indispensable ya que de no cumplirse, invalida todo el modelo y los resultados dejan de ser admisibles.

Si existen patrones generales en el conjunto de la muestra, éstos también estarán presentes en una muestra representativa. Si un nicho es tan pequeño como para no estar representado en la muestra y pese a ello tan importante como para influir en la imagen completa, podrá descubrirse utilizando métodos de sumarización.

La metodología SEMMA establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra.

Exploración.

Una vez determinada una muestra o conjunto de muestras representativas, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. El siguiente paso a seguir es la exploración visual o numérica con el fin de observar tendencias inherentes o agrupaciones. Si la exploración visual no revela tendencias claras, los analistas pueden explorar los datos mediante técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables, análisis de correspondencia y clustering, además propone la utilización de herramientas de visualización. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.

Modificación.

La tercera fase de la metodología consiste en la manipulación o modificación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo.

Por modificación de los datos entendemos la creación, selección y transformación de una o más variables para centrar el proceso de selección de modelos en una dirección particular o para aumentar

los datos para obtener claridad o coherencia.

Basándose en los descubrimientos de la fase de exploración, los analistas pueden necesitar tratar los datos para incluir información tal como la agrupación de los clientes y subgrupos significativos o introducir nuevas variables tales como un ratio obtenido comparando dos variables anteriormente definidas. Puede que los analistas también necesiten buscar valores extremos y reducir el número de variables para limitarlas a las más significativas. Además, dado que la Minería de Datos es un proceso dinámico e iterativo, a menudo es necesario modificar los datos cuando los datos anteriormente extraídos sufran algún cambio.

Modelo.

Crear un modelo de datos implica la utilización de una solución de minería que busque automáticamente una combinación de datos que prevean de forma fiable un resultado deseado.
Después de que se haya accedido a los datos y éstos se hayan modificado, los analistas pueden utilizar técnicas de modelización para construir modelos que expliquen patrones. Las técnicas de modelización en la Minería de Datos incluyen redes neuronales, modelos basados en árboles, modelos logísticos y otros modelos estadísticos tales como análisis de series temporales y análisis de supervivencia.

Estimación.

El siguiente paso en todo proyecto de minería de datos consiste en estimar el modelo para su posterior evaluación. Un método común para evaluar un modelo es aplicarlo a la porción de los datos que se dejaron de lado durante la etapa de muestreo. Si el modelo es válido debería funcionar para esta muestra reservada, de igual modo que funciona para la muestra utilizada para construir el modelo.
De forma similar, los analistas pueden probar el modelo utilizando datos conocidos. Por ejemplo, si se sabe qué clientes tuvieron altos índices de retención y el modelo se ha definido para predecir la retención, los analistas pueden comprobar si el modelo señala a esos clientes. Además, la aplicación práctica del modelo, por ejemplo hacer envíos parciales en una campaña marketing directo, ayudan a probar su validez.

Iteración.

Aunque estimar los modelos de datos es el último paso en la metodología SEMMA, estimar la eficacia de modelos de datos a menudo no es el paso final en una implementación real de SEMMA.

Como SEMMA es un ciclo, los pasos internos se suelen realizar iterativamente dentro del conjunto del

proyecto de minería de datos.

1.11.3. Comparación entre las metodologías CRIPS y SEMMA.

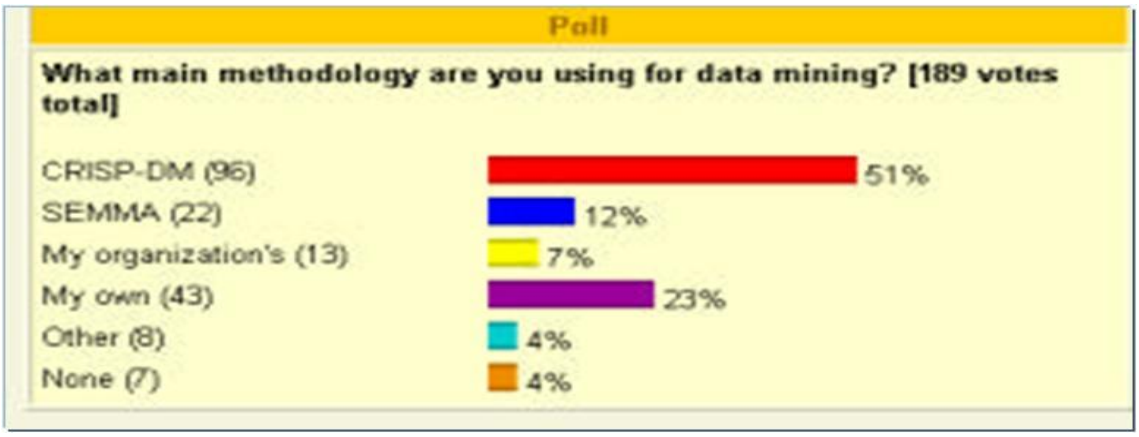

Figura 1. 1 Comparación de las metodologías en cuanto a su uso (15).

SEMMA	**CRIPS**
Orientado al desarrollo del proceso de MD	Orientado a los objetivos empresariales
Se inicia analizando los datos	Se inicia analizando los objetivos del negocio
Ligada a productos SAS	Metodología abierta y gratuita
	Orientado a una metodología de gestión de proyectos

Figura 1. 2 Comparación en cuanto a las características que presentan cada metodología.

1.12. Principales Tareas de la Minería de Datos.

Las dos principales tareas de la Minería de Datos pueden ser la predicción o la descripción. La predicción utiliza algunas variables o campos de datos para predecir el comportamiento futuro de otras variables. La descripción se centra en encontrar patrones que describan el comportamiento de los datos al usuario.

Clasificación: Es la encargada de agrupar a todas las herramientas que permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Más concretamente cada instancia o registro de la base de datos pertenece a una determinada clase la cual se indica mediante el valor de un atributo denominado clase de la instancia, el objetivo fundamental es predecir cuál sería la clase de nuevas instancias de las que se desconoce la clase; clasifica un dato dentro de una de las clases categóricas predefinidas. Se usan árboles de decisión y sistemas de reglas o análisis de discriminantes.

Regresión: Esta tarea es el aprendizaje de una función real que asigna a cada instancia un valor real de tipo numérico, tiene como objetivo inducir un modelo para poder predecir el valor de la clase, dados los valores de los atributos; el propósito de este nuevo modelo es hacer corresponder un dato con un valor real de una variable. Se usan árboles de regresión, redes neuronales artificiales y regresión lineal. La principal diferencia respecto a la clasificación consiste en que el valor a predecir es numérico.

Clustering (Agrupamiento): Esta tarea de Minería de Datos se encarga de establecer grupos de objetos que presenten características similares.
La Minería de Datos con clustering identifica patrones de comportamiento comunes en un conjunto de datos que los usuarios no podrían derivar de forma lógica a partir de una observación casual. Un cluster (grupo) es entonces una colección de objetos que son similares entre ellos y distintos de los objetos que pertenecen a otros grupos.

Reglas de asociación: Es una técnica importante en la Minería de Datos y consiste en encontrar las asociaciones interesantes en forma de relaciones de implicación entre los valores de los atributos de los objetos de un conjunto de datos. Es una de las tareas reinas de la Minería de Datos y que ha evolucionado conjuntamente, desde mediados de los 90, con la propia Minería de Datos. Surge debido a la necesidad de analizar grandes cantidades de datos recolectados en grandes almacenes y así identificar relaciones no explícitas entre los atributos.

Correlación: Es una técnica que buscara el grado de similitud de los valores de dos atributos numéricos. El grado de similitud se mide por el coeficiente de correlación r (rÎ [-1... 1]): si r es positivo los atributos tienen un comportamiento similar (ambos crecen o ambos decrecen al mismo tiempo), si r es negativo cuando un atributo crece el otro decrece, si r es cero no existe relación entre ambos atributos. El objetivo de esta tarea es de poder describir de forma concisa relaciones existentes entre

atributos del conjunto de ejemplos.

1.13. Métodos de la Minería de Datos.

Los métodos de Minería de Datos pueden ser usados para entender los datos espaciales, descubrir relaciones entre ellos, reorganizarlos en las bases de datos y determinar sus características generales de manera simple y concisa.

Estos métodos pueden ser de aprendizaje supervisado y no supervisado.

Los algoritmos supervisados o predictivos predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se le induce una relación entre esta etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usado un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en este caso hay que recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas (16).

Redes Neuronales.

Como su nombre lo indica simula el sistema nervioso real en forma abstracta. Estas deben ser entrenadas para que den solución a los problemas. Esta enseñanza se realiza repitiendo sistemáticamente entradas clásicas, con sus respectivas salidas o respuestas. Es un modelo predecible, no lineales que aprenden a través del entrenamiento. Son usadas para reconocimiento de patrones, clasificaciones de voz e imagen, procesamiento de lenguaje natural, predicción y optimización (17).

Esta técnica de inteligencia artificial, en los últimos años se ha convertido en uno de los instrumentos de uso frecuente para detectar categorías comunes en los datos, debido a que son capaces de detectar y reconocer complejos patrones, y características de los datos.

La teoría de las redes neuronales ha brindado una alternativa a la computación clásica, para aquellos

problemas, en los cuales los métodos tradicionales no han entregado resultados muy convincentes o poco convenientes. Las aplicaciones más exitosas de estas son:

- Procesamiento de imágenes y de voz.
- Reconocimiento de patrones.
- Planeación.
- Interfaces adaptativas para sistemas hombre/maquina.
- Predicción.
- Control y optimización.
- Filtrado de señales.

Una de las principales características de las redes neuronales, es que son capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente. Además esta técnica posee dos formas de aprendizaje: supervisado y no supervisado (18).

Las ventajas de las redes neuronales son:

- Aprendizaje adaptativo. Capacidad de aprender a realizar tareas basadas en un entrenamiento o una experiencia inicial.
- Auto organización. Una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa del aprendizaje.
- Generalización. Facultad de las redes neuronales de responder apropiadamente cuando se les presentan datos o situaciones a los que no habían sido expuestas anteriormente.
- Tolerancia a fallos. La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo gran daño. Con respecto a los datos, las redes neuronales pueden aprender a reconocer patrones con ruido, distorsionados o incompletos.
- Operación en tiempo real. Los computadores neuronales pueden ser realizados en paralelo, y se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.
- Fácil inserción dentro de la tecnología existente. Se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilita la integración modular en los sistemas existentes (18).

Sin embargo, aunque son clasificadores muy precisos, no son comúnmente utilizados en Minería de Datos porque producen modelo de aprendizaje inexplicables y requieren de mucho tiempo de

entrenamiento. Las redes neuronales pueden predecir con precisión, pero son como una caja negra, lo que quiere decir que son estudiadas desde el punto de vista de las entradas que recibe y las salidas o respuestas que produce, sin tener en cuenta su funcionamiento interno (19).

Arboles de decisión.

Se pueden aplicar a casi todo. La técnica basada en árboles de decisión es quizás el método más fácil de utilizar y entender. Está técnica se encuentra dentro de una metodología de aprendizaje supervisado. Su representación es en forma de árbol en donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos.

Los árboles de decisión son un método excelente para ayudar a realizar elecciones apropiadas entre muchas posibilidades. Su estructura permite seleccionar una y otra vez distintas opciones para explorar las diferentes alternativas posibles de decisión.

Son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación, presenta otras ventajas como son:

- Facilita la interpretación de la decisión de búsqueda adoptada.
- Proporciona un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
- Explica el comportamiento respecto a una determinada tareas de decisión.
- Reduce el número de variables independientes.

A pesar de todas estas ventajas también tiene sus deficiencias y una de ellas es que puede llegar a ser más lento, pues analiza todas las posibilidades existentes. Además de que una mala elección de la partición (especialmente en las partes superiores del árbol) generará un peor árbol (20).

Algoritmos Genéticos.

Los algoritmos genéticos imitan la evolución de las especies mediante la mutación, reproducción y selección, como también proporcionan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como es el caso de las redes neuronales. Además los algoritmos genéticos son inspirados en el principio de la supervivencia de los más aptos. Es una técnica matemática de búsqueda y optimización que encuentra soluciones a un problema basándose en los principios que rigen la evolución de las especies a nivel genético molecular. Estos algoritmos requieren de un conjunto de datos para realizar su proceso de aprendizaje (21).

Clustering (Agrupamiento).

Los algoritmos de Clustering permiten clasificar un conjunto de elementos de muestra en un determinado número de grupos basándose en las semejanzas y diferencias existentes entre los componentes de la muestra. Estos agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de Minería de Datos, dando como resultado un sistema híbrido.

El tipo de modelo de aprendizaje es no supervisado lo cual quiere decir que esta técnica se encarga de descubrir patrones y tendencias en los datos. Un problema relacionado con el análisis del clúster es la selección de factores en tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos. Otro problema de gran importancia y que actualmente despierta un gran interés es la fusión de conocimiento, ya que existen múltiples fuentes de información sobre un mismo tema, los cuales no utilizan una categorización homogénea de los objetos. Para poder solucionar estos inconvenientes es necesario fusionar la información a la hora de recopilar, comparar o resumir los datos.

El resultado de un análisis usando técnicas de clustering es cierto número de clusters que forman una partición o una estructura de particiones, del conjunto global de datos en el espacio de Minería de Datos. El modelo de Minería de Datos representa de una manera apropiada y fácil los datos a analizar (22).

1.14. Algoritmos de la Minería de Datos.

Los algoritmos de la Minería de Datos pueden ser clasificados en distintas formas en dependencia de la función que realicen o a la fase que correspondan:

1.14.1. Algoritmos de Clusterizado.

Los métodos de agrupamiento deben definir una función útil de clasificación sobre un conjunto Xi (donde i=1,…, N) cuando N generalmente no es determinada y el número de grupo que se va a formar es desconocido por lo que los algoritmos de clustering utilizan una técnica basada en dos pasos donde un bucle exterior que considera los posibles números de grupos y un bucle interior que ajusta de la mejor manera posible los datos a ese número fijo de grupos (23).

- **Algoritmo vecinos más cercanos**: Éste dado un número k de grupos tiene como objetivo encontrar la mejor k-partición de forma que los patrones de cada grupo de la partición estén más cercano entre sí que los patrones de los otros grupos. Una vez determinada la partición se puede intentar representar cualquier nuevo patrón en función del más cercano. Existen varias variaciones de este algoritmo que se pueden agrupar en 4 grandes grupos:

- Método de las K-medias: Es el más sencillo de los algoritmos de agrupamiento habituales. Sean p vectores de n características o vectores característicos Xj (j=1,…, p) que pueden agruparse en N clases cada uno de sus miembros Ni (i=1,…, N):

  Se elige una serie de valores del espacio como centros, a partir de las cuales se empezara a generar clases o grupos, cada vez que se presenta un patrón se calcula su distancia a todas las medias y se le asigna la clase cuya media sea la más cercana.

  Se recalcula entonces la media de esta clase como el baricentro de todos los puntos que pertenecen a ellas, incluido el último asignado de la forma siguiente:

  Siendo Xj (j=1,…, Ni) patrones asignados a Mt; y se repite la operación tantas veces como puntos se quiera clasificar o hasta que la media en el paso t+1 sea igual a la del paso t.
- Método de K-NN o K vecinos más cercanos: es uno de los algoritmos más antiguos, surgió cuando se observó que el fijarse en un solo patrón provoca que la existencia de un único punto defectuoso desvíe la clasificación sin remedio. Durante la clasificación se calcula la distancia entre los patrones de entrada y los ejemplos almacenados. Se buscan los k ejemplares más cercanos y se asignan al patrón de entrada la clase más abundante entre estos k ejemplos.
- Algoritmo LVQ (Learning Vector Quantization): estos sólo almacenan un número controlable de patrones. El entrenamiento en este algoritmo se realiza en varias etapas. En primer lugar se determina el número de ejemplo a almacenar generalmente con el método de K-Medias anteriormente mencionado u otro procedimiento de clustering. A partir de los ejemplares se asignan cada patrón de entrenamiento al ejemplar más cercano, penalizando si es la clase incorrecta y beneficiando si es de la correcta.
- Método de las distancias encadenadas (Chain-map): Consiste en elegir un vector característico al azar Xi de los p que se tiene y colocarlo en la primera posición de una lista. Después se coloca en posición siguiente de la lista el vector más cercano al primero. Se elige el siguiente más cercano al último de la lista, y así sucesivamente, quedando esta de la siguiente manera Xi(0), Xi(1), Xi(2), …, Xi(p-1) donde X(1) es el vector más cercano al X(0), X(2) es el más cercano a X(1) y así sucesivamente. Una vez obtenido este valor se calcula las distancias euclídias entre ellos y se representa gráficamente, donde se puede distinguir la distancia entre

ellos (23).

- **Método de clusterizado de Montaña**: este pertenece a las técnicas avanzadas de clusterizado, consiste en crear una rejilla donde las regiones entre las intersecciones de las líneas son posibles candidatos a clusters, con centro en la intercepción de la línea. Después se crea una función montaña que representa la densidad de datos de cada punto de la rejilla. Se selecciona el que presenta mayor altura de todos c1, se realiza una substracción de la montaña original con una montaña de centro en c1 y distribución Gaussiana obteniéndose una nueva montaña, luego se repite este procedimiento hasta que no quede ningún punto sin clasificar o la montaña que se obtenga tenga la altura menor que un umbral definido (23).

1.14.2. Algoritmos de Reglas de Asociación.

Los algoritmos de aprendizaje de reglas de asociación se basan en la búsqueda de reglas que cumplan unos requisitos mínimos de confianza, soporte o cobertura.

- **Algoritmo Apriori:** es un algoritmo muy simple y utilizado. Se basa en la búsqueda de los conjuntos de ítems con una determinada cobertura, para esto primeramente se construye los conjuntos formados por un ítem que superar la cobertura mínima. Este conjunto de conjuntos se utiliza para construir el conjunto de conjuntos de dos ítems, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de ítems con la cobertura requerida (23).

1.14.3. Arboles de Decisión

Son unos de los algoritmos más empleados en la Minería de Datos. Se basan en la partición del conjunto de ejemplo según ciertas condiciones que es aplicada a los valores de las características, su potencia descriptiva viene limitada por las condiciones o reglas con las que se dividen el conjunto de entrenamiento. La construcción de un árbol de decisión se realiza de forma recursiva, primeramente se selecciona un atributo como nodo principal o raíz del árbol y a partir del mismo se divide el conjunto de observaciones en dos o más sub-series de datos según el valor del atributo y se repite recursivamente para cada rama. Entre los algoritmos de árboles de decisión podemos encontrar CART, ID3, C4.5 (C5.0), SLIQ y M5.

- **CART:** Realiza particiones binarias con una estrategia de poda basada en un criterio de coste complejidad. Estas particiones binarias son el resultado de evaluar una condición que tiene dos únicas respuestas. La formulación de la regla de partición se realiza a partir de un conjunto estándar de preguntas.

- **ID3:** Es conocido también como TDIDT se basa en la entropía como función de impureza. La entropía o valor de información se define como la medida de incertidumbre que hay en un sistema, es decir ante una determinada situación la probabilidad de que ocurra cada uno de los posibles resultados. A cada nodo se le asocia aquel atributo con mayor decrecimiento en la función de impureza que aun no se haya considerado en la trayectoria desde la raíz. Este algoritmo tiene entre sus inconvenientes su predisposición a favorecer indirectamente a aquellos atributos con muchos valores, los cuales no tienen porque ser necesariamente los más útiles.
- **C4.5:** Este algoritmo y su extensión C5.0 es una extensión del ID3 que incluye varias mejoras como la de construir árboles de decisión cuando algunos de los ejemplos presentan valores desconocidos en algunos atributos, puede trabajar con atributos que presenta valores continuos, tolerancia a datos con ruido y genera reglas a partir de árboles.
- **SLIQ**: Es un algoritmo creado para enfrentar problemas con grandes cantidades de datos, para la construcción del árbol Tmax utiliza la misma función de impureza que el algoritmo CART. El esquema utilizado en la fase de poda se fundamenta en el principio de longitud de descripción mínima (MDL). MDL establece que el coste total de la codificación de unos datos D mediante un modelo M viene dado por la suma del coste de bit de codificar los datos dado un modelo M y el coste de codificar el modelo M.
- **M5:** Es un algoritmo desarrollado e implementado en WEKA y es un árbol de regresión donde el final de cada rama es la clase donde se representa mediante el promedio del valor de las observaciones que han llegado hasta ella (de regresión) o mediante un modelo de combinación lineal (árbol de modelizado). Este algoritmo tiene mecanismos para trabajar eficientemente con valores inexistentes, ruidos e incluso con valores nominales que son convertidos previamente en valores numéricos binarios (23).

1.14.4. Generadores de Reglas

Como los árboles de decisión en aplicaciones reales tienden a ser muy grandes y difíciles de interpretar se ha tratado de convertir estos en otras formas de representación como las reglas inducidas. Algunos de estos algoritmos son AQ, CN2, RIPPER, INDUCT, PART, FOIL, CLINT.

- **AQ:** Este tipo de algoritmo tiene sus raíces e influencias en métodos de ingeniería eléctrica utilizados para la simplificación de circuitos eléctricos. La estrategia seguida por el mismo es de abajo-hacia-arriba donde inicialmente cada uno de los patrones de entrenamiento se considera

un complejo, luego estos complejos son examinados, eliminando selectores de forma selectiva y garantizando la consistencia del complejo resultante, de esta forma en cada etapa se construye un complejo y mediante la combinación de los complejos generalizados se construye un recubrimiento completo que cubre todos los patrones de una determinada clase. Por lo que su objetivo se basa en encontrar un conjunto de recubrimiento compacto que cubra todos los posibles casos. AQ permite encontrar un conjunto completamente consistente de reglas con todos los datos de entrenamiento pero no puede clasificar correctamente con ruido ni considerar ninguna estrategia para evitar el sobreentrenamiento.

- **CN2:** Este algoritmo se crea como una extensión del AQ permitiendo el tratamiento de ruido y sobreentrenamiento. Este retiene un conjunto de complejos durante la búsqueda, de forma que estos complejos cubren un gran número de casos de una clase, aunque también pueden cubrir casos de otras clases. De forma adicional este algoritmo realiza un proceso de especialización por lo que en cada paso de especialización se le añaden nuevas preguntas o se elimina todo el complejo. En la búsqueda de los mejores complejos se utiliza dos tipos de heurística: de significancia y de bondad donde significancia es el umbral por debajo del cual no se considera un complejo para ser seleccionada como mejor complejo y la bondad la cual es una medida de la cualidad del complejo utilizada para establecer un orden entre los complejos candidatos a la inclusión final en el recubrimiento.
- **RIPPER:** Conjunto con el CN2 y C4.5 han sido los métodos básicos de este tipo de algoritmos. Este es muy parecido a C4.5 ya que genera una serie inicial de reglas, lo que luego en C4.5 son depurada y en este algoritmo son combinadas. Este genera reglas muy simples que luego se recombinan y reemplazan en otras más complejas y más complejas.
- **INDUCT:** Se basa en una versión más sencilla nombrada PRISM que usa AQ e ID3 así como su extensión C4.5 para generar reglas diferentes para cada clase a clasificar. Usa distribución binomial para determinar la bondad de una regla, lo que permite crear mejores reglas con datos con cierto porcentaje de ruido. Este ha sido mejorado con el algoritmo RDR para incluir reglas con excepciones.
- **PART:** Es un algoritmo generador de reglas basado en subárboles el cual está implementado en WEKA, está basado en técnicas para generar árboles y reglas, de forma que genera subárboles que luego son convertidos a reglas. Es bastante robusto a ruidos y valores ausentes.
- **FOIL:** Se basa en técnicas de aprendizaje con lógica de predicados ILP, esta lógica inductiva de predicado es especialmente útil cuando se disponen de una base de conocimiento que al

aplicarla se puede alimentar al sistema con una serie de reglas que ayudan a obtener nuevos patrones de comportamiento de los datos.

- **CLINT:** Se basa en la construcción de un árbol que explica los ejemplos negativos, identificar las cláusulas que cubran los casos negativos, borra las cláusulas de la base de conocimiento y recomponer la estructura de conocimiento estudiando las cláusulas positivas que habían sido cubiertas por la cláusula anterior (23).

1.15. Herramientas de la Minería de Datos.

Hoy en día son muchas las herramientas de software para la Minería de Datos encargadas de la extracción de información significativa dentro de grandes volúmenes de datos almacenados en las bases de datos, entre las herramientas de software se encuentran SAS Systems, Weka y Clementine.

1.15.1. SAS Systems.

SAS INC: Institute es la compañía independiente de software analítico de negocios más grandes del mundo, entregando la tecnología que su empresa necesita para cambiar la manera de hacer negocios. El avance de la empresa es debido al buen trato y servicios que les son proporcionados a los usuarios, y a la reinversión realizada en investigación y desarrollo, con el propósito de mantener sus productos al margen de la tecnología.

El sistema se encuentra hoy en día instalado en más de 45.000 sitios en 113 países. El producto original fue desarrollado inicialmente para analizar información proveniente de investigaciones agrícolas, en un mainframe IBM en la Universidad del Estado Carolina del Norte. Debido a los altos costos de los recursos tecnológicos en la década de los 70´s, los mainframe eran usados exclusivamente por oficinas del gobierno, centros de investigación y grandes empresas. Con la introducción de máquinas más pequeñas, poderosas y baratas; hoy en día la computación se encuentra al alcance de muchas más personas.

En la medida que cambió la industria, también lo hizo el Sistema SAS. En la década de los 80's el Sistema SAS fue totalmente reescrito en lenguaje C, a fin de incorporarle la MultiVendor Architecture (MVA), que le permite soportar nuevas plataformas al poco tiempo de liberadas. Debido a la MVA el 90% del código SAS es transportable a cualquier plataforma. Esto significa que como máximo es necesario rescribir únicamente un 10% del código para implementar el Sistema SAS en una nueva plataforma de hardware.

En 1993 la estrategia MVA permitió a SAS Institute liberar en forma casi simultánea nuevas versiones del Sistema SAS para plataformas con 13 sistemas operativos diferentes: MVS, CMS, VSE, OpenVMS para VAX y AXP, OS/2 2.0, Windows 3.1, Windows NT, AIX, HP UX, RISC/ULTRIX, Solaris y ConvexOS.

A través de años de desarrollo, el Sistema SAS se ha transformado en un completo Sistema de Entrega de Información. Ha ido creciendo hasta integrar más de 35 aplicaciones modulares; que permiten a las organizaciones un completo control sobre sus datos, desde el acceso, manejo y análisis, hasta su presentación. El Sistema SAS incluye herramientas para desarrollo de aplicaciones orientado a objetos, capacidades avanzadas de procesamiento cliente/servidor, nueva tecnología de visualización de datos, y acceso ilimitado a los datos (24).

Todas estas características han llevado consigo que SAS Systems sea reconocido por todo el mundo como líder en el software de la minería de datos (25).

1.15.2. WEKA.

La herramienta consiste en un conjunto de librerías en Java que contiene una colección de algoritmos de Minería de Datos los cuales permiten realizar tareas como pre procesamiento y filtrado, agrupamiento, reglas de asociación y visualización. Es un software desarrollado bajo la licencia GPL como código abierto e incluye interfaz gráfica compuesta por diversos entornos, desarrollada por un grupo de investigadores de la Universidad de Waikato de Nueva Zelanda. Se destaca por la cantidad de algoritmos que presenta así como la eficacia de los mismos. Aunque la herramienta está implementada en Java no presenta problemas de portabilidad mientras que el sistema disponga de la maquina virtual adecuada.

Weka tiene 4 entornos de trabajo, el primero Simple CLI es un entorno de consola para con java invocar directamente los paquetes de Weka, el segundo es una interfaz gráfica conocida como Weka Explore en la cual se pueden ejecutar y configurar los algoritmos con los que cuenta esta herramienta, otro es el Experimenter el cual es un entorno centrado en la automatización de tareas de manera que se facilite la relación de experimentos a gran escala y el último KnowledgeFlow que permite generar proyectos de Minería de Datos mediante la generación de flujos de información.

Weka emplea el formato ARFF (Atribute-Relation File Format) como soporte de datos, en el que cada

uno de los ficheros consta de una lista de instancias con los mismos atributos. Los tipos de datos que Weka permite son los numéricos, cadenas de caracteres, nominal y fecha.

1.15.3. Clementine.

Clementine es una herramienta de Minería de Datos que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales, de manera que no es preciso ser un experto en Minería de Datos.
Esta herramienta de software es la solución líder en Minería Datos que le ayuda a las organizaciones a comprender el comportamiento de las personas y a predecir qué es lo que harán. Al utilizar Clementine, los analistas y usuarios de negocios podrán acceder a datos de varias fuentes para producir, evaluar, y desplegar modelos analíticos rápida y fácilmente. La arquitectura abierta y escalable del producto le permite obtener el máximo provecho de la infraestructura actual, haciendo de la Minería de Datos un proceso efectivo en toda su empresa.

Admite la integración con herramientas de modelado y Minería de Datos disponibles en proveedores de bases de datos como OracleZData Miner, IBM DB2 Intelligent Miner y Microsoft Analysis Services 2005. Podrá generar, puntuar y almacenar modelos dentro de la base de datos. Esto permite combinar las capacidades analíticas y la facilidad de uso de la herramienta con la potencia y el rendimiento de una base de datos, al mismo tiempo que se saca partido de los algoritmos nativos de bases de datos proporcionados por estos proveedores.

Actualmente es considerada la herramienta de Minería de Datos más avanzada del mercado (26).

1.16. Conclusiones Parciales del capítulo.

En este capítulo se abarcó de manera detallada todo lo referente al proceso de Minería de Datos así como una explicación bastante amplia sobre las metodologías más usadas a nivel mundial, las tareas que presenta la Minería de Datos, así como los métodos más usados, algoritmos que se pueden emplear para darle solución a un problema determinado así como las herramientas que dan solución a un problema de este tipo.

Capítulo 2. Propuesta para la utilización de la Minería de Datos.

La Minería de Datos ha surgido del potencial del análisis de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoye a la toma de decisiones y que pueda construir una experiencia a partir de las millones de transacciones detalladas que registra una corporación en sus sistemas informáticos. La Minería de Datos parece ser más efectiva cuando los datos tienen elementos que pueden permitir una interpretación y explicación en concordancia con la experiencia humana.
En este capítulo se va a realizar la propuesta para una futura implementación en el proyecto D´TIC de la técnica de Minería de Datos para lograr una mejor recuperación de la información.

2.1. Metodología propuesta.

Para la realización de una futura implementación en el proyecto D´TIC se escogió después de una comparación entre las metodologías expuestas en el capítulo anterior que la metodología por la cual se va a regir el proyecto es la CRoss-Industry Standard Process for Data Mining (CRIPS-DM).

2.1.1. CRISP.

CRIPS es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Minería de Datos.

Los orígenes de esta se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR(Dinamarca), Ag(Alemania), SPSS(Inglaterra), OHRA(Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD(Knowlage Discovery in Data Base) [Reinartz, 1995], [Adraans, 1996], [Brachman, 1996], [Fayyad, 1996] la realización de una guía de referencia de libre distribución denominada CRIPS-DM(CRoss-Industry Standard Process for Data Mining).

Ella está dividida en cuatro niveles de abstracción organizados de forma jerárquica en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Minería de Datos en una serie de seis fases.

La sucesión de fases no es necesariamente estricta. Cada fase es estructurada en varias tareas generales de segundo nivel, las cuales se proyectan en tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún

momento se propone como realizarlas.

1. Fase de comprensión del negocio o problema.

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la Minería de Datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Minería de Datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

- **Determinar los objetivos del negocio**. Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito. Los problemas pueden ser diversos como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del proceso de Minería de Datos, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.
- **Evaluación de la situación**. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de Minería de Datos, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de Minería de Datos?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Minería de Datos.
- **Determinación de los objetivos de Minería de Datos**. Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de Minería de Datos. Finalmente esta última tarea de la primera fase de CRISP-DM, tiene como meta

desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso (27).

2. Fase de comprensión de los datos.

La segunda fase, fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de Minería de Datos, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

Las principales tareas a desarrollar en esta fase del proceso son:

- **Recolección de datos iniciales**. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.
- **Descripción de los datos**. Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.
- **Exploración de datos**. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.
- **Verificación de la calidad de los datos**. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos (27).

3. Fase de preparación de los datos.

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Minería de Datos que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. Una descripción de las tareas involucradas en esta fase es la siguiente: Selección de datos. En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud, corrección de los mismos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

- **Limpieza de los datos**. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.
- **Estructuración de los datos**. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.
- **Integración de los datos**. La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
- **Formateo de los datos**. Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres

especiales, máximos y mínimos para las cadenas de caracteres).

4. Fase de modelado.

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

Una descripción de las principales tareas de esta fase es la siguiente:

- **Selección de la técnica de modelado**. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de Minería de Datos existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales y técnicas de visualización.
- **Generación del plan de prueba**. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de Minería de Datos como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.
- **Construcción del Modelo**. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado

tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

- **Evaluación del modelo**. En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Minería de Datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas).

5. Fase de evaluación.

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Las matrices de confusión Edelstein, 1999 son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

Las tareas involucradas en esta fase del proceso son las siguientes:

- **Evaluación de los resultados**. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar este, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.
- **Proceso de revisión**. El proceso de revisión, se refiere a calificar al proceso entero de Minería de Datos, a objeto de identificar elementos que pudieran ser mejorados.
- **Determinación de futuras fases**. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros

parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de Minería de Datos.

6. Fase de implantación.

En esta fase (**Figura 2.8**), y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Minería de Datos no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

Las tareas que se ejecutan en esta fase son las siguientes:

- **Plan de implantación**. Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación. Monitorización y Mantenimiento. Si los modelos resultantes del proceso de Minería de Datos son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
- **Informe Final**. Es la conclusión del proyecto de Minería de Datos realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el mismo. Revisión: En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

2.2. Tarea de Minería de Datos propuesta.

Todas las tareas que se explican en el capítulo anterior resuelven un problema determinado dentro de la realización de un proyecto de Minería de Datos, en el caso de D´TIC específicamente en la sección de Vigilancia Tecnológica a la cual esta aplicada esta tesis investigativa, se necesita de la utilización de

una técnica que sea capaz de agrupar la información de tres maneras diferentes:

- Alertas de prensa.
- Alertas Tecnológicas.
- Riesgos y beneficios de las TIC´s.

Se decidió que se va a utilizar una tarea descriptiva la cual se centra en encontrar patrones que describan el comportamiento de los datos al usuario, específicamente la tarea de agrupamiento.

2.2.1. Agrupamiento o Clustering.

La tarea de Clustering consiste en la división de los datos en grupos de objetos similares. El representar los datos por una serie de clusters, conlleva a la perdida de detalles pero consigue la simplificación de los mismos. Clustering es una tarea descriptiva. Desde el punto de vista práctico, juega un papel muy importante en la aplicación de Minería de Datos, tales como la exploración de datos científicos, recuperación de información y minería de texto, aplicaciones sobre base de datos especiales, aplicaciones Web, marketing, diagnóstico médico y análisis de ADN en biología computacional.

Esta tarea se utiliza para el análisis de los datos, que resuelve problemas de clasificación. Su objetivo es la distribución de casos en grupos, de manera tal que el grado de asociación entre los miembros del mismo grupo es fuerte y débil entre los miembros de los distintos grupos. De esta manera cada grupo se describe en términos de los datos recogidos. Clustering es una tarea de descubrimiento.

En resumen: Clustering intenta encontrar grupos naturales de componentes sobre la base de cierta similitud.

2.3. Método de Minería de Datos propuesto.

Los métodos de Minería de Datos se utilizan para tener un mejor entendimiento de los datos, dígase: descubrir relaciones entre ellos, reorganizarlos en las bases de datos y determinar sus características generales de manera simple.

Para una futura implementación y basándose en la tarea anteriormente seleccionada se escoge un método de aprendizaje no supervisado específicamente el método de Agrupamiento o Clustering.

2.3.1. Agrupamiento o Clustering.

La técnica de agrupación o Clustering consiste en agrupar un conjunto de datos basándose en la similitud de los valores de sus atributos. El Clustering identifica regiones densamente pobladas, denominadas clusters. De esta manera se busca maximizar la similitud de las distancias en cada

cluster y minimizar la similitud entre clusters. Esta técnica ha sido estudiada en las áreas de la estadística, base de datos especiales y Minería de Datos (28).

Cluster analysis es una expresión acuñada por Tryon (1939). Consiste en diferentes algoritmos de clasificación que organizan una cantidad de información y la convierte en conjuntos comprensibles y manejables denominados clusters. Dado un conjunto de objetos (documentos) y la descripción de un conjunto X, la agrupación o clustering debe dividir el conjunto de objetos en dos partes: los que pertenecen a X y los que no, para lo que será necesario en primer lugar establecer las características que son relevantes para describir los objetos que estarán en X (similitud intra-clustering) y en segundo lugar qué características distinguen los objetos de X de aquellos que no pertenecen a él (inter-clustering) (29).

Existen dos tipos de técnicas de clusters análisis:

- Métodos jerárquicos: cuyos algoritmos reconstruyen la jerarquía completa de los objetos analizados, tanto en orden ascendente como en orden descendente.
- Métodos divisorios o particiones: cuyos algoritmos preveen que el usuario haya definido previamente el número de grupos en los cuales se dividen los objetos analizados.

Clustering aplicada al ámbito de recuperación de información consigue crear de forma automática clasificaciones de documentos considerando ciertas similitudes en su contenido. Para poder crear clusters los documentos se representan como vectores de términos. Hay que tener en cuenta que el tamaño de los vectores es igual al del vocabulario del conjunto recuperado. Para cada documento su vector define un punto en un espacio multidimensional. Las distancias entre los puntos y su posición relativa son indicadores de la similitud entre los documentos. El resultado de un análisis usando técnicas de clustering es cierto número de clusters que forman una partición o una estructura de particiones, del conjunto global de datos en el espacio de Minería de Datos. El modelo de Minería de Datos representa de una manera apropiada y fácil los datos a analizar.

2.4. Algoritmo de Minería de Datos propuesto.

Los algoritmos de la Minería de Datos pueden ser clasificados en distintas formas en dependencia de la función que realicen. Después de un análisis de todos los algoritmos expuestos en el capítulo anterior se decidió que a partir de las tareas y métodos propuestos, se va a utilizar un algoritmo

clusterizado, particularmente el algoritmo vecinos más cercanos del cual se va a utilizar específicamente el método K-means.

2.4.1. Vecinos más cercanos.

Este algoritmo usa razonamiento basado en memoria para las predicciones. Identifica los vecinos más cercanos (valores similares para igual atributo) y observa cómo se comporta la variable de salida. Parte de un conjunto de datos modelo, que representa el mecanismo de clasificación, se determina la cantidad de vecinos que participan en la clasificación (K). Es permitido ponderar atributos para expresar su importancia técnica.

Tiene como objetivo encontrar la mejor k-partición de forma que los patrones de cada grupo de la partición estén más cercanos entre sí que los patrones de los otros grupos (17).

2.4.2. K-means.

- Este método es el más sencillo de los algoritmos de agrupamiento habituales.
- K-means es un método iterativo que busca formar k clusters, con k predeterminado antes del inicio del proceso.
- Es un método particional de clustering donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido.
- Cada grupo está representado por su centro. El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático (**Figura No. 2.10**).

$$V = \sum_{i=0}^{k} \sum_{j \in S_i} \left| x_j - \mu_i \right|^2$$

Figura 2. 1 Error cuadrático.

Donde existen k grupos Si, i=1,2,…, k y μi es el punto medio o centroide de todos los puntos Xj ϵ Si.
K-means comienza particionando los datos en k subconjuntos no vacios, calcula el centroide de cada partición como el punto medio del cluster y asigna cada dato al cluster cuyo centroide sea el más próximo. Luego vuelve a particionar los datos iterativamente, hasta que no haya más datos que cambien de cluster de una iteración a otra (28).

Para calcular el centroide más cercano a cada punto se debe utilizar una función de distancia. Para los datos reales se suele utilizar la distancia euclídea. Para los datos categóricos se debe establecer una

función específica de distancia para ese conjunto de datos. Algunas de las opciones son utilizar una matriz de distancia predefinida o una función heurística. El algoritmo no garantiza que se obtenga un óptimo global. La calidad de la solución final depende principalmente del conjunto inicial de grupos. Debido a esto, se suelen realizar varias ejecuciones del algoritmo con distintos conjuntos iniciales, de modo de obtener una mejor solución.

Dado k, el método k-means se implementa en cuatro pasos:

- Particionar los objetos en k subconjuntos no vacios.
- Computar los centroides de los clusters de la partición corriente. El centroide es el centro (punto medio) del cluster.
- Asignar cada objeto al cluster cuyo centroide sea más cercano.
- Volver al paso dos, para cuando no haya más reasignaciones.

El método es ampliamente utilizado en la explotación de datos, en la cuantificación de vectores, para cuantificar variables reales en k rangos no uniformes y para reducir el número de colores e una imagen.

En resumen: K-means divide los datos en grupo de objetos basándose ampliamente en la información contenida en los datos que describen estos objetos y las relaciones que existen entre ellos.

2.5. Herramienta de Minería de Datos propuesta.

Hoy en día son muchas las herramientas encaminadas a la extracción de información desconocida en las bases de datos siguiendo un algoritmo determinado. Después de un análisis de varias de ellas, se decidió utilizar para una futura implementación en D´TIC la herramienta Weka por todas las características y ventajas que presenta.

2.5.1. Weka.

Weka es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información. Weka desarrollado bajo la licencia GPL se ha convertido en una alternativa muy interesante.

Weka se puede utilizar de 3 formas distintas:

- **Desde la línea de comandos:**

Cada uno de los algoritmos incluidos en Weka se puede invocar desde la línea de comandos de MS-

DOS como programas individuales. Los resultados se muestran únicamente en modo texto.

- **Desde una de las interfaces de usuario:**

Weka dispone de 4 interfaces de usuarios distintos, que se pueden elegir después de lanzar la aplicación completa. Las interfaces son:

Simple CLI (Command Line Interface): interfaz en modo texto, además de intérprete de comandos o consola.

Explorer: interfaz gráfico principal, proporciona acceso a las distintas funcionalidades a través de menús y formularios de datos.

Experimenter: interfaz gráfico con posibilidad de comparar el funcionamiento y rendimiento de diversos algoritmos de aprendizaje, además de distribuir la carga de trabajo entre varias máquinas (experimentos grandes). También automatiza el proceso de ejecución de varios filtros y clasificadores con diferentes parámetros sobre un conjunto de datos y proporciona estadísticas de dicho proceso.

KnowledgeFlow: interfaz gráfico que permite interconectar distintos algoritmos de aprendizaje en cascada, creando una red. Similar al funcionamiento interno del programa. Permite crear una secuencia o circuito que recoge todo el experimento.

- **Creando un programa Java:**

La tercera forma en la que se puede utilizar el programa Weka es mediante la creación de un programa Java que llame a las funciones que se desee. El código fuente de Weka está disponible, con lo que se puede utilizar para crear un programa propio.

Entrada de Datos en Weka.

- Tabla relacional en formato ARFF (attribute-relation file format):

- Cabecera con el nombre de la relación de los datos:
  - ✓ @relation <nombre>, donde <nombre> es una cadena de caracteres.

- Declaración de atributos:
  - ✓ @atribute <nombre> <tipo>, donde <nombre> es una cadena de caracteres y <tipo> puede tomar valores: NUMERIC, INTEGER, DATE y STRING.

- Conjunto de datos:

✓ A partir de la sentencia @data (una instancia por línea con los valores de los atributos separados por comas).

➢ Formato CSV (comma-separated value):

- Pueden generarse a partir de una hoja de cálculo o una consulta a una base de datos.
- Weka los convierte al formato ARFF automáticamente.
- ¡Atención a las comas!

➢ También podemos importar datos de una página web o emplear una consulta a una base de datos.

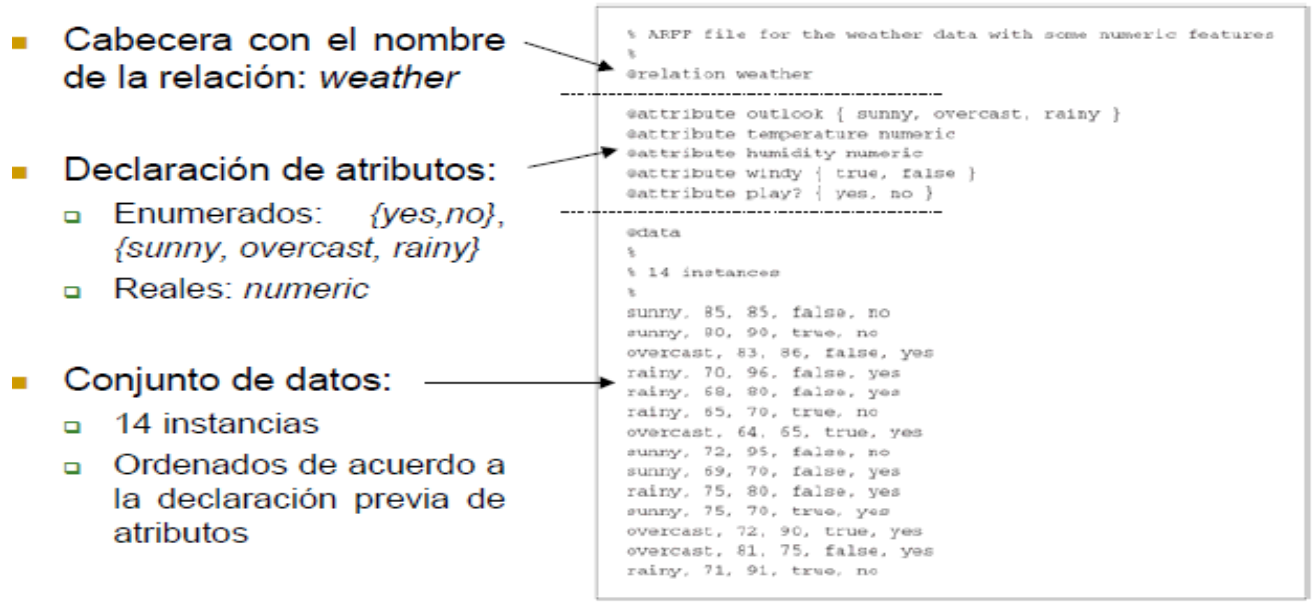

Figura 2. 2 Entrada de Datos (32).

Operaciones básicas con Datos en Weka.

➢ **Selección de atributos:**

- all, none, invert (invertir selección).

➢ **Borrado de atributos:**

- Elimina el atributo seleccionado así como los valores que le hacen referencia presentes en el conjunto de datos.

➢ **Edición:**

- Presenta los datos en forma de tabla.
- Permite modificar instancias, eliminarlas, ordenarlas en base a unos criterios y hacer búsquedas.

➢ **Ver:**

- Características de los atributos: máximo, mínimo, media, desviación típica.
- Distribución por clases (32).

Pestaña Visualizar en Weka.

➢ Permite visualizar el conjunto inicial de datos (no el resultado de aplicar un modelo).
➢ Representación a partir de una matriz o gráfico de valores para cada par de atributos.
➢ Opciones:

- PlotSize, para modificar el tamaño de los gráficos.
- PointSize, para modificar el tamaño de los puntos.
- Jitter, permite separar los puntos o representaciones solapadas (desplazamiento aleatorio).
- Colour, colorea los puntos en función de los valores de los mismos (diferencias entre los atributos nominales y los numéricos).
- Select Attributes, mostrar sólo los gráficos de los atributos seleccionados.
- SubSample %, mostrar el porcentaje indicado de la muestra.
- Podemos ampliar cualquier gráfico pinchando sobre él (32).

Aplicación de Filtros en Weka.

➢ Weka ofrece un conjunto amplio de algoritmos de filtrado.
➢ Se utilizan para transformar los datos y son aplicables a los atributos y las instancias.
➢ Tenemos dos grupos diferenciados de filtros:

- Supervisados (la aplicación incorrecta de estos filtros puede dar lugar a resultados erróneos – el filtro afecta tanto a los datos de entrenamiento como a los de prueba).

✓ Aplicar un filtro supervisado:

- Para evitar los problemas mencionados anteriormente, cargamos el filtro desde la pestaña

Classify.

- Dentro del apartado Classifier, pinchamos en Choose y dentro de meta elegimos FilteredClassifier.
- Pinchamos en el cuadro de texto Classifier y establecemos el filtro supervisado deseado en las opciones.

- No Supervisados.

✓ Aplicar filtro no supervisado:

- Abrimos el fichero de datos correspondiente.
- Dentro de la pestaña Preprocess, pinchamos en Choose dentro del apartado Filter.
- Seleccionamos el filtro deseado de la lista desplegable.
- Pinchamos en el cuadro de texto de Filter para establecer las opciones del filtro (podemos obtener información detallada de estas opciones seleccionando More).
- Aplicamos el filtro (botón Apply) (32).

2.5.2. Weka con Clustering.

Como la tarea, técnica y algoritmo elegido para darle solución al problema existente en D´TIC es el de agrupamiento, se realizará una pequeña introducción al clustering o agrupamiento con Weka.

Los algoritmos de clustering permiten clasificar un conjunto de elementos de muestra en un determinado número de grupos basándose en las semejanzas y diferencias existentes entre los componentes de la muestra.

Vamos a observar cómo se trabaja en la interfaz de usuario de Explorer, ya que permite el acceso a la mayoría de las funcionalidades integradas en Weka de una manera sencilla.

Como se puede observar existen 6 sub-entornos de ejecución en esta interfaz:

- Preprocess: Incluye las herramientas y filtros para cargar y manipular los datos.
- Classification: Acceso a las técnicas de clasificación y regresión.
- Cluster: Integra varios método de agrupamiento.
- Associate: Incluye unas pocas técnicas de reglas de asociación.
- SlectAttributes: Permite aplicar diversas técnicas para la reducción del número de atributos.
- Visualize: En este apartado podemos estudiar el comportamiento de los datos mediante técnicas de visualización (33).

Vamos hacer énfasis en el entorno con el cual se va a trabajar que es el de Cluster.

Pulsando la tercera pestaña, llamada Cluster, en la parte superior de la ventana, accedemos a la sección dedicada al Clustering. El funcionamiento es sencillo, se elige un método clustering, se selecciona las opciones pertinentes y con el botón Start empieza el funcionamiento.

Una opción propia de esta sección es la posibilidad de ver de una forma gráfica la asignación de muestras de clusters. Esto se puede conseguir activando la opción Store cluster for evaluation, y ejecutando el experimento y seguidamente, en la lista de resultados, pulsando el botón secundario sobre el experimento en cuestión y marcando la opción Visualize cluster assignments con esto obtendremos una ventana similar a las del modo explorador para mostrar gráficas en el que nos mostrará el clustering realizado.

2.6. Conclusiones parciales del capítulo.

En este capítulo se realizó la propuesta para una futura implementación en D´TIC de la técnica de Minería de Datos, teniendo en cuenta las características de la metodología, tarea, método, algoritmo y herramienta seleccionada, mirando siempre que cumpla con el perfil de nuestra facultad.

Capitulo 3. Validación de la propuesta.

En este capítulo está presente la validación del problema planteado, en el cual de acuerdo con los resultados alcanzados se van a proponer la utilización de la técnica de Minería de Datos en la sección Vigilancia Tecnológica en el proyecto D´TIC para que alcance un nivel de calidad elevado para satisfacer las necesidades de los usuarios del MIC.

3.1. Métodos de validación existentes.

La validación es la parte más importante dentro de esta investigación, pues es justamente donde se establece si el documento creado se ajusta a las restricciones descritas en el esquema utilizado para su construcción, esta es la confirmación por examen y la provisión de evidencia objetiva de que se cumplen los requisitos particulares para un uso especifico propuesto. Controlar el diseño de documentos a través de esquemas aumenta su grado de fiabilidad, consistencia y precisión, facilitando su intercambio entre aplicaciones y usuarios. En este proceso es donde se comprueba que tanto la precisión de los datos, el conjunto de métodos y técnicas utilizadas se pueden aplicar a una organización.

Existen varios métodos para llevar a cabo la validación entre los que se encuentran:

- Métodos de expertos.
- Estudio de casos.
- Método Delphi.

3.2. Validación de la propuesta.

3.2.1. Validación utilizando el método Delphi.

El método Delphi consiste en poner en práctica un grupo de cuestionarios a un grupo de expertos en la materia.

En este método se seleccionan un grupo de expertos los cuales deben llegar a un consenso en las respuestas que den acerca de una serie de preguntas que se les plantean Se pretende extraer y maximizar las ventajas mediante el criterio emitido por un conjunto de expertos en el tema. Este modelo excluye las discusiones cara a cara entre los miembros del grupo.

La calidad de los resultados utilizando el método Delphi dependen de:

- La elaboración del cuestionario.
- Elección de los expertos consultados.

Este método cuenta con cuatro fases:

- Formulación del problema.
- Elección de expertos.
- Elaboración y lanzamiento de los cuestionarios.
- Desarrollo práctico y explotación de resultados.

Contiene las presentes características:

- Anonimato: En esta no debe existir contacto alguno entre los participantes, pero el que realiza la encuesta si puede identificar a cada participante y sus respuestas.
- Iteración: En esta se pueden hacer tantas rondas como sean necesarias para obtener datos concisos.
- Retroalimentación controlada: Aquí los resultados totales de la primera ronda no se entregan a los participantes, solo una parte seleccionada de la información es la que se muestra.
- Resultados estadísticos: Esta respuesta se puede representar estadísticamente, es decir como promedio.

Con la utilización de este método se desea tener la opinión de varios expertos en el tema de Minería de Datos, ya que así será realizada la validación de la propuesta planteada en el capítulo anterior, mostrando su efectividad para una futura implementación en D´TIC. Se seleccionaron 12 especialistas en los temas de la Minería de Datos y recuperación de información, a los cuales se les aplicaron pregunta reflejadas en el cuestionario que se presenta en los anexos.

3.2.2. Fase Exploratoria.

En esta fase se realizan los cuestionarios, en este caso se realizó solo uno ya que no fue necesario un número mayor. Lugo se pasa a aplicarlos a un conjunto de expertos que son los que interactúan con él y responden las preguntas planteadas. En esta ronda de preguntas se recopilan también las opiniones que cada uno de los especialistas pueda emitir respecto a lo encuestado. Todos estos objetivos son necesarios para la validación de la propuesta. A continuación se realiza la ronda de preguntas, los objetivos y el número correspondiente a la misma.

Ronda de preguntas.

En esta se evalúan algunos de los objetivos más necesarios y cada uno de ellos responde a una

pregunta. Para mayor entendimiento del cuestionario dirigirse a anexo.

Objetivos	1	2	3	4
Necesidad de poner en práctica la propuesta.	x			
Mejora que trae consigo la propuesta.		x		
Necesidad de la recuperación de la información.			x	
Aporte que le brindará a D´TIC.				x

Figura 3. 1 Tabla de objetivos.

Análisis de cada uno de los objetivos medidos en el cuestionario aplicado para proporcionar un mayor entendimiento de lo medible:

Necesidad de poner en práctica la propuesta.

A este primer objetivo se le otorga respuesta en la primera pregunta, a la cual los especialistas en el tema respondieron de manera positiva, pues llegaron a la conclusión que con la puesta en práctica de la propuesta se eliminan las deficiencias existente a la hora de la recuperación de la información en la sección Vigilancia Tecnológica del proyecto D´TIC. De lo antes mencionado surge la gráfica con los valores emitidos por dichos expertos, en la cual se pueden observar las coincidencias establecidas entre los criterios.

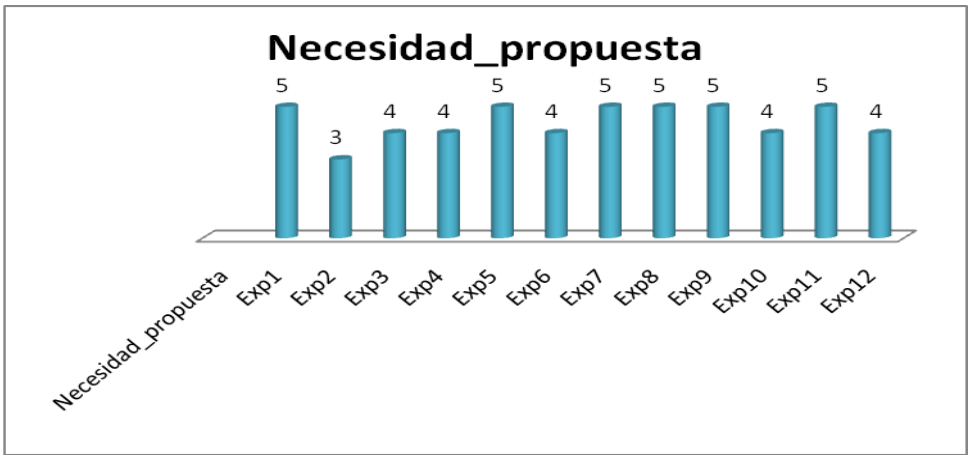

Figura 3. 2 Evaluación del primer objetivo a los expertos.

Mejora que trae consigo la puesta en práctica de la propuesta.

Este objetivo obtiene respuesta con la información recopilada de la segunda pregunta en la cuan se evaluó la opinión de aceptación sobre la medida en que se considera que la propuesta definida va a garantizar una mejora en la implementación de la técnica de Minería de Datos., asumiendo además que con esto se optimice la recuperación de la información facilitando así una mejor gestión por parte de los usuarios del MIC. La respuesta emitida fue satisfactoria ya que concordaron en un 100% que la puesta en práctica de dicha propuesta mejora la gestión de información para los usuarios del MIC. A continuación está el gráfico correspondiente a la pregunta realizada:

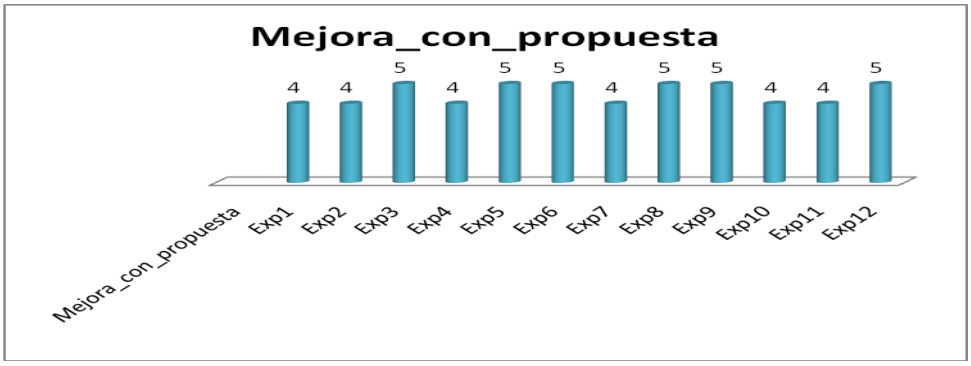

Figura 3. 3 Evaluación del segundo objetivo a los expertos.

Necesidad de la recuperación de la información.

A este objetivo se le da respuesta en la realización de la segunda pregunta, mediante la cual se van a obtener opiniones de la necesidad que existe actualmente de recuperar la información en D´TIC. La respuesta emitida fue buena obteniendo que existe gran necesidad de recuperación de información. En

la gráfica que se muestra a continuación se encuentra el criterio emitido por los expertos:

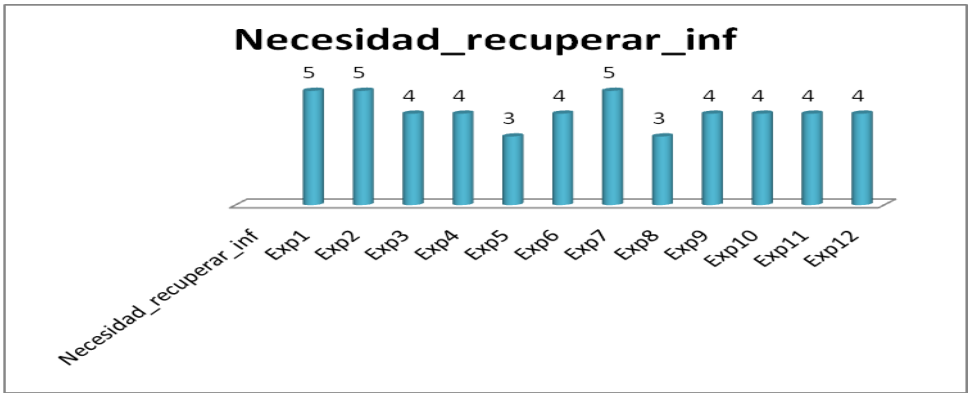

Figura 3. 4 Evaluación del tercer objetivo a los expertos.

Aporte que le brinda a D´TIC.

A este objetivo se le da respuesta en la puesta en práctica de la cuarta pregunta del cuestionario, en la cual se midieron los siguientes puntos: muy útil (5), bastante útil (4), útil (3), poco útil (2) e inútil (1). En la grafica que se presenta a continuación se muestra la opinión de los especialistas:

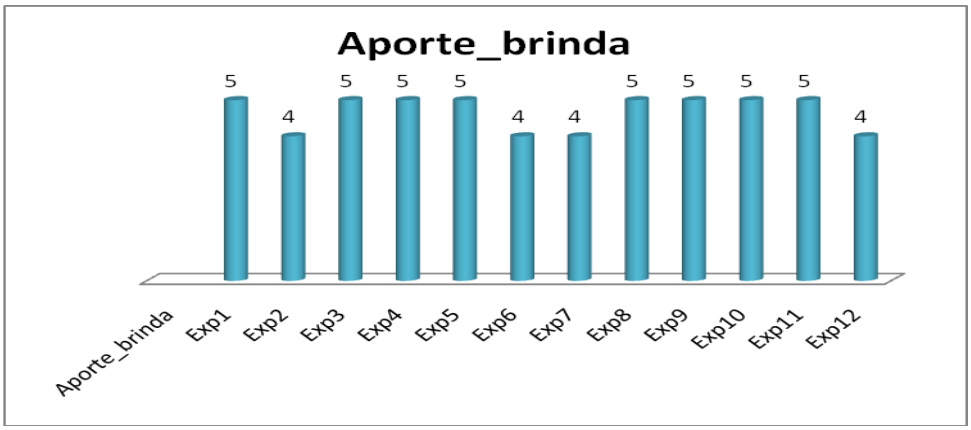

Figura 3. 5 Evaluación del cuarto objetivo a los expertos.

3.2.3. Fase Final.

Es importante realizar un porcentaje de las respuestas establecidas por los especialistas en el tema, ya que le mismo indica la positividad o la negatividad de las respuestas de cada uno de los objetivos

sometidos a evaluación. En todos los objetivos planteados los resultados fueron positivos ya que todos vieron que con el establecimiento de la propuesta se mejora el funcionamiento en D´TIC. Todo lo antes planteado se encuentra reflejado en la tabla de los porcientos que a continuación se muestra, dejando bien claro cada uno de los objetivos medidos y sus respectivos valores.

Al responder el conjunto de preguntas los especialistas tienen un grado de concordancia, el cual se calcula mediante el coeficiente de concordancia de Kendall, utilizando para ello la herramienta de software "SPSS 13.0 for Windows", el que arrojó como resultado 0.756, el cual si da un valor cercano a uno es porque los especialistas tienen un alto nivel de concordancia a todas las preguntas planteadas.

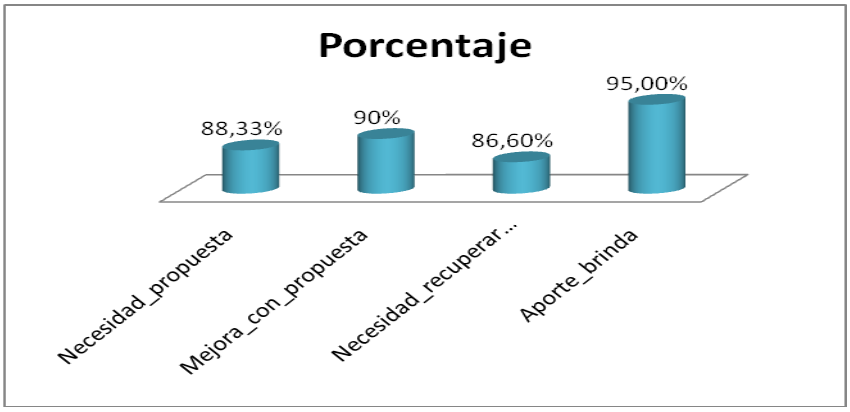

Figura 3.7 Tabla de Porcentaje.

Con esta validación utilizando el método Delphi mediante el criterio de los expertos se llegó a la siguiente conclusión:

- La propuesta va a mejorar la gestión de información a los usuarios de D´TIC.
- Con la ayuda de esta propuesta se hará más fácil la recuperación de información en D´TIC.

3.3. Conclusiones parciales del capítulo.

En este capítulo se establecen los métodos de evaluación que se pueden utilizar para la validación de la propuesta definida en el capítulo anterior además de especificar el empleado para la realización de la misma. Se realizaron los cálculos estadísticos correspondientes a los resultados obtenidos luego de haber aplicado una encuesta. De manera general se graficaron las respuestas y se plasmaron las recomendaciones obtenidas.

Conclusiones generales.

La extracción del conocimiento es una necesidad fundamental en todas las empresas y organismos tanto nacional como mundial, D´TIC no se ha quedado atrás en este sentido, la necesidad de la recuperación de la información se ha vuelto un proceso importante en este organismo. La propuesta de solución cuenta con una metodología, una tarea, un método, un algoritmo y una herramienta que posibilitan una futura implementación de la técnica de Minería de Datos para solucionar los problemas existentes.

Por lo tanto se puede concluir:

- El estudio de la técnica de Minería de Datos, de las tendencias en el ámbito nacional y extranjero de la misma permitió elegir la metodología, la tarea, el método, el algoritmo y la herramienta para guiar el proceso para un futuro desarrollo.

- Se realizó un estudio de de los diferentes métodos de validación existente, permitiendo elegir uno para realizar la validación de la propuesta planteada.

Recomendaciones.

Para solucionar los problemas existente, al realizar la implementación de la técnica de Minería de Datos en el proyecto D´TIC es necesario utilizar la propuesta contenida en esta investigación. Esto va a ser más efectivo si se lleva a cabo desde el inicio de la implementación, es decir de la manera que ha sido realizada la propuesta en el presente trabajo, permitiendo así obtener una mejor gestión de la información para los usuarios del MIC.

Por lo antes explicado se recomienda:

- La continuación del presente trabajo investigativo.
- Realizar un estudio en a profundidad del algoritmo propuesto.
- Trabajar en el desarrollo de herramientas que viabilicen una buena recuperación de la información y gestión del conocimiento que sea capaz de solucionar todas las necesidades que presentan los usuarios del MIC.
- Tener referencias de nuevas herramientas libres que han surgidos con el fin de mejorar los temas relacionados con la Minería de Datos y la recuperación de información.

Referencias Blibliográficas.

1. **Valencia, Carlos Mario Cordona.** *Utilidad Práctica Derivada De Aplicar Minería de Datos en Algunas Empresas de Medelín.* 2005.

2. **Navarro, Miguel Ángel Montero.** *Extracción de conocimiento en bases de datos astronómicas, Memoria del periodo de investigación.* Sevilla : s.n., Junio de 2009.

3. **Alfonso, Édgar.** Vivir la UNAB. [En línea] 01 de Marzo de 2004. [Citado el: 10 de enero de 2010.] http://www2.unab.edu.co/vivir/n120/a10.htm/.

4. Dataminnig (Minería de datos). [En línea] 2007. [Citado el: 11 de enero de 2010.] http://www.sinnexus.com/business_intelligence/datamining.aspx/.

5. *DEADALUS - MINERÍA DE DATOS.* [En línea] [Citado el: 14 de enero de 2010.] http://www.daedalus.es/mineria-de-datos/.

6. Minería de Datos. [En línea] 28 de enero de 2009. [Citado el: 20 de enero de 2010.] http://mineriadedatos.blogspot.es/.

7. **Molina, Luis Carlos.** *"Data Mining: Torturando a los datos hasta que confiesen".* s.l. : FUOC, 2002.

8. **Vallejos, Sofia J.** *Minería de Datos.* Corrientes - Argentina : s.n., 2006.

9. **Nadinic, Mladen W.** *DATA MINING Y DATA WAREHOUSING.* 2008.

10. *MINERÍA DE DATOS.*

11. **María Isabel Ángeles Larrieta, Angélica María.** *Minería de Datos: Conceptos, características, estructura y aplicaciones.*

12. *Documento Básico Daedalus. Minería Web.* 2002.

13. **Félix, Luis Carlos Molina.** Data mining: torturando a los datos hasta que confiesen. [En línea] noviembre de 2002. [Citado el: 21 de enero de 2010.] http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.

14. DEADALUS - PROCESO DE MINERÍA DE DATOS. [En línea] [Citado el: 22 de enero de 2010.] http://www.daedalus.es/mineria-de-datos/proceso-de-mineria-de-datos/.

15. **Oporto, Samuel Díaz.** *El Proceso de la Minería de Datos.*

16. **María N. Moreno García, Luis A. Miguel Quintales, Francisco J. García Peñalvo y M. José Polo Martín.** *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS.* Salamanca : s.n.

17. Universidad Nacional de Colombia. [En línea] [Citado el: 28 de enero de 2010.] http://www.virtual.unal.edu.co/cursos/sedes/manizales/4060029/lecciones/cap8-5.html/.

18. **Tanco., Fernando.** *INTRODUCCION A LAS REDESNEURONALES ARTIFICIALES.*

19. **Bot, Romina Laura.** *Data Mining utilizando Redes Neuronales.* Buenos Aires : s.n., 2005.

20. **Enrique Bonsón Ponte, Tomás Escobar Rodríguez, María del Pilar Martín Zamora.** SISTEMAS DE INDUCCIÓN DE ÁRBOLES DE DECISIÓN: UTILIDAD EN EL ANÁLISIS DE CRISIS BANCARIAS. [En línea] [Citado el: 30 de enero de 2010.] http://ciberconta.unizar.es/Biblioteca/0007/arboles.html/.

21. **Zauschkevich, Juan Andrés Conrads.** Minería de Datos: Algoritmos Genéticos y su Aplicación en la Fermentación Vínica. [En línea] 2004. [Citado el: 10 de febrero de 2010.] http://dcc.puc.cl/investigacion/tesis/mineriadedatosal/.

22. Técnicas más usadas en la Minería de Datos. [En línea] 03 de Octubre de 2007. [Citado el: 12 de febrero de 2010.] http://gamoreno.wordpress.com/2007/10/03/tecnicas-mas-usadas-en-la-mineria-de-datos/.

23. **Ascasibar, Martínez de Pison.** *Optimización mediante técnicas de minería de datos del ciclo de recorrido de una línea de galvanizado.* 2003.

24. Historia de SAS. [En línea] 1976. [Citado el: 15 de febrero de 2010.] http://www.sas.com/offices/latinamerica/andean/history.html/.

25. SAS®ayuda a la SUNAT a detectar el fraude aduanero y obtener mayores recaudos impositivos . [En línea] [Citado el: 20 de febrero de 2010.] http://www.sas.com/offices/latinamerica/andean/news/sunat.html/.

26. **Rocha, Andres.** SPSS Clementine. [En línea] 21 de mayo de 2010. [Citado el: 23 de mayo de 2010.] http://psicoandres99.blogspot.com/2010/02/spss-clementine-111-full-espanol-crack.html/.

27. **Arencibia, José Alberto gallardo.** *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.*

28. **Perversi, Ignacio.** *APLICACIÓN DE MINERÍA DE DATOS PARA LA EXPLORACIÓN Y DETECCIÓN DE PATRONES DELICTIVOS EN ARGENTINA.* 2007.

29. **Marcos Mora, Mari Carmen.** Browsing y clustering: dos técnicas en auge para la recuperación de información. [En línea] 2004. [Citado el: 14 de abril de 2010.] http://www.documentaciondigital.org/.

30. **Diego Guevara, Marilin Jaramillo, Katty Landacay.** Mineria De Datos Secuenciales. [En línea] [Citado el: 17 de abril de 2010.] http://www.slideshare.net/marilynsilvana/mineria-de-datos-secuenciales/.

31. Efficient K-Means Clustering. [En línea] 27 de marzo de 2008. [Citado el: 20 de marzo de 2010.] http://www.mathworks.com/matlabcentral/fileexchange/19344-efficient-k-means-clustering-using-jit/.

32. **Monreal, Fausto Andrés.** *WEKA.*

33. **José Hernández Orallo, Cèsar Ferri Ramírez.** *Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software.* València : s.n., 2006.

34. **Morate, Diego García.** *MANUAL DE WEKA.*

Bibliografía

1. **Dr. David L. Olson, Dr. Dursun Delen**. Advanced Data Mining Techniques. Estados Unidos : s.n.

2. **Nikhil R. Pal, Lakhmi Jain**. Advanced Techniques in Knowledge Discovery and Data Mining. 2004.

3. **Paolo Giudici, Silvia Figini**. Applied Data Mining for Business and Industry. s.l. : John Wiley & Sons, 2009 .

4. **János Abonyi, Balázs Feil**. Cluster Analysis for Data Mining and System Identification. Berlin : s.n., 2007.

5. **Julio Ponce, Adem Karahoca**. Data Mining and Knowledge Discovery in Real Life Applications. Croatia : I-Tech, 2009.

6. **McCue, Dr. Colleen**. Data Mining and Predictive Analysis. AMSTERDAM : ELSEVIER, 2007.

7. **Wang, John**. Data Mining: Opportunities and Challenges. Hershey : IGP, 2003.

8. **Mattison, Rob**. Data Warehousing and Data Mining for Telecommunications. Boston, London : s.n., 1997.

9. **Chakrabarti, Soumen**. M I N I N G T H E WEB D I S C O V E R I N G KNOWLEDGE FROM HYPERTEXT D A T A. Bombay : s.n., 2003.

10. **Bramer, Max**. Principles of Data Mining. s.l. : Springer.

11. **Ying Zhao, George Karypis**. Clustering in Life Sciences. Minnesota : s.n.

12. **Eui-Hong (Sam) Han, George Karypis**. Centroid-Based Document Classification: Analysis & Experimental Results. 2000.

13. **Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR).** CRISP-DM 1.0. Estados Unidos : SPSS, 2000.

Glosario de Términos.

- KDD - Knowledge Discovery in Data Base.
- MD - Minería de Datos.
- VT - Vigilancia Tecnológica.
- D´TIC - Centro virtual de recursos de las tecnologías informáticas en Cuba.
- MIC - Ministerio de la informática y las comunicaciones.
- TIC - Tecnologías de la Información y las Comunicaciones
- CRIPS – DM-CRoss-Industry Standard Process for Data Mining.
- SEMMA - Sample, Explore, Modify, Model, Assess.
- DCD - Depresión cortical diseminada.
- UCI- Universidad de las Ciencias Informáticas.
- MDL- Longitud de descripción mínima.
- ARFF- Atribute-Relation File Format.