Universidad de las Ciencias Informáticas.

Facultad 6.

Título: "Proceso de análisis y gestión del conocimiento a partir de los datos obtenidos en la conducción de los de Ensayos Clínicos del Centro de Inmunología Molecular, aplicando técnicas de Minería de Datos"

Trabajo de Diploma para optar por el título de

Ingeniero Informático

Autores:

Danay Perera Baró

Enrique Ramírez Alonso

Tutor:

M.Sc. Maypher Román Durán

Ciudad de la Habana, Junio 2009

"Año del 50 Aniversario del Triunfo de la Revolución"

"El éxito de los hombres no se mide por su éxito inmediato, sino por su éxito definitivo:- no se mide por el dinero que acumularon, sino por el resultado de sus obras."

José Martí

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 firmamos la presente a los _____días del mes de _____del año 2008.

Danay Perera Baró

_____

Firma del Autor.

Enrique Ramírez Alonso

_____

Firma del Autor

M.Sc. Maypher Román Durán

_____

Firma del Tutor.

DATOS DE CONTACTO

M.Sc. Maypher Román Durán

Graduado en el 2004 de Ingeniería Informática en la CUJAE.

Profesor Asistente.

Máster en Informática Aplicada.

4 años de trabajo en temas de Calidad del Software, Ingeniería de Software y Sistemas de Bases de Datos.

maypher@uci.cu

Agradecimientos:

Agradecer a nuestros padres porque son nuestra razón de ser, por su apoyo incondicional, en fin por todo su amor y por prepararnos como futuros profesionales.

A nuestros familiares por apoyarnos durante estos cinco años de la carrera.

Agradecer a las especialistas del CIM Carmen y Patricia por su contribución al desarrollo de esta investigación y por ayudarnos siempre que lo necesitamos.

A nuestro tutor Maypher por su ayuda incondicional y paciencia. Muchas gracias.

A las amistades de estos cinco años por los momentos vividos, se han convertido en parte de nuestra familia.

A todos los que están aquí hoy brindado su apoyo.

Muchas gracias.

Dedicatoria:

A mi mamá, por ser la luz de mis ojos, mi guía, mi razón de ser, por hacerme ver el lado positivo de las cosas por muy difícil que fuese el camino, gracias por animarme cada vez que pensaba que todo estaba perdido, por hacerme ver que con esfuerzo, mucha fe y dedicación, todos salimos adelante.

A mi papá por su preocupación y cariño.

A mis hermanas/os por todo su amor y apoyo en estos cinco años de la carrera. Sé que cuento con ellos siempre.

A Dayana por ser más que una amiga, una hermana. Por apoyarme siempre y por ser tan incondicional, a ella le debo estar aquí.

A Yoendris y Yuya por ser tan buenos conmigo, y estar ahí siempre que los necesité.

A mi tío Tomás, que aunque no esté físicamente se sentiría muy contento con este resultado.

A mis amigas/os compañeros de muchas alegrías y tristezas, que estuvieron en los días de festejo pero también en los de mucho trabajo.

A mis amigos de Santiago, los de la vocacional y los que conocí en la universidad. Los voy a querer siempre.

A todos los que pusieron su granito de arena para hacer este sueño realidad.

Danay.

A mis padres por ser mi fuente de inspiración.

A mi hermana Ariadnis por su amor incondicional.

A mi tía Marta Rosa por ser la mejor tía del mundo.

Enrique

Resumen

El volumen de datos que se acumula continuamente, y la necesidad de encontrar métodos que permitan descubrir conocimiento dentro de estos datos, han convertido a la Minería de Datos en una disciplina de importancia estratégica para la planeación y la toma de decisiones. El desarrollo de la investigación se rige por la metodología más utilizada actualmente en los procesos de KDD: CRISP-DM 1.0 y se apoya en la herramienta de libre distribución WEKA 3.6.0 que goza de gran prestigio por las prestaciones que ostenta.

En Cuba muchos científicos se dedican al estudio de fármacos para combatir el cáncer, el Centro de Inmunología Molecular (CIM) es uno de los centros dedicados a esta labor. En el CIM se han realizado varios Ensayos Clínicos para probar fármacos, tal es el caso de la primera vacuna contra el cáncer de pulmón, llamada *CIMAVAX EGF,* por lo que dicho centro dispone de un cúmulo de información de análisis realizados a estos pacientes.

En el presente trabajo se emplean un grupo de técnicas de Minería de Datos como la clasificación; con el objetivo de predecir el tiempo de supervivencia de un paciente con cáncer de pulmón, para posteriormente encontrar patrones ocultos y reglas que los caractericen; a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, color de piel, estadio clínico, clasificación histológica, peso). Se espera con estos resultados mejorar la conducción de los Ensayos Clínicos en el CIM.

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

Índice

Índice de Figuras

Índice de Tablas

Introducción

La informática en el mundo ha tenido un crecimiento vertiginoso. El auge de las comunicaciones con las nuevas tecnologías de la información no permite que ninguna esfera económica o social pueda pensar en el desarrollo si no es con la presencia de esta herramienta tan importante. Una sociedad que aplique la informatización en todas sus esferas y procesos será más eficaz y competitiva.

El almacenamiento digital de información ha sido una necesidad desde los inicios de la computación. Desde el surgimiento de las tarjetas perforadoras hasta los Sistemas Gestores de Bases de Datos el volumen de información que se genera ha aumentado considerablemente.

El desarrollo de nuevas técnicas de almacenamiento y recuperación de la información es una de las soluciones que en la actualidad resuelve de una manera u otra el problema del registro y análisis de la información para cualquier institución. Una forma muy valiosa de análisis de la información es la Minería de Datos (MD).

La MD es una de las etapas de lo que se ha venido llamando el proceso de extracción de conocimientos a partir de datos. Este proceso consta de varias fases e incorpora diferentes técnicas de Aprendizaje Automático, la Estadística, las Bases de Datos, los Sistemas de Toma de Decisiones, la Inteligencia Artificial y otras áreas de la informática y de la gestión de información. Entre las múltiples definiciones que identifican a la MD se encuentran:

"…el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en Bases de Datos, Data-Warehouses, o cualquier otro medio de almacenamiento de información." [Servente, 2002]

"…el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos" [Fayyad, 1996].

"…el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos" [Witten, 2000].

"…término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos" [Molina, 2006].

De manera general, puede afirmarse que la MD constituye un proceso para descubrir conocimiento a partir de los datos, apoyada en técnicas y herramientas, a fin de que su uso ayude a tomar decisiones más seguras, que reporten algún tipo de beneficio a las organizaciones.

Las herramientas de MD son utilizadas para resolver situaciones donde el volumen de datos es muy grande o complejo por la cantidad de variables que se manipulan, o donde los especialistas no están disponibles para el análisis de los datos y la extracción de conocimiento. Se pueden clasificar en dos grandes grupos: técnicas de verificación, en las que el sistema se limita a comprobar hipótesis suministradas por el usuario, así como los métodos de descubrimiento, en los que se ha de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción.

En el mundo la MD se ha convertido en un factor indispensable en las empresas e instituciones para la toma de decisiones y la creación de modelos que permiten un aumento de sus ofertas, una reducción de sus costos y una mejor planificación de su estrategia de trabajo. Anualmente se desarrolla en la Ciudad de México el Congreso Internacional sobre Minería de Datos y Sistemas de Información donde se reúnen investigadores e industriales para compartir e intercambiar ideas y prever tendencias en los sistemas de información y aquellos avances que tendrán impacto en nuestra sociedad futura.

En Cuba el uso de las herramientas y técnicas de la MD todavía es joven. El Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) desde su creación en el año 2004 se dedica a las investigaciones teóricas y aplicadas en este campo y sus aplicaciones están dirigidas a áreas tales como la biometría, la recuperación de información, el procesamiento de información de texto, entre otras. Otros centros e instituciones cubanos se dedican a la investigación y aplicación de técnicas de MD, tal es el caso del Centro de Estudios de Reconocimiento de Patrones y Minería de Datos (CERPAMID), el Instituto de Cibernética, Matemática y Física (ICIMAF) y el Instituto Superior Politécnico José Antonio Echeverría (CUJAE), centro en el que se han empleado con éxito herramientas y metodologías de MD para apoyar la gestión docente del mismo.

La Universidad de las Ciencias Informáticas (UCI) nació como un proyecto de la Revolución Cubana con el objetivo de informatizar el país y desarrollar la industria del Software para contribuir al desarrollo económico del mismo. En la misma se han realizado algunas investigaciones de Minería de Datos, tal es el caso de dos estudiantes de la CUJAE que aplicaron técnicas de MD a la Base de Datos de Akademos, obteniendo resultados satisfactorios.

La Facultad 6, situada en esta universidad, cuenta con varios proyectos que almacenan grandes volúmenes de datos, entre ellos se encuentra el proyecto “alasClínicas” el cual constituye un aporte al Centro de Inmunología Molecular (CIM) de una nueva versión del sistema OpenClínica. Dicho proyecto está siendo desarrollado por la facultad en colaboración con el CIM, con el objetivo de mejorar la calidad de la gestión de los datos referente a los Ensayos Clínicos que se realizan en dicho centro dado que actualmente el centro no cuenta con un sistema que le permita gestionar toda esta información y se recogen los datos en Bases de Datos Microsoft Access o con el programa estadístico SPSS, realizándose una entrada doble de los mismos.

Dado el gran volumen de datos acumulado, y la incapacidad de los especialistas del CIM de identificar patrones de comportamiento y extraer conocimiento oculto en los datos almacenados para apoyar sus decisiones surge la necesidad de aplicar MD a dicho proyecto, por lo que el **problema a resolver** es el ¿Cómo apoyar el proceso de gestión y toma de decisiones de los administrativos y especialistas del CIM a partir del estudio de los datos recopilados en los ensayos clínicos realizados en el centro, a través de la identificación de patrones de comportamiento útiles presentes en estos ensayos?

Este problema presenta como **objeto de estudio** la gestión del conocimiento a través de técnicas de MD teniendo como **campo de acción** las técnicas de MD para la gestión del conocimiento en proyectos de gestión biomédica.

Por lo que el **objetivo general** consiste en aplicar técnicas de MD sobre los datos recopilados en la conducción de los Ensayos Clínicos realizados en el CIM a los pacientes con cáncer de pulmón, con el fin de predecir el tiempo de vida de dichos pacientes.

De este objetivo general se derivan los siguientes **objetivos específicos:**
1. Integrar los datos del proyecto que serán empleados en la MD.
2. Transformar los datos almacenados que serán empleados en la MD.
3. Aplicar técnicas de MD a los conjuntos de datos seleccionados para el estudio.
4. Evaluar la calidad de los resultados obtenidos a partir de las técnicas de MD aplicadas.

Después de obtener toda la información relacionada con el tema del trabajo para guiar la investigación se plantea la siguiente **idea a defender**:

Si se aplican técnicas de MD a la información almacenada sobre los ensayos clínicos del CIM, entonces se mejorará el proceso de gestión del gran volumen de información que dichos ensayos generan, facilitando la toma de decisiones para los ejecutivos y especialistas del centro.

Tareas investigativas:

- ✓ Realización de estudios bibliográficos sobre las técnicas, herramientas y metodologías empleados en la MD.
- ✓ Realización de estudios a los datos de la conducción de los Ensayos Clínicos del CIM.
- ✓ Integración y recopilación de los datos del proyecto que serán empleados en el estudio.
- ✓ Selección, limpieza y transformación de los datos del proyecto que serán empleados en la MD.
- ✓ Selección y aplicación de técnicas de MD a los conjuntos de datos seleccionados para el estudio.
- ✓ Evaluación e interpretación de los patrones obtenidos a partir de las técnicas de MD aplicadas.
- ✓ Difusión, a los especialistas del centro, de los resultados obtenidos en el estudio.
- ✓ Validación de la propuesta.

El presente trabajo está estructurado de la siguiente manera:

Capítulo 1: Fundamentación teórica: se realiza un estudio detallado del estado del arte de las herramientas, técnicas y metodologías que se utilizan en la MD, con el propósito de definir la más adecuada para la solución del problema.

Capítulo 2: Solución propuesta: se describen los pasos y actividades propuestos por la metodología y herramienta seleccionadas para realizar el proceso de MD, aplicadas a los datos recopilados en la conducción de los EC realizados en el CIM a los pacientes con cáncer de pulmón.

Desarrollo.

Capítulo 1: Fundamentación Teórica.

Introducción

Se presenta una descripción detallada sobre diversos aspectos de la Extracción de Conocimientos en Bases de Datos, especificando varias características de la Minería de Datos, su aplicación en diferentes esferas, así como técnicas que se usan para realizar dicho proceso. Además ofrece una descripción de las principales herramientas y metodologías usadas en el proceso de Minería de Datos con el fin de seleccionar la herramienta y metodología más adecuada para dar solución al problema.

1.1 KDD. Extracción de Conocimientos en Bases de Datos.

KDD (Knowledge Discovery in Databases) es un proceso no trivial de identificación válida, reciente, potencialmente útil de patrones comprensibles ocultos en los datos [Fayyad_Shapiro, 1996].
Su objetivo es procesar automáticamente grandes cantidades de datos crudos, identificar los patrones más significativos y relevantes, y presentarlos como conocimiento apropiado para satisfacer las metas del usuario.
Las metas que persigue KDD son [Vallejos]:

- ✓ Procesar automáticamente grandes cantidades de datos crudos.
- ✓ Identificar los patrones más significativos y relevantes.
- ✓ Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

Técnicas de KDD [Bressán, 2003]:

Método de Clasificación:

- ✓ Es el más usado de todos los métodos de KDD.
- ✓ Agrupa los datos de acuerdo a similitudes o clases.
- ✓ Existen numerosas herramientas disponibles que son automatizadas.

Método Probabilístico.

- ✓ Utiliza modelos de representación gráfica.

- ✓ Se basa en las probabilidades e independencias de los datos.
- ✓ Puede usarse en los sistemas de diagnóstico, planeación y sistemas de control.

Método Estadístico.

- ✓ Usa la regla del descubrimiento y se basa en las relaciones de los datos.
- ✓ Es usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados.

Método Bayesian de KDD.

- ✓ Es un modelo gráfico que usa directamente los arcos para formar una gráfica acíclica.
- ✓ Se usa muy frecuentemente las redes de Bayesian cuando la incertidumbre se asocia con un resultado que puede expresarse en términos de una probabilidad.
- ✓ Este método es usado para los sistemas de diagnóstico.

Los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados. Incluyen además, la evaluación y posible interpretación de los mismos para convertirlos en conocimiento [Hernández, 2004], tal y como se muestra en la Figura 1.

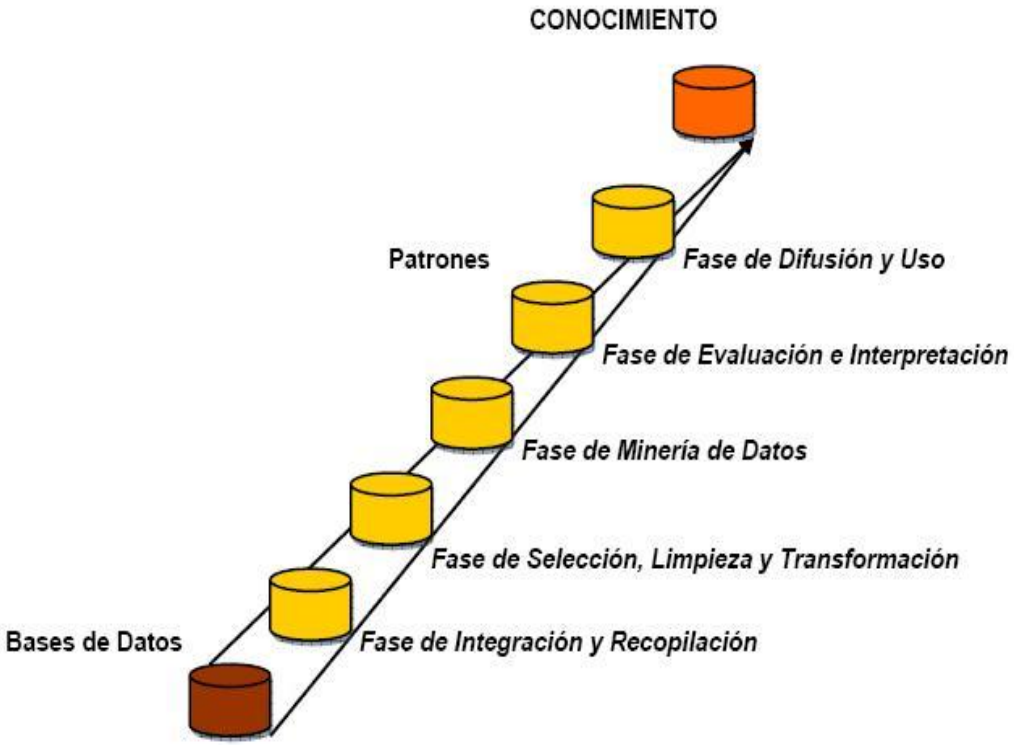


Figura 1: Proceso KDD [Hernández, 2004]

1. **Integración y recopilación.**

   Se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas; se transforman todos los datos a un formato común, y se detectan y resuelven las inconsistencias.

2. **Selección, limpieza y transformación.**

   Se eliminan o corrigen los datos incorrectos, y se decide la estrategia a seguir con los datos incompletos; además, se consideran únicamente aquellos atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería.

3. **Minería de Datos.**

   Se aplica la tarea, la técnica, la herramienta y la metodología seleccionada para la obtención de reglas y patrones.

4. **Evaluación e Interpretación.**

   Se evalúan los patrones y se analizan por expertos, y si es necesario, se vuelve a las fases anteriores para una nueva iteración.

5. **Difusión y Uso.**

   Se hace uso del nuevo conocimiento y se difunde para la comprensión de los interesados.

El desarrollo de las fases antes descritas es iterativo e interactivo con el usuario. Iterativo pues la salida de alguna de las fases puede hacer volver a pasos anteriores y porque son necesarias varias iteraciones para extraer conocimiento de alta calidad. Interactivo porque el usuario debe ayudar en la preparación de los datos y a la validación del conocimiento extraído.
Señalar además que las dos primeras fases se engloban bajo el nombre de preparación de datos. Además, aproximadamente el 60 por ciento del esfuerzo total para realizar este proceso, se emplea durante la etapa de preparación de los datos, como se muestra en la figura.

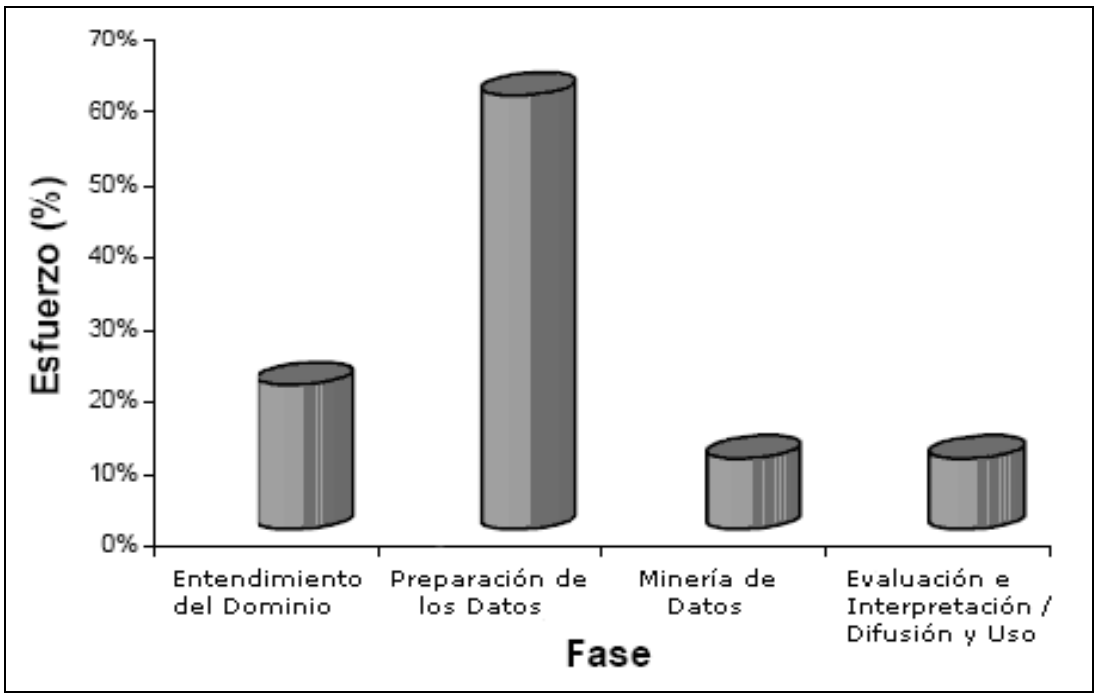


Figura 2: Esfuerzo requerido por fases en un proceso de KDD [Molina López]

Como se muestra en la figura 2, una de las fases del proceso KDD es la Minería de Datos, fase que se describe a continuación dado que es el objetivo de la investigación.

1.2 Minería de Datos

En la actual sociedad de la información, donde cada día a día se multiplica la cantidad de datos almacenados casi de forma exponencial, la minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización.

Desde hace varias décadas, científicos de varias ramas entre las que se destacan la Inteligencia Artificial, las Bases de Datos, la Estadística y la Investigación de Operaciones vienen desarrollando las bases tecnológicas que permiten hoy hacer realidad una nueva ciencia que recibe el nombre de Minería de Datos. La Minería de Datos, usando varias técnicas que incluyen los algoritmos inductivos, los algoritmos de agrupamiento, las técnicas de visualización inteligente de información, las redes neuronales, los métodos heurísticos y muchos otros son realidades que permiten utilizar las ventajas de la Minería de Datos en aplicaciones tan disímiles como la economía, el marketing, el deporte y la educación.

1.2.1 ¿Qué es la Minería de Datos?

La Minería de Datos es el proceso analítico diseñado para explorar grandes cantidades de datos con el objetivo de determinar patrones de comportamiento consistentes o relaciones entre las diferentes variables para aplicarlos a nuevos conjuntos de datos. Constituye el proceso de descubrir conocimientos interesantes y estructuras significativas a partir de grandes cantidades de datos almacenados en Bases de Datos, Data-Warehouses o en otro medio de almacenamiento.

La Minería de Datos puede ser dividida en [Facena, 2003]:

- ✓ Minería de Datos Predictiva (MDP): usa primordialmente técnicas estadísticas.
- ✓ Minería de Datos para el Descubrimiento de Conocimiento (MDDC): usa principalmente técnicas de inteligencia artificial.

1.2.2 Fases de la 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 [Vallejos, 2006]:

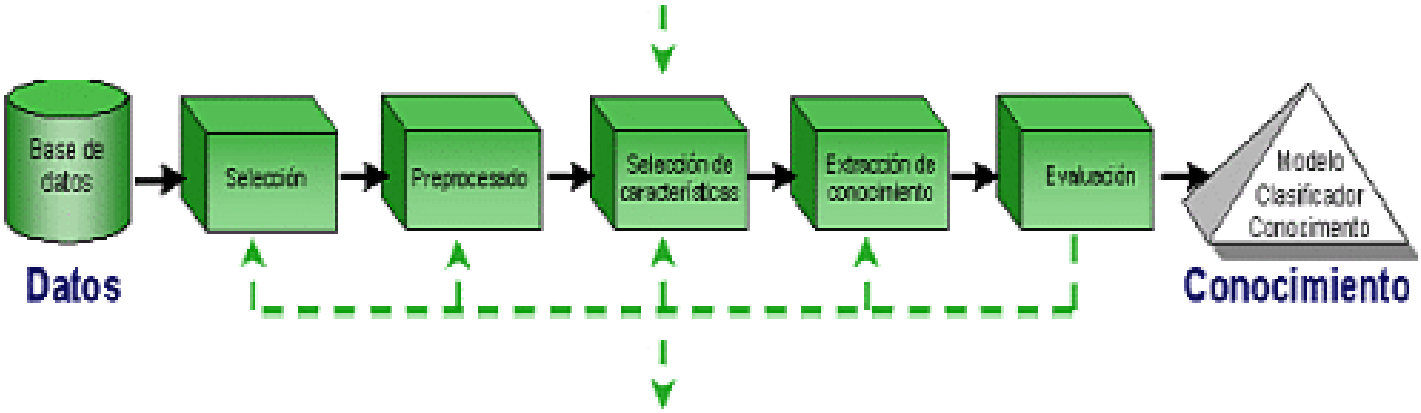


Figura 3: Proceso de Minería de Datos [Vallejos]

Selección y preprocesado de datos:

El formato de los datos contenidos en la fuente de datos (Bases 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 preprocesado 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).

Selección de variables:

Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad ingente 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 elección de los mejores atributos del problema.
- ✓ Y aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

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 preprocesado diferente de los datos.

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

1.2.3 Técnicas en Minería de Datos

Existen varias técnicas de Minería de Datos la mayoría basadas en reglas de decisión o reglas de inducción a través de árboles de decisión, entre ellas se encuentran [Bressán, 2003]

Técnicas de Visualización: son aptas para ubicar patrones en un conjunto de datos, puede usarse al comienzo de un proceso de Minería de Datos para determinar la calidad de los datos.

Redes neuronales artificiales: son modelos predecibles, no lineales que aprenden a través del entrenamiento.

Reglas de Asociación: establecen asociaciones en base a los perfiles de los clientes sobre los cuales se realiza la Minería de Datos.

Algoritmos Genéticos: son técnicas de optimización que usan procesos tales como combinaciones genéticas y mutaciones, etc.

Redes Bayesianas: buscan determinar relaciones causales que expliquen un fenómeno según los datos contenidos en una base de datos. Se han usado principalmente para realizar predicciones.

Los algoritmos de la Minería de Datos se clasifican en [Bressán, 2003]:

Supervisados o predictivos: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

No supervisados o del descubrimiento del conocimiento: con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

1.2.4 Modelos y tareas de Minería de Datos

La Minería de Datos tiene como objetivo analizar los datos para extraer conocimiento, que se representa a través de patrones o reglas obtenidos a partir de los datos.

Los **modelos** constituyen la forma de representar el conocimiento, y su construcción está determinada por la tarea de minería de datos escogida, el tipo de técnica empleada, y el algoritmo implementado para realizarlo. A partir de las características que engloban, se clasifican en dos tipos: predictivos y descriptivos. [Orallo-Ramírez, 2004]

Los *modelos predictivos* estiman o predicen valores futuros de la variable objetivo del análisis, partiendo de otros datos que se consideran influyentes en su comportamiento.
Los *modelos descriptivos* posibilitan explorar las propiedades de los datos que se examinan e identificar patrones que explican, resumen o caracterizan los mismos. [Cano, 2005]

Una de las **tareas** que produce modelos predictivos es la clasificación, otras como el agrupamiento y la asociación dan lugar a modelos descriptivos; además, cada tarea puede ser realizada utilizando distintas técnicas y algoritmos.

Un sistema de Minería de Datos actual realiza una o más de las siguientes tareas [Orallo-Ramírez, 2004]:

- ✓ **Descripción de clases**: provee una clasificación concisa y resumida de un conjunto de datos y los distingue unos de otros. La clasificación de los datos se conoce como caracterización, y la distinción entre ellos como comparación o discriminación.
- ✓ **Asociación**: es el descubrimiento de relaciones de asociación o correlación en un conjunto de datos. Las asociaciones se expresan como condiciones atributo-valor y deben estar presentes varias veces en los datos.
- ✓ **Clasificación**: analiza un conjunto de datos de entrenamiento cuya clasificación de clase se conoce y construye un modelo de objetos para cada clase. Dicho modelo puede representarse con árboles de decisión o con reglas de clasificación, que muestran las características de los datos. El modelo puede ser utilizado para la mayor comprensión de los datos existentes y para la clasificación de los datos futuros.
- ✓ **Predicción**: esta función de la minería predice los valores posibles de datos faltantes o la distribución de valores de ciertos atributos en un conjunto de objetos.
- ✓ **Clustering**: identifica clúster en los datos, donde un clúster es una colección de datos “similares”. La similitud puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos. La Minería de Datos trata de encontrar clúster de buena calidad que sean escalables a grandes bases de datos y a Data-Warehouses multidimensionales.
- ✓ **Análisis de series a través del tiempo**: analiza un gran conjunto de datos obtenidos con el correr del tiempo para encontrar en él regularidades y características interesantes, incluyendo la búsqueda de patrones secuenciales, periódicos, modas y desviaciones.

En dependencia del tipo de búsqueda empleado para obtener conocimiento, las tareas mencionadas se pueden clasificar en directas o indirectas. La clasificación es una tarea directa, pues se conoce claramente lo que se busca. El agrupamiento y la asociación son indirectas, y se emplean para descubrir patrones que describan los datos sin un objetivo concreto definido. [Orallo-Ramírez, 2004]

1.2.5 Algoritmos supervisados de clasificación para la Minería de Datos

Para extraer y recuperar la información de manera supervisada se pueden utilizar los siguientes **algoritmos**:

- ✓ Árboles de decisión:

Son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos. [Facena, 2003]

Los árboles de decisión son una manera de representar una serie de reglas que culminan en una clase o valor. Los modelos de árboles de decisión son comúnmente usados en la minería de datos para examinar los datos e inducir las reglas para realizar predicciones. [Inteligencia_Negocios]

Los árboles de decisión manejan datos no numéricos muy bien y pueden clasificarse en dependencia del tipo de variables que predicen; si es usado para predecir variables nominales, recibe el nombre de árbol de clasificación, mientras que si su uso está determinado en la predicción de variables numéricas, se denomina árbol de regresión o predicción. [Inteligencia_Negocios]

- ✓ Modelos Lineales de redes neuronales:

Buscan determinar relaciones causales que expliquen un fenómeno según los datos contenidos en una base de datos. Se han usado principalmente para realizar predicciones. [Facena, 2003]

Las redes neuronales son de un interés particular para la minería de datos ya que ofrecen un significativo modelo para problemas grandes y complejos, donde puede haber cientos de variables predictivas que interactúan entre sí. Las redes neuronales pueden ser usadas en problemas de clasificación cuando la variable de salida es clasificada como categórica, o pueden usarse para regresiones cuando la variable de salida es continua. [Inteligencia_Negocios]

- ✓ Vecinos próximos (IB1-IBK):

Consiste en métodos de aprendizaje basados en ejemplos, los de entrenamiento se almacenan tal cual se usa una función de distancia para determinar que patrón del conjunto de entrenamiento está más cerca del patrón a clasificar. [Porta, 2005]

1.2.6 Alcance de la Minería de Datos

Provee las siguientes capacidades [Pinto, 2007]:

- ✓ **Predicción automatizada de tendencias y comportamientos**. La Minería de Datos automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- ✓ **Descubrimiento automatizado de modelos previamente desconocidos**. Las herramientas de Minería de Datos barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores en la carga de datos.

1.2.7 Aplicaciones de la Minería de Datos

Uno de los campos en que la Minería de Datos se está viendo cada día más utilizada es en la medicina. En una encuesta realizada por el portal para el análisis de datos KDnuggets, en junio del 2007, sobre las diversas áreas en las que se emplea la Minería de Datos; aparecen las aplicaciones en la medicina en el doceavo lugar, con un 9.4% de empleo [Acosta]. En la figura que se muestra aparecen los resultados de dicha encuesta.

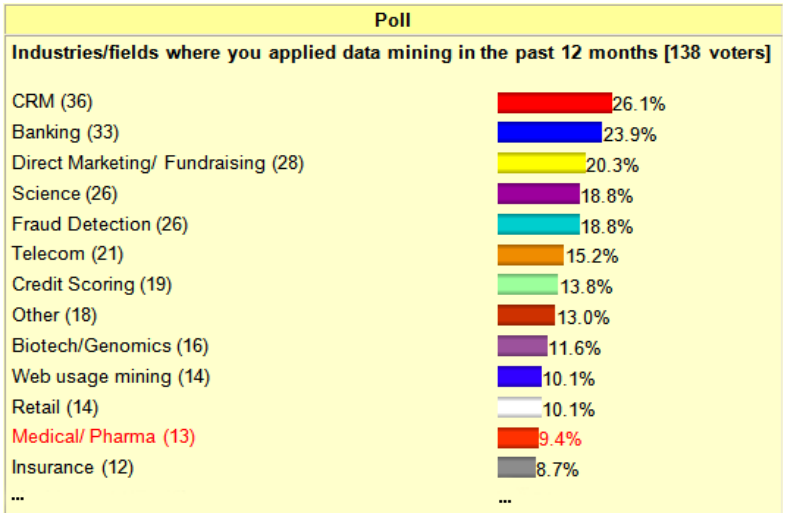


Figura 4: Resultados de la encuesta por el portal para el análisis de datos KDnuggets [Acosta].

Algunas de las aplicaciones en este campo son las siguientes:

Medicina: Se utiliza la Minería de Datos para realizar estudios epidemiológicos, para analizar el rendimiento de campañas de información y prevención, para predecir el comportamiento de diferentes tipos de tumores. Además está contribuyendo a la aparición de una nueva forma de diagnóstico y tratamiento de enfermedades, usando grandes bases de datos para asociar síntomas a enfermedades y recomendar los tratamientos. [Orallo]

Diagnóstico de Accidentes Cerebro-vasculares Agudos

Los accidentes cerebro vasculares agudos (ACVAs) amenazan la vida de miles de personas cada año. Las causas de esta dolencia pueden ser muy diversas y para su tratamiento es importante y necesario disponer de una estimación rápida y precisa del origen del problema. Utilizando la Minería de Datos se ha desarrollado un sistema de soporte a la decisión para el diagnóstico de las causas de estos accidentes. [Daedalus]

La Minería de Datos es muy utilizada para el desarrollo de sistemas de diagnóstico médico; una de las técnicas más utilizadas para representar el conocimiento son los árboles de decisión. [Daedalus]

Economía: Las aplicaciones más importantes son el estudio de mercados financieros, concesiones de préstamo, estudio de productos, detección de fraudes, entre otros. [Facena, 2003]

Distribución de energía: Hoy en día muchas compañías gasistas, eléctricas y petroleras utilizan la Minería de Datos para realizar previsiones de consumo que ayuden a mejorar la gestión de los productos. [Facena, 2003]

Compañías de telefonía: Para realizar previsiones de ocupación de líneas, de demanda de servicios, de ancho de banda utilizado, entre otros. [Facena, 2003]

Detección de fallos: Algunas empresas están incorporando técnicas de Minería de Datos para gestionar la cadena de producción. El objetivo es prever y detectar fallos antes de que interfieran en el proceso de fabricación. [Facena, 2003]

Aspectos climatológicos: predicción de tormentas, etc. [Facena, 2003]

1.3 Herramientas de Minería de Datos

Las herramientas de minería de datos son utilizadas para resolver problemas del mundo real en la ingeniería, la ciencia y los negocios. Son utilizadas para resolver situaciones donde el volumen de datos es muy grande por la cantidad de variables que se manipulan, o la extracción de conocimiento se hace compleja.

A continuación se describen las herramientas de Minería de Datos más utilizadas.

1.3.1 Weka

Es una colección de aprendizaje automático de algoritmos para tareas de minería de datos, desarrollada por la Universidad de Waikato, en Nueva Zelanda. Es un software con código abierto, desarrollado en Java. Es una de las primeras aplicaciones open source.

Los algoritmos pueden ser llamados desde un código java propio. Weka contiene herramientas para el pre-procesamiento de los datos, la clasificación, regresión, reglas de asociación y visualización. Presenta una interfaz fácil de utilizar en modo Explorer. [Calderón, 2005] *Ver Anexo 1*

La versión 3.6.0 incluye las siguientes características:

- ✓ Diversas fuentes de datos (ASCII, JDBC).
- ✓ Interfaz visual basado en procesos/flujos de datos (rutas).
- ✓ Distintas herramientas de minería de datos: reglas de asociación (a priori, Tertius), agrupación/segmentación/conglomerado (Cobweb, EM y k-medias), clasificación (redes neuronales, reglas y árboles de decisión, aprendizaje Bayesiona) y regresión (Regresión lineal, SVM).
- ✓ Manipulación de datos (pick & mix, muestreo, combinación y separación).

- ✓ Combinación de modelos (Bagging, Boosting ...)
- ✓ Visualización anterior (datos en múltiples gráficas) y posterior (árboles, curvas ROC, curvas de coste).
- ✓ Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas.

Algunos de los entornos de trabajo que posee Weka son [Hernández-Ferri, 2006]:

- ✓ Explorer: Entorno visual que ofrece una interfaz gráfica para el uso de los paquetes. *Ver Anexo 2*
- ✓ Experimenter: Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala.
- ✓ KnowledgeFlow: Permite generar proyectos de minería de datos mediante la generación de flujos de información.
- ✓ Simple CLI: Entorno consola para invocar directamente con java a los paquetes de Weka.

1.3.2 Yale

YALE es una herramienta creada en la universidad de Dortmund bastante flexible para el descubrimiento del conocimiento y la minería de datos. Puesto que YALE está escrito enteramente en Java, funciona en las plataformas o sistemas operativos más conocidos. Es un software de código abierto GNU y con licencia GPL. [Vilches, 2007]

Desde la perspectiva de la visualización YALE ofrece representaciones de datos en dispersión en 2D y 3D; representaciones de datos en formato SOM (Self Organizing Map); coordenadas paralelas y grandes posibilidades de transformar las visualizaciones de los datos. [Vilches, 2007]

Recientemente fue lanzada la última versión, la cual incluye características como las de implicar nuevos formatos de entrada de datos con operadores para Microsoft Excel y SPSS. [Vilches, 2007]

1.3.2.1 RapidMiner

Es la última versión de Yale, la cual incluye características como las de implicar nuevos formatos de entrada de datos con operadores para Microsoft Excel y SPSS. Cubre un amplio rango de minería de datos, además de ser una herramienta flexible para aprender y explorar la minería de datos. La interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área. [Vilches, 2007]

1.3.3 Clementine

Es una de las herramientas más populares, creada por SPSS, proveedor mundial de software de análisis profético y soluciones. Permite que los expertos en procesos de negocios, análisis de datos y modelado colaboren en la exploración de los datos y la construcción de modelos. Es la primera herramienta de minería de datos que brinda una interface visual al usuario, con una descripción global del proceso. [Martin]

La arquitectura abierta y escalable de Clementine permite llevar a cabo diferentes procedimientos dentro de la base de datos, incluyendo el acceso a algoritmos incrustados. Esto puede ayudar a maximizar su base de datos para un mejor rendimiento y mayor velocidad.

La sencilla interfaz visual del flujo de trabajo de Clementine no requiere conocimientos de programación, lo que reduce la curva de aprendizaje y hace que el poder de la analítica sea accesible tanto para expertos como para novatos. [Martin]

1.3.4 Enterprise Miner

Herramienta creada por SAS *Institute*, es un software de análisis, con una arquitectura cliente/servidor, basado en cliente Java, puede ser desarrollado tanto en la plataforma de Windows como en Linux/Unix. Tiene una interfaz de usuario muy fácil de usar y muy parecida a otras herramientas como Clementine y Weka. Soporta el proceso de minería de datos para crear modelos descriptivos y predictivos, sobre la base del análisis de las grandes cantidades de datos que tienen las empresas. [Martin]

Incluye:

- ✓ Fuentes de datos (ASCII, Oracle, Informix, Sybase e Ingres).
- ✓ Interfaz visual.
- ✓ Distintas herramientas de minería de datos: redes neuronales y reglas.
- ✓ Manipulación de datos (pick & mix, combinación y separación). [Martin]

1.3.5 Fundamentación de la herramienta seleccionada

Después de un estudio de las herramientas, Yale y Weka son las herramientas más usadas en el mundo en la actualidad, por lo que la herramienta escogida para el desarrollo del proceso de Minería de Datos fue Weka, dado que la misma es una de las más utilizadas actualmente y su distribución es gratuita. Además de que está implementada en código abierto, y sus algoritmos pueden ser llamados desde un código de java propio. WEKA permite realizar manipulaciones sobre los datos aplicando

filtros para la normalización y selección de atributos, conjuntamente se pueden importar datos en varios formatos. Las ventajas encontradas en esta herramienta son:

- ✓ De libre distribución.
- ✓ Multiplataforma.
- ✓ Tiene muchos algoritmos de regresión/clasificación.
- ✓ Incluye meta-algoritmos de aprendizaje.
- ✓ Tiene preprocesado de datos.
- ✓ Incorpora herramientas para la visualización de los datos y resultados.
- ✓ Se distribuye también su código fuente JAVA.
- ✓ Se pueden añadir nuevas clases de clasificadores y filtros.
- ✓ Tiene versiones de consola y con interfaz gráfico.

1.4 Metodologías de la Minería de Datos

Para la realización de los proyectos de Minería es necesario contar con una metodología que guíe el proceso. De esta manera diversas empresas han especificado y propuesto procesos de modelado con el objetivo de guiar al desarrollador a través de una serie de pasos dirigidos a obtener buenos resultados.

El instituto de Sistemas de Análisis eStadísticos (SAS) fue el desarrollador de la metodología SEMMA (Sample, Explore, Modify, Model, Assess) para la realización de proyectos de Minería. Por otra parte, en 1999 varias empresas europeas como la NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para desarrollar la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Estas metodologías son las más utilizadas en la actualidad para realizar proyectos de Minería de Datos. [Chapman, 2000]

La utilización de una metodología estructurada para la realización de proyectos de Minería de Datos presenta las siguientes ventajas:

- ✓ Posibilita la realización de nuevos proyectos de Minería con características similares.
- ✓ Facilita la planificación y dirección del proyecto.
- ✓ Permite realizar un mejor seguimiento del proyecto.
- ✓ Permite realizar el proyecto de forma organizada paso a paso.

- ✓ Facilita la comunicación y distribución de tareas entre los distintos miembros del equipo de desarrollo del proyecto.

A continuación se describen algunas de las metodologías usadas en la Minería de Datos.

1.4.1 CRISP-DM 1.0

CRISP-DM 1.0 [Chapman, 2000] es producto de la experiencia de varias empresas que se dedican a la Minería de Datos (SPSS, Daimler-Chrysler y NCR; entre otras).
Actualmente se encuentra en la versión 1.0. La metodología expone un modelo de referencia y una guía para el usuario. El modelo de referencia presenta una vista general de las fases y tareas con sus salidas. La guía del usuario provee orientaciones y consejos más detallados para el desarrollo de cada fase y tarea. [Chapman, 2000]

El modelo da referencia enuncia las siguientes seis fases generales [Chapman, 2000]:

Análisis del problema: Fase inicial enfocada a entender los objetivos y requerimientos desde una perspectiva de negocio; para luego definirlos en términos de un problema de Minería de Datos y diseñar un plan para satisfacerlos.

Comprensión de los datos: Se hace una recolección y exploración inicial de los datos para familiarizarse con ellos e identificar problemas de calidad. Además, se trata de descubrir o estimar las relaciones más evidentes para formular las primeras hipótesis sobre información oculta en ellos.

Preparación de los datos: Esta fase cubre todas las actividades necesarias para construir la colección de datos que finalmente será minada a partir del grupo inicial. Incluye la colección, exploración, limpieza, transformación y construcción de datos.

Modelado: Durante esta fase se aplican varias técnicas de modelado. Comúnmente existen varias técnicas para resolver un problema de Minería de Datos del mismo tipo. Incluye la evaluación desde el punto de vista de precisión de los modelos.

Evaluación: Al llegar a esta fase se tendrán los modelos de mayor calidad desde la perspectiva de la precisión. Se impone una evaluación de los modelos y de los pasos que se siguieron para su construcción, a fin de determinar si responden apropiadamente a los objetivos de negocio que se determinaron en la primera fase. Es de vital importancia analizar si alguna regla del negocio no fue tomada en cuenta con el suficiente peso.

Despliegue: En dependencia de los requerimientos y objetivos, la fase de despliegue puede ser tan simple como generar un reporte, o tan compleja como emprender un proceso de KDD de mayor envergadura. En ocasiones, son los clientes y no los desarrolladores quienes implementan esta fase; deben comprender cómo desarrollarla. Se hace imprescindible documentar y presentar los resultados de manera que todos los puedan entender.

Cada una de estas seis macro-fases es descompuesta en un conjunto de tareas generales; donde se especifican además, los resultados esenciales que se deben obtener al concluir cada una, y se describe cómo realizarlas.

En la figura se muestran las relaciones existentes entre estas fases:

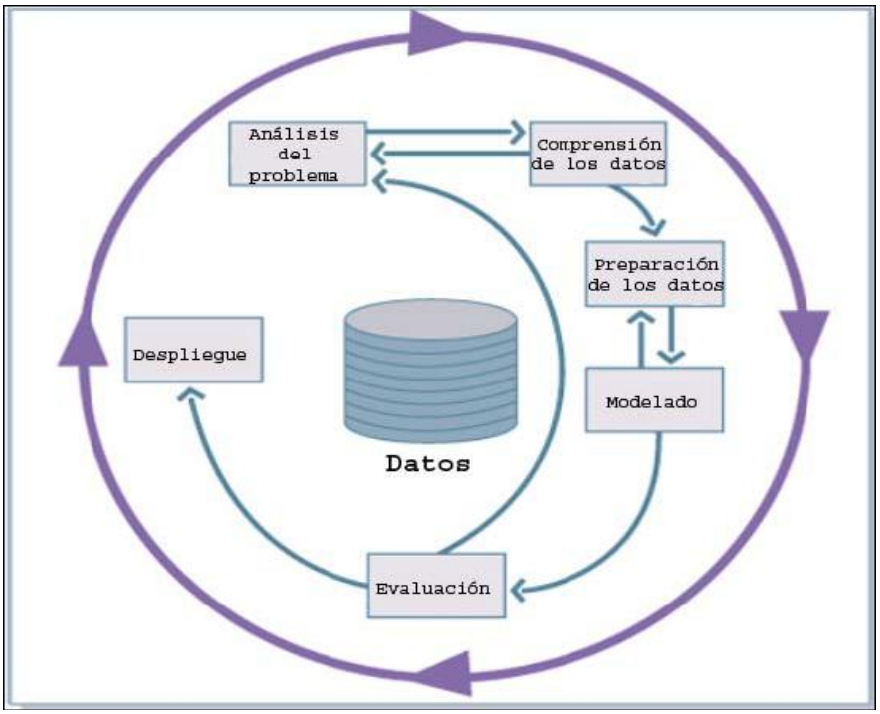


Figura 5: Fases del modelo de referencia CRISP-DM 1.0 y sus principales relaciones [Chapman, 2000].

El proceso de modelo de esta metodología es cíclico.

1.4.2 Semma

La metodología Semma consta de 5 fases, las cuales se describen a continuación [Gondar]:

Muestreo: Extracción de la población sobre la que se va a realizar el análisis con el objeto de seleccionar una muestra representativa y asociando a esta un nivel de confianza.

Exploración: Estudio de las bases de datos con el fin de simplificar al máximo el problema y hacer más eficiente el modelado.

Modificación: Modificación o transformación de variables para crear (en su caso) variables más aptas para utilizar en sus análisis.

Modelado estadístico: El objetivo del modelado estadístico consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales(tales como análisis discriminante, métodos de agrupamiento y análisis de regresión), así como técnicas basadas en datos, tales como redes neuronales, técnicas adaptativas, lógica difusa, árboles de decisión, reglas de asociación y computación evolutiva.

Evaluación: Evaluación del modelado de Minería de Datos contrastándolo con otros métodos o con nuevas poblaciones muestrales.

1.4.3 Análisis comparativo de las metodologías CRISP-DM y Semma.

Las metodologías Semma y CRISP-DM comparten la misma esencia, estructurando el proceso de Minería de Datos en fases que se encuentran interrelacionadas entre sí, siendo el mismo un proceso iterativo e interactivo. La metodología Semma se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proceso de Minería de Datos donde la metodología Semma comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global se

puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Otra diferencia significativa entre la metodología Semma y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología Semma sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proceso de Minería de Datos siendo su distribución libre y gratuita.[Gondar]

1.4.4 Fundamentación de la metodología seleccionada

Después de un estudio comparativo entre las metodologías más utilizadas, dígase CRISP-DM y Semma se seleccionó CRISP-DM como metodología de desarrollo a utilizar en el proceso de Minería de Datos. Las ventajas encontradas en esta metodología son:

- ✓ Concibe el proyecto de Minería de Datos de forma global y estrechamente relacionado al negocio en cuestión.
- ✓ Fue diseñada de forma neutra a la herramienta que se utilice para el desarrollo del proyecto.
- ✓ Es de distribución libre y se encuentra en constante perfeccionamiento por parte de la comunidad internacional.
- ✓ Presenta una precisa y sólida distribución de tareas de carácter general con sus resultados, así como una guía para su desarrollo.
- ✓ Muchas de las metodologías que podemos encontrar en la actualidad se basan en este estándar.
- ✓ Es la que cuenta con mayor aceptación por parte de los desarrolladores de procesos de extracción de conocimientos a partir de datos.

CRISP-DM es producto de la experiencia de varias empresas que se dedican a la Minería de Datos (SPSS, Daimler-Chrysler y NCR; entre otras) y no de un simple estudio teórico.

1.5 Ensayos Clínicos

Los ensayos clínicos constituyen un tipo de estudio clínico que conforma una de las fases del proceso de desarrollo de fármacos para prevenir, detectar o tratar una enfermedad en seres

humanos y ayudan a los médicos a descubrir si estos nuevos tratamientos son mejores que los actuales.

En la medida que se avanza en el desarrollo de los productos, se avanza en las fases de los Ensayos Clínicos. Esto trae aparejado el incremento del número de pacientes a estudiar y con ello el de hospitales. Para la recogida de datos de estos ensayos se hace uso de los Cuadernos de Recogida de Datos (CRD), los cuales guardan grandes volúmenes de datos.

Con los ensayos clínicos se pretende [Cabrera, 2008]:

- ✓ Mejorar las condiciones de salud de la población.
- ✓ Prestar servicios de salud con calidad y seguridad.
- ✓ Mayor beneficio para los pacientes al disponer de nuevos fármacos.
- ✓ Disponer de alternativas terapéuticas para enfermedades que son causas importantes de mortalidad y morbilidad.

Tipos de Ensayos Clínicos

De acuerdo con el Instituto Nacional del Cáncer, hay diferentes tipos de ensayos clínicos para el cáncer, los cuales incluyen [Cabrera, 2008]:

- ✓ Los ensayos de prevención están diseñados para prevenir el desarrollo del cáncer en las personas que no han tenido cáncer anteriormente.
- ✓ Los ensayos de prevención están diseñados para prevenir el desarrollo de un tipo nuevo de cáncer en las personas que ya han tenido cáncer.
- ✓ Los ensayos para la detección temprana están diseñados para encontrar el cáncer especialmente en sus etapas más tempranas.
- ✓ Los ensayos de tratamiento están diseñados para probar terapias nuevas en las personas que tienen cáncer.
- ✓ Los estudios de la calidad de vida están diseñados para mejorar la comodidad y la calidad de la vida de las personas que tienen cáncer.
- ✓ Los estudios conductuales para evaluar las formas de modificar las conductas que causan el cáncer, como el uso del tabaco.
- ✓ Los estudios genéticos para abordar el modo en el que la composición genética afecta la detección, diagnóstico y tratamiento del cáncer. [Cabrera, 2008]:

Fases de los ensayos

Todos los ensayos clínicos pasan por tres etapas obligatorias, estas son [Cabrera, 2008]:

Los ensayos en **fase I** representan la primera vez que un nuevo tratamiento o terapia se prueba en un pequeño número de pacientes (a veces tan sólo una docena). Los ensayos en fase I evalúan la metodología de administración (intravenosa u oral), la frecuencia y la dosificación.

Los ensayos en **fase II** se concentran en el efecto del tratamiento como agente anticancerígeno. Los ensayos en fase II requieren un número ligeramente superior de pacientes que los ensayos en fase I.

Los ensayos en **fase III** son los que comúnmente se denominan como ensayos de tratamiento contra el cáncer. En estos, un gran grupo de pacientes permite a los doctores y científicos que comparen los resultados del nuevo tratamiento contra el estándar de tratamiento actual. Si el nuevo tratamiento prueba ser efectivo, se puede convertir en el nuevo estándar para el cuidado.

En Cuba muchos científicos se dedican al estudio de fármacos para combatir el cáncer, el Centro de Inmunología Molecular (CIM) es uno de los centros dedicados a esta labor, el mismo fue inaugurado el 5 de diciembre de 1994, y tiene como principal misión obtener y producir nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles e introducirlos en el Sistema Nacional de Salud Pública. Hacer la actividad científica y productiva económicamente sostenible y realizar aportes importantes a la economía del país. [CIM]

En el CIM se han realizado varios Ensayos Clínicos para probar fármacos, tal es el caso de la primera vacuna terapéutica contra el cáncer de pulmón, llamada *CIMAVAX EGF.* La vacuna provoca una respuesta inmune y es muy segura porque no provoca eventos adversos severos que pongan en peligro la vida del paciente. [CIM]

Según los estudios realizados, entre los pacientes a los que se les aplicó la nueva vacuna se registró además un incremento de la sobrevida de hasta cinco meses, así como una mejora de la calidad de vida al disminuir los síntomas de la enfermedad como la falta de aire o el dolor, además de registrarse en los enfermos ganancia de peso. [CIM]

Aunque hasta el momento los estudios de la vacuna se han centrado en los enfermos de cáncer de pulmón en un estadio avanzado el producto podría ser potencialmente útil en tumores de cabeza y cuello, cerebro, estómago, mamas, próstata, colorrectal, ovarios o vejiga. [CIM]

Conclusiones

A partir de los resultados del estudio realizado se llegaron a las siguientes conclusiones fundamentales:

- La Minería de Datos es una herramienta eficaz para dar respuestas a preguntas complejas de Inteligencia de Negocios.
- Es una buena manera de convertir datos en información, y esta a su vez en conocimiento, para la correcta toma de decisiones.
- Las herramientas disponibles permiten automatizar gran parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos.
- Las metodologías exponen un modelo de referencia y una guía para el usuario con orientaciones y consejos más detallados para el desarrollo de cada fase y tarea.
- Se seleccionó como herramienta a usar Weka 3.6.0 y metodología CRISP-DM 1.0 para dar solución al problema planteado.

Capítulo 2: Solución propuesta

Introducción

En este capítulo se realiza una descripción de los pasos y actividades propuestos por la metodología seleccionada CRISP-DM 1.0 para realizar el proceso de MD, así como una descripción de los resultados de cada una de las fases que propone dicha metodología, usando la herramienta seleccionada Weka para dar solución al problema planteado. Los datos seleccionados para realizar el proyecto de Minería de Datos corresponden a la información de la Base de Datos del proyecto de Ensayos Clínicos.

Esta metodología posee 6 fases, para su primera fase, dígase **análisis del problema,** se debe tener en cuenta los objetivos y requerimientos del proyecto desde una perspectiva de negocio para luego definir un problema y darle solución al mismo aplicando las técnicas de MD. Después de definido el problema, se siguen con las restantes fases de la metodología.

1. Análisis del problema

1.1 Comprensión del negocio

En esta primera fase se deben analizar con detalle los factores determinantes en el proyecto por lo que se deben conocer los objetivos y requisitos del negocio, y a partir de estos definir el problema que permita alcanzar los objetivos del proyecto y producir resultados que cumplan con las expectativas de los clientes.

Situación actual

El Centro de Inmunología Molecular (CIM) es uno de los centros dedicados al estudio de fármacos para combatir el cáncer. Tiene como principal misión obtener y producir nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades e introducirlos en el Sistema Nacional de Salud Pública.

En este centro actualmente se llevan a cabo varios EC los cuales contienen gran volumen de información tales como datos del ensayo, imágenes y los Cuadernos de Recogida de Datos que constituyen formularios para la recogida de datos, toda esta información es necesaria para cumplir con las buenas prácticas clínicas exigidas por todas las agencias reguladoras a nivel mundial.

Actualmente el centro no cuenta con un sistema que le permita gestionar toda esta información y se recogen los datos en Bases de Datos Microsoft Access o el programa estadístico SPSS con el objetivo de facilitar el análisis final de los resultados de cada ensayo, realizándose una entrada doble de los datos.

En las BD se encuentran recogido gran cantidad de datos del EC para el fármaco EGF, vacuna terapéutica contra el cáncer de pulmón. Sobre estos datos no se ha hecho ningún estudio que permita conocer información interesante sobre los pacientes a los que se les ha aplicado dicho producto.

Después de realizar entrevistas a especialistas del CIM; se detectó que el caso de estudio a realizar en el proyecto de Minería de Datos es el siguiente:

Predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).

Personas claves dentro del negocio

Las personas claves dentro del negocio son las que tienen relación con alguna que otra actividad dentro del proyecto de Ensayos Clínicos, estos son:

- Especialistas del CIM con los roles: Gerente de Datos, Investigador Promotor o Monitor.
- Médicos de los hospitales del país con los roles: Coordinador de la Investigación Clínica o Investigador Principal.

Actores del Negocio

- Especialistas del CIM con el rol: Investigador Promotor

Trabajadores del Negocio

- Especialistas del CIM con los roles: Gerente de Datos o Monitor
- Médicos de los hospitales del país con los roles: Coordinador de la Investigación Clínica o Investigador Principal

Resultados del Proyecto

Los resultados de la investigación deben presentarse a los especialistas del CIM con los siguientes resultados:

- Descripción y resultados obtenidos en cada una de las fases de CRISP-DM 1.0.
- Informe con los resultados obtenidos de la investigación donde se muestren las predicciones del Modelo de Minería de Datos como resultado del proceso de KDD.

- Soporte digital de las soluciones, orígenes de datos y modelos de minería.

Objetivos del Negocio:

A partir de las entrevistas realizadas y la valoración de la situación actual, el objetivo del negocio propuesto para dar solución al caso de estudio, es el siguiente:

Predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).

Preguntas del negocio

Las siguientes preguntas del negocio son las que motivan el objetivo del negocio:

1. ¿Existe relación entre las variables demográficas y las variables surrogadas, de manera que intervengan en el tiempo de supervivencia de un paciente?
2. ¿Tiene influencia en los resultados del ensayo la variable *evaluación de la respuesta?*
3. ¿Cuál es la dosis tolerable para el paciente con cáncer de pulmón y su tiempo de vida teniendo en cuenta que la variable *evaluación de la respuesta* es *resp completa, resp parcial, enfermedad estable y progresión*?
4. ¿Con las técnicas de Minería de Datos se puede dar solución a estas interrogantes?
5. ¿Cuál de los algoritmos que existen permite obtener un modelo que se ajuste mejor al problema?
6. ¿Es posible predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).

Criterios de éxito del Negocio

Los criterios para lograr el éxito de la investigación desde el punto de vista del objetivo del negocio son:

1. Obtener un modelo de conocimiento y comprobar que las conclusiones obtenidas son válidas y que pueden ser utilizadas.

2. Desarrollar el caso de estudio utilizando la herramienta Weka para la minería de datos.
3. Realizar un proyecto de Minería de Datos guiado por la metodología CRISP-DM y la documentación de cada una de las fases.
4. Interpretar los resultados de la relación que existe entre las variables demográficas y la surrogada en el tiempo de vida de los pacientes con cáncer de pulmón.

1.2 Evaluación de la situación

✓ Listado de recursos

A continuación se listan los recursos disponibles para la investigación, incluyendo el personal, datos, recursos computacionales y software.

Tabla 1: Recursos Personales del Proyecto

Personales	
Personal para el proyecto de Minería de Datos	2 desarrolladores
Asesores	1 asesor

Tabla 2: Recursos de Hardware

Hardware	
Microprocesador	Intel(R) Core (TM) 2 Duo CPU 2.20 Ghz
Memoria RAM	1 GB
Capacidad en Disco Duro	110 GB

Tabla 3: Recursos de Software

Software	
Sistema Operativo	Microsoft Windows
Motor de Base de Datos	Postgresql
Herramientas para la exploración de los datos	EMS Postgresql
Herramientas de integración, limpieza y transformació de los datos.	EMS Postgresql y Weka
Herramientas de Minería de Datos	Weka

Tabla 4: Fuente de Datos y de Conocimientos

Fuente de Datos y de Conocimientos	
Documentación del Proyecto de Ensayos Clínicos	Documentación Escrita
Documentación sobre Minería de Datos, herramientas Metodologías de desarrollo.	Documentación Escrita

✓ Requerimientos y Restricciones

A continuación se describen los requerimientos y restricciones que podrían implicar la carencia de recursos para terminar algunas tareas en el proyecto en el tiempo planificado.

Tabla 5: Requerimientos del Proyecto

Requerimientos	
Confiabilidad	Proteger la información de publicidad y de acceso no autorizado a la misma.
Seguridad	No hacer pública información privada del proyecto de Ensayos Clínicos.
Documentación	Estará documentada cada una de las fases de la metodología seleccionada, dígase CRISP-DM 1.0.
Fecha de entrega	Junio 2007.
Modalidad de resultados	Entrega de la documentación de forma digital y presentación de los resultados.

Tabla 6: Restricciones del Proyecto

Restricciones	
Seguridad de los Datos	Los datos de los Ensayos Clínicos no pueden ser divulgados, y a los mismos sólo puede acceder personal autorizado.

✓ Terminología

El glosario de terminología tiene como objetivo fundamental recopilar las palabras claves tanto del Negocio como de la Minería de Datos para una mejor comprensión y familiarización con el proyecto.

Tabla 7: Glosario de términos del Negocio y de la Minería de Datos

Glosario de términos	
Cuadernos de Registro de Datos (CRD)	Formulario diseñado para anotar las variables recogidas durante un ensayo clínico.
Centro de Inmunología Molecular (CIM)	Centro dedicado al desarrollo de biomoléculas y otros fármacos para el tratamiento de enfermedades, principalmente el cáncer.
Ensayos Clínicos	Tipo de estudio clínico en el que se evalúan nuevos productos o tratamientos médicos a través de su aplicación a seres humanos.
Variable Surrogada	Se define como variable surrogada “evaluación de la respuesta”.
Variables Demográficas	Se definen como variables demográficas sexo, edad, raza, estadio, clasificación histológica, peso.
Weka	Herramienta para el proyecto de MD.
CRISP-DM 1.0	Metodología para el proyecto de MD.
Minería de Datos	Proceso para descubrir patrones ocultos en los datos.

Dadas las características actuales de la conducción de los EC del CIM el éxito de la Minería de Datos a dicho proyecto permitirá estimar el tiempo de supervivencia de los pacientes que estén involucrados en los ensayos, posibilitando que se conozca el efecto del producto aplicado al mismo, y el tiempo de vida del paciente con dicho producto.

1.3 Objetivos de la minería

El objetivo de la Minería define las metas del proyecto en términos técnicos, teniendo en cuenta las preguntas del negocio, este es:

Obtener reglas que permitan predecir el tiempo de supervivencia de un paciente para descubrir la influencia de la variable surrogada “evaluación de la respuesta” y las demográficas “sexo”, “raza”, “estadio, “clasificación histológica”, “peso”.

Criterios de éxito de la minería:

Obtener las predicciones con un valor de certeza igual o superior al 80%.

1.4 Plan de Proyecto

Con el plan de proyecto se garantiza que se cumplan los objetivos de la minería y con estos los del negocio. En dicho plan se listan las etapas para ser ejecutadas en el proyecto, su duración, recursos requeridos, entradas, salidas, y dependencias. Cada vez que una tarea nueva sea consultada se debe revisar el plan de proyecto.

Una buena planificación es de gran importancia para cualquier empresa, dado que permite conocer detalles que podrían influir en la buena calidad del proyecto y entrega en tiempo del mismo encaminado a un logro exitoso de sus objetivos.
A continuación se describen las tareas planteadas que permitirán el correcto cumplimiento de los objetivos trazados.

	Nombre de tarea	Duración	Comienzo	Fin	Nombres de los recursos
1	Diseño Teórico Metodológico	3 días	lun 24/11/08	mié 26/11/08	Resp: Danay Perera
2	Estudio del Estado del Arte	30 días	mié 26/11/08	mar 06/01/09	Resp: Enrique Ramírez
3	Estudio de los datos almacenados del proyecto SIMDECC.	15 días	lun 12/01/09	vie 30/01/09	Resp: Danay Perera
4	Reunión con los especialistas del CIM.	1 día?	jue 26/03/09	jue 26/03/09	Resp: Danay Perera;Enrique Ramírez
5	Reunión con los especialistas del CIM.	1 día?	vie 10/04/09	vie 10/04/09	Resp: Danay Perera;Enrique Ramírez
6	Aplicar las técnicas de Minería de Datos a los datos obtenidos en la tarea anterior.	13,5 días	lun 13/04/09	jue 30/04/09	Resp: Enrique Ramírez;Danay Perera
7	Reunión con los especialistas del CIM para validar la solución.	0,5 días	vie 22/05/09	vie 22/05/09	Resp: Danay Perera;Enrique Ramírez
8	Evaluar los patrones obtenidos y si es necesario, se vuelve a las fases anteriores para una nueva iteración.	5 días?	vie 22/05/09	jue 28/05/09	Resp: Danay Perera;Enrique Ramírez

Figura 6: Tareas planteadas que permitirán el correcto cumplimiento de los objetivos trazados.

2. Comprensión de los datos

2.1 Recopilar los datos iniciales

Los datos empleados para el análisis fueron específicamente los de cáncer de pulmón los cuales pertenecen al período 2005 – 2009.

En el centro los datos fueron recogidos utilizando el programa estadístico SPSS, de los que se seleccionaron las siguientes tablas significativas:

Tabla 8: Tablas del Fase III utilizadas para el proyecto de Minería

EGF Pulmón Fase III 081	
Inclus Evaluac Inicial	Almacena las variables demográficas (sexo, edad, raza, estadio clínico, clasificación histológica, peso).
Inclus Interr Fallec Evalni Inmun	Almacena la *evaluación de la respuesta,* dígase *resp completa, resp parcial, enfermedad estable y progresión.*

Inclus Interr Fallec Evalni	Almacena la fecha de muerte, entre otras variables que son de menor importancia dado que no son objeto en el proyecto de Minería.

2.2 Describir los datos

Con la descripción de los datos se pretende examinar sus propiedades, dígase formato, la cantidad de datos a utilizar, atributos de cada tabla usada, chequeo de los tipos de atributos y el rango de valores, se deben realizar actividades de análisis de correlación de atributos, análisis estadísticos, y revisar si los datos contienen entradas de texto libre.

A continuación se muestra los datos por tablas utilizados para el desarrollo del proyecto, solo se tiene en cuenta los campos que tienen mayor importancia por la relación que tienen con las variables objeto de la minería.

Tabla 9: Atributos utilizados de la tabla Inclus Evaluac Inicial de la Fase III.

Inclus Evaluac Inicial: EGF Pulmón Fase III 081.		
Atributo	**Descripción**	**Tipo de Dato**
V1i	Almacena el sexo del paciente, si es 1 es masculino y si es 2 femenino.	Numérico
edad	Almacena la edad del paciente.	Numérico
V4i	Almacena la raza del paciente, si es 1 es blanca, 2 es negra, 3 es mestiza y 4 es amarilla.	Numérico
V22	Almacena el estadio del paciente	Numérico
V26	Almacena la clasificación histológica.	Cadena
V140	Almacena el peso del paciente	Numérico

Tabla 10: Atributos utilizados de la tabla Inclus Interr Fallec Evalni de la Fase III.

Inclus Interr Fallec Evalni: EGF Pulmón Fase III 081.		
Atributo	**Descripción**	**Tipo de Dato**
V1f	Almacena la fecha de muerte del paciente.	Fecha

Tabla 11: Atributos utilizados de la tabla Inclus Interr Fallec Evalni Inmun de la Fase III.

Tabla 7: EGF Pulmón Fase III 081		
Atributo	**Descripción**	**Tipo de Dato**
V32	Almacena la *evaluación de la respuesta,* dígase *resp completa, resp parcial, enfermedad estable y progresión.*	Numérico

2.3 Explorar los datos

Este paso es necesario para comprender los datos y así poder tomar las decisiones adecuadas antes de generar los modelos. Para ello se podrán realizar consultas, reportes y vistas para consolidar el objetivo final del proyecto de minería.

Los resultados más significativos del reporte exploratorio de los datos por campos son:

- ✓ El promedio de la edad de los pacientes es de 59.
- ✓ El promedio del peso de los pacientes es de 64.
- ✓ El 59% de los pacientes es de sexo masculino y el 32% es de sexo femenino.
- ✓ El 47% de los pacientes es de raza blanca, el 10% es de raza negra y el 7% es de raza mestiza.
- ✓ El 29% de los pacientes presenta estadio IIIB y el 13% presenta estadio IV.
- ✓ El 13% de los pacientes tiene adenocarcinoma (clasificación histológica) y el 27% no tiene adenocarcinoma.
- ✓ El 4% de los pacientes tiene como "evaluación de la respuesta" respuesta completa, el 16% respuesta parcial, el 20% enfermedad estable y no hay pacientes con progresión.

2.4 Verificar la calidad de los datos

En este paso se debe revisar la calidad de los datos, que no contengan errores, que estén completos, si son usuales estos problemas en los datos y verificar si hay valores que no existen en los datos para decidir qué hacer con ellos.

Utilizando los resultados de la exploración de los datos se pudo comprobar que en general existen campos que contienen valores nulos, al analizar dichos campos se encontraron los siguientes problemas:

El 35% de los pacientes tiene el valor “edad” = 0.

El 55% de los pacientes tiene el valor “peso” = 0.

El 4% de los pacientes tiene el valor “sexo” = 0.

El 34% de los pacientes tiene el valor “raza” = 0.

El 58% de los pacientes tiene el valor “estadio” = 0.

El 60% de los pacientes tiene el valor “clasificación histológica” = 0.

El 40% de los pacientes tiene el valor “evaluación de la respuesta” = 0.

El 65% de los pacientes tiene el valor “fecha de entrada” = null.

El 53% de los pacientes tiene el valor “fecha de muerte” = null.

3. Preparación de los datos

Esta fase contiene todas las actividades para obtener uno o más conjuntos de datos. Las tareas incluyen la selección de las tablas, registros, y atributos para aplicarles transformaciones y limpieza de sus datos con el objetivo de que los datos tengan la calidad suficiente para ser usados en las herramientas de modelación.

Se realizó la integración y transformación de grandes volúmenes de datos con postgresql.

3.1 Selección de datos

Es necesario comprender bien los datos y analizar la influencia e importancia de los mismos de acuerdo a los objetivos de minería, para hacer una buena elección de los datos a utilizar para el análisis. Se pueden excluir o incluir datos de acuerdo a nuevos análisis reconsiderando el criterio de selección de los datos.

Los atributos seleccionados para el proyecto de Minería coinciden con los descritos en la fase de Comprensión de los Datos, por lo que no se realizará una descripción detallada de los campos en esta tarea.

3.2 Construir los datos

Las transformaciones sobre los datos que incluye esta tarea son: actualizar valores de columnas, crear nuevas columnas, introducir nuevos registros que se componen de valores agregados u ordenados, en caso que lo exija la tarea de minería que se desarrolla.

Atributos derivados

Para un correcto cumplimiento del objetivo planteado en la fase de Análisis del problema fue necesario adicionar un nuevo campo (tiempo_vida) en la BD dado que el mismo es la variable objetivo para la MD. Para la construcción de dicha variable se hizo necesario la creación de otro campo, el cual no constituye una variable objetivo, pero fue fundamental para llenar los registros del campo tiempo_vida.

Los valores de los registros del campo tiempo_vida van a tomar los siguientes valores:

rango1: Pertenece a los pacientes que su tiempo de vida sea de 1 a 300 días.

rango2: Pertenece a los pacientes que su tiempo de vida sea de 300 a 1000 días.

3.3 Limpieza de datos

En este paso se pueden extraer y transformar los datos para elevar la calidad de los mismos al nivel que exigen las técnicas de análisis seleccionadas. Se obtiene como salida un reporte de los datos limpios. En esta tarea se describen las decisiones y acciones realizadas para solucionar los problemas descritos en la calidad de los datos durante la tarea de verificar la calidad de los datos en la fase anterior de “Comprensión de los datos”.

Se realizaron consultas para eliminar todo los valores nulos encontrados en la **Verificación de la calidad de los datos,** además de eliminaron los campos de “fecha de entrada” y “fecha de muerte”, que solo se utilizaron para la construcción del campo tiempo_vida.

3.4 Integrar los datos

En esta tarea se analizan los datos que son necesarios para el proyecto y se combinan en el caso de que se encuentren en diversas fuentes. Se integra la información que se extrae de tablas diferentes y

se crea una o más tablas que contienen información útil sobre los mismos objetos. Puede además que se generen nuevos registros o columnas que generalicen la información de múltiples tablas.

Toda la información necesaria para realizar la investigación se encontraba en formato SPSS el cual fue transformado a una BD en postgresql, dado que la misma contiene funcionalidades que sirvieron de ayuda en el desarrollo de la investigación.

Los atributos seleccionados para realizar el proyecto de Minería, correspondiente a los datos de los pacientes se encontraban en varias tablas dentro de la BD, como se analizó en el apartado de la fase **Comprensión de los datos**, **Recopilar los datos iniciales.**

Con el objetivo de agrupar estos datos en una sola tabla se hizo una consulta, donde se obtuvo la tabla *ensayos_clínicos* con todas las variables objetivo de la minería.

En la figura 6 se muestra una parte de los datos de la tabla *ensayos_clínicos.*

edad	sexo	piel	peso	estadio	cla_histologica	eva_respuesta	tiempo_vida
77	1	1	67	2	1	3	rango1
76	1	1	61	1	2	3	rango2
73	1	1	67	1	1	3	rango2
73	1	2	77	1	2	3	rango1
72	1	1	53	1	2	3	rango1
71	1	1	72	2	2	2	rango2
70	2	4	60	1	1	3	rango2
69	1	1	46	1	2	3	rango1
69	1	1	72	1	1	2	rango1
65	2	1	46	1	1	3	rango1
65	1	1	84	2	2	3	rango1
64	1	2	96	2	2	3	rango1
63	1	1	55	1	2	3	rango1
63	1	1	54	1	2	2	rango1
62	2	1	66	1	2	1	rango1
60	1	1	65	1	1	3	rango1
60	2	1	36	2	1	2	rango1
59	1	1	91	2	2	3	rango2
59	2	1	48	2	1	2	rango2
59	2	1	43	1	1	3	rango1
59	1	1	59	2	2	2	rango1
59	1	4	51	1	1	3	rango1
58	1	1	68	2	2	2	rango2
58	1	1	47	1	2	2	rango1
58	1	1	48	1	2	2	rango1

Figura 7: Tabla que contiene la integración de los datos.

A continuación se presenta la relación de la nueva variable tiempo_vida descrita anteriormente, con las variables significativas para la minería que se describieron en la fase de **Comprensión de los Datos**.

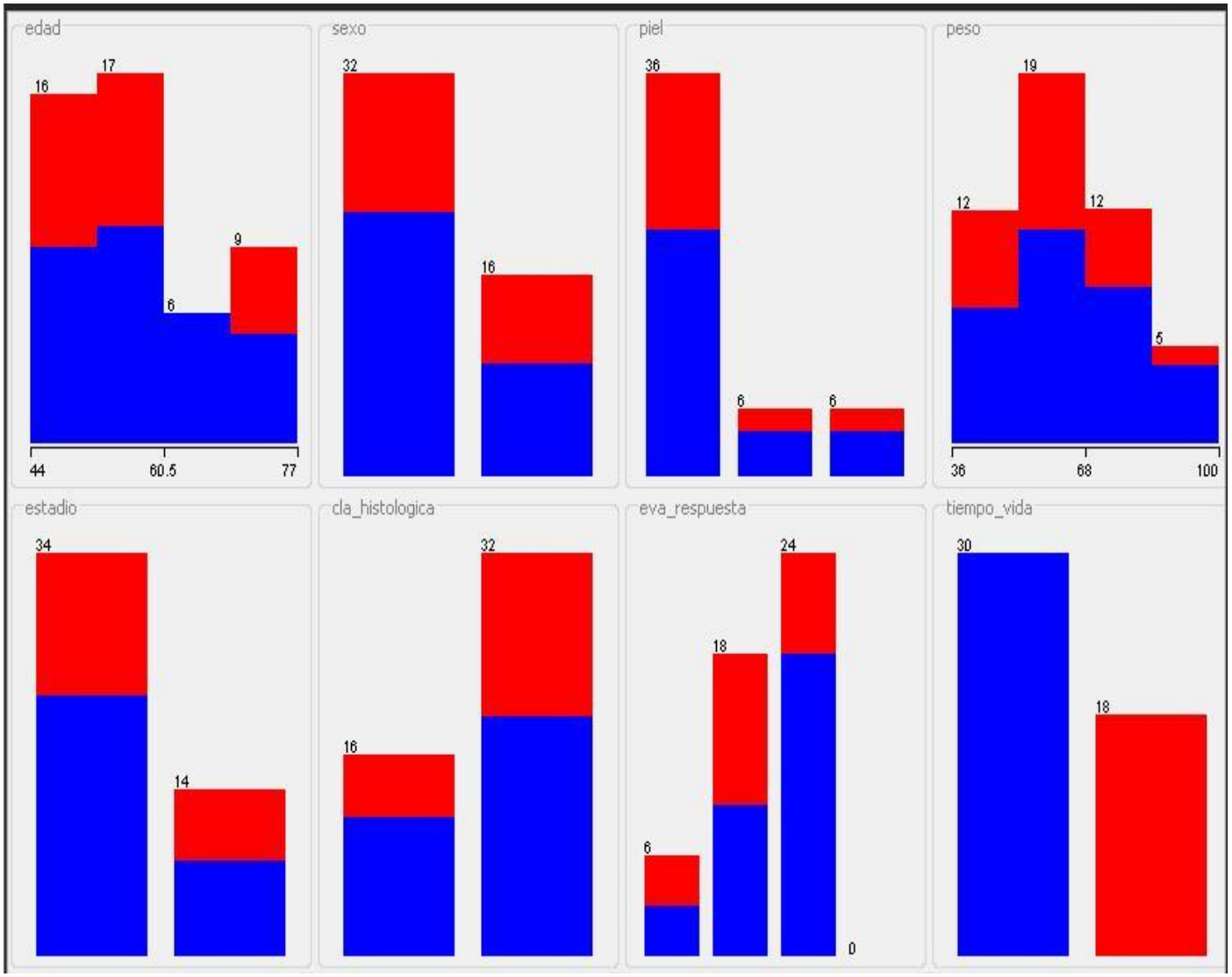


Figura 8: Comportamiento de las variables significativas con respecto al tiempo de vida.

Desde esta ventana podemos conocer varios detalles del dataset que se cargó en la herramienta. Por ejemplo, la figura indica que tenemos 8 atributos. Esta muestra un histograma con información sobre la distribución de las variables significativas con respecto a la variable tiempo_vida, reflejando con el uso

de colores la distribución de clases de cada uno de los atributos. A continuación se describe el significado de cada una de estas variables.

Tabla 12: Distribución de la variable edad con respecto a la variable tiempo_vida.

Variable Edad			
Edad	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
44-52	16	9	7
52-60	17	10	7
60-68	6	6	-
68-77	9	5	4

Tabla 13: Distribución de la variable sexo con respecto a la variable tiempo_vida.

Variable Sexo			
Sexo	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
Masculino(m)	32	21	11
Femenino(f)	16	9	7

Tabla 14: Distribución de la variable piel con respecto a la variable tiempo_vida.

Variable Piel			
Piel	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
Blanca(b)	36	22	14
Negra(n)	6	4	2
Mestiza(m)	6	4	2

Tabla 15: Distribución de la variable peso con respecto a la variable tiempo_vida.

Variable Peso			
Peso	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
36-52	12	7	5
52-68	19	11	8
68-84	12	8	4

84-100	5	4	1

Tabla 16: Distribución de la variable estadio con respecto a la variable tiempo_vida.

Variable Estadio			
Estadio	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
IIIB	34	22	12
IV	14	8	6

Tabla 17: Distribución de la variable clasificación_histológica con respecto a la variable tiempo_vida.

Variable Clasificación_histológica			
Clasificación_histológica	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
Adenocarcinoma	16	11	5
No Adenocarcinoma	32	19	13

Tabla 18: Distribución de la variable evaluación_de_la_respuesta con respecto a la variable tiempo_vida.

Variable Evaluación_de_la_respuesta			
Evaluación_de_la_respuesta	**Cantidad de Pacientes**	**Rango 1**	**Rango 2**
Respuesta Completa(rc)	6	3	3
Respuesta Parcial(rp)	18	9	9
Enfermedad Estable(ee)	24	18	6
Progresión(p)	0	-	-

Tabla 19: Distribución de la variable de la variable tiempo_vida.

Variable Tiempo_vida			
	Cantidad de Pacientes	**Rango 1**	**Rango 2**
Tiempo_vida	48	30	18

3.5 Formatear los datos

Las transformaciones de estructuración y formato se refieren a modificaciones sintácticas hechas a los datos que no cambian su significado, que pueden ser requeridas para las herramientas de modelación usadas. Algunas herramientas tienen exigencias sobre el orden de los atributos, tales como que el primer campo sea un identificador único o que el último campo sea el campo de salida que el modelo va predecir.

Fue necesario realizar transformaciones en los nombres de los campos de la tabla dado que los que contenía no eran entendibles.

4. Modelado

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas con sus parámetros medidos en valores óptimos. Existen muchas técnicas aplicables al mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto frecuentemente es necesario regresar a la fase de la preparación de datos.

4.1 Seleccionar las técnicas de modelado

Descripción de los posibles algoritmos a utilizar.

Para el modelado se hace uso de la herramienta Weka en su versión 3.6.0, así como sus técnicas para aplicarlas al objetivo planteado anteriormente.
Weka cuenta con varios entornos de trabajo, entre ellos está el Explorer el que se utilizo para dar solución al objetivo de la investigación. Este entorno cuenta con 6 sub-entornos de ejecución (Ver anexo 2) los cuales se describen brevemente a continuación.

- ✓ **Preprocess**: Incluye las herramientas y filtros para cargar y manipular los datos.
- ✓ **Classification**: Acceso a las técnicas de clasificación y regresión.
- ✓ **Cluster**: Integra varios métodos de agrupamiento.
- ✓ **Associate**: Incluye una pocas técnicas de reglas de asociación.
- ✓ **Select Attributes**: 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.

En este trabajo se evaluarán 6 algoritmos clasificadores de diferentes tipos con el objetivo de determinar cuál es el mejor para predecir el tiempo de supervivencia de un paciente con cáncer de pulmón sobre la base de comparar los resultados de sus clasificaciones en una misma muestra.
Se considera un mejor clasificador aquel que tiene mayor precisión en la predicción.

Weka dispone de una gran variedad de algoritmos para clasificar. Se aplicaran los siguientes algoritmos a los datos sobre los cuales se va a trabajar:

- ✓ OneR
- ✓ PART
- ✓ DecisionTable
- ✓ LMT
- ✓ ADTree
- ✓ C4.5 (J48)

Se han elegido estos algoritmos porque constituyen los más representativos. Se han sometido a estudio los datos con todos los algoritmos inicialmente propuestos y como se verá algunos consiguen excelentes soluciones mientras que otros dan peores aproximaciones. Sin embargo presentan la gran ventaja de que se pueden usar para cualquier tipo de datos.
Se realizará un estudio más detallado del algoritmo J48 dado que es uno de los algoritmos más utilizados en la práctica, por lo que resulta interesante su estudio.

A continuación se describirán cada uno de los posibles algoritmos a utilizar lo que mostrará más información para seleccionar el más adecuado.

Clasificador 1R (OneR)

Es un algoritmo sencillo que sin embargo funciona de forma parecida a complejos árboles de decisión. La idea es hacer reglas que prueban un solo par atributo-valor. Se prueban todos los pares atributo-valor y se selecciona el que ocasione el menor número de errores. [Weka]

En Weka, las prestaciones del clasificador **OneR** son las siguientes:
- Nombre de la clase: weka.classifiers.OneR.
- No puede manejar instancias ponderadas por pesos.
- No puede procesar datos categóricos.

- No puede ser actualizado de forma incremental (soportar añadir nuevos datos sin reclasificar a los anteriores).
- Produce reglas sencillas basadas en un solo atributo. Tiene un solo parámetro: el número mínimo de instancias que deben ser cubiertas por cada regla generada (6 por defecto). [Weka]

PART

Evita el paso de optimización global que se usa en las reglas del C4.5, genera una lista de decisión sin restricciones usando el procedimiento de divide y vencerás. Además construye un árbol de decisión parcial para obtener una regla. Para poder podar una rama (una regla) es necesaria que todas sus implicaciones sean conocidas. [Weka]

El PART evita la generalización precipitada, y usa los mismos mecanismos que el C4.5 para construir un árbol. La hoja con máxima cobertura se convierte en una regla y los valores ausentes de los atributos la instancia se divide en piezas. [Weka]

En Weka, las prestaciones del algoritmo PART son las siguientes [Weka]:

- Nombre de la clase: weka.classifiers.j48.PART.
- Puede manejar instancias ponderadas por pesos.
- Puede procesar datos categóricos.
- No puede ser actualizado de forma incremental (soportar añadir nuevos datos sin reclasificar a los anteriores).
- PART forma regla a partir de árboles de decisión parcialmente podados construidos usando los heurísticos de C4.5. Las opciones disponibles para este algoritmo son un subconjunto por tanto de las disponibles para J4.8. Al igual que podíamos reducir el tamaño del árbol de decisión J4.8 usando poda de error reducido, se puede reducir el número de reglas de PART.

Sin embargo el podado de bajo error reduce la precisión del árbol de decisión y reglas resultante porque reduce la cantidad de datos que se usan en el entrenamiento. Con grandes cantidades de datos no es necesario tener esta desventaja en cuenta. [Weka]

DecisionTable

Más que un árbol, la tabla de decisión es una matriz de renglones y columnas que indican condiciones y acciones. Las reglas de decisión, incluidas en una tabla de decisión, establecen el procedimiento a seguir cuando existen ciertas condiciones. Este método se emplea desde mediados de la década de los cincuentas, cuando fue desarrollado por General Electric para el análisis de funciones de la

empresa como control de inventarios, análisis de ventas, análisis de créditos y control de transporte y rutas.

La tabla de decisión está integrada por cuatro secciones: identificación de condiciones, entradas de condiciones, identificación de acciones y entradas de acciones de la siguiente tabla. [Weka]

La identificación de condiciones señala aquellas que son relevantes. Las entradas de condiciones indican qué valor, si es que lo hay, se debe asociar para una determinada condición. La identificación de acciones presenta una lista del conjunto de todos los pasos que se deben seguir cuando se presenta cierta condición. Las entradas de acciones muestran las acciones específicas del conjunto que deben emprenderse cuando ciertas condiciones o combinaciones de éstas son verdaderas. [Weka]

En Weka, las prestaciones de las tablas de decisión son las siguientes [Weka]:

- Nombre de la clase: weka.classifiers.DecisionTable.
- Puede manejar instancias ponderadas por pesos.
- Puede procesar datos categóricos.
- No puede ser actualizado de forma incremental (soportar añadir nuevos datos sin reclasificar a los anteriores).
- La tabla se genera seleccionando un subconjunto de atributos representativos.

Esto se hace utilizando una búsqueda del primer atributo mejor. Por defecto se prueban al menos 5 grupos de atributos en busca de la mejor solución, aunque es configurable. También se puede variar el número de agrupaciones de atributos que se hacen. Esto mejora significativamente el rendimiento. [Weka]

LMT

El algoritmo LMT adapta la idea de los trabajos de Quinlan a problemas de clasificación. Para resolver problemas de clasificación en estadística, el análogo a la regresión linear es regresión logística linear, de manera que el LMT construye los árboles de clasificación con funciones de regresión linear logística en las hojas del árbol. [Tejeda, 2004]

Construye un árbol de decisión, a partir de un subconjunto de los datos de entrenamiento, con ramificaciones binarias en atributos numéricos, ramificaciones múltiples en los nominales, y modelos de regresión logística en las hojas. El algoritmo asegura que en estos últimos solamente se incluyen atributos relevantes. [Sánchez, 2006]

ADTree

ADTree constituye un árbol de decisión alternativo. Es un método de clasificación proveniente del aprendizaje automático conocido en inglés como Alternating Decision Tree (ADTree). Las estructuras de datos y el algoritmo son una generalización de los árboles de decisión. El ADTree fue introducido por Yoav Freund y Llew Mason en 1999. [Sánchez, 2006]

C4.5 (J48)

Se trata de una versión posterior del ID3. Los árboles de decisión extienden el ID3 para que pueda trabajar con atributos numéricos. El C4.5 acaba con muchas de las limitaciones del ID3. Permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas en función de un umbral. Los árboles son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases. [Weka]

El C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (depth-first). [Weka]

Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. [Weka]

La implementación en Weka de este árbol de decisión de aprendizaje es el algoritmo J48.

El algoritmo J48 es una versión implementada en WEKA del método C4.5, como árbol de decisión para tareas de clasificación fundamentalmente, muy simple y potente. El procedimiento para generar el árbol consiste en seleccionar un atributo como raíz, y crear una rama con cada uno de los posibles valores de dicho atributo. Con cada rama resultante (nuevo nodo del árbol), se realiza el mismo proceso. En cada nodo se debe seleccionar un atributo para seguir dividiendo, y para ello se selecciona aquel que mejor separe los ejemplos de acuerdo a las clases. [Molina, 2006]

Dentro de las opciones que J48 soporta están: [Macías, 2008]

- La poda de árboles
- La especificación de factores de confianza para la poda
- La especificación de un mínimo de instancias en las hojas

• La poda de árboles con error reducido
• La especificación del número de datos en podas con error reducido
• El uso de particiones binarias en atributos nominales

En Weka, las prestaciones del algoritmo C4.5 son las siguientes:
- Nombre de la clase: weka.classifiers.j48.J48.
- Puede manejar instancias ponderadas por pesos.
- Puede procesar datos categóricos.
- No puede ser actualizado de forma incremental (soportar añadir nuevos datos sin reclasificar a los anteriores).
- Weka permite utilizar árbol podado o no podado, se puede impedir el aumento de los subárboles, lo que desemboca en algoritmos más eficientes. También se puede fijar el umbral de confianza para el proceso de poda, y el número mínimo de instancias permitido en cada hoja.
- Además de los procesos estándar de C4.5, se permite una opción que disminuye el error de poda, realizándose una poda del árbol de decisión que optimiza el rendimiento en un conjunto fijo. Se puede fijar el tamaño de este grupo: el conjunto de datos se divide por igual en el número de grupos fijado, y la última parte se usa como conjunto fijo. También permite la construcción de árboles binarios. [Weka]

Es importante señalar que los árboles de decisión presentan una gran ventaja respecto a otras técnicas de clasificación. Esta ventaja consiste en poder representar al conocimiento obtenido mediante el uso de reglas de decisión.

A continuación se muestra una tabla donde se realiza una comparación entre los algoritmos descritos anteriormente a partir de los resultados arrojados por cada uno de ellos con el conjunto de datos seleccionado durante la fase de **Preparación de los datos.**

Tabla 20: Comparación entre varios algoritmos de clasificación.

Parámetros	OneR	PART	DecisionTable	LMT	ADTree	J48
Instancias clasificadas correctamente	32	41	34	36	41	42
Instancias clasificadas incorrectamente	16	7	14	12	7	6

Precisión	0.685	0.855	0.719	0.761	0856	0.875

Después de analizados los resultados obtenidos y por ser el de mejor precisión se selecciona como técnica usada en el modelado al Algoritmo de Árboles de Decisión: J48.

Esta técnica tiene como objetivo obtener modelos de clasificación que permita predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada “evaluación de la respuesta”, basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).

4.2 Construcción de los modelos

En este paso se crean los modelos necesarios en el proyecto. Para ello se ejecuta la herramienta de modelación con el conjunto de datos preparados. En las salidas se debe mostrar el modelo real generado por la herramienta de modelación, mostrar reportes para la interpretación de modelos y de las reglas producidas por este.

A continuación se describe el modelo que se obtuvo a partir de ejecutar la técnica de análisis de información declarada en pasos anteriores sobre el conjunto de datos seleccionado durante la fase de **Preparación de los datos.**

En el modelo de árboles de decisión que se muestra a continuación como resultado de ejecutar el algoritmo J48, las reglas se construyen de arriba a abajo y de izquierda a derecha de manera escalonada desde el nodo raíz hasta los nodos hojas.

El nodo ubicado más a la izquierda en la representación constituye la raíz del árbol. Los nodos hojas por su parte son aquellos a los que le sigue el valor alcanzado por el tiempo de vida (variable a predecir).

eva_respuesta = rc
| *sexo = m: rango2 (3.0/1.0)*
| *sexo = f: rango1 (3.0/1.0)*
eva_respuesta = rp
| *cla_histologica = adeno: rango1 (5.0/1.0)*
| *cla_histologica = noadeno*
| | *estadio = IIIB*

| | | edad <= 55: rango2 (4.0)
| | | edad > 55: rango1 (3.0)
| | estadio = IV: rango2 (6.0/2.0)
eva_respuesta = ee
| estadio = IIIB
| | peso <= 56: rango1 (7.0)
| | peso > 56
| | | peso <= 73
| | | | edad <= 65
| | | | | edad <= 51: rango2 (2.0)
| | | | | edad > 51: rango1 (2.0)
| | | | edad > 65: rango2 (3.0)
| | | peso > 73: rango1 (4.0)
| estadio = IV: rango1 (6.0/1.0)
eva_respuesta = p: rango1 (0.0)

El conjunto de patrones que se presentan en el modelo fueron obtenidos con una precisión de 0.875, lo que equivale decir que se clasificaron correctamente el 87,5% del total de casos. La herramienta arrojó los siguientes resultados.

Correctly Classified Instances	*42*	*87.5 %*
Incorrectly Classified Instances	*6*	*12.5 %*
Kappa statistic	*0.7333*	
Mean absolute error	*0.1792*	
Root mean squared error	*0.2993*	
Relative absolute error	*38.1206 %*	
Root relative squared error	*61.8208 %*	
Total Number of Instances	*48*	

Para un mayor entendimiento del modelo se muestra el árbol obtenido a partir de aplicar la técnica de Árboles de Decisión: J48. Los nodos representan atributos, las ramas representan valores de dichos atributos y los nodos finales representan los valores de la clase. Cada camino del árbol representa una regla.

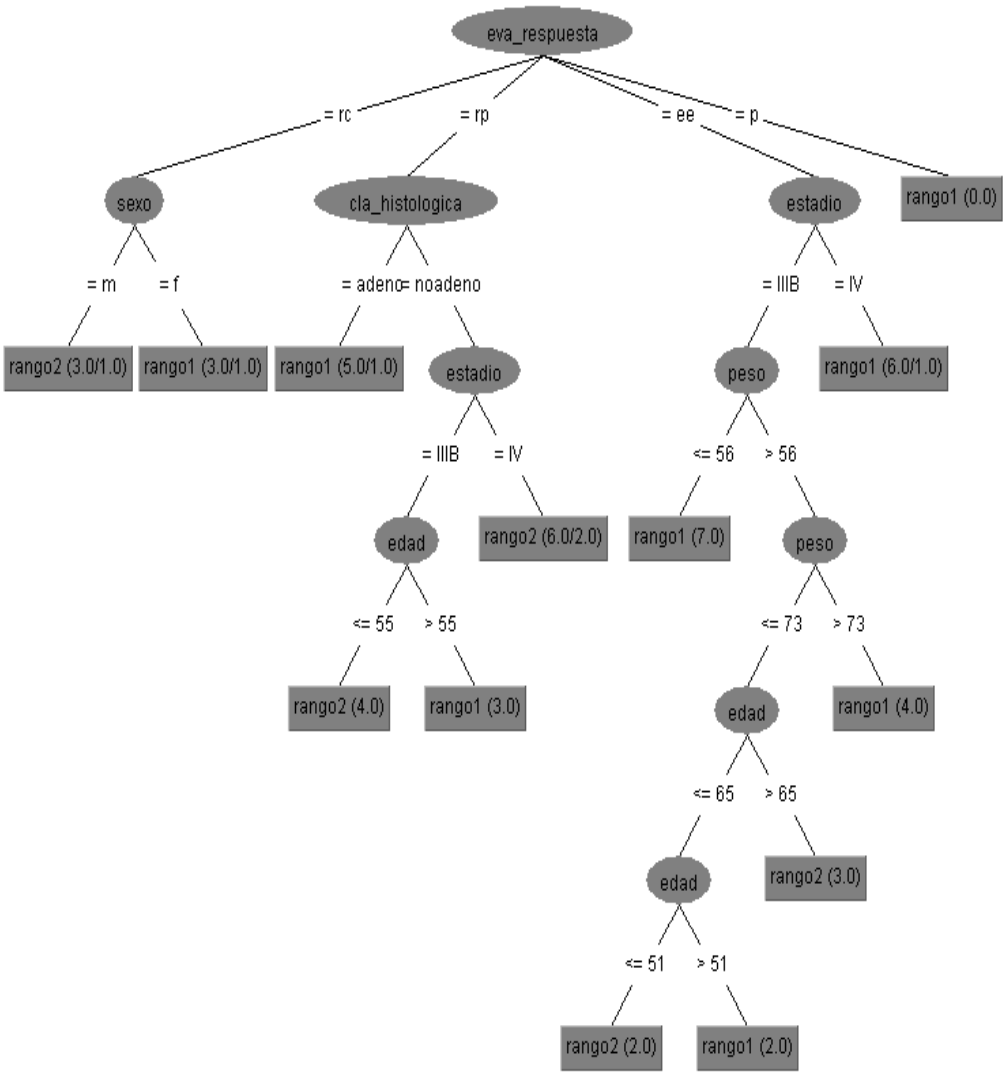


Figura 9: Modelo obtenido a partir de la aplicación del algoritmo utilizado.

Entre las observaciones que se pueden realizar según los patrones más relevantes obtenidos en el anterior modelo se encuentran las siguientes:

- Si un paciente tiene “evaluación de la respuesta” respuesta completa y el sexo es masculino entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” respuesta completa y el sexo es femenino entonces su tiempo de vida es de 1 a 300 días.

- Si un paciente tiene “evaluación de la respuesta” respuesta parcial y su “clasificación histológica” es adenocarcinoma entonces su tiempo de vida es de 1 a 300 días.
- Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IIIB y su “edad” es menor o igual que 55 entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IIIB y su “edad” es mayor que 55 entonces su tiempo de vida es de 1 a 300 días.
- Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IV entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB y su “peso” es menor o igual que 56 entonces su tiempo de vida es de 1 a 300 días.
- Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es menor o igual que 51 entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es mayor que 51 y menor o igual que 65 entonces su tiempo de vida es de 1 a 300 días.
- Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es mayor que 65 entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 73 entonces su tiempo de vida es de 300 a 1000 días.
- Si un paciente tiene “evaluación de la respuesta” progresión entonces su tiempo de vida es de 1 a 300 días.

5. Evaluación

En esta fase se evalúa el modelo escogido, desde el punto de vista del cumplimiento de los objetivos del negocio. Se debe revisar el proceso teniendo en cuenta los resultados obtenidos, para repetir alguna fase en caso que se hayan cometido errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase y de la precisión del mismo, se procede al despliegue de éste en caso de requerirse.

5.1 Evaluación de los resultados

Este paso determina si el modelo resuelve los objetivos de negocio e intenta determinar si hay alguna razón del negocio por la que este modelo es deficiente. Por otra parte, la evaluación también debe analizar otros resultados generados durante la minería. Los resultados de la minería cubren los modelos que están relacionados con los objetivos del negocio.

Resultados obtenidos a partir del uso de la herramienta Weka, específicamente con el algoritmo de árboles de clasificación J48, que permite evaluar el modelo obtenido. Uno de los resultados es el que se muestra a continuación.

```
=== Confusion Matrix ===

 a  b   <-- classified as
27  3 |  a = rango1
 3 15 |  b = rango2
```

Matriz de confusión: Para cada clase real registra el número de casos en los cuales el clasificador ha predicho una clase. La suma de la diagonal (Traza) corresponde al número total de aciertos.
De la anterior matriz se puede afirmar que hubo 42 aciertos en la clasificación y quedaron mal clasificados 6 de los 48 casos.

Para el cálculo de la precisión del modelo se tiene en cuenta Verdaderos Positivos (Instancias clasificadas correctamente) y Falsos Negativos (Instancias clasificadas incorrectamente), por lo que se hace uso de la siguiente ecuación para obtener dicho resultado.

$$Presición = \frac{(Verdaderos \quad Positivos)}{(Verdaderos \quad Positivos) + (Falsos \quad Positivos)}$$

Figura 10: Ecuación que utiliza el algoritmo de árboles de decisión: J48 para la precisión. [Macías, 2008]

Por lo antes expuesto el algoritmo arrojó el siguiente resultado.

Correctly Classified Instances	*42*	*87.5 %*
Incorrectly Classified Instances	*6*	*12.5 %*

Se analizó el modelo en detalle con el objetivo de obtener otros resultados de interés. Otro de los resultados fue el siguiente:

Kappa statistic *0.7333*

WEKA calcula el «Kappa statistic (Coeficiente de Kappa)» para mostrar la concordancia entre los datos de prueba y la clasificación hecha por el modelo. Cuando todas las instancias son clasificadas correctamente se obtiene la máxima concordancia, es decir, Kappa statistic = 1.
Para el caso del árbol de decisión obtenido, el «Coeficiente de Kappa» resultó ser igual a 0.7333, lo cual indica un alto nivel de concordancia entre los datos de prueba y la clasificación hecha por el modelo.

Descripción de la precisión por clases.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  Class
                 0.9       0.167      0.9       0.9        0.9      rango1
                 0.833     0.1        0.833     0.833      0.833    rango2
Weighted Avg.    0.875     0.142      0.875     0.875      0.875
```

Figura 11: Datos obtenidos del modelo obtenido.

En el detalle de precisión por clase, se observa que la clase “Rango 1” presenta una Tasa de Verdaderos Positivos (TP Rate) del 90% y la clase “Rango 2” del 83.3%. Esto quiere decir que, dada una instancia perteneciente a la clase “Rango 1”, el árbol de decisión clasifica a dicha instancia como “Rango 1” el 90% de las veces y “Rango 2” el 83.3% de las veces. En el mismo sentido, se tiene que la Tasa de Falsos Positivos (FP Rate) para las instancias clasificadas por el modelo es del 16.7% para la clase “Rango 1” y del 10% para la clase “Rango 2”.
Con las medidas anteriores, se obtiene la precisión del modelo para las clases “Rango 1” y “Rango 2”, la cual es del 90% y 83.3 % respectivamente. Este resultado se obtiene usando la ecuación planteada anteriormente en este epígrafe. Este porcentaje nos indica la proporción de aciertos del modelo obtenido.

Una medida más que ofrece WEKA es la llamada «F-Measure», la cual representa la media armónica entre la precisión y el Recall. Entre más cercana sea a 1, mayor será la confiabilidad del modelo en la clase. Para obtener los resultados del «F-Measure» y el Recall el algoritmo usa las siguientes ecuaciones:

$$F - Measure = \frac{2*(Presición*Recall)}{(Presición) + (Recall)}$$

$$Recall = \frac{(Verdaderos \quad Positivos)}{(Verdaderos \quad Positivos) + (Falsos \quad Negativos)}$$

Figura 12: Ecuación que utiliza el algoritmo de árboles de decisión: J48 para el cálculo de «F-Measure» y el Recall. [Macías, 2008]

Para la clase “Rango 1” se tiene una confiabilidad (F-Measure) del 90% y la clase “Rango 2” del 83.3%.
De esta forma, se tiene que el árbol de decisión obtenido clasifica de manera aceptable a las instancias pertenecientes a las clases “Rango 1” y “Rango 2”. En otras palabras, se cuenta con un conjunto de reglas de decisión que clasifican con un buen grado de confiabilidad a las instancias.

En la tabla 21 que se muestran las reglas obtenidas a partir de los modelos de Árboles de Decisión generados y el valor de precisión de cada una de ellas.

Tabla 21: Reglas obtenidas a partir de los datos seleccionados para la minería así como la probabilidad de que ocurra.

Reglas	Precisión
Si un paciente tiene “evaluación de la respuesta” respuesta completa y el sexo es masculino entonces su tiempo de vida es de 300 a 1000 días.	0.833
Si un paciente tiene “evaluación de la respuesta” respuesta completa y el sexo es femenino entonces su tiempo de vida es de 1 a 300 días	0.9
Si un paciente tiene “evaluación de la respuesta” respuesta parcial y su “clasificación histológica” es adenocarcinoma entonces su tiempo de vida es de 1	0.9

	a 300 días.	
	Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IIIB y su “edad” es menor o igual que 55 entonces su tiempo de vida es de 300 a 1000 días.	0.833
	Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IIIB y su “edad” es mayor que 55 entonces su tiempo de vida es de 1 a 300 días.	0.9
	Si un paciente tiene “evaluación de la respuesta” respuesta parcial, su “clasificación histológica” no es adenocarcinoma, su “estadio” es IV entonces su tiempo de vida es de 300 a 1000 días.	0.833
	Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB y su “peso” es menor o igual que 56 entonces su tiempo de vida es de 1 a 300 días.	0.9
	Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es menor o igual que 51 entonces su tiempo de vida es de 300 a 1000 días.	0.833
	Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es mayor que 51 entonces su tiempo de vida es de 1 a 300 días.	0.9
	Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 y menor o igual que 73 y su edad es mayor que 65 entonces su tiempo de vida es de 300 a 1000 días.	0.833
	Si un paciente tiene “evaluación de la respuesta” enfermedad estable, su “estadio” es IIIB, su “peso” es mayor que 56 entonces su tiempo de vida es de 1 a 300 días.	0.9
	Si un paciente tiene “evaluación de la respuesta” progresión entonces su tiempo de vida es de 1 a 300 días.	0.9

Propuesta de mejora del modelo

A continuación se describen las propuestas de mejoras para el modelo obtenido basado en el criterio de éxito del modelo.

Tabla 22: Propuesta de mejoras en el modelo Arboles de Decisión: J48. Predicción del tiempo de supervivencia de los pacientes con cáncer de pulmón.

Propuestas de mejora
Modelo: Arboles de Decisión: J48. Predicción del tiempo de supervivencia de los pacientes con cáncer de pulmón.
Objetivo del experimento: Predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).
Criterios de éxito del modelo: Obtener las predicciones con un valor de certeza igual o superior al 80%.
Propuesta de Mejoras: ✓ Realizar los modelos de Minería de Datos utilizando el algoritmo de árboles de decisión J48 con un mejor refinamiento de los datos, para dicho refinamiento se incluirán otros datos pertenecientes a otros ensayos realizados en el CIM. ✓ Utilizar como entrada otros atributos para encontrar otras relaciones entre los mismos.

Resumen de la evaluación de los resultados

A continuación se muestra una tabla con la estimación de cumplimiento del objetivo del negocio basado en los criterios de éxito.

Tabla 23: Estimado de cumplimiento de los criterios de éxito del negocio.

Criterios de éxito del negocio	Cumplimiento estimado
Obtener un modelo de conocimiento y comprobar que las conclusiones obtenidas son válidas y que pueden ser utilizadas.	100%
Desarrollar el caso de estudio utilizando la herramienta Weka para la minería de datos.	100%
Realizar un proyecto de Minería de Datos guiado por la metodología CRISP-DM y la documentación de cada una de las fases.	100%
Interpretar los resultados de la relación que existe entre las variables	100%

demográficas y la surrogada en el tiempo de vida de los pacientes con cáncer de pulmón.	

Se estima que fue cumplido el objetivo del negocio correspondiente al descubrimiento de patrones ocultos en los datos; que permitan predecir el tiempo de supervivencia de los pacientes incluidos en los Ensayos Clínicos de cáncer de pulmón, a partir de la variable surrogada “evaluación de la respuesta”, basado en las relaciones que se establecen entre las variables de control (sexo, edad, raza, estadio clínico, clasificación histológica, peso).

Aprobar modelos

Después de la evaluación del modelo con respecto a los criterios de éxito del negocio, los modelos generados que satisfacen los criterios seleccionados se convierten en modelos aprobados. Para esto se debe entender el resultado de la minería de datos para poder interpretarlos y comprobar que los mismos cumplen las metas iniciales del negocio, comprobar que estos resultados son novedosos y útiles.

De acuerdo a los resultados obtenidos en el proyecto para dar cumplimiento al objetivo del negocio; el modelo realizado es aprobado, dado que el mismo permitió encontrar patrones significativos en los datos para predecir el tiempo de supervivencia de un paciente con cáncer de pulmón. *Ver anexo 3*

5.2 Revisar el proceso

El modelo resultante en este punto parece ser satisfactorio ya que cumple con las necesidades del negocio. Aun así es apropiado ahora hacer una revisión más detallada del compromiso de minería de los datos en cada etapa del proceso para determinar si algún factor o tarea importante se excluyó de alguna manera. Esta tarea es en esencia igual que una Revisión de Control de Calidad. Debe señalarse las actividades en que se ha fallado y deben ser repetidas.

En el proyecto no se propone repetir ningún paso, ya que después de un nuevo análisis no se han encontrado fallas, ni se ha omitido ninguna variable que pudiera limitar el éxito de los resultados. Las propuestas de mejoras en los modelos se dejan para próximas iteraciones o proyectos.

5.3 Determinar los próximos pasos.

En esta tarea se decide como proceder según los resultados de la evaluación y la revisión del proceso. Aquí se decide si finalizará el proceso de minería y se pasará a la fase de despliegue o si es apropiado realizar otras iteraciones o realizar nuevos proyectos de minería. La salida debe mostrar las posibles acciones posteriores y las razones a favor y en contra de cada opción, fundamentando la decisión tomada de cómo proceder de acuerdo al análisis realizado.

De acuerdo a la revisión del proceso de Minería se decide finalizar el proyecto y pasar a la fase de despliegue. A continuación se proponen las acciones a realizar para posteriores iteraciones o desarrollos de proyectos de Minería.

Acción 1: Realizar los modelos de Minería de Datos utilizando el algoritmo de árboles de decisión J48 con un mejor refinamiento de los datos; para dicho refinamiento se incluirán otros datos pertenecientes a otros ensayos realizados en el CIM.

Acción 2: Utilizar como entrada otros atributos para encontrar otras relaciones entre los mismos. Esta actividad requiere poco esfuerzo. Puede encontrar patrones desconocidos y de utilidad para el CIM.

6. Despliegue

Esta tarea toma los resultados de la evaluación y concluye una estrategia para el despliegue de los resultados del negocio. Entre sus actividades se encuentran: resumen de los resultados desplegados, decidir para cada resultado diferente el conocimiento o la información proporcionado a sus usuarios, decidir para cada resultado obtenido del modelo cómo estos podrán ser utilizados dentro de los sistemas de la organización.

6.1 Producir el informe final

Al final del proyecto, el equipo que desarrollo el proyecto de minería de datos preparan un informe final. Depende del plan del despliegue, si este informe es solamente un resumen del proyecto y de sus experiencias o si este informe es una presentación final de los resultados de la minería incluyendo el despliegue.

Al final del proyecto el informe final reúne todos los detalles del mismo. Así como identificar los resultados obtenidos, el informe debe también describir el proceso, y hace cualquier recomendación para el trabajo futuro.

El presente documento se considera el informe final de la investigación.

6.2 Revisión del proyecto

Determina finalmente que fue positivo o negativo, qué hechos fueron bien ejecutados y qué necesidades se imponen para ser mejorado.

Documentación de las experiencias

Resume las experiencias importantes durante el desarrollo del proyecto. En proyectos ideales, la documentación de la experiencia cubre también cualquier informe que haya sido escrito por los miembros del proyecto individual durante las fases del proyecto y sus tareas.
Sus actividades incluyen entrevistas a las personas significativas implicada en el proyecto y las preguntas acerca de sus experiencias durante el desarrollo del mismo. Se debe conocer si los usuarios del negocio trabajan con los resultados del proyecto de minería, para verificar si estos están satisfechos y saber qué se pudo hacer mejor.

Durante el desarrollo del proyecto, la Minería de Datos permitió encontrar tendencias significativas o relevantes en la información almacenada; para a partir del comportamiento pasado y actual poder tomar decisiones sobre el comportamiento futuro.

La importancia y aplicaciones de los procesos de Búsqueda de Conocimiento en Bases de Datos están cobrando un auge cada vez mayor en el mundo empresarial. Las aplicaciones de estos procesos fueron expuestas con anterioridad en el documento.

La metodología CRISP-DM constituye una guía detallada paso a paso para realizar proyectos de KDD.

Presenta una precisa y sólida distribución de tareas de carácter general con sus resultados, así como una guía para su desarrollo. Muchas de las metodologías que podemos encontrar en la actualidad se basan en este estándar, la misma cuenta con mayor aceptación por parte de los desarrolladores de procesos de extracción de conocimientos a partir de datos.

Las herramientas para el desarrollo del proyecto son eficientes y pueden manipular datos complicados,

los resultados son muy fáciles de interpretar a través de los visores de resultados que contienen para cada algoritmo.

Los usuarios del negocio expusieron que el presente trabajo investigativo puede servirles de guía para posteriores proyectos de minería con el objetivo de descubrir nuevos patrones de conocimiento ocultos en los datos que se manejan en el CIM.

Conclusiones

Se concluye que:

- La metodología CRISP-DM constituye una guía valiosa para el desarrollo de proyectos de Descubrimiento de Conocimiento en Bases de Datos.
- La herramienta Weka es eficiente para realizar proyectos de minería de datos, ya que es un programa que implementa numerosos algoritmos de aprendizaje para realizar un exhaustivo análisis.
- Las técnicas clasificación resultaron útiles para obtener 12 reglas de clasificación que permitieron predecir el tiempo de vida de un paciente con cáncer de pulmón, así como las variables que influían en dicho tiempo de vida.

Conclusiones Generales

Al término de la presente investigación se concluye que:

- Se integraron los datos a partir de cada una de las tablas significativas teniendo en cuenta los atributos que daban cumplimiento al objetivo del negocio.
- Se construyó el campo tiempo_vida dado que constituye la variable objetivo para el proyecto de Minería de Datos.
- Se obtuvo un dataset con un total de 48 instancias a partir de transformar el conjunto de datos obtenidos en la fase de **Preparación de los Datos.**
- La clasificación utilizando el algoritmo de árboles de decisión J48, permitió obtener reglas que clasificaron correctamente al 87,5% de las instancias analizadas. Este resultado permitió dar cumplimiento al criterio de éxito planteado en la fase de **Comprensión de los Datos.**
- Durante el desarrollo de esta investigación se cumplieron satisfactoriamente los objetivos planteados en cada una de las etapas del proyecto.

Recomendaciones

1. Utilizar los resultados del presente trabajo en nuevos proyectos de Minería de Datos para el Centro de Inmunología Molecular en la búsqueda de nuevos patrones de conocimiento utilizando el resto de los Ensayos Clínicos de dicho centro con otros productos y otras indicaciones.

2. Fomentar el desarrollo de proyectos de Minería de Datos en la Universidad de las Ciencias Informáticas con otros proyectos.

Referencias Bibliográficas

[Acosta] Acosta Sánchez, Rolando; Rosete Suárez, Alejandro; Rodríguez Díaz, Alfredo; Brito Sarasa, Raycos. *Minería de Datos para la predicción de causas de diabetes. Preprocesado de datos.* Uciencia 2008. Instituto Superior Politécnico José Antonio Echeverría (Cujae).

[Bressán, 2003] Bressán, Griselda E. "Almacenes de datos y Minería de Datos", Trabajo monográfico de adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, julio 2003 Argentina.
Disponible en:
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm

[Calderón, 2005] Calderón Orozco, Daryna Marycruz, *"Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio "*, Tesis de grado para obtener el título de Ingeniero en Estadística Informática, Escuela Superior del Politécnico del Litoral Instituto de Ciencias Matemáticas, 2005 Guayaquil Ecuador. Disponible en: http://www.dspace.espol.edu.ec/bitstream/123456789/3983/1/6509.pdf

[Chapman, 2000] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. "CRISP-DM 1.0: Stepby-step data mining guide". USA: SPSS Inc., CRISP-DM Consortium, 2000.

[Cabrera, 2008] Dra. Niviola Cabrera Cruz, "*Retos y posibilidades de los Ensayos Clinicos Controlados para los pacientes*", Ministerio de Salud Pública de Cuba, junio 2008 Costa Rica.
Disponible en:
http://www.eventos.bvsalud.org/agendas/BVS-COR/public/documents/Niviola_RETOS%20_POSIBILIDADES_EC-150413.pdf

[Cano, 2005] José Ramón Cano, Francisco Herrera y Manuel Lozano; *"Extracción de modelos predictivos e interpretables en conjuntos de datos de tamaño grande mediante la selección de conjuntos de entrenamiento";* III Taller Nacional de Minería de Datos y Aprendizaje**;** Departamento de Informática y Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada y Jaén; 2005.

Disponible en: http:// www.lsi.us.es/redmidas/CEDI/papers/717.pdf

[CIM] Centro de Inmunología Molecular. Disponible en: http://www.cim.sld.cu/

[Daedalus] Aplicación de minería de datos para el diagnóstico de accidentes cerebrovasculares agudos (ACVAs). Daedalus. Sector Medicina.
Disponible en: http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/DAEDALUS-MD19-Accidentes_Cardiovasculares.pdf

[Facena, 2003] Facena-Unne; "Minería de Datos"; Teleprocesos y Sistemas Distribuidos; Licenciatura en Sistemas de Información; Octubre 2003. Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/SDataMining.pdf

[Fayyad, 1996] Fayyad, U. et al., "Advanced in Knowledge Discovery and Data Mining," MIT Press, MA, 1996.

[Fayyad_Shapiro, 1996]. Fayyad, D. M.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press,1996. Disponible en: http://elvex.ugr.es/etexts/spanish/kdd/KDD.html

[Gondar] Gondar N., José E. Introducción al Data Mining. José Huerta: Consultoría de Información. Disponible en: http://www.josebhuerta.com/datamining.htm#3

[Hernández, 2004] Hernández Orallo, J.; Ramírez Quintana, M. J.; Ferri Ramírez, C. Introducción a la minería de datos. Madrid, Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación: Ed. PEARSON EDUCACIÓN, S.A, 2004. ISBN 84-205-4091-9. Comandos de weka: http://weka.sourceforge.net/wekadoc/index.php/Category:Weka_3.5.1

[Hernández-Ferri, 2006] Hernández Orallo José, Ferri Ramírez César. Introducción al Weka. Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software. Universidad Politécnica de Valencia, marzo 2006.
Disponible en: http://users.dsic.upv.es/~jorallo/docent/doctorat/weka.pdf.

[Inteligencia_Negocios] *Entradas Etiquetadas Árboles de Decisión.* Inteligencia de Negocios. Consultor en Tecnologías de Información, con 10 años de experiencia. Monterrey, N.L., México. Disponible en: http://inteligencianegocios.wordpress.com/tag/arboles-de-decision/

[Macías, 2008] Macías Rodríguez, Miguel. *Técnicas de Minería de Datos para la Retención de clientes en el Sector Asegurador.* Trabajo presentado para el XV Premio de Investigación sobre Seguros y Fianzas 2008. 2008.
Disponible en: http://www.cnsf.gob.mx/Eventos/Premios/2008%20Seguros/ANIVDELAREV.pdf

[Martín] Martín Rodríguez, Diana; Socorro Llanes, Raisa; Wilford Rivera, Ingrid. Herramienta de Minería de Datos para usuarios no expertos basada en bibliotecas de Weka. Uciencia 2008. Instituto Superior Politécnico José Antonio Echeverría (Cujae).

[Molina, 2006] Molina López, J. M.; García Herrero, J. Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA. Madrid, Universidad Carlos III, 2006.

[Orallo] Hernández Orallo, José. *Parte III: Minería de Datos.* Departamento de Sistemas Informáticos y Computación Universidad Politécnica de Valencia.
Disponible en: http://users.dsic.upv.es/~jorallo/cursoDWDM/dwdm-III-1.pdf

[Orallo-Ramírez, 2004] Hernández Orallo, J.; Ramírez Quintana, M. J.; Ferri Ramírez, C. Introducción a la minería de datos. Madrid, Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación: Ed. Pearson Educación, S.A., 2004. Disponible en:
http://users.dsic.upv.es/~flip/LibroMD/

[Pinto, 2007] Pinto Rivera, Juan. *Data Mining y Data Warehouse.* Universidad de Santiago de Chile. Facultad de Administración y Economía. Departamento de Gestión y Política Públicas. 24 de mayo, 2007. Disponible en: http://www.gestionpublica.cl/biblioteca/documentos/55812datawarehouse.pdf

[Porta, 2005] Porta Zamorano, Jordi. Clasificación de patrones: Métodos supervisados. Escuela Politécnica Superior Universidad Autónoma de Madrid. Abril del 2005. Disponible en: http://www.iula.upf.edu/materials/050418porta_5.pdf

[Sánchez, 2006] Sánchez Tarragó, Dánel. *Pronóstico de supervivencia de infarto cerebral aterotrombótico usando aprendizaje automatizado.* VI Congreso Internacional de Informática en Salud. Departamento de Informática del Instituto Superior de Ciencias Médicas de Villa Clara. Noviembre, 2006.
Disponible en: http://www.informatica2007.sld.cu/Members/danel/pronostico-de-supervivencia-de-infarto-cerebral-aterotrombotico-usando-aprendizaje-atomatizado/2006-11-15.5808751092

[Servente, 2002] Servente M. "Algoritmos TDIDT aplicados a la Minería de Datos Inteligente", Prof. Dr. Ramón García Martínez (Dir.). Universidad de Buenos Aires, Facultad de Ingeniería. Tesis de Grado en Ingeniería Informática, 2002. Disponible en: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf

[Tejeda, 2004] Tejeda Gamero, Eduardo José. *Análisis de las Redes Neuronales Fuzzy ART como Motores de Inferencia.* EPIS - UNSA – Arequipa. Mayo, 2004. Disponible en: http://www.ucsp.edu.pe/~etejada/docs/tesis-eduardo.pdf

[Vallejos, 2006] Vallejos, Sofia J. "Mineria de Datos", Trabajo de Adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, 2006 Argentina.
Disponible en:
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf

[Vilches, 2007] Vilches González, Erika; Escobar Broitman, Iván A. *Minería de Datos.* Septiembre 2007. Disponible en: http://www.erikavilches.com/km/mineria_datos.pdf

[Weka] *Minería de Datos*.
Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/03-04/18.mem.pdf

[Witten, 2000]Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. EE.UU, San Diego: Morgan Kaufmann Publishers, 2000.

Bibliografía

- ✓ "Centro de Aplicaciones de Tecnologías de Avanzada". CENATAV. Disponible en: http://www.cenatav.co.cu/es/. Fecha de Acceso: 20 de noviembre del 2008.

- ✓ "Centro de Estudios de Reconocimiento de Patrones y Minería de Datos". CERPAMID. Disponible en: http://www.cerpamid.co.cu/. Fecha de acceso: 20 de noviembre del 2008.

- ✓ "Instituto de Cibernética, Matemática y Física ".ICIMAF. Disponible en: http://www.icmf.inf.cu/.Fecha de acceso: 20 de noviembre del 2008.

- ✓ "Instituto Superior Politécnico José Antonio Echeverría".CUJAE. Disponible en: http://www.cujae.edu.cu/index.html.Fecha de acceso: 20 de noviembre del 2008.

- ✓ Bressán, Griselda E. "Almacenes de datos y Minería de Datos", Trabajo monográfico de adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, julio 2003 Argentina. Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm

- ✓ Calderón Orozco, Daryna Marycruz, *"Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio "*, Tesis de grado para obtener el título de Ingeniero en Estadística Informática, Escuela Superior del Politécnico del Litoral Instituto de Ciencias Matemáticas, 2005 Guayaquil Ecuador. Disponible en: http://www.dspace.espol.edu.ec/bitstream/123456789/3983/1/6509.pdf

- ✓ Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. "CRISP-DM 1.0: Stepby-step data mining guide". USA: SPSS Inc., CRISP-DM Consortium, 2000.

- ✓ Dra. Niviola Cabrera Cruz, "*Retos y posibilidades de los Ensayos Clinicos Controlados para los pacientes*", Ministerio de Salud Pública de Cuba, junio 2008 Costa Rica.

Disponible en: http://www.eventos.bvsalud.org/agendas/BVS-COR/public/documents/Niviola_RETOS%20_POSIBILIDADES_EC-150413.pdf

- ✓ José Ramón Cano, Francisco Herrera y Manuel Lozano; *"Extracción de modelos predictivos e interpretables en conjuntos de datos de tamaño grande mediante la selección de conjuntos de entrenamiento";* III Taller Nacional de Minería de Datos y Aprendizaje**;** Departamento de Informática y Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada y Jaén; 2005.

- ✓ "Proceso de Minería de Datos". DAEDALUS - Data, Decisions and Language. Disponible en: http://www.daedalus.es/AreasMDFases-E.php. Fecha de acceso: 23 de enero del 2009.

- ✓ "Data Mining & Knowledge Discovery in Databases (KDD)". IEEE Transactions on Knowledge and Data Engineering, December 1996. Disponible en: http://elvex.ugr.es/etexts/spanish/kdd/KDD.html. Fecha de Acceso: 20 de febrero del 2009.

- ✓ "Proyecto de Inteligencia Artificial". Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Escuela de Ingeniería de Sistemas. Disponible en: http://www.cruzrojaguayas.org/inteligencia/. Fecha de Acceso: 20 de febrero del 2009.

- ✓ "Algoritmo de Asociación de Microsoft". MSDN Library. Disponible en: http://msdn2.microsoft.com/es-es/library/ms174916.aspx. Fecha de acceso: 10 de marzo del 2009.

- ✓ "Algoritmo de clústeres de Microsoft". MSDN Library. Disponible en: http://msdn2.microsoft.com/es-es/library/ms174879.aspx. Fecha de acceso:.10 de marzo del 2009

- ✓ "Algoritmo de Árboles de decisión de Microsoft". MSDN Library. http://msdn2.microsoft.com/es-es/library/ms175312.aspx. Fecha de acceso: 15 de marzo del 2009.

- ✓ Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR),Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler),Colin Shearer (SPSS) y Rüdiger Wirth (DaimlerChrysler).

"Metodología CRISP-DM 1.0 para Minería de Datos". Dataprix. Disponible en: http://www.dataprix.com/modelo_crisp-dm. Fecha de acceso: 25 de marzo del 2009.

- ✓ "Que es un ensayo clínico". Centro de Información Cardiovascular. Texas Haert Institute. Ultima modificación enero del 2009. Disponible en: http://www.texasheartinstitute.org/HIC/Topics_Esp/FAQ/clinical_trials_span.cfm. Fecha de acceso: 16 de abril del 2009.

- ✓ "Acerca de los Ensayos Clínicos: Información procedente del Instituto Nacional del Cáncer". University of Virginia. Disponible en: http://www.healthsystem.virginia.edu/uvahealth/adult_breast_sp/clinical.cfm. Fecha de acceso: 16 de abril del 2009.

- ✓ "Ensayos Clínicos". National Coalition for Cancer Survivorship. Disponible en: http://www.canceradvocacy.org/espanol/resources/trials.html. Fecha de acceso: 16 de abril del 2009.

- ✓ "Caracterización del Centro de Inmunología Molecular (CIM)". Eumed. Biblioteca virtual. Disponible en: http://www.eumed.net/libros/2009a/514/Caracterizacion%20del%20Centro%20de%20Inmunologia%20Molecular.htm. Fecha de acceso: 16 de abril del 2009.

- ✓ "Registra Cuba vacuna contra cáncer de pulmón". Juventud Técnica Digital. Disponible en: http://www.juventudtecnica.cu/Juventud%20T/ciencias/2008/paginas/cancer.html. Fecha de acceso: 16 de abril del 2009.

- ✓ "Que son las bases de datos?".Maestros del web. Publicado 26 de octubre del 2007. Disponible en: http://www.maestrosdelweb.com/principiantes/%C2%BFque-son-las-bases-de-datos/. Fecha de acceso: 22 de abril del 2009.

- ✓ "Naive Bayes método supervisado de clasificación". IEspaña. Disponible en: http://supervisadaextraccionrecuperacioninformacion.iespana.es/bayes.html

- ✓ "Tutorial de redes neuronales universidad tecnológica de Pereira Facultad de Ingeniería Eléctrica". Capítulo 2.Redes Comnpetitivas.LVQ.
  Disponible en: http://ohm.utp.edu.co/neuronales/Capitulo2/Competitivas/LVQ.htm

- ✓ Maenzas, Matías- Mansilla, Natalia."Descubrimiento de Conocimiento a partir de Datos". Trabajo Especial: Aplicación de KDD. Construcción de un Filtro Anti-Spam. Disponible en:http://www.exa.unicen.edu.ar/catedras/dbdiscov/mansilla-maenza.pdf

- ✓ Porta Zamorano, Jordi. "Clasificación de patrones: Métodos supervisados". Escuela Politécnica Superior Universidad Autónoma de Madrid. Departamento de Lingüística Computacional, Real Academia Española. Abril 2005.
  Disponible en: http://www.iula.upf.edu/materials/050418porta_4.pdf

- ✓ Martín Rodríguez, Diana; Socorro Llanes, Raisa; Wilford Rivera, Ingrid. Herramienta de Minería de Datos para usuarios no expertos basada en bibliotecas de Weka. Uciencia 2008. Instituto Superior Politécnico José Antonio Echeverría (Cujae).

- ✓ Acosta Sánchez, Rolando; Rosete Suárez, Alejandro; Rodríguez Díaz, Alfredo; Brito Sarasa, Raycos. *Minería de Datos para la predicción de causas de diabetes. Preprocesado de datos.* Uciencia 2008. Instituto Superior Politécnico José Antonio Echeverría (Cujae).

- ✓ Sánchez Tarragó, Dánel. *Pronóstico de supervivencia de infarto cerebral aterotrombótico usando aprendizaje automatizado.* VI Congreso Internacional de Informática en Salud. Departamento de Informática del Instituto Superior de Ciencias Médicas de Villa Clara. Noviembre, 2006.
  Disponible en: http://www.informatica2007.sld.cu/Members/danel/pronostico-de-supervivencia-de-infarto-cerebral-aterotrombotico-usando-aprendizaje-atomatizado/2006-11-15.5808751092

- ✓ Servente M. “Algoritmos TDIDT aplicados a la Minería de Datos Inteligente”, Prof. Dr. Ramón García Martínez (Dir.). Universidad de Buenos Aires, Facultad de Ingeniería. Tesis de Grado en Ingeniería Informática, 2002. Disponible en: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf

- ✓ Tejeda Gamero, Eduardo José. *Análisis de las Redes Neuronales Fuzzy ART como Motores de Inferencia.* EPIS - UNSA – Arequipa. Mayo, 2004. Disponible en: http://www.ucsp.edu.pe/~etejada/docs/tesis-eduardo.pdf

- ✓ Vallejos, Sofia J. "Mineria de Datos", Trabajo de Adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, 2006 Argentina.
  Disponible en:
  http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf

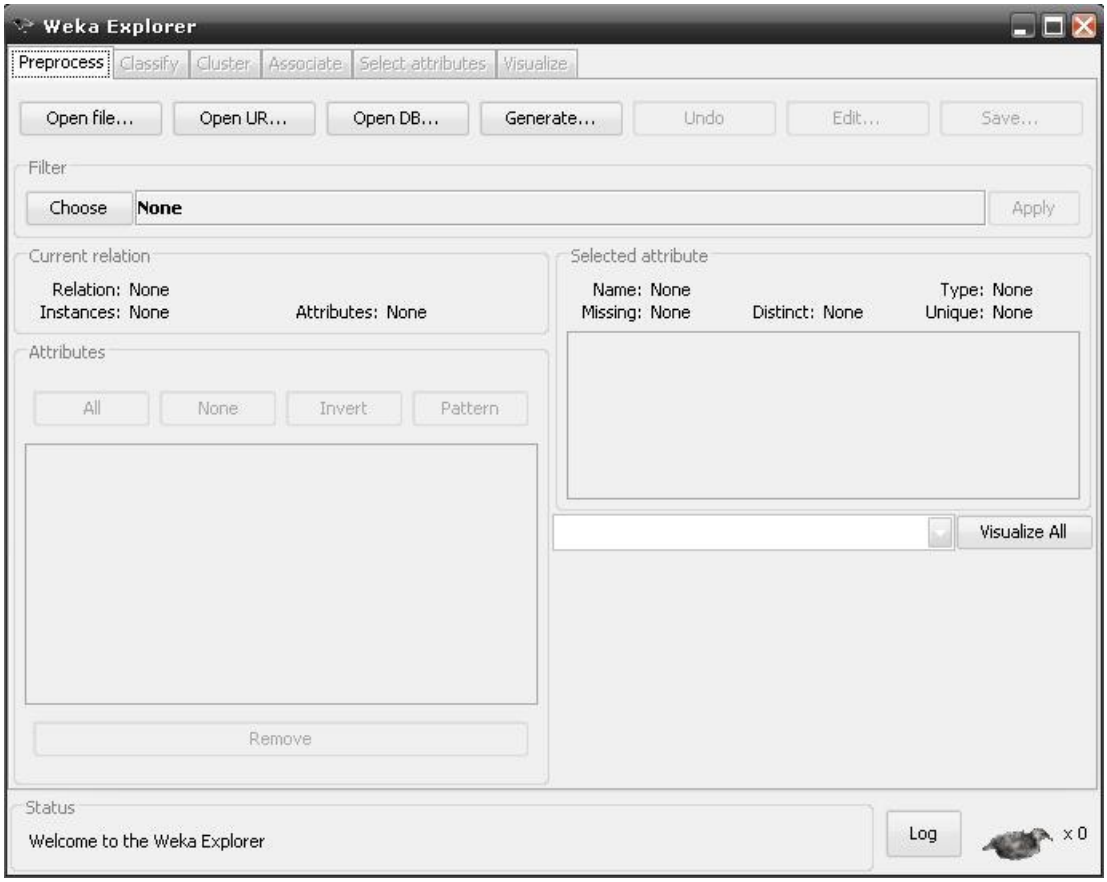
- ✓ Vilches González, Erika; Escobar Broitman, Iván A. *Minería de Datos.* Septiembre 2007. Disponible en: http://www.erikavilches.com/km/mineria_datos.pdf

- ✓ *Minería de Datos*.
  Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/03-04/18.mem.pdf

- ✓ Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. EE.UU, San Diego: Morgan Kaufmann Publishers, 2000.

Anexo 1: Herramienta utilizada para la Minería de Datos. Weka

Anexo 2: Entorno de trabajo Explorer de la herramienta Weka.



Anexo 3: Opinión de los clientes sobre el trabajo realizado.

Centro de Inmunología Molecular

Opinión del Cliente: MSc. Carmen E. Viada González y Lic. Patricia Lorenzo-Luaces

Tesis para optar por el título de Ingeniero en Ciencias Informáticas: "Proceso de análisis y gestión del conocimiento a partir de los datos obtenidos en la conducción de los de Ensayos Clínicos del CIM, aplicando técnicas de Minería de Datos.".

Autores: Danay Perera Baró y Enrique Ramírez Alonso

Este proyecto está siendo desarrollado por la Facultad 6 en colaboración con el CIM, con el objetivo de mejorar la calidad de la gestión de los datos referente a los Ensayos Clínicos que se realizan en dicho centro dado que actualmente el centro no cuenta con un sistema que le permita gestionar toda esta información y se recogen los datos en Bases de Datos Microsoft Access o con el programa estadístico SPSS, realizándose una entrada doble de los mismos.

Dado el gran volumen de datos acumulado, y por no contar con herramientas que permitan identificar patrones de comportamiento y extraer conocimiento oculto en los datos almacenados para apoyar sus decisiones surge la necesidad de aplicar Minería de Datos (MD) a dicho proyecto, por lo que el problema a resolver es apoyar el proceso de gestión y toma de decisiones de los administrativos y especialistas del CIM a partir del estudio de los datos recopilados en los ensayos clínicos realizados en el centro, a través de la identificación de patrones de comportamiento útiles presentes en estos ensayos.

Este problema presenta como objeto de estudio la gestión del conocimiento a través de técnicas de MD teniendo como campo de acción las técnicas de MD para la gestión del conocimiento en proyectos de investigación biomédica. Por lo que el objetivo general consiste en desarrollar un proceso para el

análisis y búsqueda de patrones y comportamientos presentes en los datos recopilados en la conducción de los EC realizados en el CIM a los pacientes con cáncer de pulmón que recibieron EGF, mediante el empleo de técnicas de MD. Luego extender esta aplicación al resto de los EC del CIM con otros productos y en otras indicaciones.

Los autores de este trabajo han demostrado profesionalismo en la concepción de esta investigación, en la ejecución del mismo así como en la elaboración de su trabajo de tesis de grado. Por otro lado, han mostrado una alta independencia y buen desempeño en el desarrollo de este trabajo.

Queremos hacer notar que los autores han adquirido gran experiencia en las investigaciones clínicas y en particular en los ensayos clínicos lo cual puso de manifiesto en la culminación de su tesis en tiempo, lo cual prueba su alta capacidad intelectual en el planteamiento del problema, análisis de los datos con técnicas como la Minería de Datos, confección del trabajo de tesis y la defensa de los resultados. Además tiene excelentes cualidades humanas mostradas en su relación con sus dos clientes lo que prueba su receptividad y criterio propio ante las sugerencias hechas durante el desarrollo de esta Tesis. También mantuvieron una magnifica comunicación con el equipo de investigación demostrando buen dominio del idioma inglés y alta preparación. Por su profesionalismo, independencia y capacidad de trabajo proponemos la calificación de excelente.

MSc. Carmen E. Viada González
MSc. Bioestadística
Centro de Inmunología Molecular

Glosario de Términos

Ensayos Clínicos: es un estudio que permite a los médicos determinar si un nuevo tratamiento, medicamento o dispositivo contribuirá a prevenir, detectar o tratar una enfermedad.

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

KDD [Knowledge Discovery in Databases]: su término en inglés significa Extracción de Conocimientos en Bases de Datos.

Aprendizaje Automático: es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan crear programas capaces de generalizar comportamientos a partir de una información no estructurada.

Inteligencia Artificial: es la ciencia que enfoca su estudio a lograr la comprensión de entidades inteligentes.

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

Data_Warehouses: es una colección de datos en la cual se encuentra integrada la información de la Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales.

CENATAV: Sus siglas significan Centro de Aplicaciones de Tecnologías de Avanzada y es un centro orientado a las investigaciones teóricas y aplicadas en el área del Reconocimiento de Patrones y la Minería de Datos.

CERPAMID: Sus siglas significan Centro de Estudios de Reconocimiento de Patrones y Minería de Datos y 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.

ICIMAF: Sus siglas significan Instituto de Cibernética, Matemática y Física, es un Instituto de Investigaciones de alto nivel teóricas y aplicadas en ramas de la Cibernética, la Matemática y la Física.

CUJAE: Sus siglas significan Instituto Superior Politécnico José Antonio Echeverría, es una universidad comprometida con su patria que se propone mediante la formación integral y continua de profesionales, la universalización de la enseñanza, la actividad científico-técnica y la extensión universitaria.

CIM: Sus siglas significan Centro de Inmunología Molecular y tiene como principal misión obtener y producir nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades.

Redes Neuronales: son utilizadas para la predicción, la minería de datos, el reconocimiento de patrones y los sistemas de control adaptativo. Constituyen una parte muy importante en el estudio y desarrollo de la inteligencia artificial y el de la vida artificial.

Métodos Heurísticos: Un método heurístico es un procedimiento para resolver un problema de optimización mediante una aproximación intuitiva, en la que la naturaleza intrínseca del problema se usa de manera inteligente para obtener una buena solución.

MDP: Minería de Datos Inductiva.

MDDC: Minería de Datos para el descubrimiento de Conocimiento

Árboles de Decisión: Representan reglas donde atributos independientes determinan los valores finales. En estos árboles cada nodo representa una propiedad que puede tomar diversos valores, cada uno de los cuales genera una rama.

Naives bayes: clasificador probabilístico en el proceso de MD. Establecer regiones de decisión mucho más complejas que las de dos semiplanos, como lo hace el Perceptrón de un solo nivel.

LVQ: Esta red es un híbrido que emplea tanto aprendizaje no supervisado, como aprendizaje supervisado para clasificación de patrones.

Vecinos Próximos: Método de aprendizaje basado en ejemplos**.**

Weka: Herramienta de MD.

CRISP-DM: Metodología de MD.

CRD: Cuaderno de Recogida de Datos.

SPSS (Statistical Package for the Social Sciences): es un programa estadístico informático muy usado en las ciencias sociales y las empresas de investigación.

SPSS: Es una compañía dedicada a la creación de software para la minería de datos.