Universidad de las Ciencias Informáticas

"Facultad 2"

Título: Sistema de Gestión por Competencias (SGC)

"Módulo de Tratamiento de Datos"

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

Autores:
Leodán Suárez Izquierdo.
Jose Andrés Esquivel Pérez.

Tutor:
Lic. Darían Horacio Grass Boada.

"Año del 50 Aniversario del Triunfo de la Revolución"

Ciudad de la Habana, Cuba. Junio de 2009.

Declaración de autoría.

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

Para que así conste firmo la presente a los ___ días del mes de ________ del año ______.

Leodán Suárez Izquierdo	José Andrés Esquivel Pérez
Autor	Autor

Darían Horacio Grass Boada

Tutor

Opinión del tutor sobre el trabajo de diploma.

Dedicatoria.

De Leodán:

A mis padres Amelia y Guillermo, por todo el amor, dedicación y empeño que han mostrado en todos estos años. Por ser un persevero ejemplo que siempre he de seguir. Por enseñarme y solidificar los valores necesarios para crecer espiritual y profesionalmente. Por sus consejos colmados de sabiduría y precisión. Por su ilimitada capacidad para entenderme. Quiero que sepan que son las personas que más quiero en el mundo y espero que estén orgullosos de mí. Gracias por existir.

A mi hermano Linray por ser siempre mi ejemplo y fuente de inspiración a quien le debo todo lo que soy.

A mi abuela Hortensia por mimarme siempre, y ser más que una madre.

A mis abuelos que donde quiera que estén se sientan orgullosos de su nieto.

A esa personita María Isabel, por su confianza, dedicación y por encontrar en ella lo que muchos buscamos y pocos alcanzamos: el amor.

A mis suegros María y Moisés por todo el apoyo que me han brindado desde que los conocí.

A familiares y amigos en general que estuvieron al tanto de mis estudios y aportaron lo mejor de sí, para que pudiera culminarlos exitosamente.

A todos muchas gracias.

De José Andrés:

A todas las personas que quiero y respeto, mi familia, mis amigos, mi novia, a todos los que de un modo u otro me han ayudado en mi formación como hombre y profesional, quiero dedicarles este momento, decirles que sin la ayuda de cada uno de ellos nada de estos hubiera sido posible a todos.

Agradecimientos.

Compartidos:

Queremos agradecer a la Revolución Cubana por su grandeza, especialmente a nuestro comandante Fidel Castro, por ser el creador de la Universidad de las Ciencias Informáticas y paradigma de las nuevas generaciones y de su pueblo.

A nuestro tutor Darian por su apoyo, preocupación y aporte en este trabajo y por enseñarnos que el conocimiento no tiene valor calculable.

A nuestros amigos y compañeros de estudio durante los 5 años y en especial a los miembros del proyecto Daniel, Jose Rafael, Dayron y Albin.

A todos aquellos que de una manera u otra fueron y se sintieron partícipes de este trabajo, nuestros agradecimientos.

De Leodán:

Quisiera agradecer a mis padres por su apoyo incondicional, por su confianza y cariño en cada momento.

*A mi hermano por transmitirme su conocimiento y darme el cari*ño y *apoyo necesario para mi formación.*

A Norma por brindarme su apoyo y preocupación en todos estos años.

A mi querida novia por todo su amor, comprensión y por su compañía en los buenos y malos momentos.

A mis tíos Xiomara, Horacio, Silvio Luis y Susana que de alguna forma u otra siempre se han mantenido preocupados por mi superación.

A Enrique, a Kikito, a Solange, a Solangita y Moña por brindarme su apoyo y acogerme como parte de su familia todos estos años.

A todos mis familiares en general para los cuales me he convertido en un profesional.

En fin a todos muchas gracias.

De José Andrés:

Agradecer siempre a toda mi familia por el amor que me han dado durante todos estos años, que sepan que siempre los tengo muy presente a todos. A los que están aquí y a los que están un poco más lejos, a todos, gracias por la confianza y el apoyo que me dan.

A mi mamá, mis abuelos, mi papá, mis primas: Cristina, Patricia, Galays, primos: Jorgito, Maikel, a Randy y Danny quienes también considero primos, a mi hermano, mis tíos: Juanito y Carlos, a mis tías: Vivian, Fifina. A todos muchas gracias.

A mis amigos dentro y fuera de la UCI, Daniel, Michel, Jose Carlos, Boris, Lincheta, Jose Rafael, Waldemar, a todos, gracias por estar presente siempre.

A mi novia por todo el cariño y las horas.

A mi tutor Darian, al profe Fernando, a la profe Ayme, al profe Juan Antonio, a la profe Pura, que tanto confiaron, contribuyeron y apoyaron a lo largo de mi carrera, gracias por ayudarme a formarme como profesional.

A todas mis amistades y compañeros que me ayudaron durante estos años y a todos aquellos que de un modo u otro contribuyeron a mi formación como profesional y como persona, gracias.

Resumen.

La información generada producto de la interacción de los estudiantes con los diferentes métodos de evaluación docentes para la medición de los conocimientos adquiridos, es una rica fuente de conocimiento. El análisis de esos datos puede ser de gran utilidad, no solo para la extracción de información efectiva para la toma de decisiones, sino para crear planes y estrategias basados en los patrones de comportamientos presentes en los individuos evaluados.

Este trabajo esta dedicado a la elaboración una herramienta desarrollada con el objetivo de apoyar los análisis de estos históricos acumulados en variadas fuentes. Esta herramienta aplica técnicas de minería de datos, para facilitar el análisis e interpretación de un cúmulo de datos provenientes, ya sea, de distintos gestores de base de datos o de archivos de texto, mostrándose visualmente de una forma comprensible para el usuario. Con el objetivo fundamental de apoyar el proceso de toma de decisiones.

Índice General.

Índice de Tablas.

Capítulo I

Capítulo II

Capítulo III

Capítulo IV

Introducción.

En la actualidad una de las problemáticas fundamentales de diversas Instituciones, está dada en el aumento de los volúmenes de datos que se procesan. De acuerdo a Tang y MacLennan [1], el problema fundamental de las instituciones radica en que se han vuelto “ricas en datos y pobres en conocimiento”, motivado esto fundamentalmente por las dificultades para analizar muchos datos utilizando solamente métodos y herramientas tradicionales. Determinándose de este modo, un aumento de los estudios para buscar técnicas que posibiliten extraer conocimientos a partir de grandes volúmenes de datos.

Una de las tendencias más comunes en la rama educacional, que se manifiesta en el presente, son los sistemas basados en Web, tecnología cada vez más utilizada para la educación a distancia, debido en gran medida a la facilidad de utilización y disponibilidad de las herramientas para navegar por el Web.

El gran problema de la mayoría de estos sistemas de aprendizaje web, es que proporcionan no más que una serie de páginas estáticas, entre las cuales los estudiantes navegan. En este sentido tiene mucho que ver que el *análisis del proceso de aprendizaje* (tendencias, tipos de aprendizaje predominantes, asesoramiento del aprendizaje), regulador por demás de la estrategia de aprendizaje y todos los elementos que se le vinculan (estructura de planes de estudios, sitios de aprendizaje, otros), se hace casi de manera intuitiva, basado mayoritariamente en la experiencia y observaciones que los responsables de esta tarea realizan de dicho proceso. Lo cual ocasiona que este criterio, aunque acertado en un gran por ciento, sea hasta cierto punto impreciso, pues carece de un fundamento tangible.

Si bien es cierto esto, también lo es que producto del funcionamiento de estos sistemas se recogen grandes cantidades de datos, que derivan casi siempre en una acumulación, que puede ser vista como la memoria histórica del proceso. Haciéndose evidente que se constituyen reseñas de todas las situaciones que hasta el momento se han producido. Por lo que el procesado e interpretación de estos datos, puede derivar en modelos que posteriormente analizados, constituyen una fuente de conocimientos tangibles que brinda un soporte para la toma de decisiones en este ámbito. Lo cual evidencia la necesidad de métodos y herramientas adecuadas para analizar todos estos datos.

En la Facultad 2 se lleva a cabo la concepción del Sistema de Gestión por Competencias (SGC), con el fin de realizar la acreditación por competencias[1]. El mismo pertenece a la familia de los sistemas de enseñanza basados en la web y tiene como objetivo apoyar la docencia y la producción mediante la acreditación por competencias. Pero sucede que concebido de manera tradicional no esta exento de los problemas anteriormente expuestos, lo cual es sin duda alguna **una situación problémica**, que se resume en la insuficiencia para procesar datos y obtener conocimientos a partir de estos, que sirvan como soporte al proceso de toma de decisiones en esta línea estratégica de la Facultad.

De ahí que el **problema científico** a dar solución sea: ¿Cómo procesar y analizar datos, para la obtención de conocimientos, que apoyen la toma decisiones dentro del proceso educativo?

Teniéndose como **objeto de estudio** las herramientas para el procesado, análisis y evaluación de los datos, para la obtención de reglas y patrones, que deriven en conocimientos. Correspondiendo el **campo de acción** con aquellas que se relacionan directamente al proceso educativo.

La investigación persigue como **objetivo general** desarrollar una herramienta que permita el procesado, análisis y evaluación de datos, con la finalidad de obtener conocimientos, que sirvan de soporte al proceso de toma de decisiones en el ámbito docente-productivo de la Facultad.

Como **objetivos específicos** se pretende:

- Analizar el proceso de extracción de conocimientos a partir de grandes fuentes de datos (Minería de Datos[2]) en entornos educativos y definir etapas, tareas, técnicas y algoritmos, que lo componen.
- Diseñar un modelo de solución de la aplicación para el procesado, análisis y evaluación de datos.
- Implementar el modelo creado para la aplicación de procesado, análisis y evaluación de datos.

Para dar cumplimiento a los objetivos propuestos, se han perfilado las siguientes **tareas de investigación**:

[1] Evaluación de la capacidad de desempeño (conocimientos, habilidades, actitudes), del personal ante determinadas actividades de acuerdo a los estándares y calidad exigidos por el mundo productivo.

[2] Termino asociado al Proceso de Descubrimiento de Conocimiento a partir de Bases de Datos (del ingles Knowledge Discovery from Databases, KDD)

1. Desarrollar un estudio investigativo del proceso general de Minería de Datos, etapas o fases, modelos metodologías para su desarrollo, tareas, técnicas y algoritmos.
2. Realizar un estudio de los principales trabajos que vinculen la Minería de Datos con en el entorno educativo.
3. Desarrollar un estudio de las principales herramientas para Minería de Datos.
4. Identificar los elementos o requisitos fundamentales dentro del proceso de Minería de Datos a ser implementados
5. Definir componentes reutilizables de las herramientas que realizan el proceso de Minería de Datos.
6. Realizar un diseño del software.
7. Implementar del modelo de solución obtenido en el diseño.

El presente trabajo se compone de cinco capítulos, los cuales versarán acerca del estudio realizado y la descripción de la solución propuesta.

Capítulo 1: En este se realiza un Estado del Arte sobre la Minería de Datos como técnica elegida para dar solución al problema, así como las diversas herramientas existentes y las experiencias más ligadas al desarrollo de esta técnica en entornos educativos.

Capítulo 2: Se hace énfasis en una propuesta de solución a partir de un análisis crítico de la situación actual y de los flujos de trabajo involucrados en el campo de acción.

Capítulo 3: Se hace referencia a los artefactos generados en las fases de exploración y planificación del proyecto.

Capítulo 4: Se exponen las fases de implementación y pruebas para el sistema.

Capítulo 5: Se procede con un estudio de factibilidad del sistema.

Capítulo I. Fundamentación Teórica.

1.1. Introducción.

En este capítulo abordaremos fundamentalmente los temas relacionados con la Minería de Datos, su correspondencia con otras disciplinas, así como las tareas a realizar para llevar a cabo el cumplimiento de los objetivos propuestos. También se tratará sobre su relación con la esfera educacional y las principales aplicaciones de las técnicas de minería de datos en dicha rama, las técnicas antes mencionadas son mostradas y explicadas a la par con las herramientas de apoyo para llevar a cabo el proceso de minería. Se trata además sobre la metodología de desarrollo, tecnologías y herramientas utilizadas para el desarrollo de la propuesta.

1.2. Concepto de Minería de Datos.

Con el transcurso de los años, el desarrollo de nuevas tecnologías ha hecho posible el almacenamiento de información de forma creciente. El volumen de datos recogidos ha permitido satisfacer las necesidades diarias de las organizaciones, y consecuentemente estos volúmenes de datos han superado las capacidades humanas para analizar y transformar la información presente en estos, para así convertirlos en conocimiento útil que apoye la toma de decisiones. Esta misma necesidad ha motivado el empleo de técnicas y herramientas que posibiliten la extracción de información de dichos datos.

Partiendo siempre de los datos como base fundamental para realizar cualquier análisis, desde el momento que a estos se les atribuye algún significado especial, pasan entonces a convertirse en información. Y cuando se elabora o se encuentra un modelo, de forma tal que permite establecer una comparación entre la información y obtener un valor agregado que se traduzca en acciones efectivas, entonces se está en presencia de un conocimiento.

Varias son las definiciones empleadas por distintos autores para definir este proceso al que se conoce como Minería de Datos. Entre ellas:

- “…el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos” [2].

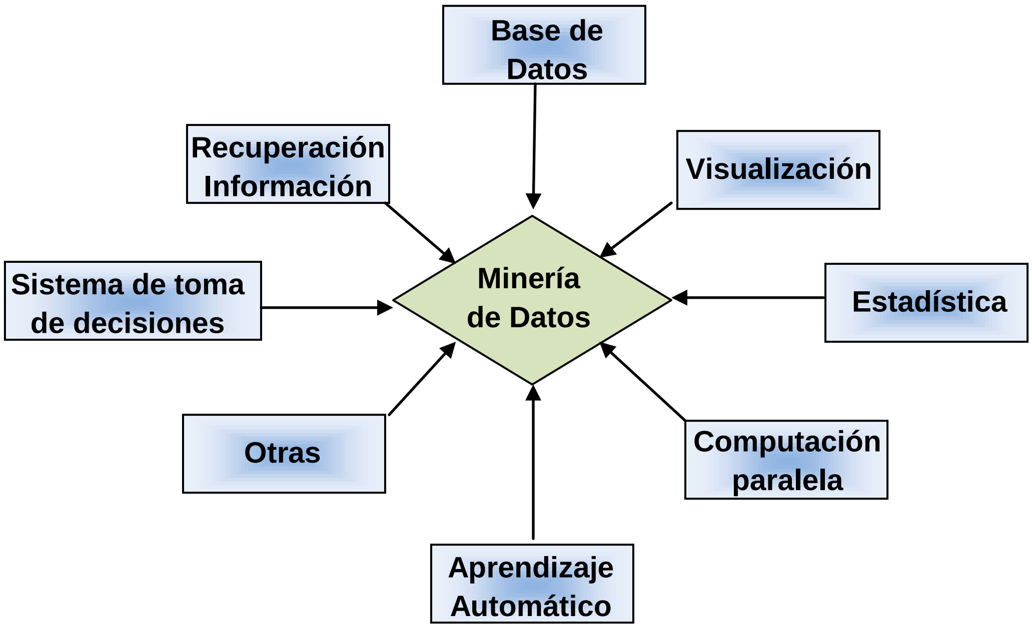
- "…es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos" [3].
- "…proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, Data Warehouses, o cualquier otro medio de almacenamiento de información" [4].
- La minería de datos posibilita que "los datos pasen de ser un producto (el resultado histórico de los sistemas de información) a ser una materia prima que hay que explotar para obtener el verdadero producto elaborado (el conocimiento)" [5].
- "…término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos" [6].

Por lo cual en sentido general y apoyado en las definiciones anteriores de distintos investigadores, que describen la tarea fundamental de la minería de datos como el descubrimiento de conocimiento (reglas, patrones) a partir de grandes volúmenes de datos, y apoyados en diversas técnicas y herramientas, de modo que su uso permita servir de soporte en el proceso de toma de decisiones estratégicas, fundamentadas en una base solida y que reportan algún tipo de beneficio a las organizaciones.

Podríamos concluir que, "dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil." [5].

1.3. Relación de la Minería de Datos con otras Disciplinas.

La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas. [5].



Podemos destacar como disciplinas más influyentes las siguientes:

- **las bases de datos:** conceptos como los almacenes de datos y el procesamiento analítico en línea (OLAP) tienen una gran relación con la minería de datos, aunque en este último caso no se trata de obtener informes avanzados a base de agregar los datos de cierta manera compleja pero predefinida (como incluyen muchas herramientas de business intelligence, presentes en sistemas de gestión de bases de datos comerciales), sino de extraer conocimiento novedoso y comprensible. [5].
- **la recuperación de información (information retrieval, IR):** consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas (recientemente digitales) y en la búsqueda por Internet. Una tarea típica es encontrar documentos a partir de palabras claves, lo cual puede verse como un proceso de clasificación de los documentos en función de estas palabras clave. Para ello se usan medidas de similitud entre los documentos y la consulta. Muchas de estas medidas se han empleado en aplicaciones más generales de minería de datos. [5].

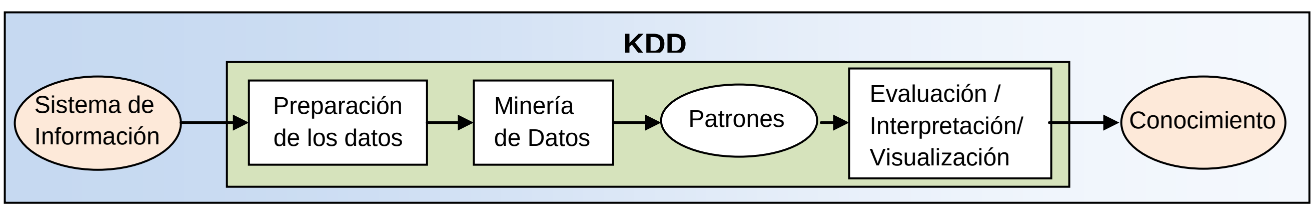
- **la estadística:** esta disciplina ha proporcionado muchos de los conceptos, algoritmos y técnicas que se utilizan en minería de datos, como por ejemplo, la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, la modelización paramétrica y no paramétrica, las técnicas bayesianas, y un largo etcétera. De hecho, algunos paquetes de análisis estadístico se comercializan como herramientas de minería de datos. [5].
- **el aprendizaje automático:** ésta es el área de la inteligencia artificial que se ocupa de desarrollar algoritmos (y programas) capaces de aprender, y constituye, junto con la estadística, el corazón del análisis inteligente de los datos. Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema. [5].
- **los sistemas para la toma de decisión:** son herramientas y sistemas informatizados que asisten a los directivos en la resolución de problemas y en la toma de decisiones. El objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico (por ejemplo en medicina). [5].
- **la computación paralela y distribuida:** actualmente, muchos sistemas de bases de datos comerciales incluyen tecnologías de procesamiento paralelo, distribuido o de computación en grid. En estos sistemas el coste computacional de las tareas más complejas de minería de datos se reparte entre diferentes procesadores o computadores. Una de las principales ventajas del procesamiento paralelo es precisamente la escalabilidad de los algoritmos, lo que lo hace idóneo para estas aplicaciones. [5].
- **otras disciplinas:** dependiendo del tipo de datos a ser minados o del tipo de aplicación, la minería de datos usa también técnicas de otras disciplinas como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc. [5].

1.4. El proceso de descubrimiento de conocimiento en bases de datos.

En muchas ocasiones se asocian distintos términos como sinónimos de minería de datos, como es el caso de análisis (inteligente) de datos, en el cual se hace más referencia a las técnicas de análisis estadísticos. Existe otro término con el cual se asocia de manera más usual a la Minería de Datos este es el caso de el "descubrimiento de conocimiento en bases de datos" (Knowledge Discovery in Databases, KDD). Este es el más frecuentemente usado y aunque a veces se usen como sinónimos presentan claras diferencias entre ellos, ya que el KDD es un proceso que consta de fases, mientras que la minería de datos es solo una de dichas fases.

Se define a KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" [2]. Esta definición nos muestra resumidas las principales características que deben estar presentes en dicho proceso:

- **válido:** hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención. [5].
- **novedoso:** que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario. [5].
- **potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario. [5].
- **comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad). [5].



Podríamos resumir entonces de acuerdo que “los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso.” [5].

El KDD es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos. [5].

1.5. Tipos de Modelos de Minería de Datos.

La Minería de Datos tiene como su objetivo analizar datos, con el fin de extraer conocimientos. El cual puede manifestarse en forma de relaciones, patrones o reglas que se infieren de los datos y que resultan novedosas, o resultan ser una descripción más concisa de los mismos. Este resultado constituye un modelo de los datos analizados. Existen formas diferentes de representar modelos y en dependencia de cada una de ellas se determina un tipo de técnica que permite inferirlos.

A partir de las características que engloban, los modelos pueden clasificarse en dos tipos: predictivos y descriptivos [5], [7].

Los *modelos predictivos* pretenden estimar valores futuros o desconocidos de variables de interés, usando otras variables o campos de la base de datos, a las que nos referiremos como variables independientes o predictivas. [5].

Los *modelos descriptivos*, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. [5].

1.6. Tareas de la Minería de Datos.

El proceso de minería de datos requiere en un principio, establecer los objetivos para el análisis de los datos disponibles [5], [7], de ahí que para su cumplimiento sean necesarias varios tipos de tareas, entre las que sobresalen por su uso [8]:

- La Clasificación.
- La Regresión (Predicción o Estimación).
- El Agrupamiento (Clustering o Segmentación).
- La Asociación.
- La Correlación.

Estas tareas se agrupan de acuerdo a los modelos que satisfacen del siguiente modo (tabla 1.1):

Tabla 1. 1: Modelos de la Minería de Datos

Modelo	Predictivo	Descriptivo
	Clasificación	Agrupamientos
	Regresión	Asociación
		Correlación

1.6.1. Clasificación.

El objetivo de esta tarea es organizar los datos en distintas clases, a partir de crear un modelo basado en su distribución. De esta forma, puede ser clasificada con precisión nueva información. Más concretamente, un algoritmo de clasificación persigue maximizar la razón de precisión de las nuevas instancias, la cual se calcula como el cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas) [5].

1.6.2. Regresión.

La regresión [6], también conocida como predicción o estimación, es otra tarea predictiva de gran importancia. Su meta es encontrar el *valor numérico* de una variable objetivo para objetos desconocidos.

1.6.3. Agrupamiento.

Al agrupamiento también se le suele llamar segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos, esta fragmentación puede presentar un notable interés, dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma seudo particularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado [9].

1.6.4. Asociación.

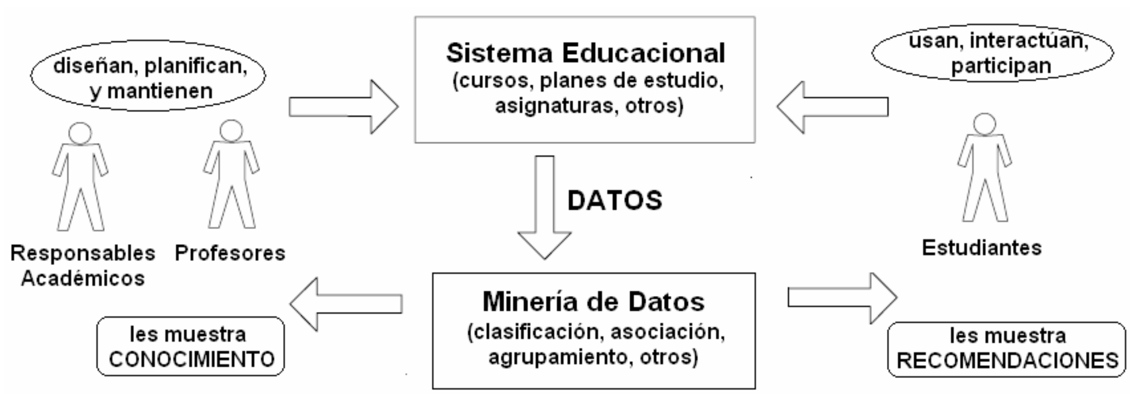
La asociación [10] es una tarea descriptiva, muy similar a la correlación, que tiene como objetivo identificar relaciones no explícitas entre atributos nominales, y se emplea frecuentemente para reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos.

1.6.5. Correlación.

La correlación se usa para examinar el grado de similitud de los valores de dos variables numéricas [11].

1.7. Minería de Datos relacionada a entornos educativos.

En la esfera educacional, eje central en torno a que gira esta investigación, las técnicas de minería de datos han resultado ser muy útiles para establecer una base de conocimientos que sirva de soporte a la toma de decisiones en este ámbito. Este proceso es un ciclo iterativo en el cual se establecen hipótesis, se prueba y se refina. En la figura se observa como los profesores y responsables académicos, se encargan de diseñar y planificar cursos, planes de estudio, asignaturas, conformando de esta manera el Sistema Educacional; los estudiantes por su parte interactúan con este y producto de ellos generan un volumen de datos, a los cuales se les aplica algoritmos de minería de datos, obteniéndose reglas y patrones, es decir conocimiento que soporta distintas decisiones que toman los responsables antes mencionados. Traduciéndose en recomendaciones para los alumnos.



Las principales aplicaciones de las técnicas de minería de datos en educación, son como sistemas de personalización, sistemas recomendadores, sistemas de modificación, sistemas de detección de irregularidades, debido a sus capacidades para : el descubrimiento de patrones de navegación regulares e irregulares, realización de clasificaciones de alumnos y de los contenidos, construcción adaptativa de planes de enseñanza, descubrimiento de relaciones entre actividades, diagnóstico incremental de los estudiantes, etc.

La aplicación de técnicas de minería de datos en educación se puede ver desde dos puntos de vista u orientaciones distintas:

- **Orientado hacia los autores.** Con el objetivo de ayudar a los profesores y/o autores de los sistemas de e-learning para que puedan mejorar el funcionamiento o rendimiento de estos sistemas a partir de la información de utilización de los alumnos. Sus principales aplicaciones son: obtener una mayor realimentación de la enseñanza, conocer más sobre como los estudiantes aprenden en el web, evaluar a los estudiantes por sus patrones de navegación, reestructurar los contenidos el sitio web para personalizar los cursos, clasificar a los estudiantes en grupos, etc. [12].
- **Orientado hacia los estudiantes.** Con el objetivo de ayudar o realizar recomendaciones a los alumnos durante su interacción con el sistema de e-learning para poder mejorar su aprendizaje. Sus principales aplicaciones son: sugerir buenas experiencias de aprendizaje a los estudiantes, adaptación del curso según el progreso del aprendiz, ayudar a los estudiantes dando sugerencias y atajos, recomendar caminos más cortos y personalizados, etc. [12]

De acuerdo con [12] en estado del arte sobre las técnicas de minería de datos empleadas en entornos virtuales de aprendizaje:

“la minería de datos aplicada a los sistemas de e-learning son: clasificación y agrupamiento, descubrimiento de reglas de asociación, y análisis de secuencias.”

Aunque es de destacar, que en la mayoría de los casos no solo se usa una de estas si no una combinación de varias.

1.7.1. Clasificación y Agrupamiento o Clustering.

La aplicación de las técnicas de *clasificación* y *agrupamiento* o *clustering* en sistemas de aprendizaje son usadas comúnmente para “agrupar a los usuarios por su comportamiento de navegación, agrupar a las páginas por su contenido, tipo o acceso y agrupar los comportamientos de navegación similares. A continuación describimos algunos trabajos de aplicación de minería de datos en e-learning.” [12].

Un empleo del agrupamiento es el realizado por Elena Gaudioso y Luis Talavera que analizan los datos obtenidos de cursos basados en sistemas e-learning y utilizan técnicas de clustering similares al modelo probabilístico de Naive Bayes para descubrir patrones que reflejan comportamientos de los usuarios. Su objetivo es utilizan la minería de datos para dar soporte a la tutoría en comunidades de aprendizaje virtual.

1.7.2. Reglas de Asociación.

Las *reglas de asociación* como uso más común en la educación “permiten descubrir relaciones o asociaciones entre distintas páginas Web visitadas”. [12].

Un trabajo que utiliza técnicas de minería de reglas de asociación y filtrado colaborativo, es el realizado por Feng-Hsu Wang para descubrir patrones de navegación útiles y proponer un modelo de navegación. El modelo de navegación consiste en dos tipos de relaciones: relaciones de asociación y relaciones de secuencia entre documentos.

1.7.3. Análisis de Secuencias.

El *análisis de secuencias* es usado para “analizar secuencias de páginas visitadas durante una sesión o en distintas sesiones de un mismo usuario.” [12].

El análisis de patrones de navegación en entornos de aprendizaje basado en web es utilizado por Karin Becker y otros dentro de una herramienta de minería de utilización web para el análisis de patrones y pre-procesado de datos de utilización de entornos de aprendizaje basados en web. Las técnicas de descubrimiento de patrones utilizadas son asociación y secuencia, y la secuencia de patrones describen accesos a páginas relaciones en un orden específico.

1.7.4. Minería de Datos en Cuba relacionadas a Entornos Educacionales.

En Cuba existen pocos trabajos de Minería de Datos relacionadas a entornos Educativos. Uno de ellos es el realizado en la Universidad de las Ciencias Informáticas (UCI) [13], cuyos principales resultados son el uso de técnicas de agrupamiento y asociación para encontrar patrones entre los resultados académico del primer año, vinculados al origen social de de estos. El otro, fue desarrollado en el Instituto Superior Politécnico José Antonio Echeverría (CUJAE) [14], con el empleo de técnicas de clasificación para predecir el resultado de los estudiantes en su primer año, en dependencia de sus características de ingreso y la especialidad que cursaban.

1.8. Herramientas utilizadas en el proceso de Minería de Datos

En la actualidad existen diferentes herramientas para el apoyo a al proceso de KDD. A continuación se refieren algunas de ellas en la tabla 1.2. [15].

Tabla 1. 2: Herramientas en la Minería de Datos.

Herramientas Propietarias	SPSS Clementine
	SAS Enterprise Miner
Herramientas Libres	YALE Rapid Miner
	WEKA

1.8.1. SPSS Clementine

SPSS Clementine [16] es una herramienta visual comercializada por SPSS que constituye uno de los sistemas más populares en el mercado, pues posibilita de forma rápida desarrollar y desplegar modelos que apoyen la toma de decisiones. Entre sus características más significativas se destaca el hecho de que a diferencia de otras herramientas que se centran en el modelado, ella apoya el ciclo completo de KDD y esta diseñada bajo la metodología CRISP-DM [17].

La herramienta permite el uso de técnicas de aprendizaje tales como: redes neuronales, árboles de decisión, agrupamiento, reglas de asociación, regresión lineal y regresión logística, entre otras. Otra de sus prestaciones consiste en un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso; y que comprende gráficos estadísticos en 3D y animados; así como visualizadores y navegadores.

1.8.2. SAS Enterprise Miner

SAS Enterprise Miner [18] es una herramienta comercial proporcionada por SAS, que perfila el proceso completo de minería de datos bajo la metodología SEMMA [19]. Entre sus características más significativas se encuentra el hecho de que posee una arquitectura distribuida y una potente interfaz gráfica de usuario, que permite acceder a todas las funcionalidades que el sistema brinda.

La herramienta integra tareas de minería con almacenes de datos e incluye técnicas para ayudar al pre-procesado de datos. Además implementa algoritmos que proveen modelos predictivos y descriptivos, tales como árboles de decisión, redes neuronales, asociación, agrupamiento, entre otros. Posee además un potente visualizador gráfico para representar los resultados mediante gráficos en dos o tres dimensiones; así como un generador automático de reportes que resume los resultados en un informe HTML.

1.8.3. YALE o Rapid Miner

YALE (Yet Another Learning Environment) [20] es una herramienta creada en la Universidad de Dortmund para el descubrimiento del conocimiento y la minería de datos. Es un entorno con muchos algoritmos de aprendizaje y otras utilidades añadidas, está desarrollada sobre el lenguaje java y funciona en los sistemas operativos más conocidos, constituyendo un software de código abierto y de libre distribución, además, se retroalimenta de las librerías de funciones de WEKA en su entorno de aprendizaje.

Incluye características como las de implicar nuevos formatos de entrada de datos con operadores para Microsoft Excel y SPSS. Desde la perspectiva de la visualización ofrece representaciones de datos en dispersión en 2D y 3D; coordenadas paralelas y grandes posibilidades de transformar las visualizaciones de los datos.

1.8.4. WEKA

WEKA (Waikato Environment for Knowledge Analysis) [21] es una de las aplicaciones de minería más populares, desarrollada por un equipo de investigadores de la Universidad de Waikato (Nueva Zelanda). Constituye un entorno de experimentación de análisis de información, formado por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, clasificación, asociación, y visualización de datos.

Entre sus principales características se encuentra el poseer una interfaz gráfica de usuario compuesta de cuatro entornos que permiten diferentes funcionalidades y formas de análisis. Una de las ventajas fundamentales de esta herramienta es que su desarrollo sobre el lenguaje java la hace multiplataforma. Además, el hecho de ser de código abierto unido a su prestigio, hace que se encuentre en constante evolución por parte de la comunidad internacional.

1.8.5. Valoración de las herramientas y propuesta.

A pesar de contar en el mercado con varias herramientas que apoyen el proceso de KDD, no existe ninguna con especialización en entornos educacionales, además de las herramientas que pertenecen al software libre y hacia las cuales va la inclinación de las políticas de la Universidad de las Ciencias Informáticas y el país, podemos señalar que si bien presentan altas prestaciones y gran eficiencia, exigen un conocimiento técnico por parte de los usuarios, es decir se encuentran orientada más bien a usuarios técnicos.

Por esto se ha optado por el desarrollo de una herramienta que permita más facilidades para usuarios sin tantos conocimientos técnicos de Minería de Datos, que además brinde prestaciones adecuadas, que se especialice en las tareas más utilizadas en los procesos de Minería de Datos en Entornos Educativos. Valorando siempre que debe incluir los elementos que resultan claves en herramientas de este tipo, obtenidos del estudio los cuales se pueden sintetizar en la representación de los conocimientos en diferentes formas, buscando una mejor interpretación de los mismos y algoritmos robustos que aseguren un correcto funcionamiento.

1.9. Metodología de Desarrollo.

Para emprender el desarrollo del proyecto se decidió utilizar XP (Programación Extrema, en inglés Extreme Programming), motivado ello porque los elementos que aportan se adaptan en gran medida a las condiciones de trabajo que se imponen para este.

La programación extrema es una metodología de desarrollo ligera (o ágil) basada en una serie de valores y de prácticas de buenas maneras que persigue el objetivo de aumentar la productividad a la hora de desarrollar programas.

Este modelo de programación se basa en una serie de metodologías de desarrollo de software en la que se da prioridad a los trabajos que dan un resultado directo y que reducen la burocracia que hay al rededor de la programación.

Una de las características principales de este método de programación, es que sus ingredientes son conocidos desde el principio de la informática. Los autores de XP han seleccionado aquellos que han considerado mejores y han profundizado en sus relaciones y en como se refuerzan los unos con los otros. El resultado de esta selección ha sido esta metodología única y compacta. Por esto, aunque no está basada en principios nuevos, sí que el resultado es una nueva manera de ver el desarrollo de software.

El objetivo que se perseguía en el momento de crear esta metodología era la búsqueda de un método que hiciera que los desarrollos fueran más sencillos. Aplicando el sentido común.

Entre las razones que motivaron el uso de esta metodología tenemos las siguientes:

- Es un proyecto pequeño. Dicha metodología esta concebida para ser usada dentro de proyectos pequeños.

- No existe un contrato previo especificando tiempo, recursos y alcance. No se dispone de un contrato con un presupuesto ni un alcance previamente definidos, puesto que es un proyecto para el uso interno de la empresa y será llevado a cabo por programadores pertenecientes a la misma.

- Poca disponibilidad de personal. El sistema será realizado por un equipo de 6 personas, teniéndose varios módulos a implementar, por lo que no se hace posible la existencia de muchos roles ni la especialización en un rol. Uno de los principios básicos de XP es la programación en equipos pequeños (2 a 12 personas) con pocos roles, pudiendo los miembros del equipo intercambiar responsabilidades en un momento determinado.

- Cambio frecuente de los requisitos. El sistema debe cambiar y ampliar sus funcionalidades de forma que sea capaz de adaptarse a cada nueva situación. Uno de los principios básicos de XP es que el cambio frecuente de los requerimientos es algo normal en el proceso de desarrollo. Esta metodología se adapta perfectamente a los proyectos cuyos requerimientos cambian a menudo.

- El cliente forma parte del equipo de desarrollo. Mediante la aplicación de XP se puede lograr una retroalimentación mayor y lograr un producto que satisfaga sus necesidades.

- El riesgo de desarrollo es elevado debido al corto tiempo de entrega planteado y a los continuos cambios de requerimientos. XP está diseñada a mitigar los riesgos en proyectos con estas características.

- Propiedad colectiva del código. XP plantea que todos los programadores pueden realizar cambios en cualquier parte del código en cualquier momento. En el proceso de desarrollo con que cuenta la empresa esta es una práctica común.

- XP enfatiza la comunicación de los programadores a través del código, utilizando líneas directivas para la codificación que están bien establecidas. Desde sus comienzos la empresa cuenta con una línea directiva para la codificación.

1.10. Plataforma de desarrollo.

Como plataforma de desarrollo se optó por el uso de J2EE (Java 2 Enterprise Edition, Java 2 edición empresarial). La cual ofrece muy buenas perspectivas en la implementación de software empresarial, para aquellos sistemas informáticos que requieran basar su arquitectura en productos de software libre. J2EE ofrece, entre otras, las siguientes ventajas:

- Soporte para múltiples sistemas operativos: al ser una plataforma Java, es posible desarrollar arquitecturas basadas en J2EE usando cualquier sistema operativo donde pueda estarse ejecutando una máquina virtual de Java, teniendo la gran ventaja de una independencia total de la arquitectura de hardware.
- Organismo de control: J2EE está controlada por un organismo formado por más de 400 empresas. Entre esas empresas se encuentran muchas de las más importantes del mundo informático, tales como Sun Microsystems, IBM, Oracle, BEA, HP, AOL, etc.
- Competitividad: muchas empresas crean soluciones basadas en J2EE que ofrecen características tales como rendimiento y precio muy diferentes. De este modo, se ha desarrollado a un nivel exponencial la plataforma y los clientes tienen la posibilidad de escoger entre una gran cantidad de opciones.

- Madurez: creada en el año 1997, J2EE ya tiene varios años de vida y una amplia cantidad de proyectos importantes a sus espaldas.
- Soluciones libres: sobre la plataforma J2EE es posible crear arquitecturas basadas por completo en productos de software libre. No solo eso, sino que los arquitectos de software disponen de muchas soluciones libres para cada una de las partes de su arquitectura.

1.11. Herramienta de Desarrollo

Para la implementación del sistema se optó el uso de NetBeans 6.5 debido al número de prestaciones que brinda y su especialización para ambientes de escritorio.

Es un IDE[3] es una herramienta de desarrollo Java, escrita puramente sobre la base de la tecnología Java, de modo que puede ejecutarse en cualquier ambiente que ejecute Java. NetBeans es un producto de código abierto, con todos los beneficios del software disponible en forma gratuita, el cual ha sido examinado por una comunidad de desarrolladores.

Que proporciona varias ventajas como es el caso de la facilidad de uso durante todo el ciclo de desarrollo y el soporte para lenguajes de modelado, lo que aumenta la productividad. Estas razones motivaron su uso.

1.12. Librerías Utilizadas.

- **IText** es una biblioteca Open Source para la creación y manipulación de archivos de tipo PDF, RTF, y HTML en Java. Distribuida bajo la Mozilla Public License con la **LGPL**[4] como licencia alternativa. El mismo documento puede ser exportado en múltiples formatos o múltiples instancias del mismo formato.

- **JFreeChart** es una biblioteca Open Source para la creación de gráficas complejas de manera sencilla en el lenguaje Java.

[3] Ambiente de Desarrollo Integrado (en ingles, Integrated Development Environment)

[4] Licencia Pública General Reducida del GNU.

Soporta varios tipos de gráficos:

- Gráficos X-Y (línea, dispersión)
- Gráficos de Pastel.
- Gráficos Gantt.
- Gráficos de Barras (vertical y horizontal, apilados e independientes).

- **XStream** es una librería del tipo Open Source de JAVA para la serialización de objetos hacia XML y la descerialización de XML hacia objetos. Tiene muchas opciones a la hora de convertir nuestro XML a una clase de java (convertidores, persistencia, json, alias, anotaciones, etc.). XStream hace uso de la API Reflection de java para hacer el mapeo del XML, así que nuestras clases tendrán que llevar los campos con sus respectivos set/get.
- **Hibernate** es una herramienta de Mapeo objeto-relacional para la plataforma Java (y disponible también para .NET con el nombre NHibernate) que facilita el mapeo de atributos entre una base de datos relacional tradicional y el modelo de objetos de una aplicación, mediante archivos declarativos (XML) que permiten establecer estas relaciones. Es software libre, distribuido bajo términos de la licencia GNU LGPL.
- **Weka** (Waikato Environment for Knowledge Analysis) - Entorno para Análisis del Conocimiento de la Universidad de Waikato) es un conocido software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. WEKA es un software libre distribuido bajo licencia GNU-GPL. Contiene un conjunto de algoritmos para análisis de datos y modelado predictivo, soporta varias tareas estándar de minería de datos, especialmente, preprocesado de datos, clustering, clasificación, regresión, visualización, y selección.

1.13. Conclusiones del Capítulo.

En el presente capítulo se realizó un estudio acerca de la Minería de Datos, sus tareas, vínculos al entorno educativo, principales herramientas y demás. Con el objetivo de arribar a conclusiones acerca del estado de las herramientas de este tipo de labor, en el área educativa, con lo cual se pretendió un entendimiento más profundo del problema y del estado de solución del mismo a nivel mundial. Se comentó acerca de las tecnologías y metodologías a emplear para el desarrollo.

Capítulo II. Características del Sistema.

2.1. Introducción.

En el presente capítulo se tiene como objetivo realizar una valoración de las principales características del producto a desarrollar, complementando las necesidades que dieron origen al mismo. Se pretende además hacer un análisis crítico de la situación actual y hacer una propuesta para dar solución al problema.

2.2. Flujo de procesos vinculados al campo de acción.

En la actualidad el proceso de toma de decisiones en el área académica dentro de la Facultad, tiene sus bases en la experiencia, informes generales y apreciaciones del proceso que hacen los responsables de esta actividad (figura No. 2.1), es decir la fundamentación que respalda dichas decisiones es en cierto grado imprecisa, aunque no dejan de ser acertadas en la mayoría de las ocasiones.

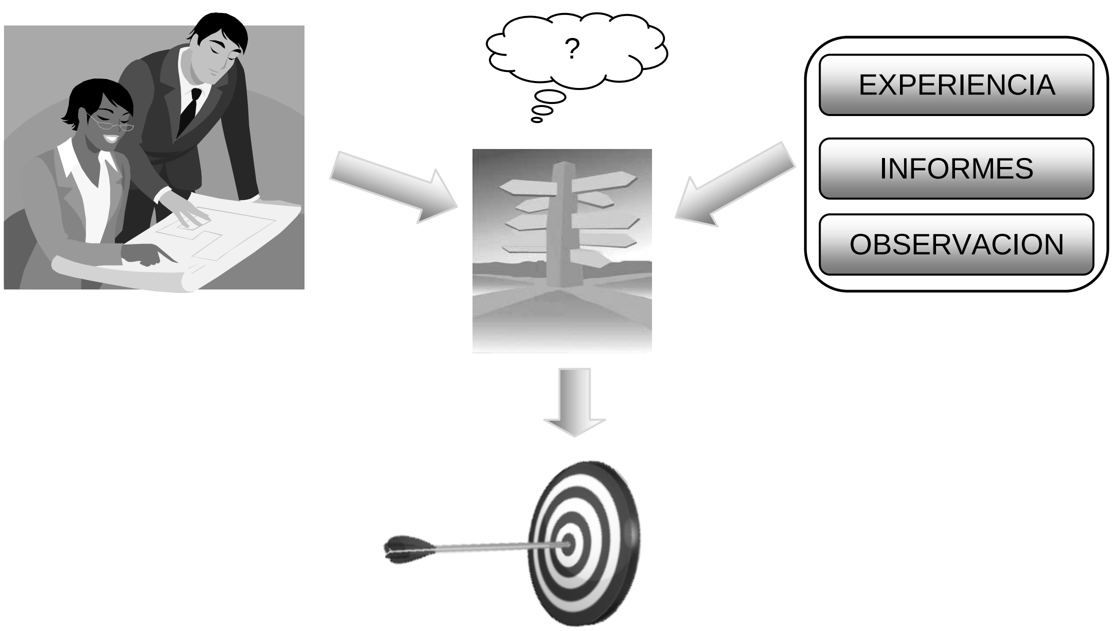


Figura. 2.1: Flujo de procesos para la toma de desiciones.

2.3. Análisis crítico del funcionamiento actual.

Esta situación se produce fundamentalmente motivada por la insuficiencia para analizar los grandes volúmenes de datos con los que se cuenta, utilizando métodos y herramientas tradicionales. Es decir en esta situación los encargados de tomar decisiones, utilizan el método tradicional de convertir los datos en conocimiento, que consiste en un análisis e interpretación realizado de forma manual. Lo cual sin duda resulta lento, caro y altamente subjetivo, en muchas ocasiones hasta impracticable cuando el volumen de crece exponencialmente, desbordando la capacidad humana de comprenderlos sin ayuda de herramientas potentes. Consecuentemente, muchas decisiones importantes se realizan, no sobre la base de la gran cantidad de datos disponibles, sino siguiendo la propia intuición del responsable.

Lo cual evidencia que no se hace uso óptimo de los recursos con que se cuentan, tales como el histórico que se va almacenando en las distintas bases de datos y otros archivos, que constituye por demás la memoria de la organización, la cual es útil para explicar el pasado, entender el presente y predecir de cierto modo la información futura. Es decir los datos de los que se disponen son una mina de conocimientos esperando a ser explotada.

2.4. Objeto de Automatización.

En el ciclo de desarrollo del problema analizado existen elementos que constituyen objetos de automatización, los cuales resultaran en una mejora práctica, revertida en mayor fiabilidad, rendimiento y facilidad de trabajo.

En este sentido se ha determinado la automatizar un proceso considerado de vital importancia para dar soporte a la toma de decisiones en determinadas líneas estratégicas de la facultad. El mismo es el proceso de extracción de reglas y patrones, que deriven en conocimiento útil y comprensible, previamente desconocido, es decir encontrar modelos inteligibles a partir de los datos. Lo cual ayudará a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la facultad.

Presentándose dos retos, por un lado el manejo de grandes volúmenes de datos procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

2.5. Características de la propuesta.

Teniendo en cuenta que en muchos casos la utilidad del conocimiento está íntimamente relacionada con la comprensibilidad del modelo inferido y conociendo que, generalmente, el usuario final no tiene por qué ser un experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados. El presente trabajo propone la implementación de un sistema capaz de extraer reglas y patrones, para la confección de modelos que sirvan como soporte al proceso de toma de decisiones.

El mismo tendrá la capacidad de realizar tareas tanto de tipo predictivo, dígase clasificación, como descriptivo, el caso de agrupamiento y extracción de reglas de asociación.

Un acercamiento inicial a los elementos que se contienen en la concepción de la solución se ilustra en la figura No.2.2.

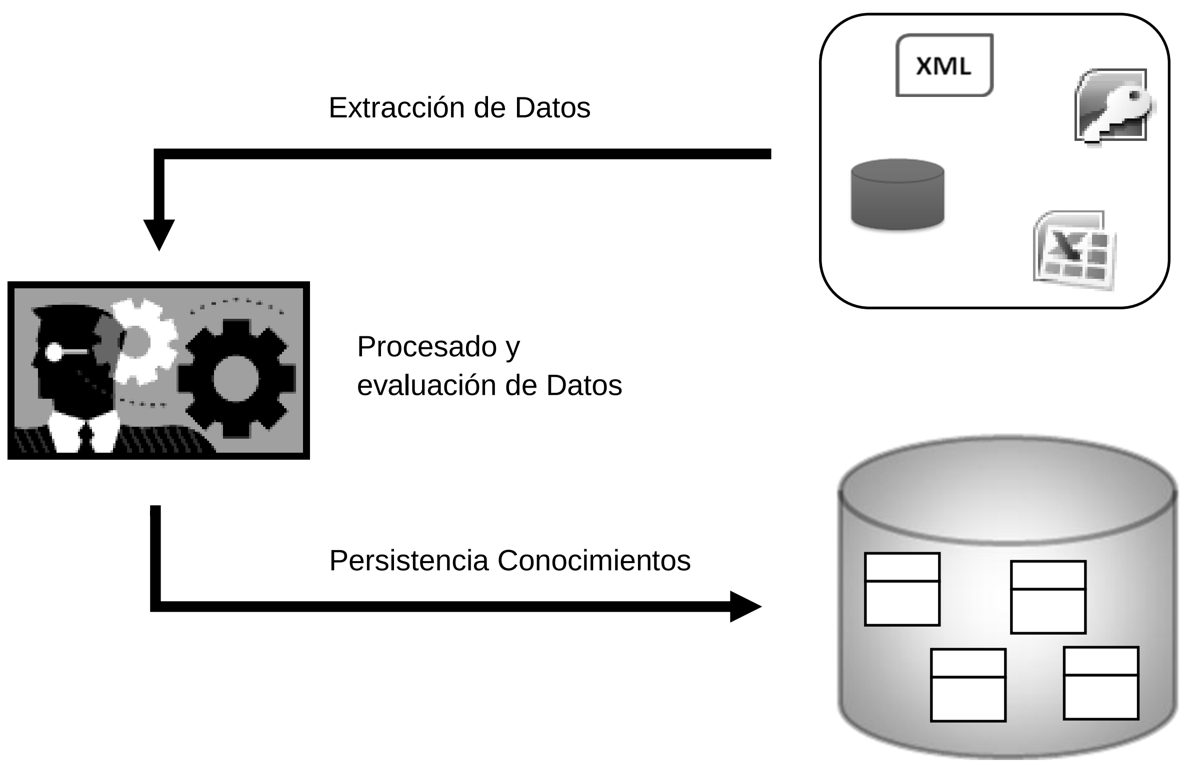


Figura. 2.2: Propuesta del Sistema.

Como se aprecia el motor para el procesado de datos y extracción de conocimientos (sistema del cual se ocupa este trabajo), concibe la extracción de datos para sus análisis de diversas fuentes y la persistencia de los resultados del proceso de análisis en un repositorio central. Una estructura más detallada de Los módulos que concebirá el sistema se ilustra en la figura 2.3.

<table>
<tr><td colspan="3">CARGA</td></tr>
<tr><td colspan="3">PREPROCESADO</td></tr>
<tr><td colspan="3">ANÁLISIS</td></tr>
<tr><td>AGRUPAMIENTO</td><td>ASOCIACIÓN</td><td>CLASIFICACIÓN</td></tr>
<tr><td colspan="3">PERSISTENCIA</td></tr>
</table>

Figura. 2.3: Módulos del sistema.

El módulo de Carga será el encargado de la extracción de los elementos para analizar, teniendo en cuanta que el origen de estos puede ser variado, es decir pueden provenir de Bases de Datos (de diferentes tipos), de archivos de texto y XML, la solución contendrá soporte en este sentido.

El de Preprocesado está encargado de la visualización inicial de las características de los datos, lo cual permite una mejor comprensión de los posibles estudios a realizar, así como de la aplicación de diferentes filtros que permitan modelar los datos hasta alcanzar el formato deseado.

El Análisis es el núcleo del proceso, encargado de aplicar las distintas técnicas concebidas para la obtención de conocimientos (agrupamiento, asociación y clasificación). Motivado por el hecho de que el usuario puede no ser experto en el proceso de Extracción de Conocimientos y teniéndose en cuenta que la información descubierta sea lo más comprensible por los usuarios, se hará uso de representaciones gráficas, conversión los patrones a lenguaje natural y otras técnicas de visualización de los datos, que proporcionaran un entendimiento superior de los resultados.

El módulo de Persistencia se encarga de que los resultados logrados no se pierdan y almacena los mismos en un repositorio fijado por el usuario, para que puedan ser consultados en otro momento que se requiera su análisis.

2.6. Personas vinculadas con el sistema.

Las personas relacionadas o vinculadas al sistema (tabla 2.1) se definen, como aquellas que obtienen cierto resultado de valor, producto de uno o varios procesos dentro del sistema. O aquellas que participan activamente dentro de dichos procesos.

Tabla 2. 1: Personas relacionadas al sistema.

Personas	Desempeño
Productor	Persona encargada de realizar los procesos de extracción de conocimientos. El cual trabaja en todo el flujo de extracción, produciendo las reglas y patrones que serán interpretadas posteriormente como conocimiento.

2.7. Conclusiones del Capítulo.

En el capítulo se abordó acerca de la solución propuesta, luego de realizar un análisis crítico del estado actual de la situación que da origen al trabajo y del flujo de proceso que interviene en el mismo. Se describieron además, los principales elementos que intervienen en la solución propuesta.

Capítulo III: Exploración y Planificación.

3.1. Introducción

En este capítulo se hace referencia a las fases de Exploración y Planificación de la metodología XP[5], detallando cada uno de los artefactos que surgen como resultado de estas dos fases.

3.2. Fase de Exploración

La metodología para desarrollo de software Extreme Programming, comienza su ciclo de desarrollo durante la fase de Exploración, que corresponde con la fase donde se identifican las historias de usuarios, que no serán más que los elementos rectores dentro del desarrollo del software. Además en la misma se explora la familiarización del equipo de trabajo con las tecnologías y herramientas que se emplearan a lo largo del desarrollo.

3.3. Historias de usuarios

Las historias de usuarios (HU, en lo adelante) son escritas por los propios clientes, tal y como ven ellos las necesidades del sistema. Son similares al empleo de escenarios, con la excepción de que no se limitan a la descripción de la interfaz de usuario. También conducirán el proceso de creación de los test de aceptación (empleados para verificar que las historias de usuario han sido implementadas correctamente). Existen diferencias entre estas y la tradicional especificación de requisitos. La principal diferencia es el nivel de detalle. Las historias de usuario solamente proporcionaran los detalles sobre la estimación del riesgo y cuánto tiempo conllevará la implementación de dicha historia de usuario.

Tabla 3. 1: HU. Carga de Datos

Historia de usuario	
Número: 1	**Nombre:** Carga de Datos
Usuario: Productor	
Prioridad en Negocio: Alta	**Riesgo en Desarrollo:** Alta

[5] Extreme Programming, metodología ágil para desarrollo de software.

Puntos Estimados: 3	**Iteración Asignada:** 1
Descripción: El usuario tendrá la posibilidad de escoger su fuente de datos de distintos orígenes, ya sean Gestores de Bases de Datos (MySQL, Postgres) o archivos de texto (*.CSV, *.ARFF,*.ARFF). No limitando su capacidad para escoger su origen de datos.	
Observaciones:	

Tabla 3. 2: HU. Preprocesado de los datos

Historia de usuario	
Número: 2	**Nombre:** Preprocesado de los datos
Usuario: Productor	
Prioridad en Negocio: Alta	**Riesgo en Desarrollo:** Alta
Puntos Estimados: 3	**Iteración Asignada:** 1
Descripción: El usuario tendrá acceso a informes detallados de los datos seleccionados para el proceso, a través de gráficos y resúmenes. Lo cual le dará un mayor campo visual a la hora de aplicar algún tipo de cambios en los mismos que atenderán a las necesidades que se plantee. Para poder variar ciertas características de su data el usuario contará con una serie de filtros que le permitirán obrar en este sentido.	
Observaciones:	

Tabla 3. 3: HU. Agrupamiento.

Historia de usuario	
Número: 3	**Nombre:** Agrupamiento
Usuario: Productor	
Prioridad en Negocio: Alta	**Riesgo en Desarrollo:** Alta
Puntos Estimados: 2	**Iteración Asignada:** 1
Descripción: El usuario tendrá a su disposición técnicas de agrupamiento que le facilitarán la obtención de modelos de este tipo. Para la visualización se deberá mostrar gráficos representativos y tablas resúmenes de los resultados.	

Observaciones:

Tabla 3. 4: HU. Asociación.

Historia de usuario	
Número: 4	**Nombre:** Asociación
Usuario: Productor	
Prioridad en Negocio: Alta	**Riesgo en Desarrollo:** Alta
Puntos Estimados: 2	**Iteración Asignada:** 1
Descripción: El usuario contará con una interfaz mediante la cual realizará la tarea de asociación y correspondientemente obtendrá representados los resultados de la misma, mediante gráficos y tablas ilustrativas.	
Observaciones:	

Tabla 3. 5: HU. Clasificación

Historia de usuario	
Número: 5	**Nombre:** Clasificación
Usuario: Productor	
Prioridad en Negocio: Alta	**Riesgo en Desarrollo:** Alta
Puntos Estimados: 2	**Iteración Asignada:** 1
Descripción: El usuario poseerá algoritmos para realizar la tarea de clasificación, teniendo como resultado final un modelo que será expresado de forma gráfica y mediante el lenguaje natural.	
Observaciones:	

Tabla 3. 6: HU. Persistencia de Datos.

Historia de usuario	
Número: 6	**Nombre:** Persistencia de Datos
Usuario: Productor	
Prioridad en Negocio: Medio	**Riesgo en Desarrollo:** Alta
Puntos Estimados: 3	**Iteración Asignada:** 1
Descripción: El usuario dispondrá del beneficio de hacer persistentes los resultados obtenidos mediante su procesado, contando con dos vías para esto una la persistencia a un servidor central y dos, la persistencia en forma de informes y de manera local en formato PDF. Con lo cual asegurará que los mismos puedan ser consultados por él u otro interesado en cualquier momento.	
Observaciones:	

3.4. Fase de Planificación.

En la fase de Planificación se hace una estimación del tiempo requerido para la implementación de cada una de las historias de usuarios planteadas, utilizando como medida para expresar el resultado el *punto*. Un punto es considerable como una semana de trabajo sin interrupciones, en la cual cada miembro del equipo desarrolla su trabajo a tiempo completo de acuerdo de acuerdo a lo planificado. Esta estimación incluye todo el esfuerzo asociado a la implementación de la historia de usuario, por ejemplo, las pruebas unitarias, la integración y refactorización del código, y la preparación y ejecución de las pruebas de aceptación. [22].

Estimación de esfuerzos por historias de usuario

Para la realización del sistema se realizo una estimación del tiempo previsto para cada una de las Historias de usuarios que se identificaron. A continuación se detalla este resultado en la tabla 3.7.

Tabla 3. 7: Estimación de esfuerzos por Historia de Usuarios.

Historia de usuario	Puntos de estimación
Carga de Datos	3

Preprocesado de los datos	3
Agrupamiento	2
Asociación	2
Clasificación	2
Persistencia de Datos	3

3.5. Plan de iteraciones

Después de haber sido descritas e identificadas las Historias de Usuario y realizada una estimación del esfuerzo para la realización de cada una de ellas, el siguiente paso es la planificación de la implementación del sistema. El mismo define las iteraciones a realizar, así como cuáles historias de usuario serán implementadas para cada iteración del sistema y las posibles fechas para estas liberaciones.

En el caso del sistema actual se definieron dos iteraciones para proceder con las Historias de usuarios precisadas. A continuación se detallan como estará estructurada cada una de estas.

Iteración 1

En esta iteración se pretende implementar tres Historias de Usuarios que garanticen una primera versión aunque aun sin terminar si que demuestran la funcionalidad del Sistema y una primera vista al cliente que dará su criterio para la incorporación de nuevos elementos y modificación, en caso de ser necesaria de los existentes. En este sentido se abarcan las Historias de Usuarios 1, 2 y 3. Como resultado final de esta iteración se contará con una primera aunque no completa versión del sistema.

Iteración 2

El objetivo de la segunda iteración es implementar todas las Historias de Usuarios pendientes y que complementen la totalidad del sistema. Para ello se implementarán las Historias de Usuario 4, 5 y 6. Como resultado de la iteración se obtendrá una primera versión del sistema, la cual será sometida a consideración del cliente para atender alguna sugerencia final del mismo.

3.5.1. Plan de duración de las iteraciones.

Como un elemento dentro del ciclo de vida de un proyecto utilizando la Metodología XP se tiene la creación del plan de duración de cada una de las iteraciones. Este plan tiene como finalidad mostrar la duración de cada iteración, así como el orden en que serán implementadas las historias de usuario en cada una de las mismas. Esto se ilustra más a detalle en la tabla 3.8

Tabla 3. 8: Plan de duración de las iteraciones.

Iteraciones	Orden de las UH a implementar	Duración de las iteraciones
Iteración 1	1. Carga de Datos 2. Preprocesado de los datos 3. Agrupamiento	8 semanas
Iteración 2	1. Asociación 2. Clasificación 3. Persistencia de Datos	7 semanas

3.5.2. Plan de entregas

A continuación se describe el plan de entregas previsto para la fase implementación. El mismo esta concebido de acuerdo a iteraciones previstas y a los módulos que se desarrollan (Tabla 3.9 y 3.10).

Tabla 3. 9: Módulos e Historias de Usuario que abarcan.

Módulos	**Historias de usuario que abarca**
Carga	Carga de Datos
Preprocesado	Preprocesado de los datos
Análisis	Agrupamiento Asociación Clasificación
Persistencia	Persistencia de Datos

Tabla 3. 10: Plan de duración de entregas.

Módulos	Final 1ra Iteración 4ta semana de Marzo	Final 2da Iteración 3ra semana de Mayo
Carga	v 1.0	v 1.0
Preprocesado	v 1.0	v 1.0
Análisis	v 0.1	v 1.0
Persistencia		v 1.0

3.6. Conclusiones del Capítulo.

En el presente capítulo se versó acerca de las fases de exploración y planificación del proyecto, realizándose una descripción de los artefactos generados durante el transcurso de las mismas. En las Historias de Usuario propuestas se tiene una relación de las principales funcionalidades a implementar en el sistema.

Capítulo IV: Implementación y Prueba.

4.1. Introducción

La implementación de software en XP[6] es un proceso que se realiza de forma iterativa, obteniendo como resultado de cada una de estas un producto funcional que debe ser sometido a pruebas y mostrado al cliente para permitir una retroalimentación por parte de los desarrolladores. El siguiente capítulo esta dedicado a detallar las dos iteraciones llevadas a cabo durante la etapa de construcción del sistema, exponiéndose cada una de las tareas designadas por Historia de Usuarios. De igual modo se expondrán las pruebas de aceptación efectuadas sobre el proyecto.

En cada una de las iteraciones se procede a la implementación de las historias de usuario seleccionadas dentro de estas. Al inicio de las mismas, se lleva a cabo una revisión del plan de iteraciones y se modifica de ser necesario [23]. Como parte de este plan, se descomponen las Historias de Usuarios en tareas de desarrollo, asignando posteriormente cada una de estas a un equipo (o una persona) responsable de su implementación [22]. Estas tareas son para el uso de los programadores, pueden escribirse utilizando un lenguaje técnico y no necesariamente deben ser entendibles para el cliente [22].

De acuerdo a la planificación realizada, se llevaron a cabo dos iteraciones de desarrollo en el sistema. Cada uno de las Historias de Usuario comprendidas en las 2 iteraciones fueron desglosadas en tareas (Tabla 4.1) desarrolladas por cada uno de los programadores.

Tabla 4. 1: Resultados Generales obtenidos de los costos del proyecto.

Historias de Usuario	Tareas
Carga de Datos	1. Programación de la lógica para extracción y conversión de datos de diferentes orígenes. 2. Creación de las interfaces para extracción de datos de diversos orígenes. 3. Desarrollo de interfaces para la visualización de los datos cargados.

[6] Extreme Programming, metodología ágil para desarrollo de software.

Preprocesado de los datos	1. Programar la lógica del preprocesado de los datos. 2. Creación de la interfaz para análisis de los datos seleccionados. 3. Creación de interfaces para la aplicación de filtros.
Agrupamiento	1. Programación de la lógica para la tarea de agrupamiento. 2. Creación de interfaces para aplicación de la tarea de agrupamiento. 3. Desarrollo de interfaces para visualización de los resultados obtenidos por la aplicación de la tarea de Agrupamiento.
Asociación	1. Programación de la lógica para la tarea de Asociación. 2. Creación de interfaces para aplicación de la tarea de Asociación. 3. Desarrollo de interfaces para la visualización de los resultados obtenidos por la aplicación de la tarea de Asociación.
Clasificación	1. Programación de la lógica para la tarea de Clasificación. 2. Creación de interfaces para aplicación de la tarea de Clasificación. 3. Desarrollo de interfaces para la visualización de los resultados obtenidos por la aplicación de los algoritmos de Clasificación.
Persistencia de Datos	1. Manejo de la persistencia hacia Bases de Datos de los resultados obtenidos por las distintas tareas de Minería de Datos, que implementa la aplicación. 2. Persistencia a formato *.PDF, de resúmenes de los resultados obtenidos en cada una de las distintas tareas. 3. Creación de interfaces para el manejo de la persistencia hacia Bases de Datos y a formato PDF.

En los epígrafes siguientes se detallarán cada una de las Historias de Usuario que comprenden el desarrollo del sistema, durante las dos iteraciones previstas. Para ello se hará énfasis en cada una de las tareas comprendidas por Historia de Usuario.

4.2. Iteración 1

Tabla 4. 2: Historias de usuarios comprendidas en la Iteración 1.

Historia de Usuario	Estimación	Real
Carga de Datos	3	3
Preprocesado de los datos	3	3
Agrupamiento	2	2
Totales:	**8**	**8**

Tabla 4. 3: Tarea #1 de la Historia de Usuario Carga de Datos.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Carga de Datos.
Nombre Tarea: Programación de la lógica para extracción y conversión de datos de diferentes orígenes.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 1.4
Fecha Inicio: 02/02/2009	**Fecha Fin:** 10/02/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se creará un grupo de clases que permitan la conexión a diferentes Bases de Datos y otras para cargar archivos de texto (.CSV, .ARFF, .XARFF). Además se implementarán las clases que sean necesarias para transformar los datos cargados al formato requerido para la posterior aplicación de los algoritmos de las librerías de WEKA.	

Tabla 4. 4: Tarea #2 de la Historia de Usuario Carga de Datos.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Carga de Datos.

Nombre Tarea: Creación de las interfaces para extracción de datos de diversos orígenes.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 11/02/2009	**Fecha Fin:** 16/02/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se desarrollará un grupo de interfaces para la extracción de datos, tanto de Bases de Datos como de archivos de texto. Además se conectarán las interfaces con las funciones ya implementadas para estos fines de extracción de datos.	

Tabla 4. 5: Tarea #3 de la Historia de Usuario Carga de Datos.

Tarea	
Número Tarea: 3	**Historia de Usuario:** Carga de Datos.
Nombre Tarea: Desarrollo de interfaces para la visualización de los datos cargados.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 17/02/2009	**Fecha Fin:** 20/02/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se desarrollarán interfaces para visualizar los datos cargados, que permitirán una confirmación visual de que los datos son los correctos. En caso de no ser los correctos se tendrá la posibilidad de retroceder al paso anterior.	

Tabla 4. 6: Tarea #1 de la Historia de Usuario Preprocesado de los Datos.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Preprocesado de los Datos.

Nombre Tarea: Programar la lógica del preprocesado de los datos.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 1.6
Fecha Inicio: 23/02/2009	**Fecha Fin:** 04/03/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán clases que contendrán las funcionalidades necesarias para todo el proceso de preprocesado. Ellas abarcarán, la aplicación de filtros y la obtención de detalles de los datos para análisis (cantidad de atributos, media o moda de cada uno de estos, distintas clases que toma un atributo, entre otras).	

Tabla 4. 7: Tarea #2 de la Historia de Usuario Preprocesado de los Datos.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Preprocesado de los Datos.
Nombre Tarea: Creación de la interfaz para análisis de los datos seleccionados.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 1
Fecha Inicio: 05/03/2009	**Fecha Fin:** 11/03/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	

Descripción: Se crearán un conjunto de interfaces que permitirán hacer un análisis de los datos cargados, estas reflejarán características generales de los datos (cantidad de atributos, cantidad de instancias, tipos de los atributos, entre otras) y además característica individuales de cada uno de sus atributos (tipo del atributo, posibles valores a tomar, media o moda, entre otras). Además se conectarán estas interfaces de manera funcional con la lógica establecida para estas funciones de análisis de los datos.

Tabla 4. 8: Tarea #3 de la Historia de Usuario Preprocesado de los Datos.

Tarea	
Número Tarea: 3	**Historia de Usuario:** Preprocesado de los Datos.
Nombre Tarea: Creación de interfaces para la aplicación de filtros.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.4
Fecha Inicio: 12/03/2009	**Fecha Fin:** 13/03/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán un conjunto de interfaces para cada uno de los filtros establecidos a usar y se conectará además la lógica de programación establecida para cada uno de ellos con dichas interfaces.	

Tabla 4. 9: Tarea #1 de la Historia de Usuario Agrupamiento.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Agrupamiento.
Nombre Tarea: Programación de la lógica para la tarea de agrupamiento.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8

Fecha Inicio: 16/03/2009	**Fecha Fin:** 19/03/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se programarán todas las clases que permitirán manejar la tarea de Agrupamiento y las auxiliares que sean necesarias.	

Tabla 4. 10: Tarea #2 de la Historia de Usuario Agrupamiento.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Agrupamiento.
Nombre Tarea: Creación de interfaces para aplicación de la tarea de agrupamiento.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.4
Fecha Inicio: 20/03/2009	**Fecha Fin:** 23/03/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán las interfaces para manejar la aplicación de la tarea de Agrupamiento, es decir interfaces para recogida de parámetros, de acuerdo al algoritmo a usar.	

Tabla 4. 11: Tarea #3 de la Historia de Usuario Agrupamiento.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Agrupamiento.
Nombre Tarea: Desarrollo de interfaces para la visualización de los resultados obtenidos por la aplicación de la tarea de Agrupamiento.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 24/03/2009	**Fecha Fin:** 27/03/3009

Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán interfaces para la visualización de los resultados obtenidos en la tarea de Agrupamiento, estas interfaces comprenderán la visualización por medio de gráficos y de representación en el leguaje natural de los resultados.	

4.3. Iteración 2

Tabla 4. 12: Historias de usuarios comprendidas en la Iteración 2.

Historia de Usuario	Estimación	Real
Asociación	2	2
Clasificación	2	2
Persistencia de Datos	3	3
Totales:	**7**	**7**

Tabla 4. 13: Tarea #1 de la Historia de Usuario Asociación.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Asociación
Nombre Tarea: Programación de la lógica para la tarea de Asociación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 30/03/2009	**Fecha Fin:** 02/04/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se programarán todas las clases que permitirán manejar la tarea de Asociación, o sea la aplicación de algoritmos, y las auxiliares que sean necesarias.	

Tabla 4. 14: Tarea #2 de la Historia de Usuario Asociación.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Asociación
Nombre Tarea: Creación de interfaces para aplicación de la tarea de Asociación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.4
Fecha Inicio: 03/04/2009	**Fecha Fin:** 06/04/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán las interfaces para manejar la aplicación de la tarea de Asociación, es decir interfaces para recogida de parámetros, de acuerdo al algoritmo a utilizar.	

Tabla 4. 15: Tarea #3 de la Historia de Usuario Asociación.

Tarea	
Número Tarea: 3	**Historia de Usuario:** Asociación
Nombre Tarea: Desarrollo de interfaces para la visualización de los resultados obtenidos por la aplicación de la tarea de Asociación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 07/04/2009	**Fecha Fin:** 10/04/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán interfaces para la visualización de los resultados obtenidos en la tarea de Asociación, estas interfaces comprenderán la visualización por medio de gráficos y de representación en el leguaje natural de los resultados.	

Tabla 4. 16: Tarea #1 de la Historia de Usuario Clasificación.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Clasificación
Nombre Tarea: Programación de la lógica para la tarea de Clasificación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 13/04/2009	**Fecha Fin:** 16/04/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se programarán todas las clases relacionadas a la aplicación la tarea de Clasificación y las auxiliares que sean necesarias para ello.	

Tabla 4. 17: Tarea #2 de la Historia de Usuario Clasificación.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Clasificación
Nombre Tarea: Creación de interfaces para aplicación de la tarea de Clasificación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.4
Fecha Inicio: 17/04/2009	**Fecha Fin:** 20/04/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán las interfaces para manejar la aplicación de la tarea de Agrupamiento, es decir interfaces para recogida de parámetros, de acuerdo al algoritmo a utilizar.	

Tabla 4. 18: Tarea #3 de la Historia de Usuario Asociación.

Tarea	
Número Tarea: 3	**Historia de Usuario:** Clasificación
Nombre Tarea: Desarrollo de interfaces para la visualización de los resultados obtenidos por la aplicación de los algoritmos de Clasificación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.8
Fecha Inicio: 21/04/2009	**Fecha Fin:** 24/0/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán interfaces para la visualización de los resultados obtenidos en la tarea de Clasificación, estas interfaces comprenderán la visualización por medio de gráficos y de representación en el leguaje natural de los resultados.	

Tabla 4. 19: Tarea #1 de la Historia de Usuario Persistencia de Datos.

Tarea	
Número Tarea: 1	**Historia de Usuario:** Persistencia de Datos
Nombre Tarea: Manejo de la persistencia hacia Bases de Datos de los resultados obtenidos por las distintas tareas de Minería de Datos, que implementa la aplicación.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 1.2
Fecha Inicio: 27/04/2009	**Fecha Fin:** 04/05/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán las estructuras necesarias para manejar la persistencia de los resultados obtenidos en las diferentes tareas, hacia una Base de Datos. Se creará además el modelo de datos que identificará a la base de datos donde se guarden los resultados (Anexo 1).	

Tabla 4. 20: Tarea #2 de la Historia de Usuario Persistencia de Datos.

Tarea	
Número Tarea: 2	**Historia de Usuario:** Persistencia de Datos
Nombre Tarea: Persistencia a formato *.PDF, de resúmenes de los resultados obtenidos en cada una de las distintas tareas.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 1.2
Fecha Inicio: 05/05/2009	**Fecha Fin:** 12/05/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se programarán las clases necesarias para manejar la persistencia de los resultados obtenidos en las diferentes tareas, hacia formato de *.PDF, esto se hará en forma de resumen es decir se guardarán resúmenes de cada uno de estos resultados.	

Tabla 4. 21: Tarea #3 de la Historia de Usuario Persistencia de Datos.

Tarea	
Número Tarea: 3	**Historia de Usuario:** Persistencia de Datos
Nombre Tarea: Creación de interfaces para el manejo de la persistencia hacia Bases de Datos y a formato PDF.	
Tipo de Tarea : Desarrollo	**Puntos Estimados:** 0.6
Fecha Inicio: 13/05/2009	**Fecha Fin:** 15/0/2009
Programador Responsable: Leodán Suárez Izquierdo – José Andrés Esquivel Pérez	
Descripción: Se crearán las diferentes interfaces para el manejo de la persistencia hacia Bases de Datos y hacia formato PDF. En ambos casos se enlazará la lógica creada para estas labores	

con las interfaces creadas.

4.4. Diagrama de Clases

El diseño de aplicaciones, en la metodología XP no requiere la representación del sistema mediante diagramas de clases utilizando notación UML, en su lugar se usan otras técnicas como las tarjetas CRC[7]. Estas determinan responsabilidades y colaboraciones de las clases. A continuación se representan las mismas.

Tabla 4. 22: Tarjeta CRC#1 Conexión a Base de Datos.

Clase: DatabaseConexion	
Responsabilidad	**Colaboraciones**
- Permitir la conexión a Base de Datos.	

Tabla 4. 23: Tarjeta CRC#2 Conversión de datos desde Base de Datos.

Clase: QueryToInstances	
Responsabilidad	**Colaboraciones**
- Convertir el elemento seleccionado de la Base de Datos (tabla o vista), al formato requerido (Instances[8]).	- DatabaseConexion

[7] Class Responsibility Collaborator ()

[8] Es el formato definido por el conjunto de librerías de WEKA, al cual se deben transformar los datos para su manipulación.

Tabla 4. 24: Tarjeta CRC#3 Carga de Fichero de Texto.

Clase: ConverterFileChooser	
Responsabilidad	**Colaboraciones**
- Carga de archivos de texto. - Conversión al formato requerido.	

Tabla 4. 25: Tarjeta CRC#4 Preprocesado de los datos.

Clase: Preprocesado	
Responsabilidad	**Colaboraciones**
- Brindar los elementos estadísticos de los datos cargados.	- ConverterFileChooser - QueryToInstances

Tabla 4. 26: Tarjeta CRC#5 Filtrado de los Datos.

Clase: Filtrado	
Responsabilidad	**Colaboraciones**
- Responsable de la aplicación de filtros.	- ConverterFileChooser - QueryToInstances - Preprocesado

Tabla 4. 27: Tarjeta CRC#6 Análisis, reglas de Asociación.

Clase: AsociacionReglas	
Responsabilidad	**Colaboraciones**
- Aplicación de la tarea de Asociación. - Proveer los resultados de la tarea de Asociación	

Tabla 4. 28: Tarjeta CRC#7 Análisis, Agrupamiento o Clustering.

Clase: Clustering	
Responsabilidad	**Colaboraciones**
- Aplicación de la tarea de Agrupamiento. - Proveer los resultados de la tarea de Agrupamiento.	

Tabla 4. 29: Tarjeta CRC#8 Análisis, Clasificación.

Clase: Clasificador	
Responsabilidad	**Colaboraciones**
- Aplicación de la tarea de Clasificación. - Proveer los resultados de la tarea de Clasificación.	

Tabla 4. 30: Tarjeta CRC#8 Persistencia hacia Bases de Datos.

Clase: RepositorioDatos	
Responsabilidad	**Colaboraciones**
- Salvar resultado de la tarea de Asociación. - Salvar resultado de la tarea de Agrupamiento. - Salvar resultado de la tarea de Clasificación.	- AsociacionReglas - Clustering - Clasificador

Tabla 4. 31: Tarjeta CRC#8 Persistencia hacia archivo PDF, Asociación.

Clase: ConvertRules2PDF	
Responsabilidad	**Colaboraciones**
- Salvar resultado de la tarea de Asociación a un archivo de extensión PDF.	

Tabla 4. 32: Tarjeta CRC#8 Persistencia hacia archivo PDF, Clasificación.

Clase: ConvertClasif2PDF	
Responsabilidad	**Colaboraciones**
- Salvar resultado de la tarea de Clasificación a un archivo de extensión PDF.	

Tabla 4. 33: Tarjeta CRC#8 Persistencia hacia archivo PDF, Agrupamiento.

Clase: ConvertClusters2PDF	
Responsabilidad	**Colaboraciones**
- Salvar resultado de la tarea de Agrupamiento a un archivo de extensión PDF.	

4.5. Pruebas

Uno de los principales procesos dentro de XP es el de prueba, el cual anima constantemente a los desarrolladores a probar constantemente. El objetivo de esta filosofía es el número de errores no detectados así como el tiempo entre la introducción de este en el sistema y su detección [24]. Lo cual eleva la calidad de los productos desarrollados.

En el caso de XP se usan las pruebas de aceptación para designar si al final de cada iteración se consiguieron las funcionalidades requeridas, además de comprobar que dicha funcionalidad sea la esperada por el cliente. Las pruebas de aceptación son pruebas de caja negra que se crean a partir de las historias de usuario [24]. El objetivo final de estas es garantizar que los requerimientos han sido cumplidos y que el sistema es aceptable [25].

A continuación se detallan las distintas pruebas unitarias empleadas para constatar el correcto funcionamiento de las funcionalidades previstas en cada una de las Historias de Usuario

Tabla 4. 34: Prueba de Aceptación #1 de la Historia de Usuario Carga de Datos.

Prueba de Aceptación	
Código: HU1_P1	**Historia de Usuario:** Carga de Datos.
Nombre: Origen de datos, Base de Datos Postgres.	
Descripción: Se prueba la conexión a una Base de Datos Postgres y se intenta cargar de ella datos procedentes de una tabla o vista.	
Condiciones de Ejecución: La Base de Datos Postgres a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos Postgres.	
Resultado Esperado: Conexión satisfactoria y todas las tablas y vistas de la Base de Datos, reflejadas en un árbol de conexión.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 35: Prueba de Aceptación #2 de la Historia de Usuario Carga de Datos.

Prueba de Aceptación	
Código: HU1_P2	**Historia de Usuario:** Carga de Datos.
Nombre: Origen de datos, Base de Datos MySQL	
Descripción: Se prueba la conexión a una Base de Datos MySQL y se intenta cargar de ella datos procedentes de una tabla o vista.	
Condiciones de Ejecución: La Base de Datos MySQL a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos MySQL.	
Resultado Esperado: Conexión satisfactoria y todas las tablas y vistas de la Base de Datos, reflejadas en un árbol de conexión.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 36: Prueba de Aceptación #3 de la Historia de Usuario Carga de Datos.

Prueba de Aceptación	
Código: HU1_P3	**Historia de Usuario:** Carga de Datos.
Nombre: Origen de datos, archivo *.ARFF.	
Descripción: Se prueba la carga de archivos de extensión ARFF.	
Condiciones de Ejecución: El formato dentro del archivo *.ARFF, debe ser correcto.	
Entrada / Pasos de ejecución: Se intenta la carga del archivo para la prueba eligiendo el mismo dentro del directorio origen.	
Resultado Esperado: El archivo *.ARFF es cargado satisfactoriamente y su contenido es mostrados en una tabla.	

Evaluación de la Prueba: Satisfactoria

Tabla 4. 37: Prueba de Aceptación #4 de la Historia de Usuario Carga de Datos.

Prueba de Aceptación	
Código: HU1_P4	**Historia de Usuario:** Carga de Datos.
Nombre: Origen de datos, archivo *.CSV.	
Descripción: Se prueba la carga de archivos de extensión CSV.	
Condiciones de Ejecución: El formato dentro del archivo *.CSV, debe ser correcto (la separación debe ser por comas (,) dentro del archivo).	
Entrada / Pasos de ejecución: Se intenta la carga del archivo para la prueba eligiendo el mismo dentro del directorio origen.	
Resultado Esperado: El archivo *.CSV es cargado satisfactoriamente y su contenido es mostrados en una tabla.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 38: Prueba de Aceptación #1 de la Historia de Usuario Preprocesado de los Datos.

Prueba de Aceptación	
Código: HU2_P1	**Historia de Usuario:** Preprocesado de los Datos.
Nombre: Eliminación de Atributos	
Descripción: Se prueba la eliminación de atributos de un conjunto de datos.	
Condiciones de Ejecución: Exista al menos 2 atributos en la relación de datos cargada.	
Entrada / Pasos de ejecución: Selección del o los atributos a eliminar.	
Resultado Esperado: Es refrescado el estado de la interfaz y desaparecen los atributos	

seleccionados anteriormente.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 39: Prueba de Aceptación #2 de la Historia de Usuario Preprocesado de los Datos.

Prueba de Aceptación	
Código: HU2_P2	**Historia de Usuario:** Preprocesado de los Datos.
Nombre: Aplicación de filtro discretizar.	
Descripción: Se hace uso del filtro Discretizar. Se convierte un atributo de tipo numérico en uno de tipo nominal.	
Condiciones de Ejecución: Exista al menos 1 atributo de tipo numérico en los datos cargados.	
Entrada / Pasos de ejecución: Selección del atributo a discretizar, se selecciona el filtro, se modifican sus parámetros y se aplica.	
Resultado Esperado: Es refrescado el estado de los datos y el atributo antes numérico ahora es de tipo nominal.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 40: Prueba de Aceptación #3 de la Historia de Usuario Preprocesado de los Datos.

Prueba de Aceptación	
Código: HU2_P3	**Historia de Usuario:** Preprocesado de los Datos.
Nombre: Aplicación de filtro Randomize.	
Descripción: Se hace uso del filtro Randomize. Ordena aleatoriamente el orden de las instancias dentro de los datos cargados.	
Condiciones de Ejecución: La cantidad de instancias en los datos cargados sea de gran tamaño donde sea significativo el orden de las instancias al analizarlas.	

Entrada / Pasos de ejecución: Selección del filtro Randomize, se modifican sus parámetros y se aplica.
Resultado Esperado: Es alterado el orden de las instancias dentro del paquete de datos en uso.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 41: Prueba de Aceptación #1 de la Historia de Usuario Agrupamiento.

Prueba de Aceptación	
Código: HU3_P1	**Historia de Usuario:** Agrupamiento.
Nombre: Aplicación del algoritmo de Agrupamiento.	
Descripción: Se procede a aplicar el algoritmo de Agrupamiento y se calibran los parámetros de acuerdo al resultado deseado.	
Condiciones de Ejecución: La estación de trabajo debe contener espacio en memoria libre, para su correcto funcionamiento, de acuerdo al volumen de los datos.	
Entrada / Pasos de ejecución: Se procede a aplicar el algoritmo de Agrupamiento y se calibran los parámetros de acuerdo al resultado deseado.	
Resultado Esperado: El algoritmo luego de transcurrido el tiempo necesario detiene su ejecución de manera no forzada y se visualizan los resultados.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 42: Prueba de Aceptación #2 de la Historia de Usuario Agrupamiento.

Prueba de Aceptación	
Código: HU4_P2	**Historia de Usuario:** Agrupamiento.
Nombre: Visualización de los resultados de la tarea de Agrupamiento.	
Descripción: Luego de haber aplicado la tarea de Agrupamiento se procede a la visualización de los resultados, mediante gráficos y exposición en lenguaje natural.	

Condiciones de Ejecución: La tarea de Agrupamiento debe haberse ejecutado correctamente.
Entrada / Pasos de ejecución: Luego de haber concluido la tarea se verifica que los resultados obtenidos se encuentren reflejados en dos tablas, una referencia un resumen y otra los detalles. Se pedirá la visualización gráfica de los mismos.
Resultado Esperado: En la tabla resumen se reflejarán datos estadísticos correspondientes con la media o la moda de cada uno de los atributos en cuestión y en la tabla detalle cada una de las instancias clasificadas para el grupo (clúster) señalado. La visualización gráfica nos mostrara un gráfico de pastel que contendrá representados los grupos de acuerdo al número de instancias de cada uno, dando sensación de parte o porciento del total de la muestra analizada.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 43: Prueba de Aceptación #1 de la Historia de Usuario Asociación.

Prueba de Aceptación	
Código: HU4_P1	**Historia de Usuario:** Asociación.
Nombre: Aplicación del algoritmo de Asociación.	
Descripción: Se procede a aplicar el algoritmo de Asociación y se calibran los parámetros de acuerdo al resultado deseado.	
Condiciones de Ejecución: La estación de trabajo debe contener espacio en memoria libre, para su correcto funcionamiento, de acuerdo al volumen de los datos.	
Entrada / Pasos de ejecución: Se procede a aplicar el algoritmo de Asociación y se calibran los parámetros de acuerdo al resultado deseado.	
Resultado Esperado: El algoritmo luego de transcurrido el tiempo necesario detiene su ejecución de manera no forzada y se visualizan los resultados.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 44: Prueba de Aceptación #2 de la Historia de Usuario Asociación.

Prueba de Aceptación	
Código: HU4_P2	**Historia de Usuario:** Asociación.
Nombre: Visualización de los resultados de la tarea de Asociación.	
Descripción: Luego de haber aplicado la tarea de Asociación se procede a la visualización de los resultados, mediante gráficos y exposición en lenguaje natural.	
Condiciones de Ejecución: La tarea de Asociación debe haberse ejecutado correctamente.	
Entrada / Pasos de ejecución: Luego de haber concluido la tarea se verifica que los resultados obtenidos se encuentren reflejados en una tabla central. Se pedirá la visualización gráfica de los mismos.	
Resultado Esperado: En la tabla del centro se verán un conjunto de reglas en forma de premisas que inducen hipótesis y reflejan un determinado grado de confianza u otra métrica determinada. La visualización gráfica nos devolverá un grafico de relaciones entre atributos que reflejarán la regla señalada.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 45: Prueba de Aceptación #1 de la Historia de Usuario Clasificación.

Prueba de Aceptación	
Código: HU5_P1	**Historia de Usuario:** Clasificación.
Nombre: Aplicación del algoritmo de Clasificación.	
Descripción: Se procede a aplicar el algoritmo de clasificación y se calibran los parámetros de acuerdo al resultado deseado.	
Condiciones de Ejecución: La estación de trabajo debe contener espacio en memoria libre, para su correcto funcionamiento, de acuerdo al volumen de los datos.	
Entrada / Pasos de ejecución: Se procede a aplicar el algoritmo de clasificación y se calibran los	

parámetros de acuerdo al resultado deseado.
Resultado Esperado: El algoritmo luego de transcurrido el tiempo necesario detiene su ejecución de manera no forzada y se visualizan los resultados.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 46: Prueba de Aceptación #2 de la Historia de Usuario Clasificación.

Prueba de Aceptación	
Código: HU5_P2	**Historia de Usuario:** Clasificación.
Nombre: Visualización de los resultados de la tarea de Clasificación.	
Descripción: Luego de haber aplicado la tarea de Clasificación se procede a la visualización de los resultados, mediante gráficos y exposición en lenguaje natural.	
Condiciones de Ejecución: La tarea de Clasificación debe haberse ejecutado correctamente.	
Entrada / Pasos de ejecución: Luego de haber concluido la tarea se verifica que el resultado se encuentre reflejado en el panel central, como una traducción al lenguaje natural del mismo, debiéndose observar también los resultados de la validación del modelo obtenido. Además se pide la visualización gráfica de los resultados.	
Resultado Esperado: En el panel central se tiene una descripción en lenguaje natural de los resultados, observándose también los resultados de la validación del modelo obtenido y la visualización gráfica nos devuelve un árbol de clasificación con el modelo obtenido.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 47: Prueba de Aceptación #1 de la Historia de Usuario Persistencia de Datos

Prueba de Aceptación	
Código: HU6_P1	**Historia de Usuario:** Persistencia de Datos.

Nombre: Persistencia de los resultados de la tarea de Agrupamiento hacia una Base de Datos Postgres.
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Agrupamiento hacia una base de Datos Postgres.
Condiciones de Ejecución: La Base de Datos Postgres a ser accedida debe tener sus servicios activos.
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos Postgres, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.
Resultado Esperado: Los resultados obtenidos de la tarea de Agrupamiento son guardados correctamente y aparece un mensaje que lo indica.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 48: Prueba de Aceptación #2 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P2	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Agrupamiento hacia una Base de Datos MySQL.	
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Agrupamiento hacia una base de Datos MySQL.	
Condiciones de Ejecución: La Base de Datos MySQL a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos MySQL, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.	
Resultado Esperado: Los resultados obtenidos de la tarea de Agrupamiento son guardados	

correctamente y aparece un mensaje que lo indica.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 49: Prueba de Aceptación #3 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P3	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Asociación hacia una Base de Datos Postgres.	
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Asociación hacia una base de Datos Postgres.	
Condiciones de Ejecución: La Base de Datos Postgres a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos Postgres, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.	
Resultado Esperado: Los resultados obtenidos de la tarea de Asociación son guardados correctamente y aparece un mensaje que lo indica.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 50: Prueba de Aceptación #4 de la Historia de Usuario Persistencia de Datos

Prueba de Aceptación	
Código: HU6_P4	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Asociación hacia una Base de Datos MySQL.	
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Asociación hacia	

una base de Datos MySQL.
Condiciones de Ejecución: La Base de Datos MySQL a ser accedida debe tener sus servicios activos.
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos MySQL, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.
Resultado Esperado: Los resultados obtenidos de la tarea de Asociación son guardados correctamente y aparece un mensaje que lo indica.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 51: Prueba de Aceptación #5 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P5	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Clasificación hacia una Base de Datos Postgres.	
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Clasificación hacia una base de Datos Postgres.	
Condiciones de Ejecución: La Base de Datos Postgres a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos Postgres, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.	
Resultado Esperado: Los resultados obtenidos de la tarea de Clasificación son guardados correctamente y aparece un mensaje que lo indica.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 52: Prueba de Aceptación #6 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P6	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Clasificación hacia una Base de Datos MySQL.	
Descripción: La prueba pretende salvar los resultados obtenidos en la tarea de Clasificación hacia una base de Datos MySQL.	
Condiciones de Ejecución: La Base de Datos MySQL a ser accedida debe tener sus servicios activos.	
Entrada / Pasos de ejecución: Se intenta acceder mediante los datos requeridos (nombre de la Base de Datos, dirección, puerto, usuario y contraseña) a una Base de Datos MySQL, se selecciona el usuario propietario de los resultados (esto se hace en caso de existir, si no se crea en el mismo momento) y se procede a salvarlos.	
Resultado Esperado: Los resultados obtenidos de la tarea de Clasificación son guardados correctamente y aparece un mensaje que lo indica.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 53: Prueba de Aceptación #7 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P7	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Agrupamiento a un archivo *.PDF.	
Descripción: La prueba pretende exportar un resumen de los resultados obtenidos en la tarea de Agrupamiento hacia un archivo de formato PDF.	
Condiciones de Ejecución: En el directorio elegido para salvar los resultados debe haber suficiente espacio físico para los mismo.	
Entrada / Pasos de ejecución: Se procede a exportar los resultados a formato PDF, sugiriendo el	

directorio donde residirá este archivo, se le da nombre y se salvan los mismos.
Resultado Esperado: Los resultados obtenidos de la tarea de Agrupamiento son guardados correctamente y aparece un mensaje que lo indica.
Evaluación de la Prueba: Satisfactoria

Tabla 4. 54: Prueba de Aceptación #8 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P8	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Asociación a un archivo *.PDF.	
Descripción: La prueba pretende exportar un resumen de los resultados obtenidos en la tarea de Asociación hacia un archivo de formato PDF.	
Condiciones de Ejecución: En el directorio elegido para salvar los resultados debe haber suficiente espacio físico para los mismo.	
Entrada / Pasos de ejecución: Se procede a exportar los resultados a formato PDF, sugiriendo el directorio donde residirá este archivo, se le da nombre y se salvan los mismos.	
Resultado Esperado: Los resultados obtenidos de la tarea de Asociación son guardados correctamente y aparece un mensaje que lo indica.	
Evaluación de la Prueba: Satisfactoria	

Tabla 4. 55: Prueba de Aceptación #9 de la Historia de Usuario Persistencia de Datos.

Prueba de Aceptación	
Código: HU6_P9	**Historia de Usuario:** Persistencia de Datos.
Nombre: Persistencia de los resultados de la tarea de Clasificación a un archivo *.PDF.	
Descripción: La prueba pretende exportar un resumen de los resultados obtenidos en la tarea de Clasificación hacia un archivo de formato PDF.	

Condiciones de Ejecución: En el directorio elegido para salvar los resultados debe haber suficiente espacio físico para los mismo.
Entrada / Pasos de ejecución: Se procede a exportar los resultados a formato PDF, sugiriendo el directorio donde residirá este archivo, se le da nombre y se salvan los mismos.
Resultado Esperado: Los resultados obtenidos de la tarea de Clasificación son guardados correctamente y aparece un mensaje que lo indica.
Evaluación de la Prueba: Satisfactoria

4.6. Conclusiones del Capítulo.

En el capítulo se trataron las tareas a desarrollar, por el equipo de desarrollo, como parte de la implementación de cada una de las Historias de Usuario propuestas en el sistema. Se abordaron de igual modo las pruebas a realizarse en la etapa de pruebas del proyecto.

Capítulo V: Estudio de Factibilidad.

5.1. Introducción

Una de las principales tareas dentro de la planificación de proyectos de software es la estimación, la cual consiste en determinar, los recursos de hardware y software, el costo, el tiempo y esfuerzo necesarios para el desarrollo de los mismos. Este capítulo estará destinado a realizar un estudio de factibilidad correspondiente al sistema propuesto, realizándose una estimación del esfuerzo necesario para llevar a cabo el mismo.

5.2. Características del proyecto.

Nombre de la entrada externa	**Cantidad de archivos referenciados**	**Cantidad de elementos de datos**	**Clasificación** (simple, media o compleja)
Entrar de datos, a partir de la Base de Datos seleccionada.	2	5	media
Entrada de datos, a partir de la lectura de archivos de texto (*.CSV, *. ARFF)	1	1	simple

Tabla 5. 1: Entradas Externas.

Nombre de la salida externa	**Cantidad de archivos referenciados**	**Cantidad de elementos de datos**	**Clasificación** (simple, media o compleja)
Mostrar datos cargados por el sistema.	1	1	simple
Mostrar metadatos de los datos cargados en el sistema.	1	1	simple

Tabla 5. 2: Salidas Externas.

Nombre de la consulta externa	Cantidad de archivos referenciados	Cantidad de elementos de datos	Clasificación (simple, media o compleja)
Mostrar miembros y características, por clúster obtenido.	1	1	simple
Mostrar reglas de asociación.	1	1	simple
Mostrar modelo de clasificación obtenido.	1	1	simple

Tabla 5. 3: Consultas Externas.

Nombre de los grupos lógicos de datos internos	Cantidad de tipos de registro	Cantidad de elementos de datos	Clasificación (simple, media o compleja)
Interacción con archivos de configuración *.properties.	3	1	simple

Tabla 5. 4: Grupos lógicos de datos internos.

Nombre de los grupos lógicos de datos de interfaz	Cantidad de tipos de registro	Cantidad de elementos de datos	Clasificación (simple, media o compleja)
Salvar conocimientos a repositorio (Base de Datos).	1	1	simple
Exporta a resúmenes a formato *.PDF	1	1	simple

Tabla 5. 5: Grupos lógicos de datos de interfaz.

5.2.1. Estimación Inicial

Al realizar la suma de todos los aportes se obtienen los puntos de función sin ajustar.

Elementos	Simple		Media		Compleja		Aportes
	Cantidad	Valor	Cantidad	Valor	Cantidad	Valor	
Entradas Externas	1	3	1	4	0	6	7
Salidas Externas	2	4	0	5	0	7	8
Consultas Externas	3	3	0	4	0	6	9
Grupos lógicos de datos internos	1	7	0	10	0	15	7
Grupos lógicos de datos de interfaz	2	5	0	7	0	10	10
Total							**41**

Tabla 5. 6: Aportes de los puntos de función sin ajustar.

5.3. CÁLCULO DE INSTRUCCIONES FUERTES, ESFUERZO, TIEMPO DE DESARROLLO, CANTIDAD DE HOMBRES Y COSTO.

El modelo COCOMO II se basa en el uso de los Puntos Función y/o Líneas de Código Fuente (SLOC) como base para medir tamaño en los modelos de estimación de Diseño Temprano y Post-Arquitectura. Los Puntos de Función procuran cuantificar la funcionalidad de un sistema de software. La meta es obtener un número que caracterice completamente al sistema. Son útiles estimadores ya que están basados en información que está disponible en las etapas tempranas del ciclo de vida del desarrollo de software. COCOMO II considera solamente UFP (del inglés: Puntos Función Desajustados).

5.3.1. Cálculo del esfuerzo nominal.

Aplicación de la ecuación del esfuerzo nominal:

$$\mathbf{PM_{nominal} = A \times (Size)^{E}}$$

$PM_{nominal}$: es el esfuerzo nominal requerido en meses-hombre.

A: Es una constante que se utiliza para capturar los efectos multiplicativos en el esfuerzo requerido de acuerdo al crecimiento del tamaño del software. El modelo la calibra con un valor de 2.94.

Size: Es el tamaño estimado del software, en Puntos de Función sin ajustar (convertibles a KSLOC). Se calcula el producto de los puntos de función sin ajustar por un factor de conversión que depende del lenguaje a utilizar en el desarrollo del sistema. Se utiliza JAVA (factor de conversión = 53 SLOC/UFP). Por tanto:

Size = 53 × 41 = 2173 SLOC

Size ≈ 2.2 KSLOC

E: Es una constante denominada *Factor Escalar*. Se calcula ponderando las variables escalares, mediante la ecuación:

E = 0.91 + 0.01 × Σ (Wi)

Donde las Wi se muestran en la siguiente tabla:

Nombre	Valor	Justificación
PREC	2.25	Existen varios proyectos similares a nivel mundial y nacional.
FLEX	2.50	Cuenta con alta flexibilidad.
TEAM	3.50	El equipo de desarrollo presenta una alta cohesión.
RESL	2.80	No se identifican riesgos críticos.
PMAT	1.80	No existe mucha experiencia en aplicaciones de tipo.
Total	**12.85**	

Tabla 5. 7: 5.7 Factor Escalar (SF).

Luego:

E = 0.91 + 0.01 × 12.85 = 1.167

Siendo:

$\mathbf{PM}_{nominal} = \mathbf{A \times (Size)^{E}} = 2.94 \times 2.2^{1.167} = 7.3782$ meses-hombre

5.3.2. Cálculo del esfuerzo ajustado

Se aplica la ecuación de cálculo del esfuerzo ajustado:

PM $_{ajustado}$ = **PM** $_{nominal}$ **× Π (ME$_i$)**

Nombre	Valor	Justificación
RUSE	1.07	Existen varios proyectos similares a nivel mundial y nacional.
RCPX	1.05	Complejidad y confiabilidad del producto, altas.
FCIL	0.95	Se utilizan herramientas facilitan el trabajo.
PDIF	1.00	Uso de memoria y almacenamiento, normal. Plataforma estable.
PERS	0.9	Alta capacidad del personal.
SCED	1.00	Se empleo el tiempo planificado para el desarrollo del sistema.
Total	**0.96**	

Tabla 5. 8: Multiplicadores de esfuerzo (EM).

Entonces:

PM $_{ajustado}$ = 7.38 × 0.96 = 7.08 meses-hombre.

5.3.3. Cálculo del tiempo de desarrollo, cantidad de hombres y costo.

Valores calibrados: **A =** 2.94; **B =** 0.91; **C =** 3.67; **D =** 0.24

F = D + 0.2 × (E - B) = 0.24 + 0.2 × (1.167 – 0.91) = 0.29

TDEV (Tiempo de desarrollo) = C × (**PM** $_{ajustado}$) ^{F} = 3.67 × $(7.08)^{0.29}$ = 6.47 meses

CH (Cantidad de Hombres) = PM $_{ajustado}$ / **TDVE** = 7.08/ 6.47 = 1.1 personas

Como la cantidad real de hombres disponibles para el desarrollo de la aplicación es ***2***, al reajustar el tiempo de desarrollo según la cantidad de hombres, resultó un tiempo equivalente a ***3.5*** meses.

Salario promedio:

Para determinar el salario promedio se tiene en cuenta que los desarrolladores del sistema han de ser estudiantes de pregrado pertenecientes a la UCI (Universidad de las Ciencias Informáticas), por lo que se toma como salario correspondiente una media de **$100.00.**

Costo = 2 × $100 × $ 7.08 = $1416

Obteniéndose como resultados generales que:

Magnitud	**Valor**
Esfuerzo	7.08 meses – hombre
Tiempo de Desarrollo Ajustado	3.5 meses
Cantidad de Hombres	2
Salario Medio	$100
Costo	1416

Tabla 5. 9: Resultados Generales obtenidos de los costos del proyecto.

5.4. Beneficios tangibles e intangibles.

El desarrollo del producto propuesto en el presente trabajo brinda como beneficio fundamental el contar con un elemento de soporte a la hora de la toma de decisiones que permite realizar análisis de grandes volúmenes de datos y extraer patrones y modelos que no son tangibles a simple vista o mediante el uso de técnicas y métodos tradicionales.

Además el proceso de incorporación de la práctica de la integración continua dentro del proceso de desarrollo de una empresa supone una serie de beneficios, entre los que se destaca la reducción a casi nulo del tiempo de integración de los diferentes módulos que conforman el proyecto. Esto se debe a que los errores de integración son detectados rápidamente, posibilitando su corrección de forma inmediata.

5.5. Análisis del Costo.

Asociado al desarrollo de un producto existe siempre un costo de producción, el cual debe ser justificado, fundamentalmente en base a los beneficios que el mismo ha de reportar. La propuesta de sistema que propone este trabajo no conlleva a grandes gastos, como así lo ha demostrado el estudio realizado para estimar su factibilidad, lo cual no permite afirmar que soportados en este elemento y en los beneficios que este sistema provee, que su implementación es factible. Elemento justificado en gran medida por el uso de plataformas, bibliotecas y herramientas libres que no requieren el pago de licencia alguna.

5.6. Conclusiones del Capítulo.

En el presente capítulo se ha realizado un análisis de factibilidad de la solución propuesta, la cual a determinado como conclusión la viabilidad de su desarrollo, basada en la comparación entre los costos de su producción y desarrollo, contra los beneficios reportados con su puesta en funcionamiento.

Conclusiones

Se culmina el presente trabajo, donde se han logrado cumplir los objetivos trazados con técnicas y metodologías para cada uno de ellos. En vista de lograr la mejor aproximación a estos, podemos concluir los siguientes resultados:

- ✓ Se hizo un análisis acerca de las metodologías usadas en el desarrollo de software haciendo una selección conveniente, concluyendo con el uso de la Metodología XP.
- ✓ Se realizó una selección adecuada de herramientas, librerías y plataformas de desarrollo de software de tipo libre, garantizando una ejecución del software en cualquier Sistema Operativo sin ningún tipo de restricción.
- ✓ Se desarrolló un estudio profundo con respecto a la técnica Minería de Datos siendo las tareas que aplica el centro de estudio de esta técnica, su aplicación en varios entornos y en especial su vinculación en el entorno educativo.
- ✓ Se obtuvo como resultado una herramienta de fácil uso para el usuario, para su uso en el tratamiento a datos de diversos orígenes aplicando algoritmos de minería de datos de carácter descriptivo y predictivos, lo cual produce un resultado a modo de información valiosa para el apoyo en el proceso de toma de decisiones.
- ✓ Se llevó a cabo un estudio de Factibilidad de la solución en cuestión concluyendo que es viable su desarrollo comparando los costos de producción con los beneficios reportados por su puesta en práctica.

Recomendaciones

Como resultado final del proceso investigativo, se obtuvo una serie de recomendaciones a tener en cuenta para desarrollos futuros, que den continuidad al proceso. A continuación se detallan las que resultaron de mayor relevancia.

- ✓ Como todo producto de software sobre la base del cual las necesidades del usuario crecen con el tiempo, se debe continuar el desarrollo del mismo con la incorporación de nuevos algoritmos que fortalezcan las tareas ya concebidas, así como la ampliación en general del sistema, proveyéndolo de mayor número de orígenes de datos y facilidades de visualización de los resultados. Todo en base al crecimiento de los requerimientos del usuario.

- ✓ Mantener una constante interacción con los usuarios de la aplicación que propicie una retroalimentación, dando así margen a la corrección de posibles errores que ni hayan sido localizados.

- ✓ Desplegar la misma en la facultad y otros centros que de la posibilidad de constatar su facilidad de uso, el correcto entendimiento de los resultados obtenidos y por consiguiente la mejora practica que supone en el proceso de toma de decisiones. Sirviendo también como medio de retroalimentación.

- ✓ Desarrollar una aplicación que sirva como cliente para obtener los conocimientos que se almacenan en el repositorio, dando la posibilidad de que estos puedan ser utilizados por otros sistemas en vísperas de ampliar la potencialidad de los mismos.

- ✓ Utilización de los resultados producidos por la aplicación para emplearlos como base de conocimientos de otros sistemas, lo cual brindaría una mayor potencialidad a los mismos.

Referencias Bibliográficas

[1] Tang, Z. and J. MacLennan (2005). Data Mining with SQL Server 2005. United States of America, Wiley Publishing, Inc.

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

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

[4] Servente, M. Algoritmos TDIDT aplicados a la Minería de Datos Inteligente. Universidad de Buenos Aires, Facultad de Ingeniería, Laboratorio de Sistemas Inteligentes. Tesis de Grado en Ingeniería Informática, 2002.

[5] 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.

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

[7] Larose, D. T. Data Mining. Methods and Models. Department of Mathematical Sciences. Central Connecticut State University. Copyright by John Wiley & Sons, Inc. All rights reserved, Hoboken, New Jersey. 2006

[8] KDnuggets Polls. Data mining methods, Mar 26 - Apr 8, 2007. <http://www.kdnuggets.com/polls/2007/data_mining_methods.htm> [Consulta en línea: 20 de enero del 2009]

[9] Martínez de Pisón Ascacibar, F. J. Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado. Universidad de la Rioja, Departamento de Ingeniería Mecánica. Memoria presentada para la obtención del grado de Doctor en Ingeniería Industrial, 2003.

[10] Dehaspe, L.; Toivonem, H. "Discovery of Relational Association Rules", en [Dzeroski & Lavrac 2001], pp: 189-212, 2001.

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

[12] Romero, C., Ventura, S. Educational data mining: a survey from 1995 to 2005. Expert Systems with Applications. 33:1. 2007.

[13] Espinosa, I., Pérez, S. Obtención de Reglas y Patrones en el Proceso Académico de la Universidad de Ciencias informáticas. Tesis de Diploma. CEIS. CUJAE, julio, 2007.

[14] Acosta, R., Vázquez, L. Obtención de patrones y reglas en el Sistema Docente del Instituto Superior Politécnico José A. Echevarría (CUJAE) utilizando Minería de Datos. Trabajo de Diploma. Coautores Brito, R.; Rosete, A. Ciudad de La Habana, Cuba. Junio, 2007.

[15] KDnuggets Polls. Data Mining / Analytic Software (Tools), May 5 - 20, 2007. <http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm> [Consulta en línea: 20 de enero del 2009]

[16] SPSS Clementine. SPSS, Statistical Product and Service Solutions. [Software]. SPSS, 2007. <http://www.spss.com/clementine/index.htm>. [Consulta en línea: 26 de enero de 2009]

[17] Chapman, P. [et al.] CRISP-DM 1.0: Step-by-step data mining guide. USA: SPSS Inc., CRISP-DM Consortium, 2000.

[18] SAS Enterprise Miner. SAS Institute (Statistical Analysis Systems). [Software]. SAS, 2007. <http://www.sas.com/technologies/analytics/datamining/miner/> [Consulta en línea: 26 de enero de 2009].

[19] SAS Enterprise Miner SEMMA. SAS Inc., 2006. <http://www.sas.com/technologies/analytics/datamining/miner/semma.html> [Consulta en línea: 26 de enero de 2009].

[20] RapidMiner 4.0, User Guide, Operator Reference, Developer Tutorial. 2007. Rapid Miner (YALE) [Software] 2007. <http://www.rapidminer.com> [Consulta en línea: 26 de enero de 2009].

[21] WEKA, Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java. [Software]. Weka 3.4.7, 2005. <http://www.cs.waikato.ac.nz/ml/weka/> [Consulta en línea: 26 de enero de 2009].

[22] Beck, K., Programación Extrema Explicada. Addison Wesley. Título original: Extreme Programming Explained, 2000.

[23] Beck K. y Fowler M., Planeando en Programación Extrema. Addison Wesley. Título original: Planning Extreme Programming, 2000.

[24] Crispin, L. y House, T. Probando la Programación Extrema. Addison Wesley. Título original: Testing Extreme Programming, 2002.

[25] Extreme Programming: A gentle introduction. Disponible en <http://www.extremeprogramming.org/>. [Consulta en línea: 23 de marzo de 2009].

Bibliografía

- ✓ 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.

- ✓ WEKA, Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java. [Software]. Weka 3.4.7, 2005. <http://www.cs.waikato.ac.nz/ml/weka/> [Consulta en línea: 26 de enero de 2009].

- ✓ Acosta, R., Vázquez, L. Obtención de patrones y reglas en el Sistema Docente del Instituto Superior Politécnico José A. Echevarría (CUJAE) utilizando Minería de Datos. Trabajo de Diploma. Coautores Brito, R.; Rosete, A. Ciudad de La Habana, Cuba. Junio, 2007.

- ✓ Tang, Z. and J. MacLennan (2005). Data Mining with SQL Server 2005. United States of America, Wiley Publishing, Inc.

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

- ✓ Valenga, F., Perversi, I., Fernández, E., Merlino, H., Rodríguez, D., Britos, P., García-Martínez, R. Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina. XIII Congreso Argentino de Ciencias de la Computación. 2007.

- ✓ Rosete Suárez, A. Minería de Datos: el camino de la academia a la realidad cotidiana. Cuba, Ciudad de la Habana, Instituto Superior Politécnico “José Antonio Echeverría” (CUJAE), Centro de Estudios de Ingeniería de Sistemas (CEIS), 2004.

- ✓ Inmon, W. H. Building the Data Warehouse, Fourth Edition. Wiley Publishing, Inc. Indianapolis, Indiana, ISBN-13: 978-0-7645-9944-6. 2005

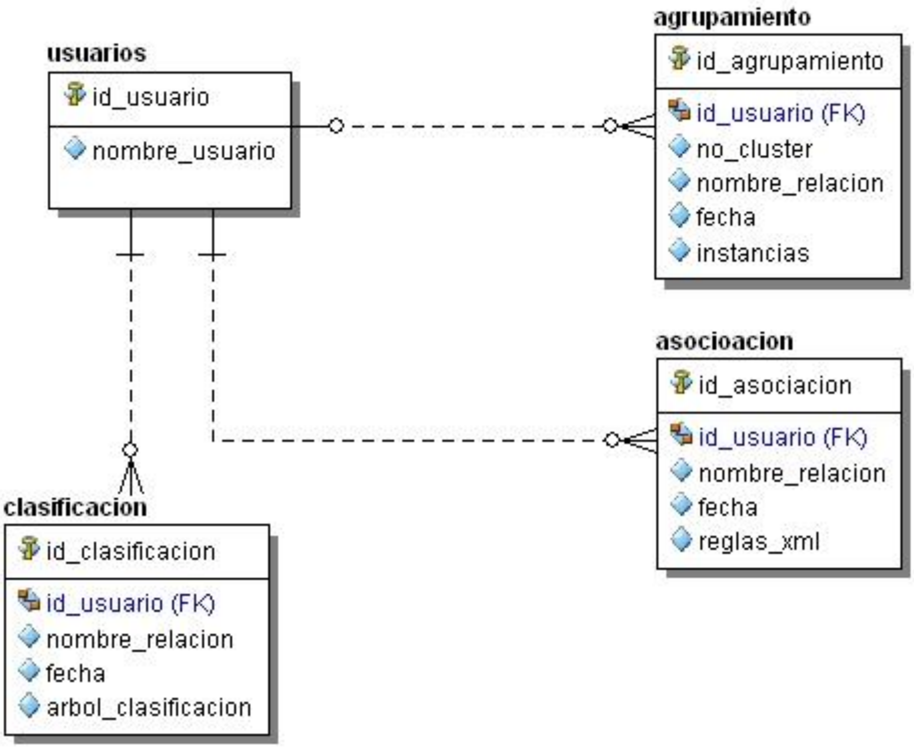
- ✓ Fayyad, U. et al., "Advanced in Knowledge Discovery and Data Mining," MIT Press, MA, 1996.

- ✓ Dunham, M. H. Data Mining. Introductory and Advanced Topics, Prentice Hall, 2003.

- ✓ Chapman, P. [et al.] CRISP-DM 1.0: Step-by-step data mining guide. USA: SPSS Inc., CRISP-DM Consortium, 2000.
- ✓ Brito, R., Rosete, A., Acosta, R. Minería de Datos aplicada a la Gestión Docente: Agrupamiento, Reporte 2008-7, 20/5, Reporte de Investigaciones, Centro de Estudios de Ingeniería y Sistemas, Facultad de Ingeniería Informática. 2008.
- ✓ Richardson, W. Clay, Professional Java™, JDK™ 5 Edition, Wiley Publishing, 2005.
- ✓ Akif, Mohammad, Java y XML, ediciones Anaya Multimedia (grupo Anaya, s.a.), 2002.
- ✓ Peak, Patrick , Hibernate Quickly, Manning Publications Co., 2006.

Anexos

Anexo 1: Modelo de Datos para el Repositorio donde se almacenarán los conocimientos obtenidos.



Glosario de Términos

KDD: Knowledge Discovery in Databases, o su traducción al español Descubrimiento de Conocimiento en Bases de Datos, es el proceso en el cual se extraen con el uso de varias técnicas conocimientos a partir de grandes volúmenes de datos, comúnmente se le asocia con el nombre de Minería de Datos.

CRC: Class Responsibility Collaborator (Clase Responsabilidad Colaboradores). Es una técnica informal, utilizada en la metodología XP, que define las responsabilidades y colaboraciones de cada clase a través de todos los escenarios. Fueron introducidas por Kent Beck y Ward Cunningham para enseñar el paradigma orientado a objeto.

WEKA: Waikato Environment for Knowledge Analysis, entorno para Análisis del Conocimiento de la Universidad de Waikato) es un conocido software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. WEKA es un software libre distribuido bajo licencia GNU-GPL.

XML: del inglés Extensible Markup Language (lenguaje de marcas extensible), es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C). Permite definir la gramática de lenguajes específicos, no es realmente un lenguaje en particular, sino una manera de definir lenguajes para diferentes necesidades.

Open Source: es el término con el que se conoce al software distribuido y desarrollado libremente. La idea bajo el concepto open source es sencilla: cuando los programadores pueden leer, modificar y redistribuir el código fuente de un programa, éste evoluciona, se desarrolla y mejora.

Data Warehouses: en español Almacén de Datos, es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. En esencia es el expediente completo de una organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.