UNIVERSIDAD MICHOACANA DE SAN NICOLAS DE HIDALGO

FACULTAD DE INGENIERÍA ELÉCTRICA

División de Estudios de Posgrado

HERRAMIENTA DE MINERÍA DE DATOS BASADA EN EL ALGORITMO C4.5 Y SU APLICACION AL INVENTARIO MULTIFASICO DE LA PERSONALIDAD MINNESOTA 2

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN INGENIERIA ELECTRICA

PRESENTA

JOEL LOAEZA VALERIO

DIRECTOR DE TESIS

JUAN JOSÉ FLORES ROMERO

Morelia, Mich. Septiembre 2009

HERRAMIENTA DE MINERÍA DE DATOS BASADA EN C4.5 Y SU APLICACION AL INVENTARIO MULTIFASICO DE LA PERSONALIDAD MINNESOTA 2

Asesor: Dr. Juan José Flores Romero Elaboró: Ing. Joel Loaeza Valerio

Agosto 2009

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

En la actualidad se lleva a cabo una generación masiva de datos y la velocidad con la que éstos se almacenan es muy superior a la velocidad con la que se analizan. En consecuencia se necesita un análisis de los datos que sea rápido y eficiente, es decir, el descubrimiento automatizado de conocimiento en bases de datos, que consiste en interpretar grandes cantidades de datos en un proceso iterativo e interactivo, para buscar relaciones o patrones que se puedan convertir en conocimiento útil y novedoso.

Entonces, como una alternativa de bajo costo que haga posible el descubrimiento automatizado de conocimiento en bases de datos, en el presente trabajo se propone la implementación de un Agente Inteligente Descubreconocimiento conformado por Agentes Inteligentes Humanos y un Agente Inteligente de Software. Además, para que el Agente Inteligente Descubreconocimiento tenga éxito en su tarea, en el presente trabajo se desarrollaron también, un algoritmo de referencia que muestra como realizar el descubrimiento automatizado de conocimiento en bases de datos y la aplicación computacional correspondiente denominada Ambiente Descubreconocimiento con el algoritmo C4.5Web (ADC4.5Web). El algoritmo de referencia para descubrir conocimiento, se desarrolló tomando como base el algoritmo propuesto por Fayyad et al (1996) para el proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases: KDD) y el modelo de referencia Proceso Estándar Industrial Híbrido para minería de datos (CRoss-Industry Standard Process for Data Mining: CRISP-DM). La herramienta de minería de datos ADC4.5Web, incluye una adecuación para web del algoritmo C4.5 (algoritmo que genera árboles de decisión a partir de ejemplos) desarrollado por el Dr. J. Ross Quinlan. ADC4.5Web tiene como objetivo proporcionar al Agente Inteligente Descubreconocimiento un ambiente de manejo simple que le permita llevar a cabo el proceso completo de KDD: preparación, análisis, interpretación y visualización de los datos.

Para evaluar la eficacia del Agente Inteligente Descubreconocimiento propuesto, así como la del algoritmo de referencia y la herramienta de minería de datos desarrollados, éstos se aplican en la obtención experimental de una alternativa reducida del Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2), el cual es un instrumento bastante extenso para el análisis psicométrico de la personalidad del ser humano. La evaluación de la versión experimental reducida del inventario se realiza comparando los resultados obtenidos con ésta y los resultados obtenidos con la versión estándar de dicho inventario, para lo cual se desarrolló adicionalmente, un programa computacional con interfaz visual que permite la aplicación de la versión estándar del inventario en forma individual o colectiva y su respectiva calificación automática (incluyendo gráficas e interpretación).

Abstract

At present there is a massive generation of data and the speed with which they are stored far exceeds the speed with which they are analyzed. Therefore is need a data analysis that is fast and efficient, ie, automated knowledge discovery in databases, which is to interpret large amounts of data in an iterative and interactive process, to find relationships or patterns that can make new and useful knowledge.

Then, as a low cost alternative that enables automated knowledge discovery in databases, this paper proposes the implementation of an Intelligent Agent Discoverknowledge comprises human intelligence agents and an Intelligent Agent Software. Furthermore, if the Intelligent Agent Discoverknowledge succeed in his task, in this study also developed an algorithm that shows how you reference the automated knowledge discovery in databases and applying computational Discoverknowledge Environment called for the algorithm C4.5Web (ADC4.5Web). The baseline algorithm to discover knowledge, was developed based on the algorithm proposed by Fayyad et al (1996) for the process of knowledge discovery in databases (Knowledge Discovery in Databases: KDD) and the Reference Model Standard Industrial Process hybrid data mining (Cross-Industry Standard Process for Data Mining: CRISP-DM). he data mining tool ADC4.5Web includes a fitness for web C4.5 algorithm (algorithm that generates decision trees from examples), developed by Ph. Dr. J. Ross Quinlan. ADC4.5Web aims to provide the Intelligent Agent Discoverknowledge a simple environment that enables it to carry out the complete KDD process: preparation, analysis, interpretation and visualization of data.

To evaluate the effectiveness of the proposed intelligent agent Discoverknowledge and the reference algorithm and data mining tool developed, they are applied in obtaining an experimental alternative reduced Minnesota Multiphase Personality Inventory 2 (MMPI-2) , which is quite extensive for a psychometric analysis of the personality of man. The evaluation of the pilot version of the inventory is reduced by comparing the results obtained with this and the results obtained with the standard version of the inventory, which was developed further, a computerized visual interface that allows the implementation of the standard version inventory individually or collectively, and their automatic qualification (including graphics and interpretation).

Capítulo 1

Introducción

Los avances en la tecnología de bases de datos y técnicas de recolección de información han permitido almacenar grandes cantidades de datos en las bases de datos. Este crecimiento explosivo de los datos ha generado la necesidad de observarlos desde diferentes puntos de vista, para encontrar otra información que pudiera estar oculta en estos datos almacenados. Lo anterior ha llevado al desarrollo de un nuevo campo de investigación denominado "Descubrimiento del Conocimiento desde los datos", siendo sus denominaciones más comunes en la actualidad "Minería de Datos" (Data Mining: DM) o "Descubrimiento del Conocimiento en Bases de Datos" (Knowledge Discovery in Databases: KDD) [Fayyad et al., 1996], de manera práctica se puede considerar a estas denominaciones como sinónimos, por lo que en lo sucesivo en este documento se usarán de manera indistinta.

El problema básico del proceso de KDD es pasar de los datos simples almacenados en las bases de datos a la obtención del conocimiento contenido en los mismos. KDD consiste en la aplicación de algoritmos, técnicas de reconocimiento de patrones, análisis estadístico y otros métodos de extracción del conocimiento, sobre grandes volúmenes de datos, que permitan descubrir relaciones ocultas entre los mismos, para que ese conocimiento obtenido se pueda aplicar a la resolución de problemas.

Existen diversas técnicas utilizadas para KDD, entre las que se destacan las técnicas basadas en estadística, las redes neuronales, el aprendizaje automatizado y las técnicas matemáticas, etc. La calidad de la minería de la información está en función de la efectividad de la técnica usada, el tipo y cantidad de los datos minados.

KDD puede ser utilizado en aplicaciones relacionadas con análisis de riesgo, detección de fraudes, análisis de mercados, finanzas, medicina, psicología, telecomunicaciones, entre otras.

Justificación

El abaratamiento de los costos y los avances tecnológicos en el almacenamiento de información, son algunos de los factores que han contribuido a la generación masiva de datos, los cuales son almacenados en bases de datos pertenecientes a empresas, instituciones, organizaciones o usuarios. Entonces, ante la existencia de una gran cantidad de datos, se han desarrollado aplicaciones computacionales

propietarias para analizarlos, las mas económicas permiten obtener información en forma de resumen, las de precio medianamente elevado, proporcionan soporte para realizar un análisis descriptivo de los datos obteniendo otros con algún valor agregado y las altamente costosas, brindan la posibilidad de realizar el proceso completo para descubrimiento de conocimiento en bases de datos (KDD) y obtener conocimiento que se pueda aplicar a otros datos, éstas últimas pueden ser de gran utilidad en las actividades de investigación o de negocios. Cabe mencionar, que también existen algunas aplicaciones computacionales de código abierto y libre distribución para realizar KDD, pero éstas no permiten la preparación de los datos directamente desde las bases de datos.

El gran potencial de las aplicaciones propietarias capaces de obtener conocimiento a partir de los datos es evidente, pero debido a su alto costo, no están al alcance de la mayoría de las empresas y usuarios del país, por lo que esta situación motiva al desarrollo de esta tesis, con el fin de obtener una herramienta implementada a partir de aplicaciones de código abierto y libre distribución, para realizar el proceso completo de KDD que permita la preparación de los datos directamente desde las bases de datos y el algoritmo de referencia que oriente al personal encargado sobre como llevar a cabo el proceso de KDD, que conformen una alternativa de KDD a bajo costo para los investigadores, científicos o expertos en algún área de actividad humana.

1.1. Objetivos

Objetivos Generales

- Proposición de un Agente Inteligente Descubreconocimiento capaz de llevar cabo el proceso completo de descubrimiento de conocimiento en bases de datos.
- Implementación de una aplicación computacional de bajo costo basada en el algoritmo C4.5 [Quinlan, 1993] (algoritmo que genera árboles de decisión a partir de ejemplos), para proporcionar una herramienta de minería de datos a los investigadores o personas relacionadas con los negocios, que les permita llevar a cabo en un ambiente web el proceso completo de descubrimiento de conocimiento en bases de datos (Minería de Datos), el que incluye la preparación, análisis, interpretación y visualización de los datos.
- Desarrollo de un algoritmo de referencia para minería de datos que oriente a los usuarios para descubrir conocimiento en forma automatizada en bases de datos (Minería de Datos).
- Evaluación de la eficacia del Agente Inteligente Descubreconocimiento propuesto, del algoritmo de referencia de minería y la herramienta computacional para descubrimiento de conocimiento desarrollados, aplicándolos a la obtención experimental de una versión reducida de la prueba psicométrica Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2)[Lucio y León, 2003].

Objetivos Específicos

- Proposición del Agente Inteligente Descubreconocimiento conformado por Agentes Inteligentes Humanos que interactúen con un Agente Inteligente de Software.
- Adecuación del algoritmo C4.5 [Quinlan, 1993] (generador de árboles de decisión a partir de ejemplos), en la implementación de la herramienta computacional de bajo costo que proporcione un ambiente web para descubrimiento de conocimiento en bases de datos (Ambiente Descubreconocimiento con C4.5 para Web: ADC4.5Web).
- Desarrollo de un algoritmo de referencia para Minería de Datos basado en el algoritmo de Minería de Datos propuesto por Fayyad et al (1996) y el modelo de referencia para Minería de Datos CRISP-DM (CRoss-Industry Standard Process for Data Mining: Proceso Estándar Industrial Híbrido para la Minería de Datos).
- Preparación de los datos minables a partir de las bases de datos administradas con el sistema manejador de bases de datos MySQL, por medio de la herramienta ADC4.5Web.
- Interpretación visual del conocimiento descubierto haciendo que ADC4.5Web genere los Árboles de Decisión en el formato gráfico JPEG.
- Implementación de la herramienta computacional para la aplicación y calificación estandarizada de la prueba psicométrica Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2).
- Obtención de una versión experimental reducida de la prueba psicométrica MMPI-2.
- Comparación de la versión reducida del MMPI-2 obtenida experimentalmente con la versión estandarizada, para evaluar la eficacia del Agente Inteligente Descubreconocimiento propuesto, así como la del algoritmo de referencia y la herramienta de minería desarrollados.
- Interpretación de los datos obtenidos en forma conjunta con los expertos del dominio (prueba psicométrica MMPI-2).

1.2. Alcances y limitaciones

- Para generar los árboles de decisión a partir de ejemplos se emplea únicamente el algoritmo C4.5[Quinlan, 1993].
- El conjunto total de ejemplos para el entrenamiento y evaluación del algoritmo C4.5 fue de 1395 casos de diagnósticos del MMPI-2.
- La herramienta computacional de Minería de Datos desarrollada es para uso general (negocios, medicina, finanzas, psicología, etc.), aunque seguramente resultará limitada en algunos casos de aplicación.

1.3. Descripción de la tesis

El capítulo 2 presenta algunos conceptos básicos sobre Agentes Inteligentes, que facilitan la comprensión de los Agentes Inteligentes Humanos y el Agente Inteligente de Software que conforman al Agente Inteligente Descubreconocimiento, capaz de llevar a cabo el proceso completo de descubrimiento de conocimiento en bases de datos, por medio de la utilización de la herramienta de minería de datos ADC4.5Web descrita en el capítulo 7. El Agente Inteligente Descubreconocimiento constituye una de las principales aportaciones del presente trabajo, por lo que la descripción detallada de sus elementos se presenta en los capítulos 3, 4, 5 y 7.

El capítulo 3 describe el aprendizaje humano y computacional, puesto que el AI Descubreconocimiento requiere ser capaz de percibir, aprender, representar conocimiento, tomar decisiones, actuar y evaluar su comportamiento.

El capítulo 4 describe el aprendizaje a partir de ejemplos que es el criterio seguido por el AI de Software (algoritmo C4.5 [Quinlan, 1993]) que conforma al AI Descubreconocimiento, donde el conocimiento adquirido se puede representar como un Árbol de Decisión.

En el capítulo 5 se describe lo que es la Minería de Datos y se propone un algoritmo de referencia para minería de datos (otra aportación importante del presente trabajo), que sirva de guía para que el AI Descubreconocimiento pueda realizar exitosamente el proceso de descubrimiento de conocimiento en bases de datos.

El capítulo 6 presenta el Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2) y su aplicación computacional correspondiente (también aportación importante del presente trabajo). Con la finalidad de emplearlo como caso de estudio para evaluar la eficacia del AI Descubreconocimiento propuesto, el algoritmo de referencia y la herramienta de minería ADC4.5Web desarrollados.

El capítulo 7 describe otra de las aportaciones más importantes del presente trabajo: la herramienta de minería de datos ADC4.5Web que facilita la interacción entre las bases de datos MySQL con los Agentes Inteligentes Humanos (descritos en los capítulos 2 y 3) y con el Agente Inteligente de Software (descrito en los capítulos 2, 3, 4 y 5), haciendo posible la conformación del Agente Inteligente Descubreconocimiento (descrito en los capítulos 2, 3, 4, 5 y 7) capaz de realizar tareas de descubrimiento de conocimiento en bases de datos, empleando como guía metodológica el algoritmo de minería de datos propuesto en el capítulo 5. ADC4.5Web hace posible que se cumpla el objetivo principal del presente trabajo: proporcionar a bajo costo una herramienta de minería de datos a los investigadores o a las personas relacionadas con los negocios.

El capítulo 8 presenta las conclusiones y aportaciones de este trabajo.

Capítulo 2

Agente Inteligente

En este capítulo se describe en forma breve qué es un Agente Inteligente, algunas de sus características y su comportamiento, que hacen posible que éstos agentes sean capaces de razonar y algunos incluso de aprender. Se describe también una de las principales aportaciones de ésta tesis: el Agente Inteligente Descubreconocimiento capaz de llevar a cabo el proceso completo de Descubrimiento de Conocimiento en Bases de Datos, conocido también como Minería de Datos.

2.1. Agente Inteligente

La inteligencia natural con la que el ser humano ha sido dotado por la naturaleza es impresionante, aún así, se puede complementar con el uso de la inteligencia artificial, aclarando que el objetivo de la IA como ciencia no es el de crear un sustituto para la inteligencia humana natural, sino crear herramientas útiles al ser humano cuando éste realiza algunas actividades de tipo cognoscitivo. Estas herramientas son los Agentes Inteligentes.

Un Agente Inteligente puede definirse como lo hacen Russell y Norvig (1996):

> "Se considera un **agente inteligente** a todo aquello que **percibe** su ambiente mediante **sensores** y que **responde o actúa** en tal ambiente por medio de **actuadores**. Los *agentes humanos* (**humanos**)poseen ojos, oídos, nariz, boca y otros órganos que les sirven de sensores, así como manos, piernas y otras partes de su cuerpo que les sirven de actuadores. En los *agentes robóticos* (**robots**), los sensores son implementados con cámaras, micrófonos, sensores infrarojos, láser, sistemas de comunicación alámbricos e inalámbricos y los actuadores son implementados por motores y mecanismos.Un *agente de software* (**softbot**), *percibe y actúa* por medio de *las cadenas de bits codificados*, es decir, la información de entrada/salida procesada por la computadora."

Dicho en forma resumida, un *agente inteligente* es toda entidad capaz de percibir y actuar. Un agente inteligente puede ser *agente individual* si habita individualmente en un ambiente o mundo, en caso contrario, será un *multiagente* si es capaz de interactuar con otros agentes que habitan en

el mismo mundo o bien si está conformado por varios agentes. En el caso de esta tesis, el agente inteligente utilizado es en realidad un multiagente, pues está conformado por agentes humanos que interactúan con un agente de software, lo cual le permite la interacción con otros agentes o multiagentes inteligentes, esto se explica en la sección 2.4.

2.2. Ambiente

El **ambiente** es el **entorno** o **mundo** en el cual habita un agente inteligente. La relación que existe entre ellos es: el ambiente aporta percepciones al agente, el que a su vez, ejerce acciones sobre el ambiente. Dicha relación es la misma en todos los casos. El ambiente puede ser artificial o real.

Algunas características de los ambientes son:

- Accesibles y no accesibles o parcialmente accesibles. Un ambiente es accesible a un agente, si el aparato sensorial de éste le permite percibirlo completamente, en caso contrario es un ambiente no accesible o parcialmente accesible al agente.
- Deterministas y estocásticos. Si el estado siguiente de un ambiente se determina por medio del estado actual y las acciones elegidas por el agente, entonces el ambiente es determinista, en caso contrario es estocástico.
- Estáticos y dinámicos. Si el ambiente no cambia mientras el agente razona que hacer, se considera estático en caso contrario si cambia o puede cambiar se considera dinámico.
- Discretos y continuos. Cuando la cantidad de percepciones es limitada y las acciones son claramente diferenciables, se dice que el ambiente es discreto, en caso contrario, el ambiente es continuo.
- Ambiente de agente individual o multiagente. Un ambiente es de agente individual si lo habita sólo un agente, en caso contrario si es habitado por varios agentes es un ambiente multiagente.

Un ambiente puede presentar una característica o una combinación de las mismas. Para el caso del presente trabajo el ambiente es el mundo real presentando todas las características mencionadas y muchas otras más, en este ambiente habita el multiagente inteligente Descubreconocimiento propuesto.

2.3. Proceder de un Agente Inteligente

Lo deseable es que el agente inteligente proceda en forma racional, por lo tanto de esto deriva que un **agente racional** es aquel que razona antes de actuar, en consecuencia, resulta necesario que el agente pueda *medir su desempeño* y de esta forma evaluar su actuación. Resulta importante que la conducta de un agente racional se base tanto en el conocimiento incluido o integrado en su construcción, como en el conocimiento adquirido por medio del aprendizaje, de tal manera que dicha

conducta sea adecuada al ambiente específico en el cual va a operar. Esto es similar al hecho de cómo la naturaleza provee a los animales con reflejos incorporados, con la finalidad de que sobrevivan hasta que sean capaces de aprender por sí mismos.

De ésto se concluye que el proceder de un agente racional consiste primeramente en percibir al ambiente, representar las percepciones como conocimiento o aprender de estas percepciones y representarlas como conocimiento, para finalmente poder razonar sobre éste y decidir como actuar, además el agente racional debe ser capaz de evaluar su desempeño para determinar si su actuación resulta apropiada o debe ser mejorada.

2.4. Agente Inteligente Descubreconocimiento

En el presente proyecto se propone que el agente inteligente capaz de realizar el proceso completo de descubrimiento automatizado de conocimiento (Minería de Datos), es decir, el *Agente Inteligente Descubreconocimiento*, sea en realidad un multiagente conformado por uno o más Agentes Inteligentes Humanos que interactúan con un Agente Inteligente de Software (agente inteligente para Minería de Datos), puesto que se requiere que el AI Descubreconocimiento habite en el mundo real, ya que está concebido para realizar tareas de investigación científica o de negocios. Como dichas tareas son propias de los humanos, el AI Descubreconocimiento debe ser capaz de interactuar con estos agentes humanos, objetivo que logra al poseer una parte humana en su conformación. Resulta recomendable que el equipo heterogéneo de personal que conforman la parte humana del AI Descubreconocimiento presenten los siguientes perfiles: computación, estadística y el área de aplicación y que al menos se cuente con una persona de cada perfil en la conformación del AI Descubreconocimiento.

El personal con perfil en computación, debe tener conocimientos sobre Minería de Datos, bases de datos relacionales, programación e interfases visuales; se requiere al menos una persona con dicho perfil. Este perfil proporciona la capacidad técnica para el manejo de tecnologías de la información y el aprendizaje automatizado. Las personas con perfil en estadística proporcionan el rigor sobre las tareas de limpieza y selección de los datos por analizar, así como en el análisis exploratorio de los mismos y en la validación de los modelos obtenidos. Se puede prescindir de elementos con este perfil y en consecuencia, el personal con perfil en computación y/o área de aplicación deberán realizar estas tareas. Las personas con el perfil del área de aplicación son requeridas para que proporcionen el conocimiento experto sobre el dominio de aplicación, siendo imprescindible su participación en la selección y exploración de los datos y en la validación de los modelos obtenidos. El equipo de personal con los tres perfiles mencionados interactuando con un AI de Software conforman al AI Descubreconocimiento como se muestra en la figura 2.1. En el siguiente capítulo se presentan diagramas del agente inteligente propuesto y se describe éste con más detalle.

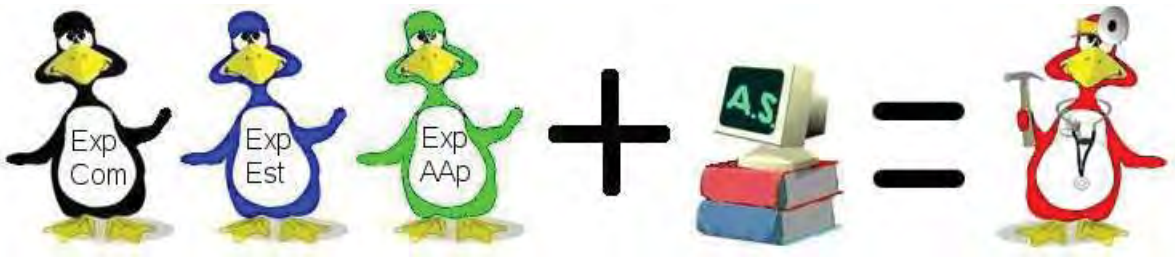


Figura 2.1: HumanoPerfilComputación + HumanoPerfilEstadística + HumanoPerfilAreaAplicación + AgenteInteligenteSoftware = AgenteInteligenteDescubreConocimiento.

El AI Descubreconocimiento propuesto, pone de manifiesto la importancia imprescindible del ser humano en el proceso de descubrimiento automatizado de conocimiento, en el cual interactúa con herramientas útiles de IA, las cuales no se pueden considerar en ningún momento como sustitutas del ser humano, puesto que carecen de autonomía y requieren al ser humano para poder operar, estas herramientas de IA son en realidad un buen complemento para el ser humano cuando éste realiza algunas actividades de tipo cognitivo.

De acuerdo con lo anterior, se puede definir que el Agente Inteligente Descubreconocimiento es aquel que se conforma por las personas que realizan actividades de investigación científica o de negocios y que cuentan con una herramienta que les auxilia en su búsqueda de conocimiento, éste el objetivo del presente trabajo de tesis: dotar de una herramienta de minería de datos a los investigadores o a las personas relacionadas con los negocios, para concebir en forma real al AI Descubreconocimiento.

Cabe mencionar, que en el ambiente de la Minería de Datos se emplea el término “minero” para referirse a las personas que realizan alguna tarea de minería, pero éste resulta ambiguo en el sentido de que no se especifica si es sólo una persona la que debe realizar el proceso de minería o deben ser más, así como tampoco se define el perfil de esta persona, aunque la idea generalizada es la de una persona con perfil en computación que emplea algoritmos o herramientas de minería. Por esta razón, en el presente trabajo se propone al AI Descubreconocimiento especificando como debe conformarse, puesto que de esto depende en buena medida que se tenga éxito o no en una tarea de minería.

2.5. Conclusión del capítulo

Se considera que los Agentes Inteligentes pueden ser herramientas útiles al ser humano, cuando éste realiza algunas actividades de tipo cognoscitivo y se aclara que no son un sustituto para la inteligencia humana natural, puesto que carecen de autonomía y requieren al ser humano para poder operar, este criterio permite concebir al Agente Inteligente Descubreconocimiento propuesto, capaz de realizar el proceso completo de descubrimiento automatizado de conocimiento (Minería de Datos) y enfatiza la importancia imprescindible del ser humano en dicho proceso, en el cual, al interactuar con el Agente Inteligente de Software conforma en realidad al Multiagente Inteligente Descubreconocimiento, que resulta del hecho de proporcionar una herramienta de minería de datos a los investigadores o a las personas relacionadas con los negocios.

Capítulo 3

Aprendizaje Humano y Aprendizaje Computacional

En el capítulo anterior se menciona que el Agente Inteligente Descubreconocimiento es en realidad un multiagente conformado por uno o más Agentes Inteligentes Humanos y un Agente Inteligente de Software. Puesto que el AI Descubreconocimiento está concebido para realizar el proceso completo de descubrimiento automatizado de conocimiento, como herramienta en las tareas de investigación científica o de negocios, en consecuencia, es necesario que éste agente posea una gran velocidad y capacidad para procesar y almacenar información y que habite en el mundo real, por lo tanto, se requiere que dicho agente sea capaz de percibir, aprender, representar conocimiento, tomar decisiones, actuar y evaluar su comportamiento. Estas características son proporcionadas al AI Descubreconocimiento por los AI Humanos y se complementan al interactuar con el AI de Software. Por esta razón se describen en este capítulo el *aprendizaje humano* y el *aprendizaje computacional* que en conjunto conforman al aprendizaje del AI Descubreconocimiento.

3.1. Aprendizaje del Agente Inteligente Humano

El AI Humano percibe el mundo por medio de su aparato sensorial obteniendo una secuencia de percepciones, de las cuales se hace distinción entre *percepciones externas* (vista, oído, olfato, gusto, tacto, etc) y *percepciones internas* (instinto, conciencia de sí mismo, sentimientos, memoria, etc). Estas percepciones se representan en un lenguaje interno de acuerdo con la idea de algunos filósofos, psicólogos y lingüistas, de que el ser humano y otros animales pueden ser considerados como máquinas para el procesamiento de información. Lo anterior es denominado en IA como representación del conocimiento, donde dicha representación puede ser almacenada como conocimiento o bien razonar a partir de ésta, de esto se concluye, que al proceso conjunto de percibir, representar el conocimiento, almacenar conocimiento y/o razonar sobre éste, es lo que se considera como *aprendizaje*. Para los AI Humanos la toma de decisiones puede obedecer al aprendizaje y por lo tanto actuar en forma racional, o a las percepciones y consecuentemente actuar en forma instintiva o emocional, o bien obedecer al aprendizaje y a las percepciones en forma conjunta. Basándose en esta toma de decisiones

el AI Humano actúa con sus actuadores externos (manos, pies, etc) y/o actuadores internos (cerebro, sistema nervioso, etc). Otro aspecto importante también, es que el AI Humano es capaz de evaluar su desempeño o comportamiento tanto de manera interna (autoevaluación) y/o externa (normas o leyes de la sociedad en la que vive), es decir, en base a un criterio propio y/o de otros, este criterio puede ser en relación a la utilidad y/o satisfacción proporcionada por su comportamiento. Tomando como referencia la evaluación de su desempeño, el AI Humano puede retroalimentarse y evolucionar obteniendo más conocimiento, para mejorar su actuación futura en el mundo que habita.

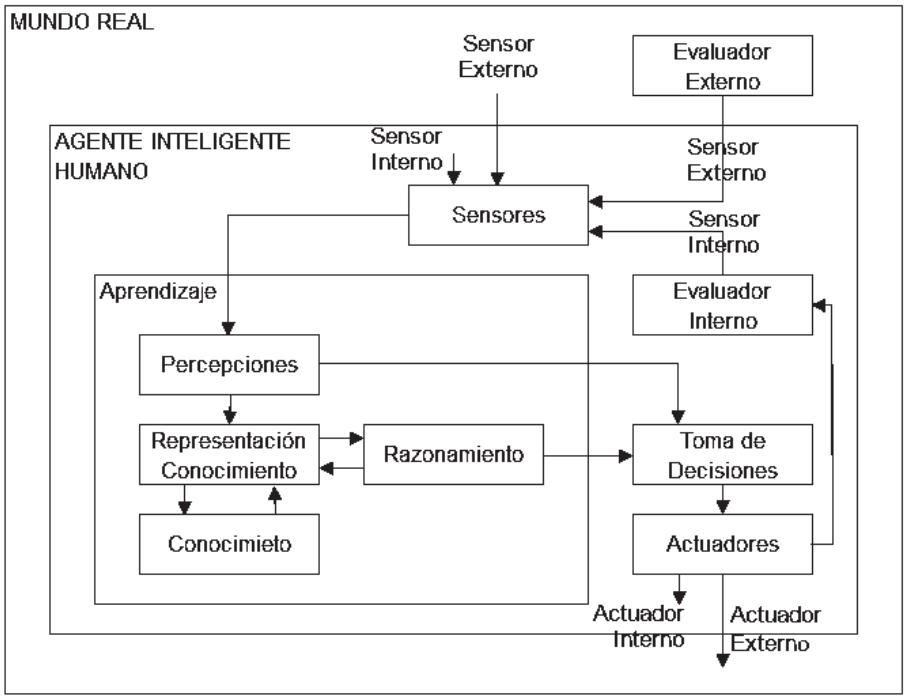


Figura 3.1: Agente Inteligente Humano.

En la figura 3.1 se muestra el diagrama del Agente Inteligente Humano. En este diagrama se puede observar que el AI Humano reside en el mundo real, al que puede percibir como ya se mencionó, por medio de su aparato sensorial, distinguiendo entre percepciones externas (del mundo real) y percepciones internas (de sí mismo). Estas percepciones se pueden representar como conocimiento y razonar sobre él o almacenar como conocimiento, el AI Humano tiene acceso a este conocimiento almacenado al que puede volver a representar y razonar sobre él, el razonamiento también puede ser representado como conocimiento y almacenado como tal, por lo que como ya se comentó anteriormente, a este proceso conjunto es a lo que se denomina **aprendizaje humano**, como se puede observar en el elemento Aprendizaje de la figura 3.1. En esta misma figura 3.1 se aprecia que el AI Humano puede tomar decisiones a partir de las percepciones y actuar en forma instintiva o emocional, o bien, el AI Humano puede tomar decisiones a partir del razonamiento y actuar en forma racional, o también, el AI Humano puede tomar decisiones a partir de las percepciones y del razonamiento y actuar en forma emocional o racional según sea el balance final entre las percepciones y el razonamiento. El AI Humano puede actuar con sus actuadores externos sobre el mundo real o actuar con sus actuadores

internos sobre sí mismo. Otro aspecto muy importante que se puede observar también en la figura 3.1, es el hecho de que el AI Humano puede evaluar su actuación o desempeño, ya sea de manera interna (autoevaluación) y/o externa (criterios externos), proporcionando consecuentemente, que el AI Humano pueda retroalimentarse y evolucionar obteniendo más conocimiento, que le permita mejorar su desempeño. La evaluación mencionada permite además, proponer las acciones respectivas de exploración, que es justamente lo que hacen los investigadores al realizar sus experimentos.

3.2. Aprendizaje del Agente Inteligente de Software

En la sección anterior se discute que el aprendizaje es un factor muy importante para el buen desempeño del AI Descubreconocimiento, por lo que el AI de Software es el complemento adecuado para el AI Humano en el proceso de descubrimiento automatizado de conocimiento. Esto implica que el AI de Software debe ser capaz de razonar y de aprender, por lo tanto, debe ser un agente inteligente para Minería de Datos.

En la figura 3.2 se muestra el diagrama del Agente Inteligente de Software (softbot), en dicho diagrama se aprecia que el AI de Software habita en una computadora, éste percibe al mundo por medio de los dispositivos de entrada (teclado, ratón, archivos en disco, etc), obteniendo una secuencia de percepciones externas (entrada de cadenas de bits procesadas), a diferencia del AI Humano, el AI de Software no posee percepciones internas pues carece de sentimientos, conciencia de sí mismo y de instinto. De manera semejante al AI Humano, estas percepciones se pueden representar como conocimiento en algún lenguaje computacional y razonar sobre él o almacenar como conocimiento, el AI de Software tiene acceso a este conocimiento almacenado al que puede volver a representar y razonar sobre él, el razonamiento también puede ser representado como conocimiento y almacenado como tal. Por lo tanto, al proceso conjunto de percibir, representar el conocimiento, almacenar conocimiento y/o razonar sobre éste y ser realizado por un programa de computadora, es lo que se denomina como **aprendizaje computacional**.

En la figura 3.2, se observa que el AI de Software toma decisiones a partir de su razonamiento aprendido, por lo que siempre decide en forma racional y basándose en éstas decisiones, actúa racionalmente (proporcionando una salida de cadenas de bits procesadas) con sus actuadores compuestos por los dispositivos de salida (monitor, impresora, archivos en disco, etc). Cabe mencionar que el AI de Software no es capaz de evaluar su desempeño ni de retroalimentarse o evolucionar por sí mismo, su complemento el AI Humano es quien evalúa su comportamiento y quien lo retroalimenta proporcionándole evolución para mejorar su desempeño.

3.2.1. Aprendizaje computacional a partir de ejemplos

El aprendizaje computacional es un subcampo de la IA que se encarga del estudio de los programas con capacidad de aprender, considerando que lo más importante del aprendizaje es la idea de que las percepciones deben servir no sólo para actuar, sino también para mejorar la capacidad del agente en su actuar futuro. El aprendizaje se produce cuando el agente interactúa con el mundo, dando lugar

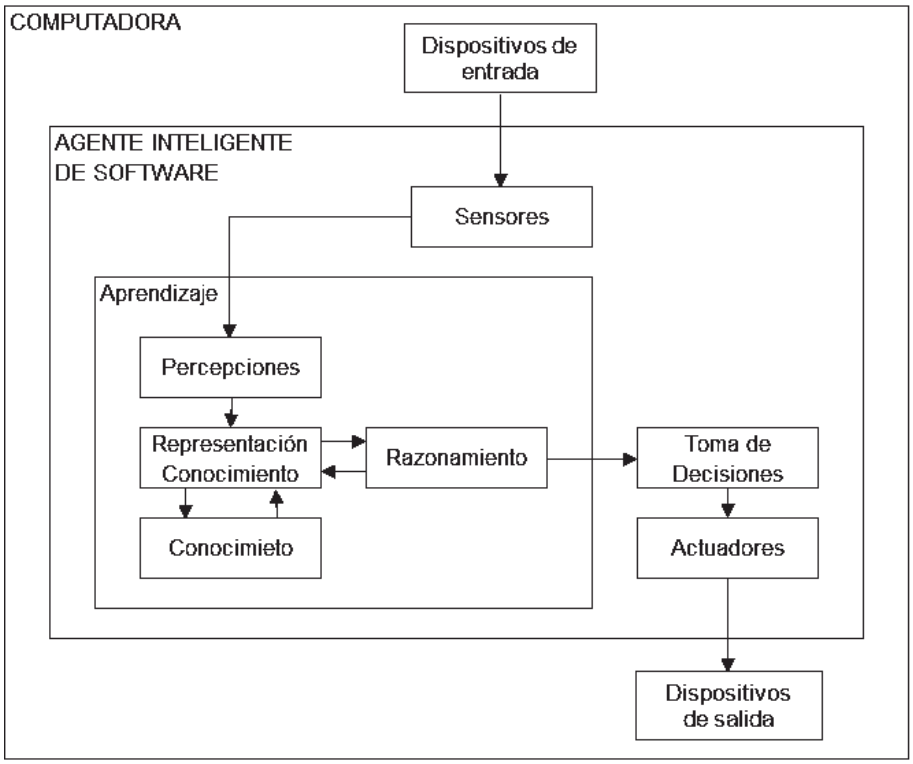


Figura 3.2: Agente Inteligente de Software.

a diferentes tipos de aprendizaje como pueden ser: aprendizaje a partir de ejemplos, aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo, aprendizaje en redes neuronales, etc. El aprendizaje a partir de ejemplos, es el criterio seguido por el AI de Software que conforma al AI Descubreconocimiento propuesto en el capítulo 2.

3.3. Aprendizaje del Agente Inteligente Descubreconocimiento

El AI Descubreconocimiento es en realidad un multiagente, pues está conformado por Agentes Inteligentes Humanos que interactúan en su proceso de aprendizaje con un AI de software, con esto el agente Descubreconocimiento obtiene su gran potencial de aprendizaje, heredando las cualidades de los dos tipos de agentes que lo conforman e inevitablemente también sus defectos. Por lo tanto, el AI Descubreconocimiento es capaz de percibir, representar conocimiento, aprender conocimiento, almacenar conocimiento, razonar sobre conocimiento, tomar decisiones, actuar y evaluar su desempeño, lo cual hace posible que éste realice tareas de investigación científica o de negocios y, posea una gran velocidad y capacidad para procesar y almacenar información, dando como resultado al agente inteligente que sirve de base para el desarrollo e implementación del sistema de descubrimiento de conocimiento automatizado, es decir, el sistema de Descubrimiento de Conocimiento en Bases de Datos también conocido como Minería de Datos.

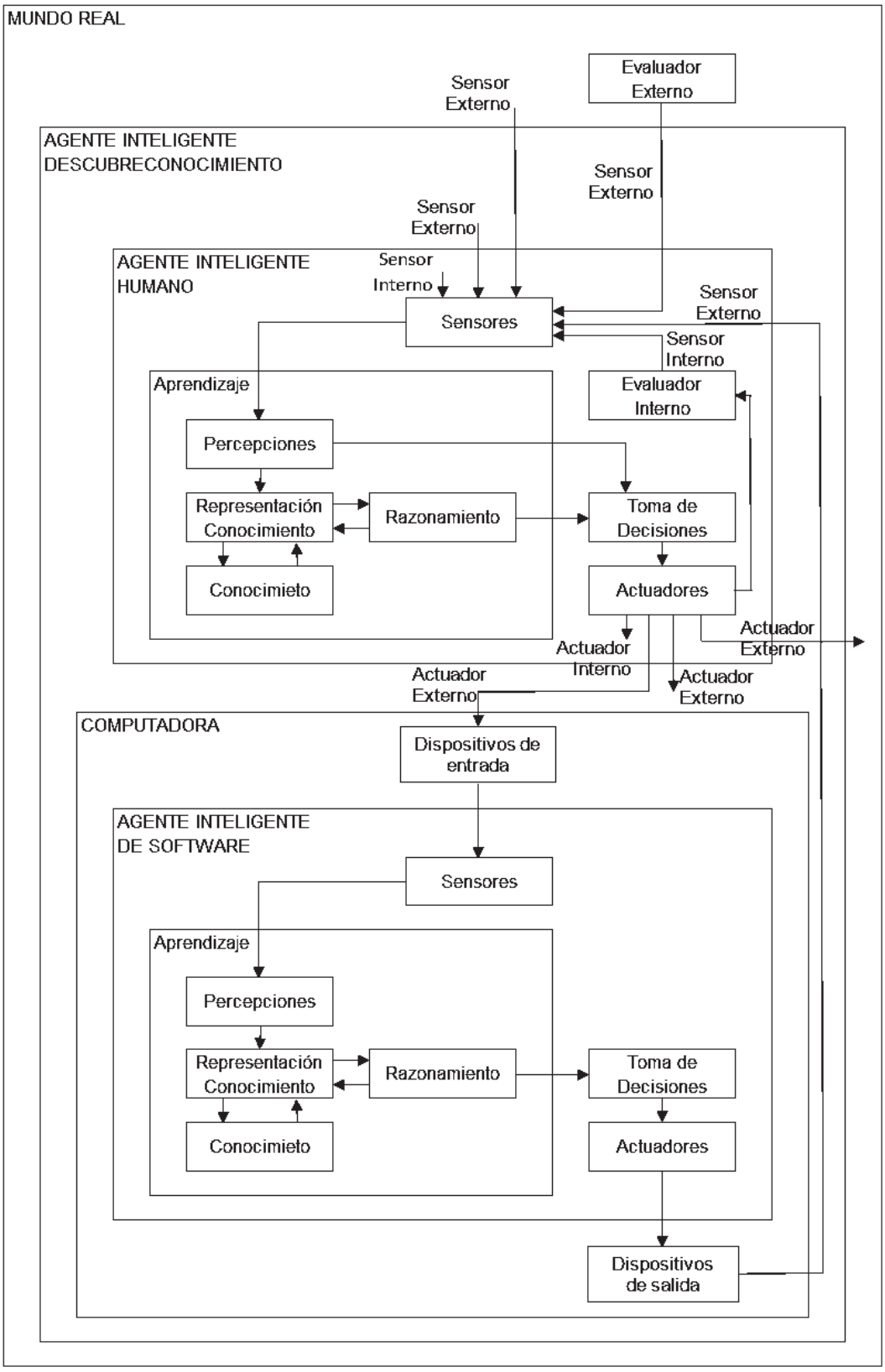


Figura 3.3: Agente Inteligente Descubreconocimiento.

En la figura 3.3 se muestra el diagrama del AI Descubreconocimiento. En este diagrama se puede

apreciar la interacción entre el AI Humano y el AI de Software, de la cual resulta el AI Descubreconocimiento. El AI Humano es el que percibe al mundo real y a sí mismo, por lo tanto, percibe también al AI Descubreconocimiento y al AI de Software. El AI Humano actúa en el mundo real, en sí mismo, en el AI Descubreconocimiento y en el AI de Software. El AI de Software percibe al AI Humano y actúa sobre éste. Con la interacción de ambos agentes se obtienen las percepciones y el actuar del AI Descubreconocimiento.

3.4. Conclusión del capítulo

El aprendizaje es el punto clave para el desarrollo e implementación del AI Descubreconocimiento propuesto. Dado que el AI Descubreconocimiento es un multiagente conformado por Agentes Inteligentes Humanos que interactúan en su proceso de aprendizaje con un AI de Software, con esto obtiene su gran potencial heredando las cualidades de los dos tipos de agentes que lo conforman aunque también sus defectos.

Por lo tanto, el AI Descubreconocimiento es capaz de percibir, representar conocimiento, aprender conocimiento, almacenar conocimiento, razonar sobre conocimiento, tomar decisiones, actuar y evaluar su desempeño, lo que posibilita que éste realice tareas de investigación científica o de negocios y además, posea una gran velocidad y capacidad para procesar y almacenar información, proporcionando al agente inteligente que sirve de base para el desarrollo e implementación del sistema de Descubrimiento de Conocimiento en Bases de Datos también conocido como Minería de Datos.

Capítulo 4

Árboles de Decisión

Como se mencionó en el capítulo anterior, el aprendizaje a partir de ejemplos es el criterio seguido por el AI de Software que conforma al AI Descubreconocimiento. Este se produce cuando el AI de Software construye un modelo a partir de un conjunto de ejemplos de entrenamiento y evaluación. Esto significa que se trata de aprendizaje inductivo (adquisición de conocimiento para aplicación general obtenido a partir de información particular), en el que el conocimiento adquirido se representa como un Árbol de Decisión (Decision Tree). Por tal motivo el aprendizaje computacional con Árboles de Decisión se describe en este capítulo.

4.1. Árbol de Decisión

En el ámbito de la Inteligencia Artificial, un Árbol de Decisión es un *modelo de predicción*, construido por un algoritmo de aprendizaje a partir de un conjunto de ejemplos de entrenamiento y evaluación, donde el modelo obtenido sirve para representar una serie de condiciones sucesivas para la resolución de un problema. El Árbol de Decisión puede recibir como entrada valores discretos o continuos. Cuando los valores de entrada son discretos, el problema a resolver se denomina *clasificación* (es el caso más común de aplicación) y cuando los valores de entrada son continuos, el problema a resolver se denomina *regresión.* De los métodos de aprendizaje conocidos, quizás los basados en árboles de decisión son los más fáciles de utilizar y de entender.

4.1.1. Datos de entrada

Los datos de entrada pueden ser numéricos o no numéricos. Para distinguir entre ambos tipos generalmente se usan los términos *datos cuantitativos* y *datos cualitativos* [Triola, 2004]. Los datos cuantitativos son números que representan mediciones o conteos. Los datos cualitativos (categóricos o de atributo) son aquellos que se pueden dividir en diferentes categorías o clases distinguiéndose por alguna característica no numérica.

Los datos cuantitativos son los que se utilizan como entrada para el algoritmo de aprendizaje que genera el Árbol de Decisión. Estos se distinguen en *datos discretos* y *datos continuos*. Los datos

discretos son aquellos en los que el número posible de valores es un número finito (ej. la cantidad de huevos que pone una gallina) y los datos continuos (numéricos) son aquellos que resultan de un número infinito de posibles valores que se pueden asociar a los puntos de una escala continua, en un rango de valores sin espacios ni interrupciones (ej. la cantidad de leche que produce una vaca, cuya medición puede tomar cualquier valor dentro de un rango o intervalo continuo).

Existe otra clasificación común de los datos denominada uso de niveles de medición: nominal u ordinal. Los *datos nominales* consisten exclusivamente en nombres, etiquetas, categorías o clases que no pueden ordenarse de acuerdo a un esquema. Los *datos ordinales* son similares a los nominales con la diferencia de que estos si se pueden ordenar en base a un esquema (ej. bajo, medio, alto).

Existen otras nomenclaturas para definir los tipos de datos, aunque para el propósito de los datos de entrada del algoritmo de aprendizaje con los tipos mencionados es suficiente y además, existe la flexibilidad para referirse a los datos nominales y ordinales como de tipo discreto.

4.1.2. Requerimientos de los datos de entrada

Enseguida se listan los requerimientos que deben cumplir los datos de entrada para generar un Árbol de Decisión de acuerdo con los autores [Mitchell, 1997],[Quinlan, 1990] y [Quinlan, 1993b].

- **Descripciones de atributo-valor**.- Los datos que se analizarán, serán una colección de ejemplos, donde cada ejemplo debe poder expresarse en términos de un conjunto fijo de atributos o propiedades. Los atributos pueden ser discretos o numéricos. Los atributos que se utilicen para describir un ejemplo deben ser los mismos para todos los casos.
- **Clases predefinidas**.- Las clases o categorías con las que se clasifican los ejemplos se deben establecer previamente, esto implica que el conjunto de datos de entrenamiento y de prueba deben estar previamente clasificados. A esto es a lo que se denomina *aprendizaje supervisado*.
- **Clases discretas y disjuntas**.- Las clases que clasifican a los ejemplos deben ser disjuntas, esto significa que un ejemplo pertenece o no a una clase, pero no a más de una a la vez. Las clases deben ser discretas y en el caso de que sean continuas se deben discretizar.
- **Datos suficientes**.- La cantidad de ejemplos debe ser suficiente, para que los patrones generados sean válidos.
- **Ejemplos de entrenamiento preferentemente sin errores y sin valores de atributo faltantes**.- Aunque los algoritmos que generan Árboles de Decisión son robustos frente a los errores en los valores de atributos y clases y a los valores de atributo faltantes, es recomendable que los datos sean lo más consistentes en la medida de lo posible.

4.1.3. Salida generada

Como ya se mencionó anteriormente, el algoritmo de aprendizaje es en el presente trabajo el AI de Software. Está implementado a partir del algoritmo C4.5, el cual genera modelos de clasificación

que pueden ser representados como Árboles de Decisión. Estos modelos se limitan a la descripción de clases como una expresión lógica cuyas primitivas son afirmaciones de los valores de atributos específicos. Un Árbol de Decisión representa una expresión lógica en forma de una disyunción de conjunciones. Los casos que requieran un modelo de naturaleza diferente no podrán ser analizados por el algoritmo C4.5.

Los Árboles de Decisión representan una estructura de datos que organiza en forma eficaz al nodo raíz y los nodos internos como descriptores (atributos) de un caso o ejemplo. El árbol se construye de tal forma que en el nodo raíz y en los nodos internos se realiza una prueba sobre el valor de los descriptores y según sea la respuesta se va descendiendo a través de las ramas, hasta llegar a donde se encuentra el valor del nodo hoja clasificador como se puede apreciar en la figura 4.1.

4.2. Elementos de un Árbol de Decisión

Los elementos de un Árbol de Decisión son los siguientes [Russell y Norvig, 2004]: raíz, nodos, ramas y hojas. El **nodo raíz** y los **nodos internos** del árbol corresponden a una *prueba del valor* de una de las propiedades y las **ramas del nodo** son identificadas mediante los *posibles valores* de la prueba. En los **nodos hoja** del árbol se especifica el *valor* que hay que producir en el caso de alcanzar dicha hoja.

En la figura 4.1 se muestra un Árbol de Decisión que permite decidir si se juega o no una partida de golf, de acuerdo a las condiciones climáticas.

El nodo raíz es el Clima y tiene tres ramas: soleado, nublado y lluvioso, si el valor de la prueba del nodo raíz Clima es soleado, entonces desciende al nodo interno Humedad, este nodo tiene dos ramas: <=75 y >75, si el valor de la prueba del nodo interno Humedad es <=75, entonces desciende al nodo hoja que especifica el valor Juega, en caso contrario, desciende al nodo hoja que especifica el valor No juega.

Si el valor de la prueba del nodo Clima es nublado, entonces desciende al nodo hoja que especifica el valor Juega. Si el valor de la prueba del nodo Clima es lluvioso, entonces desciende al nodo interno Viento, este nodo tiene dos ramas: falso y verdadero, si el valor de la prueba del nodo interno Viento es falso, entonces desciende al nodo hoja que especifica el valor Juega, en caso contrario, desciende al nodo hoja que especifica el valor No juega.

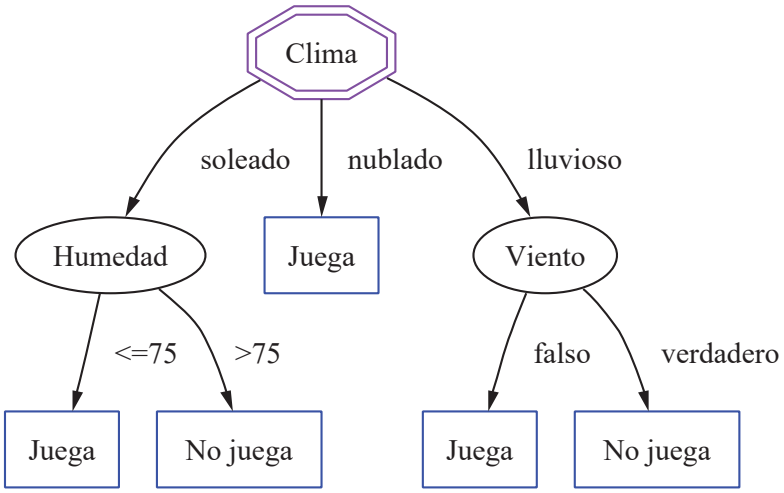


Figura 4.1: Árbol de Decisión.

4.3. Inducción de un Arbol de Decisión a partir de ejemplos

Existe un método clásico de aprendizaje inductivo a partir de ejemplos: *divide y vencerás*. Este método consiste en particionar el conjunto de ejemplos en subconjuntos sobre los que se puede trabajar más fácilmente. Este tipo de aprendizaje es el que emplean los algoritmos de la familia TDIDT (Top-Down Induction Trees: Árboles de Inducción Descendentes), a la cual pertenece el AI de Software que conforma al AI Descubreconocimiento descrito en los capítulos 2 y 3.

Para que el algoritmo de aprendizaje genere adecuadamente un Árbol de Decisión, se recomienda adoptar la siguiente metodología:

1. Reunir una gran cantidad de ejemplos.
2. Dividirla aleatoriamente en dos conjuntos: el conjunto de entrenamiento y el conjunto de prueba.
3. Emplear el algoritmo de aprendizaje con el conjunto de entrenamiento como ejemplo base para producir un Árbol de Decisión (Modelo o Representación de la regularidad existente en los datos).
4. Medir el porcentaje de ejemplos del conjunto de prueba clasificados correctamente con el Árbol de Decisión.
5. Repetir los pasos 2 a 4 en conjuntos de entrenamiento de diverso tamaño.
6. Si no se logra un entrenamiento satisfactorio del algoritmo se recomienda revisar los datos o aumentar el volumen de éstos.

4.4. Algoritmo C4.5

El algoritmo C4.5 fue desarrollado por el Dr. J. Ross Quinlan en 1993. Este algoritmo genera Árboles de Decisión a partir de ejemplos mediante particiones realizadas recursivamente. Para el caso de este trabajo de tesis, el algoritmo C4.5 es el Agente Inteligente de Software mencionado en las secciones 2.6 y 3.2.

En la figura 4.2 se muestra el pseudocódigo del algoritmo C4.5. Éste recibe como argumentos de entrada un conjunto de atributos no clasificadores (R), un atributo clasificador o clase (C) y un conjunto de datos de entrenamiento (S) y como salida, genera un modelo clasificador que se puede representar como un Árbol de Decisión. En dicha figura 4.2 se puede apreciar como opera el algoritmo C4.5.

```
Función C4.5 (R: conjunto de atributos no clasificadores,
              C: atributo clasificador,
              S: conjunto de datos de entrenamiento) retorna un Árbol de Decisión;

Inicio

        Si S está vacío, entonces

                retorna un nodo único con Valor Error;

        Si todos los registros de S tienen el mismo valor para el atributo clasificador, entonces

                retorna un nodo único con dicho valor;

        Si R está vacío, entonces

                retorna un nodo único con el valor más frecuente del atributo clasificador

        Si R no está vacío, entonces

                D <- atributo con mayor Proporción de Ganancia (D,S) entre los atributos de R;

                Sean {dj| j=1,2, .., n} los valores del atributo D;

                Sean {Sj| j=1,2, .., n} los subconjuntos de S correspondientes a los valores
                de Dj respectivamente;

                Retorna un árbol con el nodo raíz denominado como D y con las ramas
                denominadas como d1, d2, .., dn
                que descienden respectivamente a los árboles

                C4.5 (R-{D}, C, S1), (R-{D}, C, S2), .., (R-{D}, C, Sn);

Fin
```

Figura 4.2: Algoritmo C4.5.

En cada nodo, el algoritmo debe decidir cual prueba elegir para particionar los datos. Los tipos de pruebas propuestas por el Dr. Quinlan para C4.5 son tres [Quinlan, 1993]:

1. Prueba "estándar" para atributos discretos, obteniendo un resultado y una rama para cada valor posible del atributo.

2. Prueba basada en un atributo discreto, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de uno para cada valor.

3. Si un atributo A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A<=L y A>L, para lo que se debe determinar el valor límite L.

Estas pruebas se evalúan de igual forma, observando el resultado de la ganancia de información.

4.4.1. Poda de los Árboles de Decisión

Básicamente existen dos formas de modificar el método de particionamiento recursivo para producir árboles más simples:

- Poda o Pre-poda . Consiste en decidir no dividir más un conjunto de casos de entrenamiento. La ventaja al realizar esto es que no se pierde tiempo en construir una estructura que después será simplificada en el árbol final.

- Pospoda. Esta consiste en remover retrospectivamente alguna parte de la estructura construida por el particionamiento recursivo. El algoritmo C4.5 utiliza este criterio.

4.5. Teoría de la Información

Para llevar a cabo la partición de los datos, el Dr. Quinlan propone el empleo de los métodos de la Teoría de la Información, teniendo presente que ésta teoría no puede informar si determinado conocimiento es verdadero o falso, sino que sólo cuantifica numéricamente a ese conocimiento en relación a la entropía existente bajo el supuesto de que dicho conocimiento es verdadero. El algoritmo C4.5 utiliza el criterio de Ganancia de Información para realizar la partición de los ejemplos, ya que en opinión del mismo Dr. Quinlan este criterio es robusto y proporciona resultados más consistentes [Quinlan, 1988].

El algoritmo de aprendizaje que genera el Árbol de Decisión selecciona los atributos de tal manera que pueda reducir a un mínimo la profundidad del árbol final. Lo que importa es elegir el atributo que favorezca al máximo la clasificación exacta de los ejemplos. Un atributo perfecto divide los ejemplos en conjuntos que son totalmente discretos: *positivos* o *negativos*. Un atributo poco útil genera conjuntos con ejemplos positivos y negativos lo que impide su clasificación.

Para escoger un atributo, es necesario contar con una medida formal de lo que es "bueno" y lo que es "inútil", por cual el valor máximo de tal medida se obtiene cuando el atributo es perfecto y el valor mínimo cuando el atributo no sirve de gran cosa. Una medida adecuada es la cantidad de **información** que puede proporcionar el atributo [Russell y Norvig, 2004], en la teoría de la información el contenido de ésta se mide en **bits**.

Generalmente, si las respuestas posibles v_i tienen probabilidad $P(v_i)$, entonces la cantidad de información I proporcionada por el atributo es obtenida de la siguiente manera:

$I\left(P\left(v_1\right),...,P\left(v_n\right)\right) \;=\; -\;\sum_{i=1}^{n} P\left(v_i\right) log_2 P\left(v_i\right)$

La ecuación anterior representa el contenido promedio de información para los diferentes eventos, en el caso de los Arboles de Decisión se desea estimar las probabilidades de las respuestas antes de probar los atributos. Esto se hace con las proporciones de ejemplos positivos y negativos presentes en el conjunto de entrenamiento.

Si se tienen p ejemplos positivos $\left\{ ep_1, \quad ep_2, \quad ep_3, \quad \ldots, \quad ep_n \right\}$ y

n ejemplos negativos $\left\{ en_1, \quad en_2, \quad en_3, \quad \ldots, \quad en_n \right\}$, entonces:

$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) \;=\; -\frac{p}{p+n} log_2 \frac{p}{p+n} \quad -\frac{n}{p+n} log_2 \frac{n}{p+n}$

Por lo general, al realizar pruebas de un solo atributo A no producirá mucha información, pero al menos si dará parte de ésta, para medir exactamente cuánta información proporcionará basta con determinar cuánta información se necesita después de probar ese atributo. Cada atributo A, divide a los ejemplos del conjunto de entrenamiento en subconjuntos E_1, E_2,..., E_v de acuerdo a los v valores del atributo. Cada subconjunto E_i tiene p_i ejemplos positivos y n_i ejemplos negativos, por lo que para cada rama se necesitará: $I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$ cantidad de información para evaluar la prueba de partición.

Un ejemplo aleatorio tiene el valor i-ésimo del atributo A con probabilidad: $\frac{p_i+n_i}{p+n}$. Por lo que en promedio, después de probar el atributo A, la **entropía** (medida de desorden o desinformación) será:

$E\left(A\right) \;=\; \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$

Es decir, el valor en bits de información faltante para clasificar el ejemplo. La **ganancia de información** deducida de la prueba de atributo se define como la diferencia entre la necesidad original de información y la nueva necesidad:

$Ganancia\left(A\right) \;= I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) \;-\; E\left(A\right)$

Entonces, el criterio para elegir un atributo es seleccionar al atributo que tenga la mayor ganancia.

4.6. Construcción de un Árbol de Decisión con el algoritmo C4.5

La explicación continúa con un ejemplo simple ValidacionTest extraido del ambiente del MMPI-2. Se trata de clasificar como válida o no válida la aplicación del test a un paciente.

Los **ejemplos** se describen mediante los valores de los atributos y el valor del predicado *meta* o *clase*. Al valor del predicado meta se le denomina **clasificación** del ejemplo. Si el predicado de meta es válido para cierto ejemplo, entonces es un ejemplo **positivo** (P), en caso contrario, es un ejemplo **negativo** (N). En el planteamiento de este ejemplo como un problema de aprendizaje, primero se especifica cuáles propiedades o *atributos* están disponibles para describir al conjunto de ejemplos.

Se dispone de la siguiente lista de atributos y sus dominios:

1. L : escala de la mentira, tendencia del paciente a mentir al contestar el test (Bajo, Medio, Moderado, Alto).

2. F: escala de la infrecuencia, tendencia del paciente al contestar el test en forma inconsistente (Bajo, Medio, Moderado, Alto).

3. K: escala de negación de problemas, tendencia del paciente a negar sus problemas al contestar el test (Bajo, Medio, Moderado).

Considerando el conjunto de entrenamiento de la tabla 4.1 compuesto por los ejemplos X_1,...,X_{10} en el dominio del ejemplo ValidacionTest, se calculan las ganancias obtenidas con cada uno los atributos.

	Atributos			**Meta o Clase**
Ejemplo	**L**	**F**	**K**	**Válido (Positivo)/No válido (Negativo)**
X1	Moderado	Alto	Medio	No válido (N)
X2	Moderado	Medio	Bajo	Válido (P)
X3	Bajo	Medio	Medio	Válido (P)
X4	Medio	Alto	Medio	No válido (N)
X5	Moderado	Alto	Bajo	No válido (N)
X6	Bajo	Medio	Bajo	Válido (P)
X7	Bajo	Alto	Medio	No válido (N)
X8	Medio	Bajo	Medio	Válido (P)
X9	Medio	Moderado	Medio	Válido (P)
X10	Moderado	Alto	Medio	No válido (N)

Tabla 4.1: Conjunto de entrenamiento del ejemplo ValidacionTest.

Primero se calcula la información **I** obtenida con cada elemento del dominio de cada atributo y enseguida se calcula la entropía **E** correspondiente al atributo.

Para el atributo **L** se tiene:

$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} log_2 \frac{p}{p+n} \quad -\frac{n}{p+n} log_2 \frac{n}{p+n}$

Bajo: p_1=2, n_1=1, I(p_1, n_1)= 0.918 bits

Medio: p_2=2, n_2=1, I(p_2, n_2)= 0.918 bits

Moderado: p_3=1, n_3=3, I(p_3, p_3)= 1.061 bits

Alto: p_4=0, n_4=0, I(p_4, p_4)= 0 bits

$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$

Entropía(L)= $\frac{3}{10} I(p_1, n_1) + \frac{3}{10} I(p_2, n_2) + \frac{4}{10} I(p_3, n_3) = 0{,}974\ bits$

De igual forma se calcula la información y entropía para los atributos restantes:

Entropía(F)= 0.000 bits

Entropía(K)= 0.965 bits

Finalmente se calcula la ganancia de información obtenida con cada uno de los atributos:

$Ganancia(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - E(A)$

Donde I se obtiene al evaluar a la clase:

$\text{I}(5,5) = \; -\frac{5}{10} log_2 \left(\frac{5}{10}\right) \; -\frac{5}{10} log_2 \left(\frac{5}{10}\right) = 1.000$ bits

Se realizan los cálculos respectivos:

$$Ganancia\,(L) \; = 1{,}000 \; - \; 0{,}974 \; = \; 0{,}026 \; bits$$

$$Ganancia\,(F) \; = 1{,}000 \; - \; 0{,}000 \; = \; 1{,}000 \; bits$$

$$Ganancia\,(K) \; = 1{,}000 \; - \; 0{,}965 \; = \; 0{,}035 \; bits$$

Por lo tanto, el algoritmo de aprendizaje del árbol de decisión escogerá al atributo con mayor ganancia que en este caso es F, como nodo raíz y procederá a realizar el mismo proceso con los ejemplos de cada rama.

Los ejemplos positivos son aquellos en los que la meta Válido es válida (X_2,X_3,...) y los negativos son aquellos en lo que es falsa(X_1,X_4,...). En apariencia el problema de encontrar un Árbol de Decisión que corresponda al conjunto de datos de entrenamiento parece complicado, pero en realidad la solución es simple. Sólo basta con construir un Árbol de Decisón que tenga una ruta que lleve a la hoja de cada ejemplo, la ruta prueba cada uno de los atributos del ejemplo en turno y emula el valor del ejemplo (p o n, para este caso), mientras en la hoja está la clasificación del ejemplo. Si se prueba de nuevo el mismo ejemplo o uno con la misma descripción, el Árbol de Decisión devolverá la clasificación adecuada.

En la figura 4.3 se muestra el Árbol de Decisión obtenido por el algoritmo C4.5. Este representa la siguiente expresión :

si F es igual a Medio, entonces el ejemplo se clasifica como Válido, o

si F es igual a Moderado, entonces el ejemplo se clasifica como Válido, o

si F es igual a Alto, entonces el ejemplo se clasifica como No válido, o

si F es igual a Bajo, entonces el ejemplo se clasifica como Válido.

El número que aparece entre paréntesis a la derecha de la clasificación indica la cantidad de ejemplos clasificados correctamente.

```
C4.5 [versión 8] Generador de árboles de decisión        Sun Oct 19 18:17:24 2008
--------------------------------------------------------

8 casos leídos (3 atributos) desde valTest10.data

Arbol de decisión:

F = Medio: Válido (2.0)
F = Moderado: Válido (1.0)
F = Alto: No válido (4.0)
F = Baja: Válido (1.0)


Arbol guardado


Evaluación sobre los datos de entrenamiento (8 elementos):

         Antes de Poda            Después de Poda
        ---------------      ---------------------------
        Tamaño   Errores      Tamaño  Errores  Estimación

          5     0( 0.0%)         5    0( 0.0%)   (45.9%)   <<


Evaluación sobre los datos de prueba (2 elementos):

         Antes de Poda            Después de Poda
        ---------------      ---------------------------
        Tamaño   Errores      Tamaño  Errores  Estimación

          5     0( 0.0%)         5    0( 0.0%)   (45.9%)   <<


         (a)  (b)      <-clasificado como
        ----  ----
          1           (a): class Válido
               1      (b): class No válido
```

Figura 4.3: Árbol de Decisión ValidaciónTest.

En el caso mostrado en la figura 4.3, se observa que para el conjunto de datos de entrenamiento el algoritmo toma 8 ejemplos y 2 ejemplos para prueba, los que sumados conforman el total de la muestra de datos analizada:10 ejemplos descritos por 3 atributos (L, F y K).

Enseguida se muestra la evaluación para cada uno de los Árboles de Decisión generados: el árbol sin podar y el árbol podado. La evaluación se presenta en dos tablas, una obtenida a partir de los datos de entrenamiento y la otra a partir de los datos de prueba. Cada una de las tablas indica lo siguiente:

- **Tamaño del árbol**: cantidad de nodos + cantidad de hojas.
- **Errores (porcentaje de error)**: indican la cantidad de ejemplos clasificados erróneamente, donde el porcentaje de error obtiene dividiendo la cantidad de errores entre la cantidad total de ejemplos.

- **Estimación**: es un estimador del éxito del árbol obtenido expresado en porcentaje, se calcula dividiendo la cantidad de ejemplos clasificados correctamente entre la suma de la cantidad de ejemplos clasificados correctamente más la cantidad de ejemplos clasificados erróneamente.

Para el ejemplo de la figura 4.3, la evaluación sobre los datos de entrenamiento considera 8 elementos o ejemplos, el tamaño del árbol antes de la poda es 5 y los errores 0 (0.0 %). El tamaño del árbol después de la poda es 5, los errores 0 (0.0 %) y la estimación 45.9 %. La evaluación sobre los datos de prueba considera 2 elementos o ejemplos, el tamaño del árbol antes de la poda es 5 y los errores 0 (0.0 %). El tamaño del árbol después de la poda es 5, los errores 0 (0.0 %) y la estimación 45.9 %.

Finalmente se presenta la matriz de confusión sobre los datos de prueba, indicando para cada clase, la cantidad de ejemplos clasificados correctamente y los clasificados incorrectamente. En este caso un ejemplo fue clasificado correctamente como Válido y otro como No válido.

La figura 4.4 muestra el modelo clasificador generado por el algoritmo C4.5 representado gráficamente como un Árbol de Decisión.

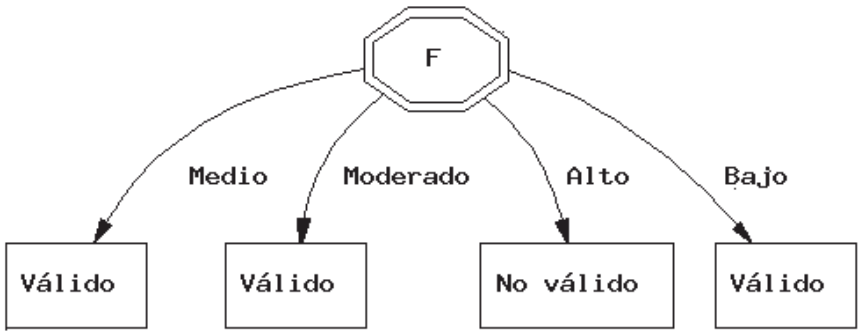


Figura 4.4: Gráfica del Árbol de Decisión ValidacionTest.

4.7. Conclusión del capítulo

El algoritmo de aprendizaje C4.5 es en el presente trabajo el AI de Software, el que en conjunto con los AI Humanos conforman al AI Descubreconocimiento. Este algoritmo genera modelos de clasificación que pueden ser representados como Árboles de Decisión. Un Árbol de Decisión es un *modelo de predicción*, construido por el algoritmo de aprendizaje como un modelo de clasificación a partir de un conjunto de datos de entrada (ejemplos de entrenamiento y evaluación). Donde el modelo obtenido sirve para representar una serie de condiciones sucesivas para la resolución de un problema. El problema a resolver se denomina *clasificación* cuando los datos de entrada son discretos y cuando los datos de entrada son continuos se denomina *regresión*. Como el aprendizaje del algoritmo C4.5 es a partir de ejemplos de entrenamiento, entonces éste es un aprendizaje inductivo, puesto que el conocimiento se obtiene de información particular para su aplicación general como un modelo de predicción.

Capítulo 5

Minería de Datos

En el capítulo 2 se propone al Agente Inteligente Descubreconocimiento capaz de llevar a cabo adecuadamente el proceso completo de descubrimiento de conocimiento, por lo que resulta necesario que este AI Descubreconocimiento siga una metodología o algoritmo de referencia, que le permita alcanzar su objetivo satisfactoriamente. Esta metodología se debe fundamentar en el proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases: KDD) conocido también como Minería de Datos (Data Mining: DM). Por esta razón, en este capítulo se describe lo que es la Minería de Datos y se propone un algoritmo de referencia para Minería de Datos, que sirva de guía para los usuarios (Agentes Inteligentes Humanos) cuando lleven a la práctica el proceso de descubrimiento de conocimiento en bases de datos, empleando la herramienta Ambiente Descubreconocimiento con C4.5 para Web (ADC4.5Web) descrita en el capítulo 7, conformando de esta manera al AI Descubreconocimiento, puesto que ADC4.5Web incluye un Agente Inteligente de Software.

5.1. Antecedentes

La mayoría de los datos que se encuentran almacenados en las bases de datos es información histórica, que generalmente representa transacciones. El análisis de estos datos se puede realizar por medio de consultas con el lenguaje *SQL* directamente sobre la base de datos en operación, es decir, junto al procesamiento transaccional en línea de dicha base de datos. Esto se conoce como *On-Line Transaction Processing* (OLTP), sin embargo, esta estrategia sólo permite generar información resumida de una forma previamente establecida y poco escalable a grandes volúmenes de datos.

Como respuesta a esta limitante han surgido los almacenes de datos (data warehouse), los cuales son repositorios de datos de diversas fuentes, debidamente integrados y organizados para hacer más fácil su análisis y poder brindar un buen soporte para la toma de decisiones. Esta tecnología permite realizar operaciones de procesamiento analítico en línea, conocidas como *On-Line Analytical Processing* (OLAP), como puede ser el resumen y permite ver la información desde diversas perspectivas.

Aunque las herramientas OLAP proporcionan soporte para realizar un análisis descriptivo y de resumen, haciendo posible la transformación de los datos en otros con valor agregado, estas herramientas no generan patrones o reglas, es decir, no generan conocimiento que se pueda aplicar a otros

datos. Los datos poseen valor por sí mismos, pero puede ser más interesante el conocimiento que se pueda inferir a partir de éstos y sobre todo, si este conocimiento se puede aplicar a otros datos.

La *Minería de Datos* (Data Mining), se distingue de las herramientas anteriores, por el hecho de que se enfoca a obtener conocimiento, el que muchas veces resulta novedoso y original, resaltando que dicho conocimiento es extraído completamente por la herramienta de minería, en el caso del presente trabajo se implementa una herramienta denominada Ambiente Descubreconocimiento con C4.5 para Web (ADC4.5Web) descrita en el capítulo 7.

5.2. Definición de Minería de Datos

Los autores [Clark y Boswell, 2000] definen a la **Minería de Datos** como "el proceso de extracción de conocimiento útil y comprensible, previamente desconocido, a partir de grandes volúmenes de datos almacenados en distintos formatos". Esto significa que el objetivo principal de la MD es encontrar modelos entendibles a partir de los datos, siendo deseable para su efectividad, que el proceso de extracción sea automatizado o semiautomatizado y que el uso de los patrones descubiertos proporcionen un buen soporte para una mejor toma de decisiones en el área de aplicación.

Para que la MD sea accesible a la mayoría de los usuarios, no se requiere que éstos sean expertos en las técnicas de MD, ni que dediquen mucho tiempo a la interpretación de los resultados, por tal razón, muchas aplicaciones de MD presentan la información descubierta de una forma clara y comprensible para los humanos, como pueden ser la conversión de patrones a lenguaje natural o técnicas de visualización de los datos, entre otras.

5.3. Tipos de datos

La MD se puede aplicar a cualquier tipo de información y las técnicas de MD que se pueden emplear varían de acuerdo a los datos. En general, la información pueden ser datos estructurados que provengan de una base de datos relacional, datos estructurados en bases de datos textuales, multimedia o temporales y datos no estructurados provenientes de repositorios o de la web.

5.3.1. Bases de datos relacionales

Una base de datos relacional es una colección de relaciones o tablas. Donde cada tabla está compuesta por un conjunto de atributos conocidos también como campos o columnas y puede contener una gran cantidad de tuplas conocidas también como registros o filas. Cada tupla representa un objeto, descrito por los valores de sus atributos y este objeto o registro se caracteriza por tener una clave única o primaria que lo identifica. Una tabla puede contener claves foráneas o ajenas, es decir, atributos o campos que hagan referencia a otra tabla.

La integridad de los datos se puede expresar por medio de las restricciones de integridad. Estas restricciones pueden ser de dominio, de identidad y referenciales. Las restricciones de dominio limitan

el valor que puede tomar un atributo con respecto a su dominio y si éste puede tomar valores nulos o no, las de identidad pueden limitar a que la clave primaria sea única y las referenciales limitan a que los valores de claves foráneas o ajenas conrrespondan sólo con un valor de la tabla referenciada.

El origen de la información para la mayoría de las aplicaciones de MD son las bases de datos relacionales. Muchas de las técnicas de MD no poseen la capacidad de trabajar con toda la base de datos, únicamente son capaces de tratar con una sola tabla a la vez. Esta limitante se resuelve en la práctica, por medio de una consulta SQL que *combine en una sola tabla* o **vista minable [Hernández et al., 2004]** la información de varias tablas que sea requerida por alguna tarea concreta de MD. Por esta razón, el formato tabular, también llamado atributo-valor, es el más utilizado por las técnicas de MD.

En el formato tabular, se pueden manejar varios tipos de datos (enteros, reales, fechas, cadenas de texto, etc), aunque, de acuerdo con las técnicas de MD más comunes, los que más interesan generalmente son dos tipos: discretos y continuos. Los tipos de datos fueron descritos en la sección 4.1.1. Debe notarse que aún considerando sólo estos dos tipos de datos, no todas las técnicas de MD poseen la capacidad de trabajar con ambos tipos.

5.4. Tipos de modelos

El objetivo de la MD es analizar la información para extraer conocimiento. Este conocimiento puede ser obtenido en forma de relaciones, patrones o reglas inferidos a partir de datos previamente desconocidos. Estas relaciones conforman el modelo de los datos analizados. Los modelos se pueden representar de muchas formas y cada una de ellas determina el tipo de técnica que se puede emplear para obtenerlos.

Los modelos se pueden resumir en dos tipos: *predictivos* o *descriptivos*. Los modelos **predictivos** estiman valores futuros o desconocidos de nuevos datos de interés. Los modelos **descriptivos**, identifican patrones que describen o resumen los datos, es decir, exploran las propiedades de los datos examinados, no predicen nuevos datos.

5.5. Tipos de problemas de Minería de Datos

De manera general los problemas se pueden clasificar en: problemas predictivos y problemas descriptivos.

5.5.1. Problemas predictivos

En el contexto de la Inteligencia Artificial, este tipo de problemas se denominan problemas de aprendizaje supervisado, ya que el agente inteligente que descubre conocimiento proporciona al sistema la respuesta deseada, inducida a partir de conocimiento previo. En este tipo de problemas, se puede hacer la siguiente distinción:

- Problemas de **clasificación**. En este tipo de problemas la variable a predecir posee un número finito de valores, esto significa que es **discreta**.
- Problemas de **predicción de valores**. La variable a predecir en este tipo de problemas es de tipo **numérica**.

5.5.2. Problemas descriptivos

Este tipo de problemas son aquellos cuya meta es encontrar una descripción de los datos de estudio. Se pueden distinguir los siguientes tipos:

- Análisis de **segmentación** o **agrupamiento**. Son problemas donde la meta es encontrar grupos homogéneos en la población de datos origen. En el contexto de la Inteligencia Artificial, se les denomina problemas de *aprendizaje no supervisado* o *clustering*.
- Análisis de **asociaciones**. En estos problemas la meta es encontrar relaciones entre los valores de atributos de una base de datos.

5.6. Técnicas de Minería de Datos

A continuación se describen brevemente algunas de las técnicas más comunes de MD.

- **Clasificación**. Esta técnica se emplea para resolver problemas predictivos. Los algoritmos que se pueden utilizar son Árboles de Decisión, Regresiones Logísticas y Redes Neuronales. Estos algoritmos emplean un conjunto de datos de entrenamiento para crear el modelo que se utilice posteriormente para clasificar datos desconocidos.
- **Predicción de valores**. Esta técnica es similar a la anterior, sólo que para la predicción de valores se requiere el uso de los algoritmos ya mencionados en forma conjunta con los de regresión lineal y regresión no lineal.
- **Clustering no jerárquico**. Esta técnica se emplea para resolver problemas descriptivos. Consiste en comparar cada registro de la base datos con todos los agrupamientos encontrados, medir la distancia de separación y asignarlo al agrupamiento más cercano. El algoritmo que más se emplea es *k-medias*.

5.7. Minería de Datos y el Proceso de Descubrimiento de Conocimiento en Bases de Datos

Existen términos que se utilizan frecuentemente como sinónimos de la Minería de Datos. Uno de ellos y el más relacionado con la MD, es el "descubrimiento de conocimiento en bases de datos" (**Knowledge Discovery in Databases: KDD**).

En [Fayyad et al., 1996] se define KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a través de los datos", donde:

- **Válido**: hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.
- **Novedosos**: significa que los patrones aporten algo desconocido tanto para el sistema y preferiblemente para el usuario.
- **Potencialmente útiles**: esto implica que la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- **Comprensibles**: esto significa que 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.

En la figura 5.1 se observa que los sistemas de KDD permiten la preparación de los datos: selección, limpieza, transformación y proyección. Hacen posible también el empleo de algoritmos de Minería de Datos para analizar a dichos datos y poder extraer patrones o modelos adecuados. También hacen posible la evaluación, interpretación y visualización de los patrones o modelos obtenidos, los que finalmente se pueden convertir en conocimiento.

Figure 5.1: Proceso KDD.

KDD es el proceso global de descubrir conocimiento útil desde las bases de datos, mientras que la MD se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos. Al ser la fase de generación de modelos, comúnmente se considera a KDD como sinónimo de Minería de Datos.

5.8. Algoritmo de referencia propuesto para el proceso KDD

Fayyad sugiere que el proceso KDD sea un proceso interactivo e iterativo [Fayyad et al., 1996], el cual permita la interacción entre el personal involucrado en el proceso de minería y que proporcione la posibilidad de retroalimentación en forma iterativa, facilitando de esta manera poder alcanzar el objetivo de la tarea de minería. En el ambiente de la industria de la información existe un modelo y guía de referencia denominado Proceso Estándar Industrial Híbrido para la Minería de Datos (CRoss-Industry Standard Process for Data Mining: CRISP-DM), que es en realidad un

consorcio de empresas entre las que se incluyen a SPSS, NCR y DaimlerChrysler. El modelo y la guía están estructurados en seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, como se aprecia en la figura 5.2. Se observa que existe retroalimentación bidireccional entre algunas fases, permitiendo la revisión parcial o total de fases anteriores.

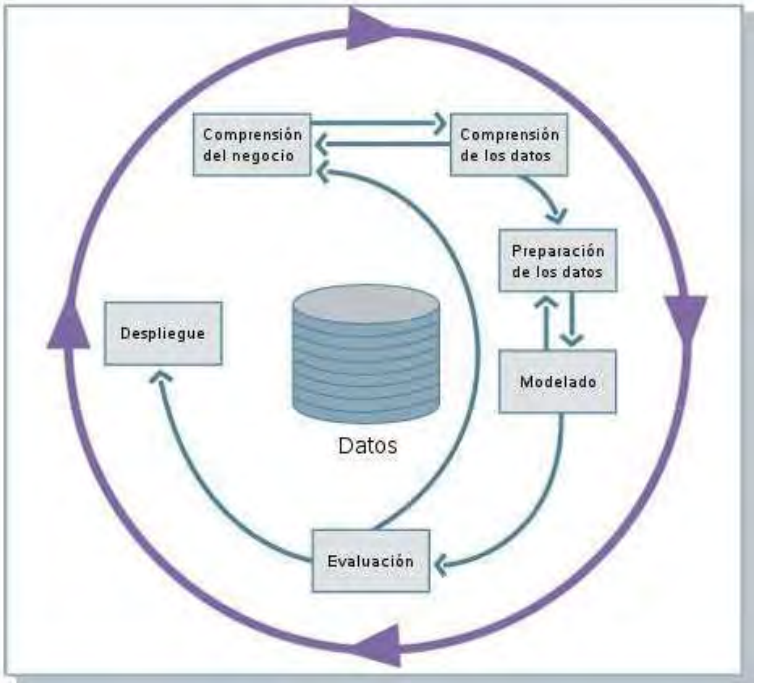


Figura 5.2: Fases del modelo de referencia CRISP-DM.

El uso del modelo CRIPS-DM se enfoca principalmente al área de los negocios y la guía de referencia de dicho modelo también, en cambio el criterio de Fayyad sobre el proceso KDD es para uso general, por lo tanto, fusionando ambos criterios, se propone un modelo de referencia de Minería de Datos para uso general, que resulta apropiado tanto para las tareas de investigación así como las de negocios. La realización del proceso de KDD resulta posible en el presente trabajo por medio de la interacción del Agente Inteligente Descubreconocimiento y la herramienta de minería desarrollada: Ambiente de Descubrimiento de Conocimiento con C4.5 Web (ADC4.5Web). El modelo propuesto es una adecuación obtenida a partir de las etapas sugeridas por Fayyad para el proceso KDD [Fayyad et al., 1996] y el modelo CRISP-DM. A continuación se presentan las etapas del modelo de referencia para el proceso KDD propuesto.

5.8.1. 1a. etapa: identificación del problema

La identificación del problema consiste en comprender el dominio de aplicación, el conocimiento relevante a usar y las metas del usuario. Es importante que el personal técnico entienda cuáles son las necesidades verdaderas del caso en estudio. Esto se hace hablando con los expertos del dominio, ya que necesitan estar informados del desarrollo del proyecto, de modo que ellos puedan contribuir.

Los expertos del dominio determinan si realmente es necesario hacer minería y si ésta debe enfocarse a un segmento o subgrupo de los datos en particular. Indican también, los aspectos relevantes del caso en estudio. Además, ellos saben en donde residen y se almacenan los datos y pueden saber cuales fuentes de datos son confiables y cuales no. Finalmente, su experiencia e intuición puede ser una fuente de conocimiento.

En esta etapa se observa que es de gran importancia la interacción entre el personal involucrado en el proceso de minería, siendo este el fundamento para concebir a los Agentes Inteligentes Humanos que interactúan con un Agente Inteligente de Software conformando al Agente Inteligente Descubreconocimiento cuando se realiza alguna tarea de minería.

5.8.2. 2a. etapa: obtención de datos

Los datos correctos están a menudo disponibles, razonablemente limpios y accesibles. Estos deben satisfacer los requisitos para resolver el problema del caso en estudio, además de estar completos como sea posible.

5.8.3. 3a. etapa: selección de datos

Se genera un conjunto de datos objetivo y se selecciona un subconjunto de variables o ejemplos para realizar el proceso de descubrimiento. En esta etapa se debe considerar la homogeneidad de los datos, las estrategias de muestreo, la posible variación de los datos en el transcurso del tiempo, etc.

5.8.4. 4a. etapa: preprocesamiento de datos

El preprocesamiento consiste en la validación, exploración, limpieza y preparación de los datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, normalización de los datos, etc. Debe verificarse también que los valores de los campos de la *vista minable* estén dentro de los límites legales, que sean razonables y que su distribución sea con fundamento.

5.8.5. 5a. etapa: transformación de los datos a la granularidad correcta

Primero se define el *enfoque de atención* o *vistas minables*, es decir, se especifica cuales tablas, campos y registros accesar. Se debe tener un mecanismo de selección aleatoria de registros, para después realizar la transformación de los datos a la granularidad correcta, es decir, realizar una transformación y reducción de los datos. Algunas formas de hacerlo son:

* Relación de datos: atributos relacionados contenidos en diferentes tablas
* Restricción de datos: en base a valores de atributos (por ejemplo, sólo aquellos datos que tengan ciertos valores)
* Proyección de datos: ignorar algún(os) atributo(s)

Esta etapa implica la búsqueda de características útiles de los datos de acuerdo al objetivo de la tarea de minería, resultando un paso crítico en el proceso de KDD, ya que se requiere tener un conocimiento e intuición adecuados acerca del problema a resolver, ya que con frecuencia esto define la diferencia entre el éxito o fracaso de la tarea de minería.

5.8.6. 6a. etapa: definición del tipo de problema de minería

En esta etapa se define el tipo de problema de minería: *predicción* o *descripción*. En base a esto, se determina la técnica de minería a utilizar, por ejemplo clasificación, agrupamiento, regresión, etc.

5.8.7. 7a. etapa: selección del algoritmo de Minería de Datos

El algoritmo de Minería de Datos se elige en base al tipo de problema de minería seleccionado, pudiendo elegir varios algoritmos si se considera necesario. Éste debe proporcionar herrramientas para su uso, una técnica de modelado y definir las especificaciones de entrenamiento del algoritmo.

5.8.8. 8a. etapa: proceso de Minería de Datos

El proceso de Minería de Datos consta de dos fases:

1. **Preparación de los datos (conjunto modelo).** El conjunto modelo son los datos usados por el algoritmo de minería para construir modelos de minería. Es necesario considerar aspectos tales como la frecuencia con que se presentan datos extremos o aislados en el conjunto modelo. Por ejemplo, si esta frecuencia es demasiado baja, entonces las predicciones hechas por el modelo, aunque es exacta, nunca podría incluir a los datos extremos o aislados. El conjunto modelo debe dividirse aleatoriamente en conjuntos de datos de entrenamiento, de prueba y de evaluación.

2. **Entrenamiento del algoritmo de minería**. Esto se lleva a cabo proporcionando los datos del conjunto de entrenamiento al algoritmo de minería, para que éste a su vez se entrene y genere ya sea un modelo o varios modelos eligiendo el mejor de ellos y enseguida verificarlo con el conjunto de prueba. Si se presentan cambios en los datos en cierto plazo de tiempo, esto trae como consecuencia la necesidad de entrenar de nuevo al algoritmo de minería con datos más recientes.

5.8.9. 9a. etapa: evaluación del modelo e interpretación de resultados

El conjunto de los datos de evaluación (parte del conjunto modelo) es usado para ver como se comporta el modelo generado por el algoritmo de minería con datos no conocidos. La comparación de los resultados reales con los resultados predichos es una buena medida de la efectividad del modelo generado. El modelo obtenido a partir de los datos, se debe evaluar e interpretar de manera conjunta

entre el personal involucrado en el proceso de minería, aprovechando la experiencia y la intuición de los expertos del dominio.

Todo esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias.

5.8.10. 10a. etapa: selección del mejor modelo

Cuando la evaluación y la interpretación de resultados de un modelo sean satisfactorias, entonces éste puede ser seleccionado como el mejor modelo que ha descubierto conocimiento novedoso, confiable y útil.

5.8.11. 11a. etapa: conocimiento descubierto

El conocimiento descubierto debe satisfacer las metas de los usuarios involucrados en la tarea de minería, pudiendo facilitarles la toma decisiones o bien incorporando el conocimiento descubierto en algún sistema real, normalmente para mejorarlo.

5.9. Diagrama del algoritmo de referencia propuesto para el proceso KDD

En la figura 5.3 se muestra un diagrama general con las etapas o fases del algoritmo de referencia propuesto para el proceso KDD completo. En dicha figura se representan gráficamente las etapas del procesoo KDD descritas en la sección anterior.

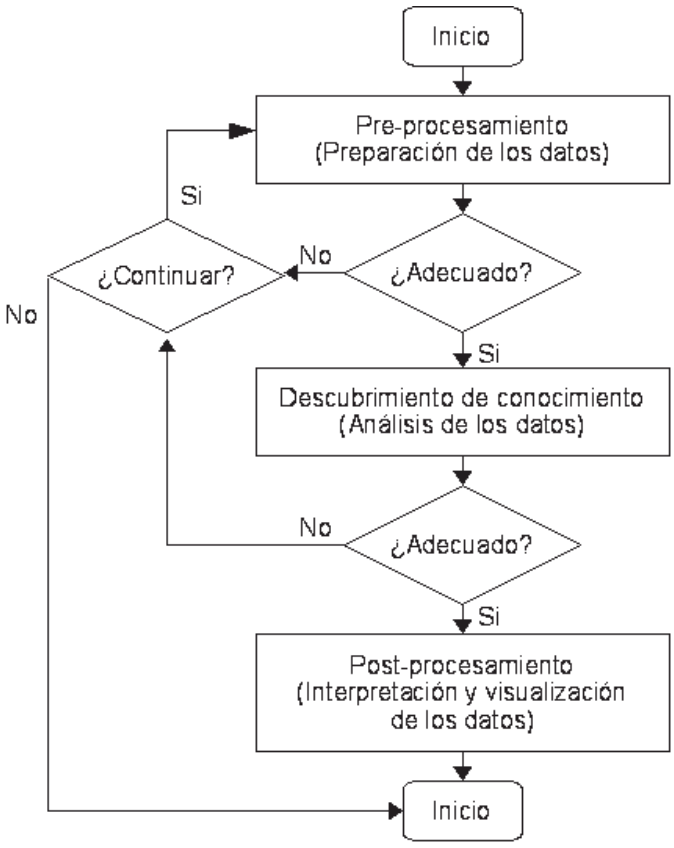


Figura 5.3: Diagrama general del algoritmo de referencia propuesto para el proceso KDD.

5.10. Algoritmo de referencia de minería aplicado al MMPI-2

MMPI-2 es el Inventario Multifásico de la Personalidad Minnesota-2 [Lucio y León, 2003], se utiliza para conocer algunos rasgos de la personalidad del ser humano. Este inventario se describe en el capítulo 6, así como también la aplicación computacional correspondiente desarrollada en el presente trabajo, que brinda la posibilidad de aplicar la versión estándar del inventario MMPI-2 tanto en forma individual como colectiva y además lo califica de manera automática, incluyendo gráficas e interpretación de los resultados.

El autor del presente proyecto que es uno de los Agentes Inteligentes Humanos que conforman al Agente Inteligente Descubreconocimiento, fue capacitado teórica y prácticamente por varios especialistas del área que se encuentran en servicio activo (psicólogos y psiquiatras), para aplicar y calificar el instrumento MMPI-2, permitiéndole conocer al detalle la **metodología** para la aplicación y calificación del test MMPI-2. Por lo tanto, el AI Descubreconocimiento tiene un conocimiento adecuado de los datos, información adicional de la estructura de dichos datos, restricciones entre campos, metas o preferencias del usuario, campos relevantes y las listas de clases, es decir, se conoce el dominio para orientarse en la búsqueda de patrones interesantes.

El proceso de minería realizado por el AI Descubreconocimiento aplicado al instrumento MMPI-2 se describe a continuación.

5.10.1. Identificación del Problema

El problema en este caso, es la obtención de una versión experimental reducida del Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2). Con la finalidad de emplearlo para evaluar la eficacia del AI Descubreconocimiento propuesto, el algoritmo de referencia y la herramienta de minería ADC4.5Web desarrollados.

5.10.2. Obtención de los datos

Se contó con el apoyo invaluable de varios profesionales del área de la Psicología quienes proporcionaron un total de 1395 ejemplos, omitiendo únicamente la información personal de los pacientes y la de dichos profesionales, para preservar su integridad. Estos datos pertenecen a la aplicación del test MMPI-2 a personas de la población en general, los cuales se consideran correctos, razonablemente limpios y completos. Por lo tanto, se afirma que estos datos satisfacen los requisitos para resolver el problema del caso en estudio.

5.10.3. Selección de datos

La selección del conjunto de datos y el enfoque de la búsqueda, fue sugerida por los especialistas en psicometría, basándose en las normas para la calificación del inventario MMPI-2. Para esto el criterio fue incluir un volumen y diversidad significativos de casos que facilitaran un buen entrenamiento del algoritmo de minería C4.5.

El MMPI-2 está compuesto por cuatro grupos de escalas: indicadores de validez, escalas clínicas o básicas, escalas suplementarias y escalas de contenido. Estas escalas permiten conocer algunos rasgos o características de la personalidad humana. Éstas a su vez están compuestas por un conjunto definido de reactivos como se aprecia en la figura 5.4.

r16	r29	r41	r51	r77	r93	r102	r107	r123	r139	r153	r183	r203	r232	r260	L
F	F	V	V	V	F	V	F	F	F	V	V	V	F	V	Medio
F	F	V	F	F	V	V	F	F	V	F	V	F	F	V	Moderado
F	V	V	V	V	V	V	V	F	V	F	V	V	F	V	Bajo
F	F	V	F	V	V	V	V	F	F	F	F	V	F	F	Moderado
F	V	F	V	V	V	V	F	V	V	V	V	V	F	V	Bajo

Figura 5.4: Los 15 reactivos que conforman a la escala L (mentira).

Cada reactivo posee un contenido (una afirmación) como se muestra en la figura 5.5. El total de reactivos que conforman al inventario MMPI-2 es de 567.

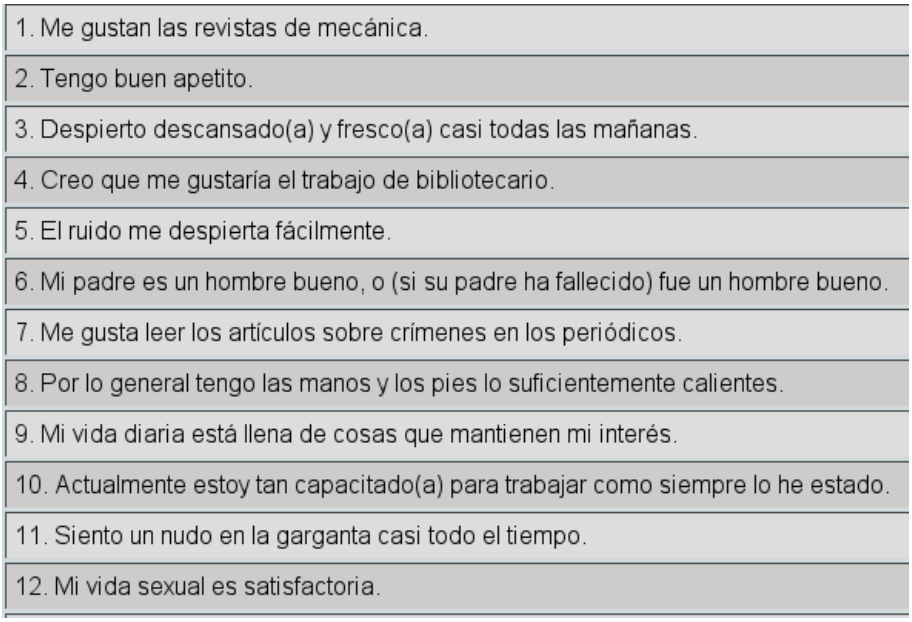


Figura 5.5: Ejemplo de algunos reactivos del MMPI-2: r1 hasta r12.

El conjunto de datos son los 1395 ejemplos almacenados como registros en las tablas de la base de datos y corresponden a cada uno de los pacientes evaluados, cada ejemplo está conformado por 567 reactivos y éstos corresponden a los campos de las tablas. Los especialistas en psicometría, sugirieron que para obtener la versión experimental reducida del inventario MMPI-2, la selección de los datos se hiciese considerando los campos del conjunto de reactivos que conforman a cada una de las escalas, así como sus respectivos campos de calificación, como se aprecia en el ejemplo de los 15 reactivos que conforman a la escala L y su calificación correspondiente mostrados en la figura 5.4. La selección de los datos se describe con más detalle en la sección 5.10.5.

5.10.4. Preprocesamiento de datos

Los datos correspondientes a las respuestas de la aplicación del test MMPI-2 fueron definidos como tipo *discreto* (enum). En la hoja de respuestas impresa para la aplicación típica del test, se tienen dos alternativas para contestar: *verdadero* o *falso*, rellenando el círculo de la opción correspondiente, pero, cuando el paciente rellena ambas opciones o bien, ninguna, es decir, el caso de una respuesta ambigua, ésta se denomina como una respuesta *no se* en la calificación del test.

Por lo tanto, para el caso del presente proyecto, los especialistas en psicometría sugirieron que no se presentara en los formularios la posibilidad de responder en forma ambigua, sino únicamente una de dos opciones *verdadero* o *falso* como se muestra en la figura 5.6, garantizando de esta manera que los valores incorrectos no estén presentes en los datos de respuesta del test.

Instrucciones: Marca tu respuesta seleccionando la opción correspondiente. Para corregir, simplemente selecciona la opción adecuada.	
31. Tengo dificultades para concentrarme en una tarea o trabajo.	○ V ◉ F
32. He tenido experiencias muy peculiares y extrañas.	○ V ◉ F
33. Raras veces me preocupo por mi salud.	◉ V ○ F
34. Nunca he tenido dificultades a causa de mi conducta sexual.	◉ V ○ F

Figura 5.6: Uno de los formularios de reactivos durante la aplicación del MMPI-2.

En el caso del campo *sexo* correspondiente a la información personal del paciente, el formulario ofrece una lista desplegable que asegura la introducción de un valor adecuado para dicho campo como se observa en la figura 5.7, quedando la responsabilidad de la elección correcta al aplicador autorizado.

Figura 5.7: Lista desplegable del campo sexo.

Los campos *clasificación* (Hs, D, ... I_s) son llenados adecuadamente por el programa que califica el test como se aprecia en la figura 5.8.

Hs	D	Hi	Dp	MfVaron	MfMujer	Pa	Pt	Es	Ma	I_s
Medio	Medio	Medio	Bajo	Medio	Bajo	Bajo	Bajo	Medio	Medio	Moderado
Medio	Medio	Medio	Bajo	Bajo	Bajo	Medio	Bajo	Bajo	Medio	Medio
Moderado	Medio	Medio	Medio	Alto	Alto	Medio	Medio	Moderado	Medio	Moderado
Alto	Alto	Alto	Moderado	Medio	Bajo	Alto	Moderado	Muy alto	Medio	Elevado

Figura 5.8: Campos clasificación de las escalas clínicas (básicas).

5.10.5. Transformación de los datos a la granularidad correcta

El criterio de enfoque de atención es el de proyección de datos, el cual fue sugerido por los especialistas en psicometría, quienes tomaron como referencia las normas de calificación del inventario MMPI-2. Por lo tanto las proyecciones realizadas al total de los 567 reactivos se hicieron en base a las siguientes escalas:

- Indicadores de validez: L, F y K.
- Escalas clínicas: Hs, D, Hi, Dp, Mf, Pa, Pt, Es, Ma, Is.

Dichas proyecciones quedaron establecidas conteniendo al conjunto de reactivos que conforman a cada una de las escalas, más los campos **sexo** y **clasificación** correspondientes (vistas minables descritas en el capítulo 7). A continuación se listan los reactivos que conforman a las escalas.

L - Mentira, tendencia del paciente a mentir al contestar el test (15 reactivos)

16, 29, 41, 51, 77, 93, 102, 107, 123, 139, 153, 183, 203, 232, 260.

Los reactivos que conforman al MMPI-2 son en total 567 y están numerados del 1 al 567, entonces, los números listados anteriormente corresponden al conjunto de 15 reactivos que conforman a la escala L, cuyo contenido se muestra en la tabla 5.1.

No. reactivo	Contenido
16	De vez en cuando pienso en cosas demasiado malas como para hablar de ellas.
29	En ocasiones siento deseos de maldecir.
41	No siempre digo la verdad.
51	No leo diariamente todos los artículos editoriales del periodico.
77	De vez en cuando dejo para mañana lo que debiera hacer hoy.
93	Algunas veces, cuando no me siento bien, soy irritable.
102	Algunas veces me enojo.
107	Mis modales en la mesa no son tan buenos en casa como cuando salgo a comer con otras personas.
123	Si pudiera entrar a un cine sin pagar y estuviera seguro(a) de no ser descubierto (a), probablemente lo haría.
139	Prefiero ganar que perder en un juego.
153	Me gusta conocer a gente importante porque eso me hace sentir importante.
183	No me agradan todas las personas que conozco.
203	En ocasiones me gusta el chisme.
232	En las elecciones, algunas veces voto por candidatos que casi no conozco.
260	A veces me río de los chistes obscenos.

Tabla 5.1: Contenido de los 15 reactivos que conforman a la escala L (mentira).

El contenido de los reactivos que conforman a las escalas restantes no se muestra, para evitar que los posibles pacientes los conozcan y se familiaricen con éstos, restando efectividad a la aplicación del instrumento MMPI-2. La consulta del contenido de todos los reactivos la pueden realizar los usuarios autorizados para el uso de la aplicación computacional MMPI-2 desarrollada en esta tesis.

F - Infrecuencia, tendencia del paciente a contestar el test en forma inconsistente (60 reactivos)

6, 12, 18, 24, 30, 36, 42, 48, 54, 60, 66, 72, 78, 84, 90, 96, 102, 108, 114, 120, 126, 132, 138, 144, 150, 156, 162, 168, 174, 180, 186, 192, 198, 204, 210, 216, 222, 228, 234, 240, 246, 252, 258, 264, 270, 276, 282, 288, 294, 300, 306, 312, 318, 324, 330, 336, 343, 349, 355, 361.

K - Negación, tendencia del paciente a negar sus problemas al contestar el test (30 reactivos)

29, 37, 58, 76, 83, 110, 116, 122, 127, 130, 136, 148, 157, 158, 167, 171, 196, 213, 243, 267, 284, 290, 330, 338, 339, 341, 346, 348, 356, 365.

Hs - Hipocondriasis, tendencia constante del paciente a sentir que padece una o más enfermedades (32 reactivos)

2, 3, 8, 10, 18, 20, 28, 39, 45, 47, 53, 57, 59, 91, 97, 101, 111, 117, 141, 143, 149, 152, 164, 173, 175, 176, 179, 208, 224, 247, 249.

D - Depresión (57 reactivos)

2, 5, 9, 10, 15, 18, 20, 29, 31, 33, 37, 38, 39, 43, 45, 46, 49, 55, 56, 68, 73, 75, 76, 92, 95, 109, 117, 118, 127, 130, 134, 140, 141, 142, 143, 146, 147, 148, 165, 170, 175, 178, 181, 188, 189, 212, 215, 221, 223, 226, 233, 238, 245, 248, 260, 267, 330.

Hi - Histeria de conversión (60 reactivos)

2, 3, 7, 8, 9, 10, 11, 14, 18, 26, 29, 39, 40, 44, 45, 47, 58, 65, 76, 81, 91, 95, 98, 101, 110, 115, 116, 124, 125, 129, 135, 141, 148, 151, 152, 157, 159, 161, 164, 166, 167, 172, 173, 175, 176, 179, 185, 193, 208, 213, 218, 224, 230, 241, 243, 249, 253, 263, 265.

Dp - Desviación psicopática (50 reactivos)

9, 12, 17, 21, 22, 31, 32, 34, 35, 42, 52, 54, 56, 70, 71, 79, 82, 83, 89, 94, 95, 99, 105, 113, 122, 125, 129, 143, 157, 158, 160, 167, 171, 185, 202, 209, 214, 217, 219, 225, 226, 243, 259, 261, 263, 264, 266, 267, 288.

Mf - Masculinidad-femineidad (balance de masculinidad en el varón) (56 reactivos)

1, 4, 19, 25, 26, 27, 62, 63, 64, 67, 68, 69, 74, 76, 80, 86, 103, 104, 107, 112, 119, 120, 121, 122, 128, 132, 133, 137, 163, 166, 177, 184, 187, 191, 193, 194, 196, 197, 199, 201, 205, 207, 209, 219, 231, 235, 236, 237, 239, 251, 254, 256, 257, 268, 271, 272.

Mf - Masculinidad-femineidad (balance de feminidad en la mujer) (56 reactivos)

1, 4, 19, 25, 26, 27, 62, 63, 64, 67, 68, 69, 74, 76, 80, 86, 103, 104, 107, 112, 119, 120, 121, 122, 128, 132, 133, 137, 163, 166, 177, 184, 187, 191, 193, 194, 196, 197, 199, 201, 205, 207, 209, 219, 231, 235, 236, 237, 239, 251, 254, 256, 257, 268, 271, 272.

Pa -Paranoia (40 reactivos)

16, 17, 22, 23, 24, 42, 81, 95, 98, 99, 100, 104, 110, 113, 138, 144, 145, 146, 162, 234, 244, 255, 259, 266, 271, 277, 283, 284, 285, 286, 297, 305, 307, 314, 315, 333, 334, 336, 355, 361.

Pt - Psicastenia, tendencia del paciente a sentir constantemente mucha tristeza (48 reactivos)

3, 9, 11, 16, 23, 31, 33, 38, 56, 65, 73, 82, 89, 94, 109, 130, 140, 147, 165, 170, 174, 175, 196, 218, 242, 273, 275, 277, 285, 289, 293, 301, 302, 304, 308, 309, 310, 313, 316, 317, 320, 321, 325, 326, 327, 328, 329, 331.

Es - Esquizofrenia (78 reactivos)

6, 9, 12, 16, 17, 21, 22, 23, 31, 32, 34, 35, 38, 42, 44, 46, 48, 65, 85, 90, 91, 92, 106, 138, 145, 147, 165, 166, 168, 170, 177, 179, 180, 182, 190, 192, 210, 218, 221, 229, 233, 234, 242, 247, 252, 255, 256, 268, 273, 274, 276, 277, 278, 279, 280, 281, 287, 290, 291, 292, 295, 296, 298, 299, 303, 307, 311, 316, 319, 320, 322, 323, 325, 329, 332, 333, 343, 355.

Ma - Hipomanía (46 reactivos)

13, 15, 21, 23, 50, 55, 61, 85, 87, 88, 93, 98, 100, 106, 107, 113, 122, 131, 136, 145, 154, 155, 158, 167, 168, 169, 182, 190, 200, 243, 263.

Is- Introversión social (69 reactivos)

25, 31, 32, 49, 56, 70, 79, 86, 100, 104, 106, 110, 112, 127, 131, 135, 158, 161, 167, 181, 185, 189, 207, 209, 215, 231, 237, 243, 251, 255, 262, 265, 267, 275, 280, 284, 289, 296, 302, 308, 321, 326, 328, 335, 337, 338, 340, 342, 344, 345, 347, 348, 350, 351, 352, 353, 354, 357, 358, 359, 360, 362, 363, 364, 366, 367, 368, 369, 370.

5.10.6. Definición del tipo de problema de minería

En este caso la herramienta ADC4.5Web sólo puede resolver problemas de clasificación.

5.10.7. Selección del algoritmo de Minería de Datos

ADC4.5Web únicamente emplea al algoritmo C4.5, la idea es que en un trabajo futuro se pueda extender para incorporar más algoritmos de minería.

5.10.8. Proceso de Minería de Datos

Esta etapa se lleva a cabo con la herramienta de minería ADC4.5Web.

5.10.9. Evaluación del modelo e interpretación de resultados

La evaluación se lleva a cabo de manera conjunta entre el personal que colabora en el proceso de minería: personas con perfil en computación y los especialistas en psicometría que emplean al MMPI-2.

5.10.10. Selección del mejor modelo

Al igual que en la etapa anterior, esta fase se lleva a cabo de manera conjunta entre el personal que colabora en el proceso de minería: personas con perfil en computación y los especialistas en psicometría que emplean al MMPI-2.

5.10.11. Conocimiento descubierto

El apartado para administración de la aplicación computacional MMPI-2 desarrollada (descrita en el capítulo 6), hace posible la aplicación y calificación del test en forma automatizada de acuerdo a las normas de MMPI-2, así como el manejo de expedientes, esta aplicación ha tenido una gran aceptación por parte de los especialistas en psicometría que emplean al MMPI-2. La incorporación del conocimiento descubierto, como alternativa simplificada para la aplicación y calificación del test es aceptada con cierto excepticismo por los especialistas cercanos al proyecto, pues es de su conocimiento que fue generada por un agente de software y no por un ser humano.

5.11. Relación con otras disciplinas

El proceso KDD ha servido para unir investigadores de áreas en principio dispersas como Inteligencia Artificial, Estadística, Técnicas de Visualización, Matemáticas, Apendizaje Automático y Bases de Datos, en la búsqueda de técnicas eficientes que ayuden a encontrar el conocimiento potencial que se encuentra inmerso en los grandes volúmenes de datos almacenados por las organizaciones diariamente.

5.12. Aplicaciones de la Minería de Datos

La integración de las técnicas de MD en las actividades cotidianas se está convirtiendo en algo habitual. Se pueden encontrar ejemplos prácticamente en todo tipo de aplicaciones: financieras, seguros, científicas (medicina, psicología, farmacéutica, astronomía, etc), políticas, económicas, demográficas, educación, procesos industriales, etc.

5.13. Conclusión del capítulo

El proceso de KDD facilita al Agente Inteligente Descubreconocimiento un análisis automatizado, bastante rápido y eficiente de grandes cantidades de datos crudos, es decir, el descubrimiento de conocimiento automatizado en bases de datos. Este análisis consiste en interpretar grandes cantidades de datos en un proceso iterativo e interactivo, para buscar e identificar relaciones o patrones que puedan ser representados como conocimiento. El conocimiento descubierto debe ser válido, novedoso y potencialmente útil para satisfacer las metas del Agente Inteligente Descubreconocimiento siempre que sea posible, para que éste incorpore el conocimiento obtenido en algún sistema real o tome decisiones a partir de los resultados alcanzados o, simplemente, registre la información conseguida y se la proporcione a quien esté interesado.

Como el proceso KDD es iterativo, posibilita la retroalimentación permitiendo regresar a algunas etapas anteriores y como es interactivo, permite la colaboración colectiva e interdisciplinaria por parte del personal involucrado en dicho proceso.

El algoritmo de referencia propuesto para el proceso KDD es de uso general y se complementa con el empleo de la herramienta de minería desarrollada en esta tesis: Ambiente Descubreconocimiento con C4.5 para Web (ADC4.5Web), esto facilita las tareas de investigación o de negocios a quienes conforman al Agente Inteligente Descubreconocimiento.

Capítulo 6

Inventario Multifásico de la Personalidad Minnesota 2

En este capítulo se describen de manera breve, algunos conceptos como son: test o prueba psicológica, rasgos, personalidad y un poco más a detalle, el caso de estudio: el Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2). Esto se hace con la finalidad de poder evaluar la eficacia del algoritmo de referencia de minería propuesto (descrito en el capítulo 5) y la herramienta de minería desarrollada (ADC4.5Web, descrita en el capítulo 7), aplicándolos a la obtención experimental de una versión reducida del inventario MMPI-2 y compararla con la versión completa del mismo.

Es necesario conocer el dominio de aplicación sobre el que se realiza la tarea de minería, que en este caso es el inventario MMPI-2, por esta razón, en este capítulo se presentan las características del MMPI-2: estructura, fundamento, normas, formas de aplicación, evaluación e interpretación. Ya que una vez conocidas dichas características, se puede formular una buena estrategia para que el Agente Inteligente Descubreconocimiento, aplique adecuadamente el algoritmo de referencia para el proceso KDD propuesto en el capítulo anterior, empleando la herramienta de minería desarrollada en el presente trabajo: Ambiente Descubreconocimiento con C4.5 Web, la que se describe en el capítulo siguiente.

6.1. Inventarios de la Personalidad

A finales de la década de los 30's las pruebas de personalidad experimentaron un auge importante, éstas intentaban medir características estables o rasgos en un individuo. Las primeras implementaciones de estas pruebas se hicieron con papel y lápiz, y éstas podían ser de selección múltiple o de elección forzada de falso-verdadero, estas tendencias han perdurado hasta estos dias y se han complementado con el desarrollo de alternativas que hacen uso de la tecnología moderna. Las pruebas de personalidad poseen una estructura definida, por lo que son conocidas como pruebas objetivas estructuradas de personalidad.

El objetivo principal de las pruebas estructuradas de la personalidad es la evaluación de los rasgos de personalidad, estados de la personalidad y otros aspectos como la autoestima [Lucio y León, 2003].

De este tipo de pruebas los inventarios de la personalidad son las más populares, de los que el más ampliamente utilizado en la actualidad es el Inventario Multifásico de la Personalidad Minnesota 2.

A continuación se definen algunos conceptos para facilitar la explicación del MMPI-2.

Prueba psicológica (test).- "Es una serie de reactivos que miden características de los seres humanos que determinan su conducta. Al existir varios tipos de conductas, existen también varios tipos de pruebas psicológicas: de habilidades, de aptitudes, de intereses, de inteligencia, de personalidad, etc" [Lucio y León, 2003].

Rasgos.- "Son disposiciones relativamente estables y duraderas, como tendencias a actuar, a pensar o a sentir que distinguen a una persona de otra" [Lucio y León, 2003].

Personalidad.- Es el patrón de rasgos (sentimientos, pensamientos y comportamiento) que persisten a través del tiempo y de las situaciones. La personalidad define los aspectos que distinguen a los individuos. De manera informal se puede decir que la personalidad es la "firma psicológica" de un individuo, ya que ésta es típica y exclusiva de dicho individuo.

6.2. Breve historia del MMPI-2

El empleo de métodos de autoinforme se remonta a finales del siglo XIX, cuando se buscó la forma de lograr que los individuos se calificaran a sí mismos de acuerdo a factores (rasgos) de personalidad. Durante la Primera Guerra Mundial se desarrolló el primer cuestionario de personalidad llamado Cuestionario de Datos Personales, lo que propició un amplio seguimiento de los cuestionarios de autoinforme durante las décadas de los 20's y los 30's. Los primeros inventarios eran en su mayoría cuestionarios obtenidos de manera lógica, los cuales se recopilaban a partir de constructos teóricos definidos vagamente, en los que se prestaba muy poca atención a cuestiones como la precisión o validez de las medidas.

A fines de la década de los 30's, el psicólogo Starke Hathaway y el psiquiatra J. C. McKinley introdujeron una perspectiva diferente para la evaluación de la personalidad: el método empírico [Butcher, 2001]. Con este criterio, dichos autores desarrollaron instrumentos para evaluar y diagnosticar a pacientes que padecían de trastornos mentales en los hospitales de la Universidad de Minnesota. Estos instrumentos se convirtieron en el Inventario Multifásico de la Personalidad Minnesota (MMPI, el cual fue publicado por Hathaway hasta 1965). Muy pronto su uso se extendió a clínicas psiquiátricas y hospitales de los Estados Unidos empleándose en diversos escenarios como casos de medicina general, casos de instituciones educativas, personas con problemas de drogadicción y alcoholismo, selección de personal para puestos importantes, entre otros.

En los últimos años de la década de los 40's y al principio de los años 50's, se tradujo y comenzó el uso del MMPI en otros países como Italia, Alemania y Puerto Rico. Actualmente se tiene conocimiento de que el instrumento MMPI se utiliza en 46 países, por lo que este instrumento ha llegado a ser el inventario de personalidad objetivo, más ampliamente utilizado e investigado en todo el mundo [Lucio y León, 2003].

Posteriormente, a fines de la década de los 60's el MMPI comenzó a presentar signos de envejecimiento, lo cual sugería una revisión del instrumento original. Sin embargo, cambiar una herramienta que aún funcionaba de manera razonable no fue fácil, por dicha razón no fue sino hasta la década de los 80's que dio inicio la revisión del instrumento, la que duró 10 años y en la que se utilizaron más de 15,000 individuos de la población general, provenientes de diversos grupos normales y clínicos de los Estados Unidos. Cuando se obtuvieron datos sustanciales se publicaron dos versiones del inventario: el MMPI-2 para adultos y el MMPI-A para adolescentes, publicados en 1989 y 1992 respectivamente [Butcher, 2001].

6.3. Uso del MMPI-2 en México

El uso del MMPI en México se remonta a fines de la década de los 60's, cuando el Dr. Rafael Núñez publicó el MMPI en español para la población mexicana. Los profesionales que emplearon esta traducción notaron serios problemas lingüísticos y deficiencias culturales, lo que produjo distorsiones y problemas en su interpretación. Sin embargo dicha traducción se empleó por varios años.

En 1995, la Dra. Emilia Lucio Gómez-Maqueo publicó la versión en español del Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2) junto con el Manual de la prueba, constituyendo una de las tres versiones en español autorizadas por la Editorial de la Universidad de Minnesota [Hathaway y McKinley, 1995]. El Manual ofrecido no sólo es una simple traducción, sino más bien una transliteración, la que incluye información de las primeras investigaciones realizadas en México con el MMPI-2 sobre una muestra de la población universitaria de la UNAM. El MMPI-A (para adolescentes) fue publicado en 1998 también por la Dra. Lucio. En el año 2003 la Dra. Emilia Lucio G.M. en colaboración con la Dra. Ivonne León publican la obra *Uso e interpretación del MMPI-2 en español*, en la que presentan el proceso de adaptación del instrumento terminado para su utilización en México, obteniendo normas para la población mexicana en general.

6.4. Fundamento del Instrumento

El MMPI se implementó mediante un esquema psicométrico con un enfoque empírico, con el cual se buscaba identificar a los reactivos que diferenciaran entre un grupo clínico (criterio) de individuos y uno de personas normales, de esta manera, con un conjunto de reactivos identificado adecuadamente se construyeron cada una de las escalas. Los reactivos fueron formulados a partir de historias clínicas de los pacientes del hospital psiquiátrico de Minnesota y de la información que los pacientes proporcionaban acerca de sí mismos y de sus síntomas, ya que Hathaway y McKinley basándose en su experiencia clínica, confiaban en que los pacientes podrían y querrían describir con honestidad sus problemas por medio de autoinformes si se les proporcionaban las condiciones adecuadas [Butcher, 2001]. Dichos autores desarrollaron también cuatro escalas a las que denominaron de validez, con el propósito de detectar actitudes con las que un individuo pretendía distorsionar sus respuestas al responder el cuestionario.

6.5. Normalización, confiabilidad y validez del MMPI-2 en México

Con el propósito de utilizar el MMPI-2 en México, es decir, en una población distinta a la que sirvió de base para la elaboración y estandarización del instrumento, fue necesario un proceso de adaptación complicado. Para el proceso de adaptación se consideró conveniente aplicar la prueba inicialmente a una muestra de estudiantes de la Universidad Nacional Autónoma de México, ya que generalmente el nivel de lectura de los estudiantes universitarios así como su actitud ante la prueba son adecuados para la aplicación del inventario. Con esto se obtuvieron las normas del instrumento para su empleo con estudiantes (universitarios), las que fueron publicadas por la Dra. Lucio en 1995 [Hathaway y McKinley, 1995].

Posteriormente, para concluir el proceso de adaptación del instrumento se trabajó en la obtención de normas para la población mexicana en general, por lo que se realizó un estudio de normalización con la versión en español del instrumento para México, con la muestra más amplia posible que se obtuvo de la población general, proveniente de diferentes regiones del país [Lucio y León, 2003]. Estas fueron publicadas por la Dra. Lucio y la Dra. León en el 2003.

En la validación de la posibilidad de utilizar el instrumento en una cultura distinta a aquella en la fue construido, debe corroborarse si en la nueva cultura el instrumento conserva su estructura factorial. Con respecto a este análisis, se puede concluir que el MMPI-2 muestra factores similares en la población mexicana y la estadounidense, indicando que es una prueba válida para su utilización con la población mexicana y que, la traducción al español que se realizó en México es confiable [Lucio y León, 2003]. También fue realizado un estudio, para la determinación de la confiabilidad test-retest del MMPI-2 versión al español para México en una muestra de estudiantes (universitarios), obteniéndose resultados con un alto nivel de confiabilidad del instrumento [Lucio y León, 2003].

6.6. Reactivos y Escalas del MMPI-2

El MMPI-2 está compuesto por cuatro grupos de escalas: indicadores de validez, escalas clínicas o básicas, escalas suplementarias y escalas de contenido. Las escalas a su vez están compuestas por un conjunto definido de reactivos y cada reactivo posee un contenido (una afirmación). El inventario está compuesto por un total de 567 reactivos. La información detallada de los reactivos, así como los grupos de éstos que conforman a las escalas se encuentra en el Manual del MMPI-2 publicado por la Dra. Lucio [Hathaway y McKinley, 1995]. Enseguida se describen brevemente las escalas que conforman al MMPI-2.

En primer lugar se tienen a los *indicadores de validez*, los que determinan si una aplicación del instrumento es válida para interpretarlo, de tal forma que se incluya la actitud del individuo al responder el inventario y la medida en que esto corresponde con otros antecedentes que se tengan sobre dicho individuo, ya que podría existir la posibilidad de que éste hubiese distorsionado o manipulado sus respuestas. Estos indicadores se muestran a continuación en la tabla 6.1.

En segundo lugar, las *escalas clínicas o básicas*, mismas que fueron desarrolladas para detectar si el paciente padece alguna psicopatología, por lo que éstas adquieren un sentido significativo al

Indicadores de validez		
? No puedo decir		Número de reactivos no contestados adecuadamente
Escalas de validez		
L	**Mentira**.- Tendencia del paciente a mentir al contestar el test.	15 reactivos
F	**Infrecuencia**.- Tendencia del paciente a contestar el test en forma inconsistente.	60 reactivos
K	**Negación**.- Tendencia del paciente a negar sus problemas al contestar el test.	30 reactivos
Adicionales		
Fp	**Infrecuencia posterior**	40 reactivos
INVAR	**Inconsistencia en las respuestas variables**	67 pares de reactivos de respuesta
INVER	**Inconsistencia en las respuestas verdaderas**	23 pares de reactivos de respuesta

Tabla 6.1: Indicadores de validez.

reflejar dicha característica. Estas escalas se muestran en la tabla 6.2, la primer y segunda columnas corresponden al nombre abreviado de las escalas, la tercer columna corresponde al nombre de la escala y la cuarta columna corresponde al número de reactivos que conforman a la escala.

Escalas clínicas			
1	**Hs**	**Hipocondriasis**.- Actitud de preocupación exagerada por la salud, con la creencia errónea de padecer una o más enfermedades físicas graves.	32 reactivos
2	**D**	**Depresión**.- Estado emocional caracterizado por tristeza, desesperanza y pérdida de interés en las actividades habituales.	57 reactivos
3	**Hi**	**Histeria conversiva**.- Trastorno psíquico de origen nervioso caracterizado por fuerte ansiedad y reacciones agudas, que puede provocar ataques convulsivos, parálisis y otros trastornos.	60 reactivos
4	**Dp**	**Desviación psicopática**.- Anomalía psíquica que causa conducta social patológicamente alterada del individuo que la padece, a pesar de la integridad de las funciones perceptivas y mentales.	50 reactivos
5	**Mf**	**Masculinidad-femineidad**.- Balance de masculinidad en el varón o balance de feminidad en la mujer.	56 reactivos
6	**Pa**	**Paranoia**.- Conjunto de perturbaciones mentales que provocan un estado de delirio, que se caracteriza por pensamientos sobre sospechas o sentimientos exagerados de trato injusto o acoso.	40 reactivos
7	**Pt**	**Psicastenia**.- Trastorno mental que produce sensaciones continuas de miedo, angustia, obsesiones, así como los trastornos que se derivan de este estado de tensión aguda.	48 reactivos
8	**Es**	**Esquizofrenia**.- Grupo de trastornos mentales en personas con alteraciones en la percepción o la expresión de la realidad, que les provocan un pensamiento desorganizado, delirios, alucinaciones, alteraciones afectivas , del lenguaje y conductuales.	78 reactivos
9	**Ma**	**Hipomanía**.- Estado afectivo parecido a la Manía pero más leve, que altera el ánimo provocando episodios de aumento de energía que pueden durar de horas a días, sin perder contacto con la realidad.	46 reactivos
0	**Is**	**Introversión social**.- Actitud típica caracterizada por la concentración del interés en los procesos internos del individuo, manifestada por timidez y tendencia a apartarse de contactos sociales.	69 reactivos

Tabla 6.2: Escalas clínicas.

En tercer lugar, se cuenta con algunas escalas especiales y suplementarias del MMPI incluidas en el perfil de *escalas suplementarias* del MMPI-2, complementando de esta forma la interpretación de las escalas clínicas del instrumento. A continuación se muestran dichas escalas en la tabla 6.3, la primer columna corresponde al nombre abreviado de la escala, la segunda columna corresponde al nombre de la escala (sólo se describen las 3 primeras, puesto que el nombre de las restantes es relativamente autoexplicativo) y la tercer columna corresponde al número de reactivos que conforman a la escala.

Escalas suplementarias		
Tradicionales		
A	**Ansiedad**.- Trastorno caracterizado por inquietud y preocupación excesivas, lo cual está fuera de la proporción del impacto del evento o circunstancia motivo de la preocupación.	39 reactivos
R	**Represión**.- Es el mecanismo de defensa que empuja hacia las zonas inconscientes de la mente, a las experiencias y apetitos que un sujeto considera inaceptables y dificíles de integrar en su personalidad.	37 reactivos
Fyo	**Fuerza del yo**.- Es el grado de mantenimiento de las funciones del *yo*, por ejemplo, percepción, pensamiento y habla.	52 reactivos
A-MAC	**Alcoholismo de MacAndrew**	49 reactivos
Adicionales		
HR	**Hostilidad reprimida**	28 reactivos
Do	**Dominancia**	25 reactivos
Rs	**Responsabilidad social**	30 reactivos
Dpr	**Desajuste profesional**	41 reactivos
GM	**Género masculino**	47 reactivos
GF	**Género femenino**	46 reactivos
EPK	**Desorden de estrés postraumático de Keane**	46 reactivos
EPS	**Desorden de estrés postraumático de Schlenger**	60 reactivos

Tabla 6.3: Escalas suplementarias.

Y finalmente, las *escalas de contenido*, siendo éstas de gran utilidad para refinar o precisar el significado de las elevaciones en la puntuación que se presentan en las escalas clínicas. A través de dichas escalas se puede obtener información acerca de sentimientos, el funcionamiento de la personalidad y de problemas pasados o actuales del individuo, ya que esta información no está disponible en otras escalas clínicas. Estas escalas se muestran en la tabla 6.4, la primer columna corresponde al nombre abreviado de la escala, la segunda columna corresponde al nombre de la escala y la tercer columna corresponde al número de reactivos que conforman a la escala.

Escalas de contenido		
ANS	**Ansiedad**	23 reactivos
MIE	**Miedos**	23 reactivos
OBS	**Obsesividad**	16 reactivos
DEP	**Depresión**	33 reactivos
SAU	**Preocupación por la salud**	36 reactivos
DEL	**Pensamiento delirante**	23 reactivos
ENJ	**Enojo**	16 reactivos
CIN	**Cinismo**	23 reactivos
PAS	**Prácticas antisociales**	22 reactivos
PTA	**Personalidad tipo A**	19 reactivos
BAE	**Baja autoestima**	24 reactivos
ISO	**Incomodidad social**	24 reactivos
FAM	**Problemas familiares**	25 reactivos
DTR	**Dificultad en el trabajo**	33 reactivos
RTR	**Rechazo al tratamiento**	26 reactivos

Tabla 6.4: Escalas de contenido.

6.7. Aplicación del MMPI-2

La aplicación de la prueba consiste en proporcionar al individuo un formato (generalmente un cuadernillo) conteniendo las frases correspondientes a cada uno de los 567 reactivos del instrumento y una hoja de respuestas, en la que el individuo debe marcar la respuesta (verdadero o falso) apropiada a su caso particular, evitando omitir respuestas o responder en forma ambigua en la medida de lo posible.

La aplicación de la prueba puede ser individual o colectiva, recomendándose que ésta se aplique en un tiempo máximo de 3 horas, ya sea en una sesión y sólo si es necesario, en dos sesiones. Debe procurarse que el ambiente de aplicación sea cómodo, con privacía y supervisado. Es requisito para el individuo que contesta la prueba, tener como mínimo la habilidad de lectura de segundo grado de secundaria. También resulta recomendable que la prueba se aplique preferentemente a la población urbana, debido a que las actividades contenidas en los reactivos, están bastante relacionadas con la vida en la ciudad. El aplicador de la prueba debe estar al pendiente de que no se presente alguna condición como pueden ser: agudeza visual limitada, dislexia, desórdenes de aprendizaje, intoxicación por abuso de sustancias, entre otras, que incapacite al individuo para contestar adecuadamente el instrumento.

Las personas encargadas de aplicar la prueba deben ser entrenadas cuidadosamente y además, es necesario que estén muy bien informadas sobre los pasos requeridos para la obtención de un rendimiento válido y útil [Lucio y León, 2003]. Al emplear el MMPI-2 se deben tomar las precauciones de seguridad éticas y profesionales propias de los instrumentos de evaluación psicológica, al respecto Lucio y León (2003) señalan:

> Cualquier aplicación del MMPI-2 debe llevarse a cabo de tal manera que garantice al participante la discreción y seguridad de que los resultados del examen serán respetados,

> protegidos y utilizados sólo para beneficio y mejoramiento de su bienestar. La aplicación o calificación negligente, la falta de cuidado para mantener los resultados seguros y a salvo, o alguna otra evidencia de insensibilidad con respecto a la aplicación de la prueba, puede desvirtuar el valor de la información obtenida por medio del MMPI-2.

6.8. Material para la aplicación del MMPI-2

El material para la aplicación de la prueba en México consiste en un solo formato de cuadernillo y dos formatos de hojas de respuesta, de las que la roja sirve para calificarse con computadora (capturando ópticamente el contenido de la hoja de respuestas) y la morada, que es para calificación manual (utilizando plantillas). También se cuenta con dos hojas de perfil, siendo el primero con las normas estadounidenses y el segundo con las normas mexicanas. Cuando se elabora un perfil se recomienda emplear el de las normas mexicanas, debido al criterio que Lucio y León (2003) indican:

> El perfil de normas mexicanas corresponde al estudio hecho en la población general, por lo que se considera más adecuado utilizar el perfil de las normas mexicanas cuando se trata de participantes individuales. También es aconsejable emplear este perfil si se realiza investigación y se desea comparar los perfiles de diferentes grupos de la población mexicana. Sin embargo, si el objetivo es comparar un grupo específico de la población mexicana con la población estadounidense, es aconsejable utilizar el perfil de las normas estadounidenses, pues así resultan más ilustrativas las diferencias.

En este trabajo se desarrolló y empleó un formato experimental de formularios web para presentar y responder los reactivos, como complemento para agilizar la aplicación y calificación del instrumento. Actualmente este formato es para uso exclusivo de los participantes en el presente proyecto, no está disponible al público, puesto que el MMPI-2 posee derechos de autor y sus propietarios se reservan la autorización del uso de cualquier material desarrollado en base a dicho inventario, entonces, para que el MMPI-2Web esté a disposición del público, se debe solicitar la autorización de los propietarios del MMPI-2.

6.9. Aplicación computacional MMPI-2Web

La aplicación computacional MMPI-2 permite aplicar la versión estándar del inventario MMPI-2 en forma individual y colectiva, además lo califica de manera automática, incluyendo gráficas e interpretación de los resultados. También facilita la administración de pacientes, diagnósticos y reactivos.

6.9.1. Instalación del MMPI-2Web

MMPI-2Web es una aplicación cliente/servidor, por lo que para la parte del servidor se necesita lo siguiente:

- Los requerimientos mínimos de hardware son: procesador Intel PENTIUM-II o superior (o equivalente), disco duro de 6 Gb o superior, memoria RAM de 128 Mb o superior, acceso a una red LAN (red de área local) y/o WAN (red de área global).
- Los requerimientos de software son: sistema operativo SuSE Linux 8.0 o equivalente, sistema manejador de bases de datos MySQL 3.23.48 o equivalente, servidor web Apache 1.3.23 con los módulos de PHP 4.1.0 incluidos.
- Se recomienda que la instalación y configuración del MMPI-2Web, así como la administración de las bases de datos, las realice un profesional con perfil en computación.

6.9.2. Ejecución del MMPI-2Web

Para la parte del cliente, la ejecución del MMPI-2Web se puede realizar con cualquier navegador web con soporte para JAVA, desde cualquier computadora, con cualquier sistema operativo, siempre y cuando se tenga acceso al servidor por medio de una red LAN o WAN. Para ejecutar la aplicación MMPI-2Web, únicamente basta escribir en la barra de direcciones del navegador web, la dirección correpondiente al archivo *main_ admin_ minnesota1.php*, misma que debe ser proporcionada al usuario por el administrador del MMPI-2Web. En la figura 6.1 se muestra un ejemplo de ruta de acceso a MMPI-2Web.

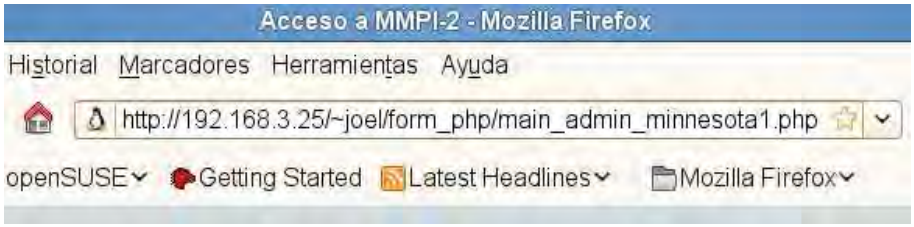


Figura 6.1: Ejemplo de ruta de acceso al MMPI-2Web.

6.9.3. Manejo del MMPI-2Web

Cuando se escribe la ruta correspondiente del archivo *main_ admin_ minnesota1.php* en el navegador web, aparece el menú de bienvenida mostrado en la figura 6.2. Este menú presenta dos opciones para el MMPI-2 y una para minería de datos. El botón ***Administración*** conduce al formulario que permite entrar al menú con las opciones para administrar datos (consultas, altas, bajas y actualizaciones) de los pacientes, diagnósticos y reactivos. El botón ***Aplicar test*** abre el formulario para ingresar a la aplicación de un test previamente registrado en la opción ***Administración***. El botón ***Minería de Datos*** se describe en el capítulo 7. Cabe mencionar que el manejo del MMPI-2Web es sencillo e intuitivo.

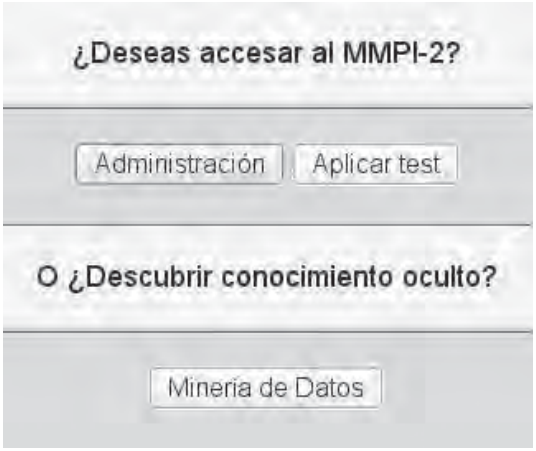


Figura 6.2: Menú de bienvenida al MMPI-2Web.

6.9.4. Formulario de acceso a la administración del MMPI-2Web

Cuando se oprime el botón ***Administración*** del menú de bienvenida al MMPI-2Web se abre el formulario que se muestra en la figura 6.3. El nombre de usuario, clave de usuario y nombre de la base de datos, deben ser proporcionados por el administrador del MMPI-2Web. Estos datos se introduc*altas/baja*sen en los campos correspondientes y se oprime el botón ***Enviar Datos***, como se indica en el formulario de la figura 6.3.

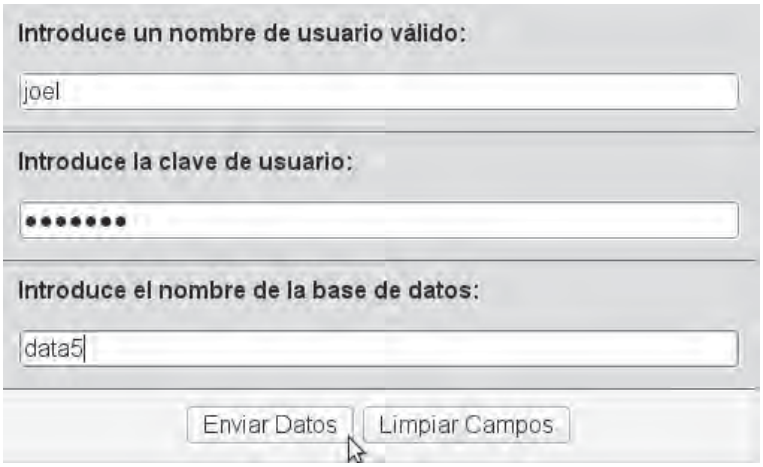


Figura 6.3: Formulario de acceso a la administración del MMPI-2Web.

6.9.5. Menú de administración del MMPI-2Web

En el formulario mostrado en la figura 6.4 se observan las opciones de administración del MMPI-2Web, divididas en los grupos consultas, altas/bajas y actualizaciones, aplicables a la información personal de pacientes, diagnósticos y reactivos del MMPI-2. El grupo minería de datos se describe en el capítulo 7.

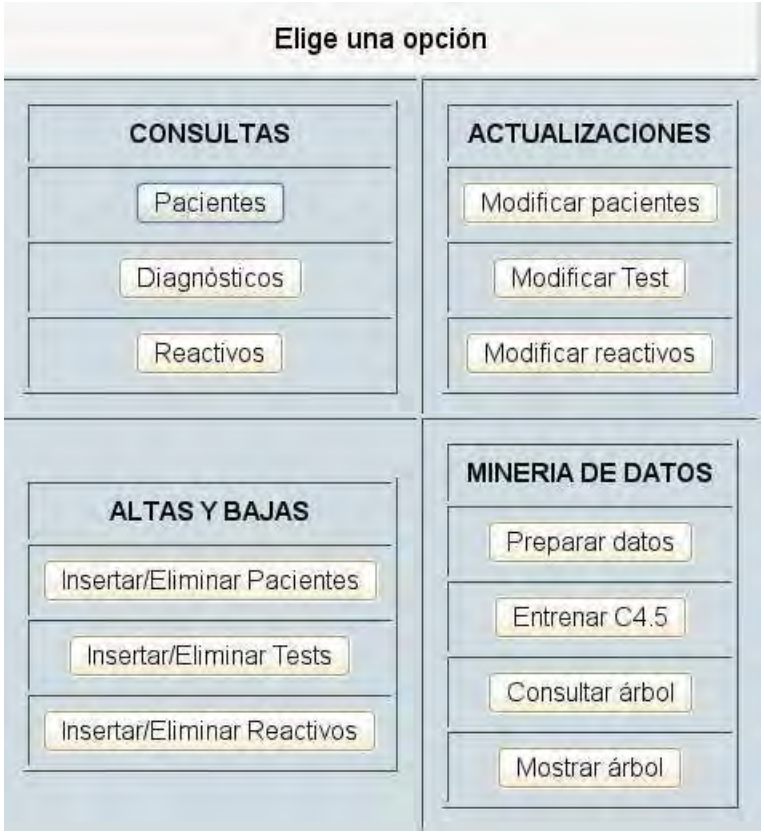


Figura 6.4: Menú de administración del MMPI-2Web.

6.9.5.1. Consultas del MMPI-2Web

Para consultar información de los pacientes se oprime el botón ***Pacientes*** del formulario mostrado en la figura 6.4, éste abre al formulario correspondiente con las siguientes opciones de consulta:

Lista de pacientes.- Esta opción permite seleccionar los campos que se desean consultar: apellidos paterno y materno, nombre, sexo y fecha de nacimiento. Con el botón ***Mostrar todos*** del formulario de la figura 6.5, se muestra la lista completa de pacientes. Se puede mostrar una lista de pacientes por sexo, seleccionando primero el sexo en la lista desplegable y oprimiendo después el botón ***Por sexo***, como se puede observar en la figura 6.5. También se puede mostrar una lista de pacientes por carrera, seleccionando primeramente la carrera en la lista desplegable y oprimiendo después el botón ***Por carrera***, como se aprecia en la figura 6.5.

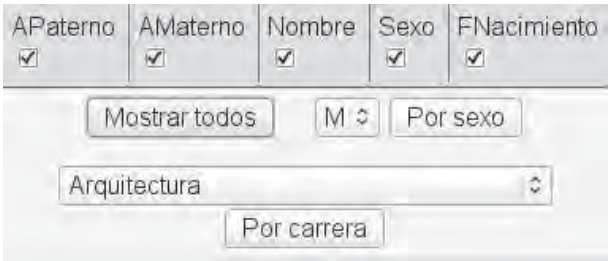


Figura 6.5: Opciones para consultar lista de pacientes.

Lista de pacientes y sus direcciones.- De manera similar a la anterior, esta opción muestra los campos del nombre del paciente y permite seleccionar los campos de dirección que se desean consultar: calle, número, ciudad, estado, colonia, código postal, teléfono y correo electrónico. Con el botón ***Ver direcciones*** del formulario de la figura 6.6, se muestra la lista completa de pacientes y sus direcciones. Se puede mostrar una lista de direcciones de pacientes por sexo, seleccionando primero el sexo en la lista desplegable y oprimiendo después el botón ***Por sexo.***, como se puede observar en la figura 6.6. También se puede mostrar una lista de direcciones de pacientes por carrera, seleccionando primeramente la carrera en la lista desplegable y oprimiendo después el botón ***Por carrera.***, como se aprecia en la figura 6.6.

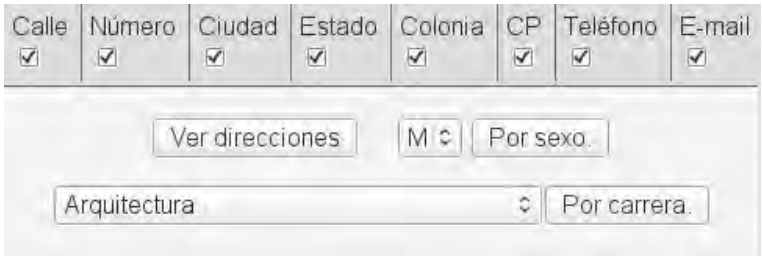


Figura 6.6: Opciones para consultar lista de pacientes y direcciones.

Consultar un paciente en particular.- Esta opción permite mostrar la información de un paciente en particular, proporcionando algún dato de su nombre o bien su nombre completo como se aprecia en la figura 6.7.

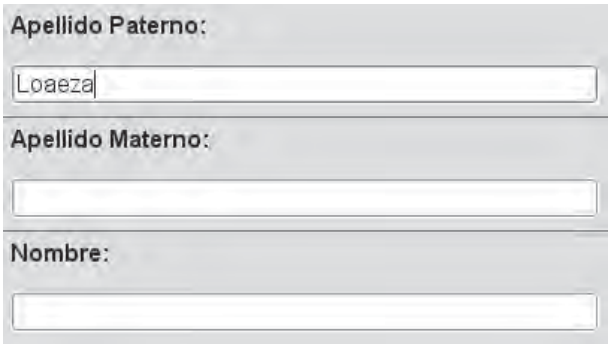


Figura 6.7: Formulario para consultar un paciente en particular.

Además, esta opción también permite seleccionar los campos que se desean consultar: apellidos paterno y materno, nombre, sexo y fecha de nacimiento. Con el botón ***Buscar paciente*** del formulario de la figura 6.8, se muestra la información del paciente seleccionado.

Consultas del MMPI-2Web

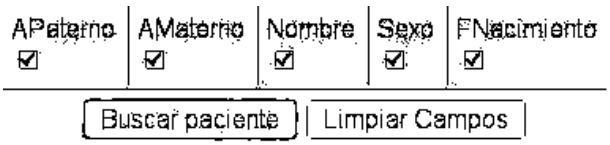


Figura 6.8: Botón *Buscar paciente*.

Con el botón ***Buscar dirección*** del formulario de la figura 6.9, se muestran los campos del nombre del paciente y además, permite seleccionar los campos de la dirección que se desean consultar: calle, número, ciudad, estado, colonia, código postal, teléfono y correo electrónico.

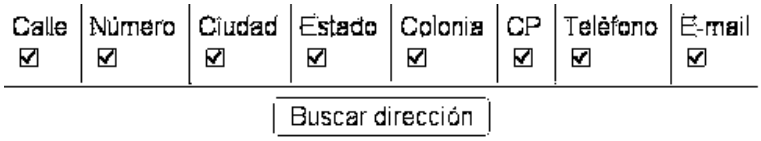


Figura 6.9: Botón *Buscar dirección*.

Lista de diagnósticos con búsqueda personalizada de pacientes.- El botón ***Diagnósticos*** de la figura 6.4, abre el formulario con las opciones para buscar un paciente del que se desea consultar un diagnóstico, como se puede observar en la figura 6.10.

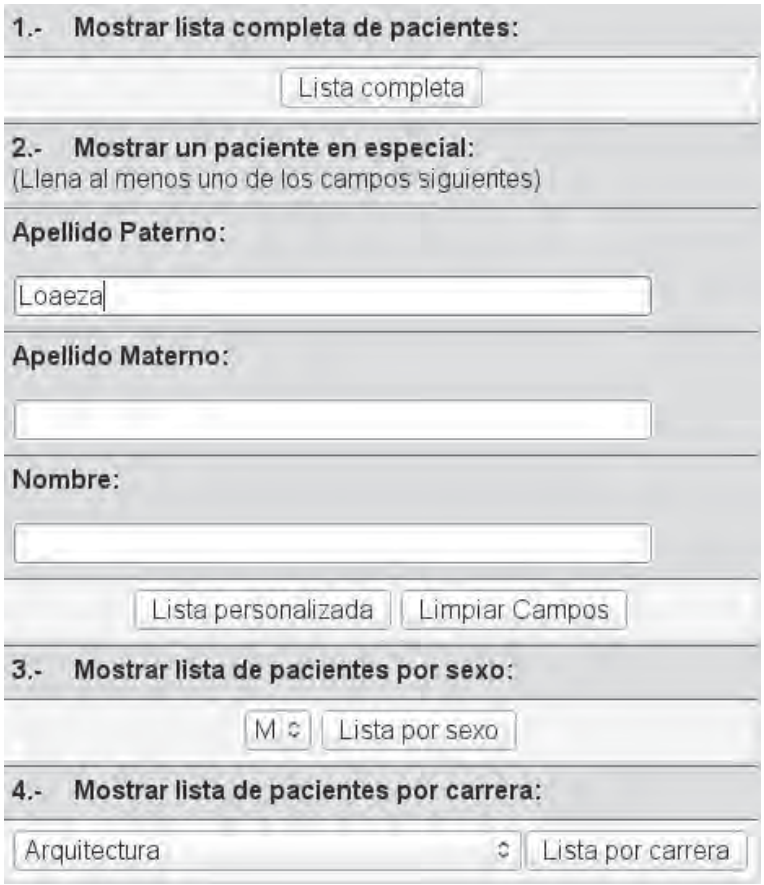


Figura 6.10: Formulario para búsqueda personalizada de pacientes.

Cuando se ha localizado al paciente requerido, se selecciona el diagnóstico de interés en la lista mostrada y se oprime el botón ***Ver diagnóstico*** correspondiente, como se indica en la figura 6.11.

A. Paterno	A. Materno	Nombre	Fecha	Diagnósticos (Selecciona uno)	Oprime un botón
Loaeza	Valeño	Joel	2003-05-04	◉	Ver diagnóstico

Figura 6.11: Botón *Ver diagnóstico.*

Lista de reactivos.- El botón ***Reactivos*** de la figura 6.4, muestra la lista de reactivos del MMPI-2.

6.9.5.2. Altas y bajas del MMPI-2Web

El botón ***Insertar/Eliminar Pacientes*** de la figura 6.4, abre el formulario con las opciones que se observan en la figura 6.12.

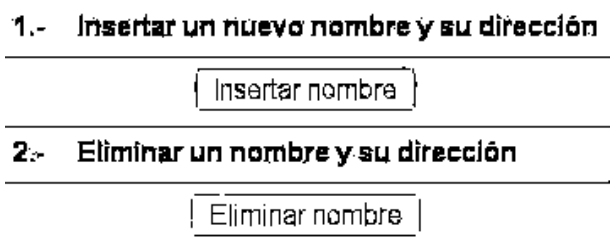


Figura 6.12: Menú insertar o eliminar paciente.

Insertar paciente.- El botón ***Insertar nombre*** de la figura 6.12, abre el formulario para insertar un paciente y sus datos personales como se aprecia en la figura 6.13.

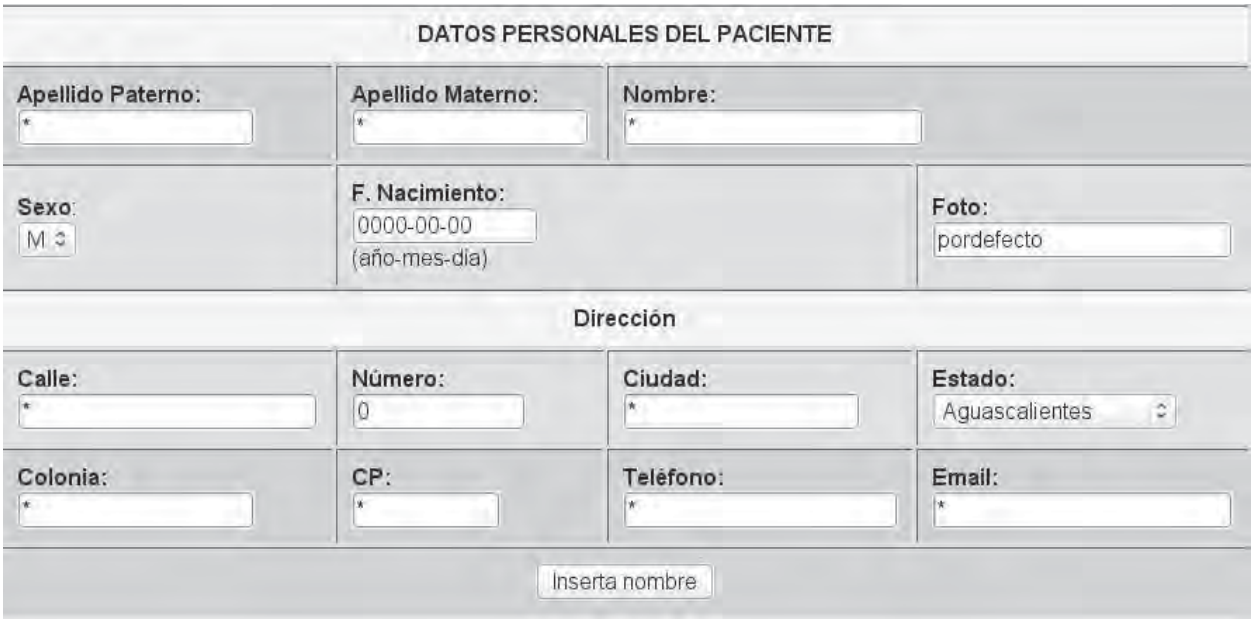


Figura 6.13: Formulario para insertar datos personales de un paciente.

Eliminar paciente.- El botón ***Eliminar nombre*** de la figura 6.12 abre un formulario similar al de la figura 6.10, para buscar al paciente que se desea eliminar, cuando aparece la lista correspondiente se oprime el botón ***Elimina nombre*** para eliminar al paciente, su dirección y diagnósticos correspondientes como se observa en la figura 6.14.

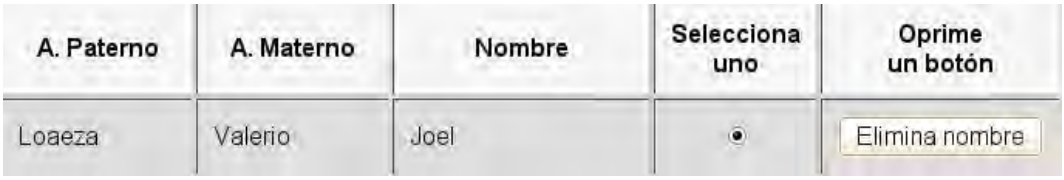


Figura 6.14: Botón *Elimina nombre*.

El botón ***Insertar/Eliminar Tests*** de la figura 6.4, abre el siguiente menú con las opciones que se observan en la figura 6.15.

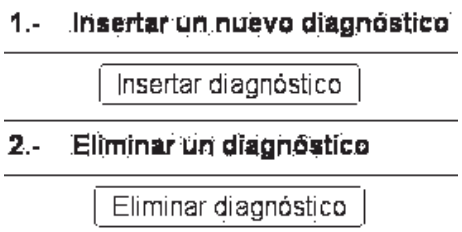


Figura 6.15: Menú insertar o eliminar test.

Insertar test.- El botón ***Insertar diagnóstico*** de la figura 6.15, abre un formulario similar al de la figura 6.10, para buscar un paciente a quien asignarle el test oprimiendo el botón ***Asignar test*** que se observa en la figura 6.16.

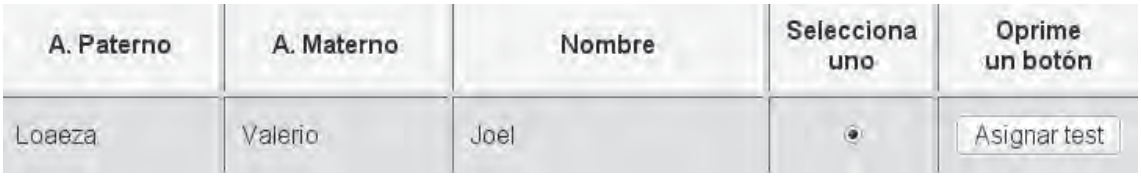


Figura 6.16: Botón *Asignar test*.

Eliminar test.- El botón ***Eliminar diagnóstico*** de la figura 6.15, abre un formulario similar al de la figura 6.10, para buscar un paciente al que se desea eliminar un test oprimiendo el botón ***Verificar diagnóstico*** en la lista que se muestra, enseguida se confirma la eliminación del test oprimiendo el botón ***Elimina diagnóstico*** que se aprecia en la figura 6.17.

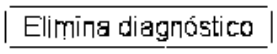


Figura 6.17: Botón *Elimina diagnóstico*.

El botón ***Insertar/Eliminar Reactivos*** de la figura 6.4, abre el siguiente menú con las opciones que se observan en la figura 6.18.

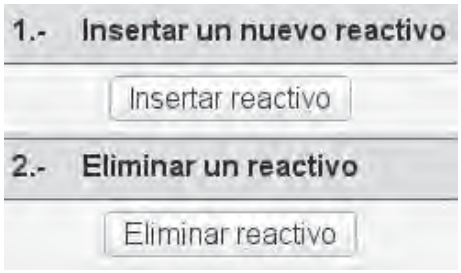


Figura 6.18: Menú insertar o eliminar reactivos.

Insertar reactivos.- El botón ***Insertar reactivo*** de la figura 6.18 abre el formulario que se muestra en la figura 6.19, para insertar un reactivo oprimiendo el botón ***Inserta reactivo***.

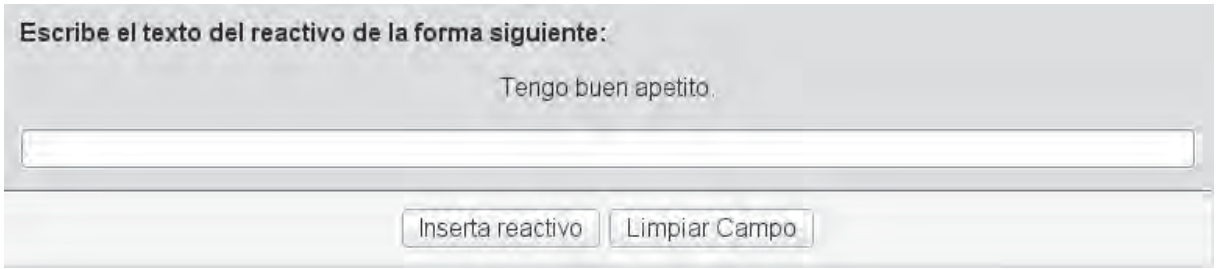


Figura 6.19: Formulario para insertar un reactivo.

Eliminar reactivos.- El botón ***Eliminar reactivo*** de la figura 6.18 muestra una lista de reactivos

en la que se puede seleccionar al que se desea eliminar, oprimiendo el botón ***Elimina reactivo*** que se observa en la figura 6.20.

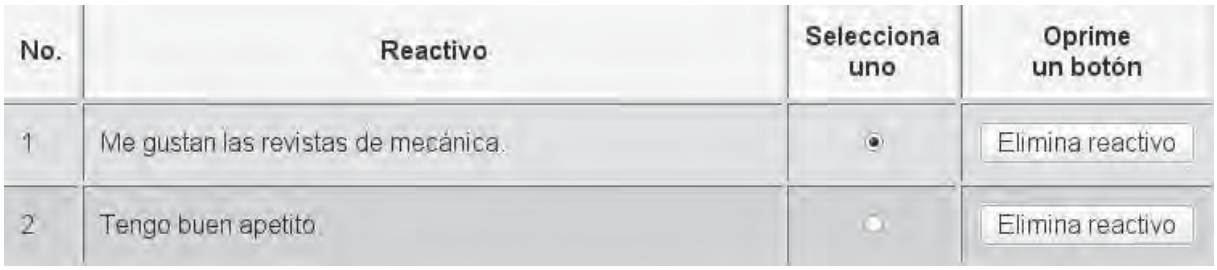


Figura 6.20: Botón *Elimina reactivo*.

6.9.5.3. Actualizaciones del MMPI-2Web

Las actualizaciones funcionan de manera similar a las *altas/bajas* descritas anteriormente.

6.9.6. Formulario de acceso para aplicar un test con MMPI-2Web

Cuando se oprime el botón ***Aplicar test*** del menú de bienvenida al MMPI-2Web mostrado en la figura 6.2, se abre el formulario que se observa en la figura 6.21. El nombre y clave de aplicador autorizado y el nombre de la base de datos, deben ser proporcionados por el administrador del MMPI-2Web. El identificador del test se obtiene del apartado ***Administración*** del MMPI-2Web. Esta información se introduce en los campos correspondientes y se oprime el botón ***Enviar Datos***. del formulario mostrado en la figura 6.21.

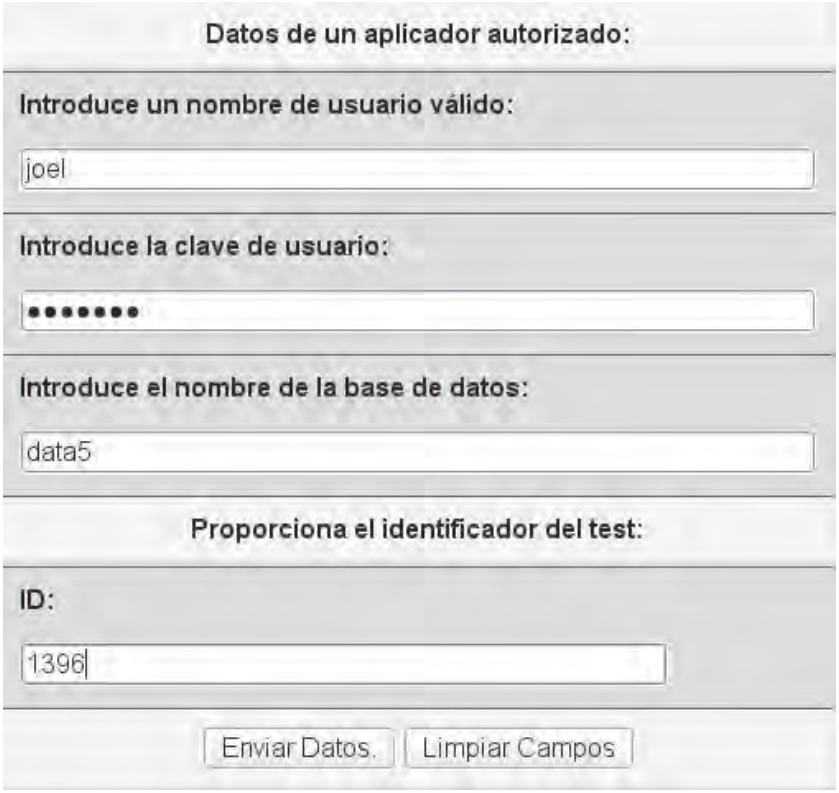


Figura 6.21: Formulario de acceso para aplicar un test MMPI-2.

Una vez que se tiene acceso para aplicar un test, aparecen los formularios correspondientes similares al que se muestra en la figura 6.22, estos son de uso sencillo e intuitivo. Nota: al finalizar el test, se debe oprimir el botón ***Regresar para guardar*** del formulario mostrado en la figura 6.22, para regresar al formulario inicial de datos personales y oprimir el botón ***Guardar test*** que se muestra en la figura 6.23, ya que esto hace que se valide y califique el test aplicado.

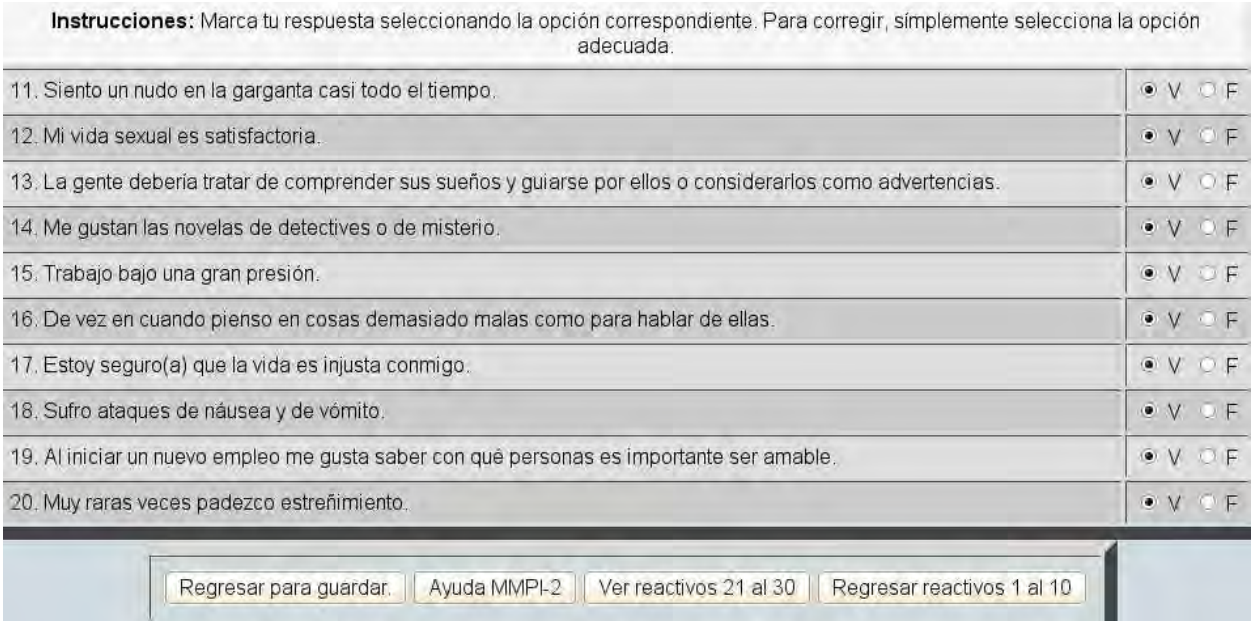


Figura 6.22: Formulario de ejemplo de aplicación del test MMPI-2.

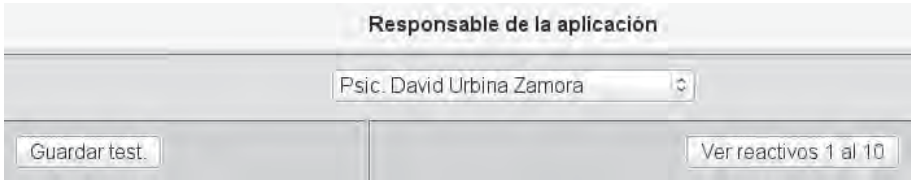


Figura 6.23: Botón *Guardar Test* del formulario de datos personales.

6.10. Calificación del MMPI-2

La calificación del instrumento MMPI-2 en su versión revisada sirve para elaborar tres perfiles: el básico (clínico), de escalas de contenido y de escalas suplementarias, los que son necesarios para su interpretación posterior por parte del profesional a cargo. En el caso del MMPI-2Web la calificación se realiza en forma automatizada y se presenta en forma concentrada como se observa la clasificación previa en la figura 6.24.

CLASIFICACION PREVIA	
L	**Bajo**
F	**Bajo**
K	**Medio**
Utilidad del Perfil	**Válido**
Validación del Test	**Válido**
Media de las Escalas Clínicas	**45.3**
Clasificación de la MEC	**Normal**
Tipo de Perfil	**Perfil normal**
Tipo de Clave	**Clave normal**
Codificación del Perfil	**0/685341279:# LKF:#**

Figura 6.24: Concentrado de calificaciones previas del MMPI-2.

En la figura 6.25 se muestra el concentrado de calificaciones de las escalas clínicas, suplementarias y de contenido.

CLASIFICACION ESCALAS CLINICAS (BASICAS)			
Hs	**Medio**	Pa	**Medio**
D	**Medio**	Pt	**Medio**
Hi	**Medio**	Es	**Medio**
Dp	**Medio**	Ma	**Medio**
Mf	**Medio**	Is	**Medio**

CLASIFICACION ESCALAS SUPLEMENTARIAS			
A	**Medio**	Do	**Medio**
R	**Medio**	Rs	**Medio**
Fyo	**Medio**	Dpr	**Medio**
A-MAC	**Medio**	GM	**Alto**
HR	**Medio**	GF	**Medio**

CLASIFICACION ESCALAS DE CONTENIDO					
ANS	**Medio**	DEL	**Medio**	BAE	**Bajo**
MIE	**Medio**	ENJ	**Medio**	ISO	**Medio**
OBS	**Medio**	CIN	**Medio**	FAM	**Medio**
DEP	**Medio**	PAS	**Medio**	DTR	**Medio**
SAU	**Medio**	PTA	**Medio**	RTR	**Bajo**

Figura 6.25: Concentrado de calificaciones de las escalas del MMPI-2.

El perfil básico se obtiene tomando en cuenta los siguientes puntos:

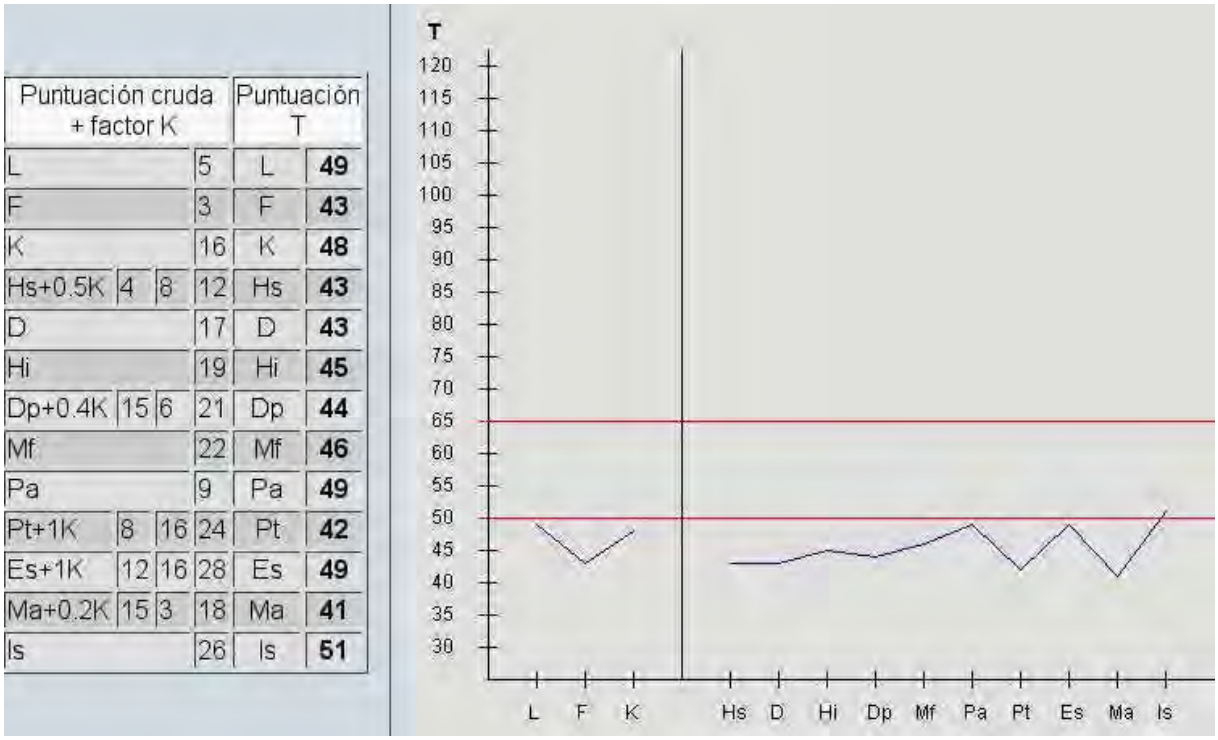
1. Se debe emplear el perfil apropiado al género del individuo (masculino o femenino).

2. En la aplicación MMPI-2Web, las puntuaciones naturales o crudas se obtienen de la siguiente manera: primero se determina el número de respuestas omitidas o con dos contestaciones, enseguida se calculan las puntuaciones naturales de cada uno de los tres indicadores de

validez (L, F y K), así como de las 10 escalas clínicas (básicas) sumando aritméticamente las respuestas que corresponden con los conjuntos de reactivos para respuestas verdaderas y falsas, de acuerdo al apéndice C del *manual para aplicación y calificación del MMPI-2* [Hathaway y McKinley, 1995].

3. Se lleva a cabo la corrección K agregando .5K a Hs, 0.4K a Dp, 1K a Pt, 1K a Es y 0.2K a Ma. Con estas puntuaciones corregidas y el resto de puntuaciones naturales (L, F, K y escalas clínicas que no requieren corrección) se obtienen las puntuaciones estándar (T) conforme al apéndice B de la obra *Uso e interpretación del MMPI-2 en español* [Lucio y León, 2003].

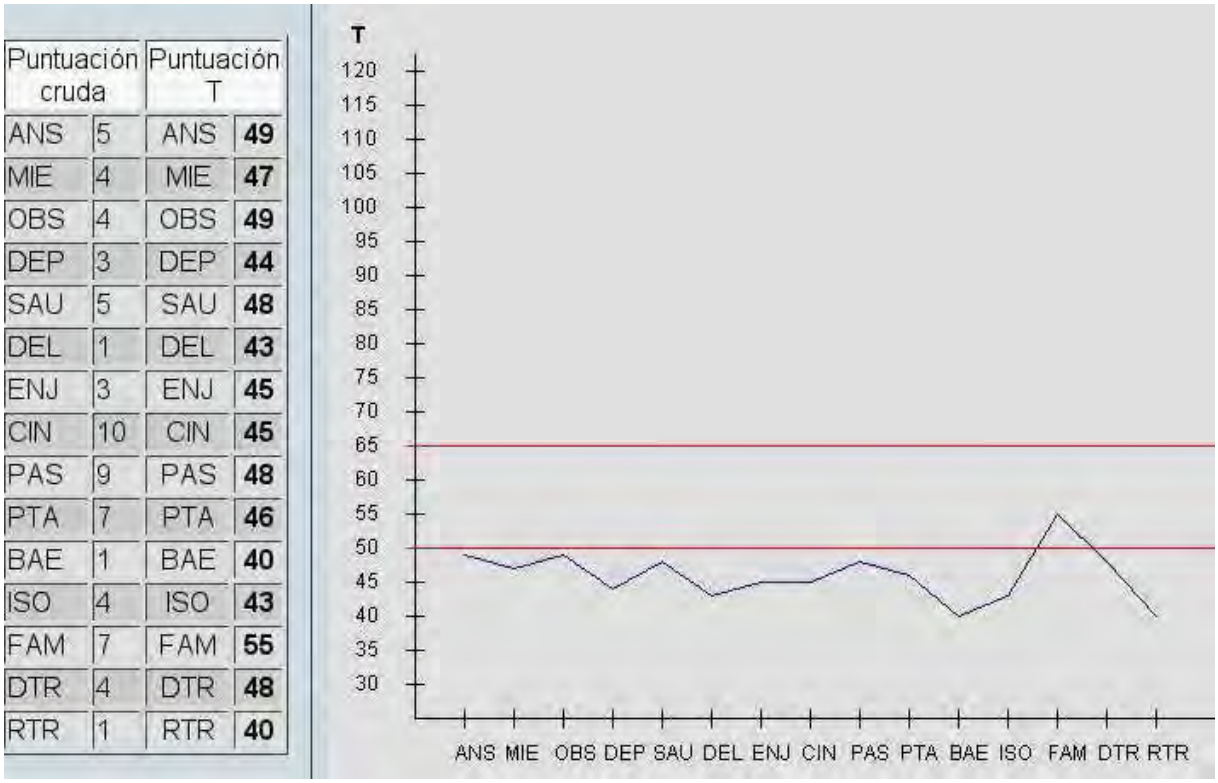
4. Con las puntuaciones estándar (T) basadas en la muestra normativa mexicana de la población general, se grafica el perfil básico como se muestra en la figura 6.26:



Puntuación cruda + factor K				Puntuación T	
L			5	L	**49**
F			3	F	**43**
K			16	K	**48**
Hs+0.5K	4	8	12	Hs	**43**
D			17	D	**43**
Hi			19	Hi	**45**
Dp+0.4K	15	6	21	Dp	**44**
Mf			22	Mf	**46**
Pa			9	Pa	**49**
Pt+1K	8	16	24	Pt	**42**
Es+1K	12	16	28	Es	**49**
Ma+0.2K	15	3	18	Ma	**41**
Is			26	Is	**51**

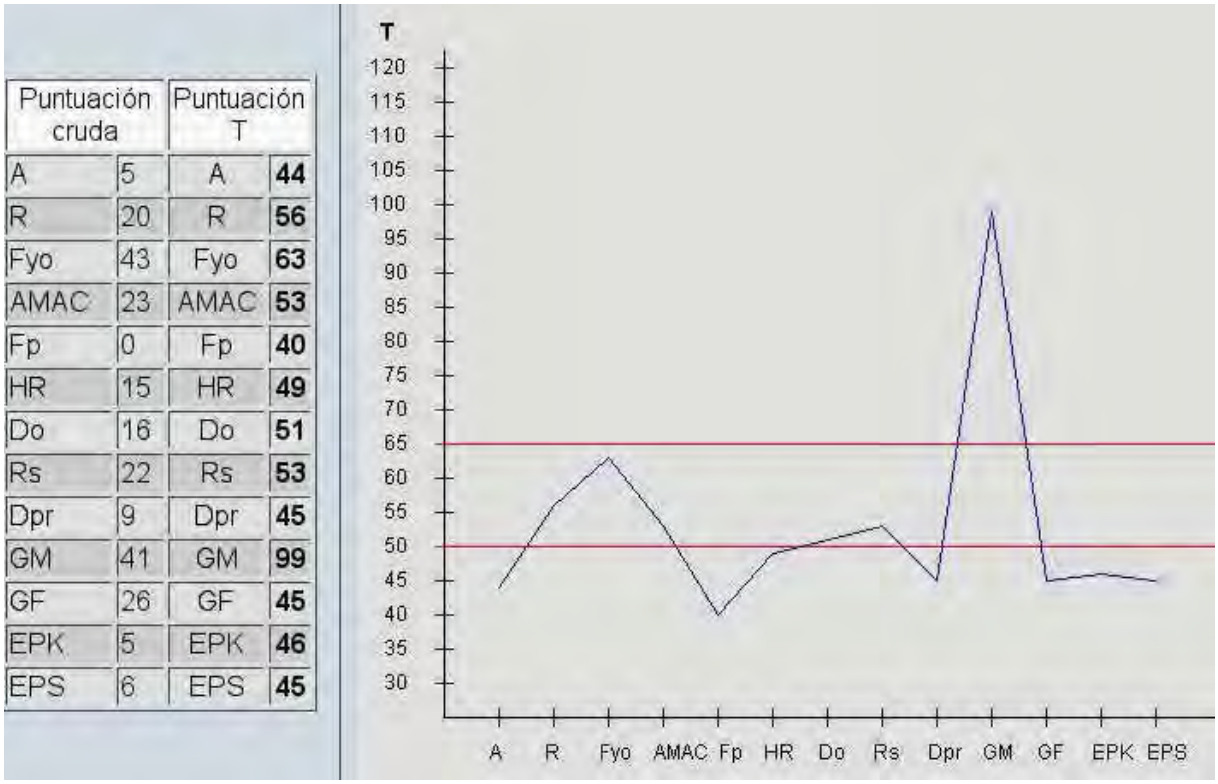
Figura 6.26: Perfil de las escalas clínicas (básicas) generado por el MMPI-2Web.

El perfil de las escalas de contenido y el de las escalas suplementarias se obtiene siguiendo el mismo procedimiento empleado para la obtención del perfil de las escalas clínicas, sólo que para éstos no se agrega ningún factor de corrección ni se codifica ninguna clave. Enseguida se muestran estos perfiles, en la figura 6.27 se observa el perfil de las escalas de contenido y en la figura 6.28 se aprecia el perfil de las escalas suplementarias.



Puntuación cruda		Puntuación T	
ANS	5	ANS	**49**
MIE	4	MIE	**47**
OBS	4	OBS	**49**
DEP	3	DEP	**44**
SAU	5	SAU	**48**
DEL	1	DEL	**43**
ENJ	3	ENJ	**45**
CIN	10	CIN	**45**
PAS	9	PAS	**48**
PTA	7	PTA	**46**
BAE	1	BAE	**40**
ISO	4	ISO	**43**
FAM	7	FAM	**55**
DTR	4	DTR	**48**
RTR	1	RTR	**40**

Figura 6.27: Perfil de las escalas de contenido generado por el MMPI-2Web.



Puntuación cruda		Puntuación T	
A	5	A	**44**
R	20	R	**56**
Fyo	43	Fyo	**63**
AMAC	23	AMAC	**53**
Fp	0	Fp	**40**
HR	15	HR	**49**
Do	16	Do	**51**
Rs	22	Rs	**53**
Dpr	9	Dpr	**45**
GM	41	GM	**99**
GF	26	GF	**45**
EPK	5	EPK	**46**
EPS	6	EPS	**45**

Figura 6.28: Perfil de las escalas suplementarias generado por el MMPI-2Web.

6.11. Codificación del perfil básico

El sistema de codificación de Welsh es el que se emplea con el MMPI-2 para reducir el número de perfiles a un número más práctico. La codificación se realiza asignando un número a cada escala clínica: 1 (Hs), 2(D), 3(Hi), 4(Dp), 5(Mf), 6(Pa), 7(Pt), 8(Es), 9(Ma) y 0(Is). Con estos números se codifican los perfiles para su uso cotidiano, evitando implicaciones psiquiátricas que pudiesen resultar confusas en ambientes no clínicos. La clave se compone escribiendo el número correspondiente de la escala más elevada seguido del símbolo del rango respectivo y así sucesivamente. En la tabla 6.5 se muestran los rangos y sus símbolos equivalentes para codificación Welsh:

Símbolos	**Rango**
!!	120 o más
!	110-119
**	100-109
*	90-99
”	80-89
’	70-79
+	65-69
-	60-64
/	50-59
:	40-49
#	30-39
	29 y menos de 29 a la derecha de #

Tabla 6.5: Rangos y sus símbolos equivalentes (codificación Welsh).

6.12. Interpretación del MMPI-2

La interpretación del instrumento la debe realizar un profesional del área de la Psicología debidamente capacitado para tal efecto. A continuación se describen brevemente algunos criterios de interpretación:

1. *Interpretación de los indicadores de validez.* Estos indicadores sirven para evaluar si la aplicación del instrumento es válida o no, además de mostrar el comportamiento del individuo hacia la prueba, es decir, si coopera con el instrumento o si distorsiona y manipula sus respuestas.

2. *Interpretación de las escalas clínicas.* Estas escalas sirven para detectar psicopatología y características psicodinámicas incluidas en los cuadros de dichas escalas.

3. *Interpretación de las escalas de contenido.* El empleo de estas escalas tiene un valor significativo para precisar o refinar la interpretación obtenida con las escalas clínicas. La interpretación basada en estas escalas reflejan los sentimientos de la persona, así como características de su personalidad y de sus problemas pasados o actuales.

4. *Interpretación de las escalas suplementarias.* Estas escalas evalúan características como la hostilidad, problemas de abuso de alcohol o desajuste profesional. Sirven también como complemento de interpretación para las escalas básicas.

Una vez que se ha interpretado la prueba en su totalidad se debe realizar un reporte con un formato adecuado.

6.12.1. Reporte de prueba generado por MMPI-2Web

MMPI-2Web genera un reporte con los siguientes elementos, primero los datos personales del paciente como se muestra en la figura 6.29.

DATOS PERSONALES DEL PACIENTE

Fecha de aplicación: **2003-05-04**			
Apellido Paterno: **Loaeza**	Apellido Materno: **Valerio**	Nombre: **Joel**	
Sexo: **M**	F. Nacimiento: **1968-07-25**	Estado Civil: **Soltero**	Ocupación: **Docente**
Escolaridad: **Profesional**	Carrera: **Ingeniería Civil**		Grado: **Titulado en Licenciatura**
Dirección			
Calle: **Prol. 20 de Noviembre**	Número: **1003**	Ciudad: **Paracho**	Estado: **Michoacán**
Colonia: **Villa Artesanal**	CP: **60250**	Teléfono: **01-423-52-5-09-51**	Email: **joel.loaeza@gmail.com**
Observaciones			
Motivo de canalización: **Voluntario**			
Observaciones de la prueba: **Normal**			
Responsable de la aplicación			
Nombre: **Psic. David Urbina Zamora**			

Figura 6.29: Concentrado de datos personales del paciente.

Después se presentan las gráficas de los perfiles correspondientes a las escalas clínicas, suplementarias

y de contenido, como las mostradas en las figuras 6.26, 6.27 y 6.28 respectivamente. Enseguida se muestra la hoja de respuestas correspondientes a la prueba aplicada, como se puede observar en el segmento de la figura 6.30.

HOJA DE RESPUESTAS

1 V	2 V	3 V	4 F	5 F	6 V	7 V	8 V	9 V	10 F	11 F	12 V	13 V	14 V	15 F	16 V	17 F	18 F	19 V	20 V
21 F	22 F	23 F	24 F	25 V	26 F	27 V	28 F	29 V	30 F	31 F	32 F	33 V	34 V	35 F	36 F	37 V	38 F	39 F	40 F
41 V	42 F	43 V	44 F	45 V	46 F	47 V	48 V	49 V	50 F	51 V	52 F	53 F	54 F	55 V	56 V	57 V	58 V	59 F	60 V

Figura 6.30: Segmento de la hoja de respuestas del test aplicado.

A continuación se presentan los concentrados de clasificación previa, de calificaciones de las escalas clínicas, suplementarias y de contenido, mostrados en las figuras 6.24 y 6.25 respectivamente. Y al final del reporte, se presenta un concentrado con las posibilidades de interpretación del test aplicado como se muestra en la figura 6.31.

POSIBILIDADES DE INTERPRETACION	
Indicadores de validez	
L (mentira)	Indica un sujeto confiado en sí mismo e independiente.
F (infrecuencia)	Convencional. Sincero. Socialmente adaptado.
K (corrección)	Suficientes recursos para el tratamiento
Escalas clínicas (escalas básicas)	
Hs (hipocondriasis)	Se manifiesta poco o ningún interés especial acerca del cuerpo o de la salud. La persona es emocionalmente abierta y equilibrada, así como realista y con capacidad de insight.
D (depresión)	Indica que se trata de una persona conforme consigo misma. Puede ser también un sujeto estable, equilibrado y realista.
Hi (histeria conversiva)	Se presenta en personas realistas y sensibles. El sujeto puede ser además equilibrado y razonable.
Dp (desviación psicopática)	Puede ser una persona sincera, confiable, tenaz y responsable.
Mf (masculinidad/femineidad)	Sujeto práctico y despreocupado. También puede ser realista y convencional.
Pa (paranoia)	El sujeto muestra un pensamiento claro y actúa racionalmente. Se presenta en personas precavidas y flexibles.
Pt (psicastenia)	Se presenta en personas puntuales y confiables. El sujeto puede ser adaptable y confiado. Indica también una persona bien organizada.
Es (esquizofrenia)	Se trata de personas adaptables, confiables y equilibradas.
Ma (hipomanía)	Puede tratarse de un sujeto sociable y amigable. Se presenta en personas responsables y realistas. Individuos entusiastas y equlibrados.
Is (introversión social)	Puede tratarse de sujetos enérgicos y activos, con entereza. La persona puede ser equilibrada, amistosa y platicadora.
Escalas suplementarias	
GM (género masculino)	Para el varón se relaciona con una gran confianza en sí mismo, perseverancia marcada y amplitud de intereses, además de carencia de temores o sentimientos de autorreferencia. Para la mujer se relaciona con una gran confianza en sí misma, así como con honestidad y disposición para probar nuevas cosas, también indica ausencia de preocupaciones y sentimientos de autorreferencia.
NOTA: Este cuadro no debe ser tomado textualmente pues presenta sólo inferencias generales acerca del significado de la elevación de las puntuaciones que deben considerarse en cada caso de acuerdo con la historia y antecedentes del sujeto.	

Figura 6.31: Concentrado de posibilidades de interpretación de la prueba aplicada.

6.13. Propósitos del MMPI-2

Originalmente el MMPI se desarrolló como un auxiliar en los programas de detección psiquiátrica en ambientes de salud mental dentro de la práctica médica general. Durante sus primeros años el instrumento se utilizó y se investigó en una variedad de ambientes médicos y de salud mental, así como en varios contextos no relacionados con la salud mental, como instituciones correccionales, programas de selección de personal, programas de abuso de alcohol y drogas, y también en investigación.

En la actualidad no es posible describir todos los ambientes y poblaciones en los que se utiliza el MMPI-2, pero a continuación se señalan algunas aplicaciones como se menciona en [Butcher, 2001]:

- Valoración de pacientes en ambientes de salud mental para auxiliar en la especificación de su estado de salud mental.
- Estimación de síntomas para determinar la necesidad de hospitalización.
- Evaluación de pacientes en la planificación previa al tratamiento.
- Valoración de efectos del tratamiento.
- Investigación epidemiológica en la que se utilizan criterios basados en la personalidad.
- Evaluación de la personalidad para puestos de seguridad pública, como policía, bombero, piloto de aerolínea y personal de plantas nucleares.
- Estudios de investigación psicológica en la que se utilizan estimaciones objetivas de la personalidad como criterio externo para el estudio de diferencias grupales.
- Investigación sobre la genética de la personalidad.
- Estudios longitudinales sobre los procesos y cambio de la personalidad.
- Evaluación de la personalidad en diferentes contextos culturales para estudiar semejanzas y diferencias entre culturas diferentes.
- Clasificación de delincuentes convictos dentro de internamiento carcelario.
- Valoración de los padres en demandas de custodia familiar.
- Estimación de factores de la personalidad para determinar si una persona que entabla una demanda por daños personales tiene los problemas de salud mental por los cuales demanda.

El MMPI-2 no está diseñado para dirigirse a todas las características que pudiesen interesar a un psicólogo, tampoco atiende a cualidades como la inteligencia, la presencia de trastorno cerebral orgánico ni predice la probabilidad de cometer un acto violento.

6.14. Conclusión del capítulo

El MMPI-2 está compuesto por un conjunto de 567 reactivos, con este material se forman los indicadores de validez, las escalas clínicas, de contenido y suplementarias, dichos reactivos deben ser respondidos como verdadero o falso en la aplicación de la prueba para poder evaluar, interpretar y emitir un reporte que permita detectar si el paciente padece alguna psicopatología y para conocer las características de personalidad del individuo en cuestión. Este conocimiento acerca del MMPI-2 permite que el AI Descubreconocimiento formule una estrategia para aplicar el algoritmo de referencia para minería de datos, propuesto para llevar a cabo el proceso de KDD y la herramienta de minería de datos ADC4.5Web desarrollados en el presente trabajo, con el fin de obtener una versión experimental reducida del MMPI-2 y que ésta se pueda interpretar y evaluar comparándola con la

versión estándar de dicho instrumento. Evaluando de esta manera la eficacia tanto del algoritmo de referencia para minería propuesto como la del AI de Software implementado en la herramienta de minería ADC4.5Web que al interactuar con los Agentes Inteligentes Humanos participantes en esta tarea de minería conforman al AI Descubreconocimiento.

Capítulo 7

Ambiente de Descubrimiento con C4.5Web (ADC4.5Web)

En este capítulo se presenta la mayor aportación de este trabajo de tesis: la herramienta de minería de datos denominada Ambiente de Descubrimiento con C4.5Web (ADC4.5Web). ADC4.5Web es una herramienta de bajo costo implementada con aplicaciones de código abierto y libre distribución (frontend[1] y backend[2] programados con PHP 4.0, JAVA y ANSI C), que integra una adecuación para ambiente web del algoritmo de minería C4.5 descrito en el capítulo 4, con acceso a la información administrada por medio del sistema manejador de bases de datos MySQL 3.23.48, donde se almacenan los datos que serán minados.

ADC4.5Web es una aplicación dinámica para web, que facilita la interacción de las bases de datos MySQL 3.23.48 con los Agentes Inteligentes Humanos y con el Agente Inteligente de Software (algoritmo de minería C4.5) para realizar tareas de descubrimiento de conocimiento, empleando como guía metodológica el algoritmo de minería de datos propuesto en el capítulo 5.

La administración de MySQL 3.23.48 con respecto a la definición y manipulación de datos, tanto del MMPI-2 como la información en general que pueda servir para ser *minada*, es una tarea que corresponde en este caso al AI Humano con perfil en computación.

La herramienta ADC4.5Web puede funcionar de manera local y remota en ambiente web, para llevar a cabo el proceso de descubrimiento de conocimiento, permitiendo de esta manera una mayor colaboración entre los usuarios interesados en aportar datos y/o conocer los resultados obtenidos en dicho proceso.

7.1. Instalación de ADC4.5Web

ADC4.5Web es una aplicación cliente/servidor, por lo que para la parte del servidor se necesita lo siguiente:

[1] Parte del software que interactúa con el usuario.
[2] Parte del software que procesa los datos.

- Los requerimientos mínimos de hardware son: procesador Intel PENTIUM-II o superior (o equivalente), disco duro de 6 Gb o superior, memoria RAM de 128 Mb o superior, acceso a una red LAN (red de área local) y/o WAN (red de área global).
- Los requerimientos de software son: sistema operativo SuSE Linux 8.0 o equivalente, sistema manejador de bases de datos MySQL 3.23.48 o equivalente, servidor web Apache 1.3.23 con los módulos de PHP 4.1.0 incluidos.
- Se recomienda que la instalación y configuración de ADC4.5Web, así como la administración de las bases de datos, las realice un profesional con perfil en computación.

7.2. Ejecución de ADC4.5Web

Para la parte del cliente, la ejecución de ADC4.5Web se puede realizar con cualquier navegador web con soporte para JAVA, desde cualquier computadora, con cualquier sistema operativo, siempre y cuando se tenga acceso al servidor por medio de una red LAN o WAN. Para ejecutar la aplicación ADC4.5Web, únicamente basta escribir en la barra de direcciones del navegador web, la dirección correpondiente al archivo *main_ admin_ minnesota1.php*, misma que debe ser proporcionada al usuario por el administrador de ADC4.5Web. En la figura 7.1 se muestra un ejemplo de ruta de acceso a ADC4.5Web.

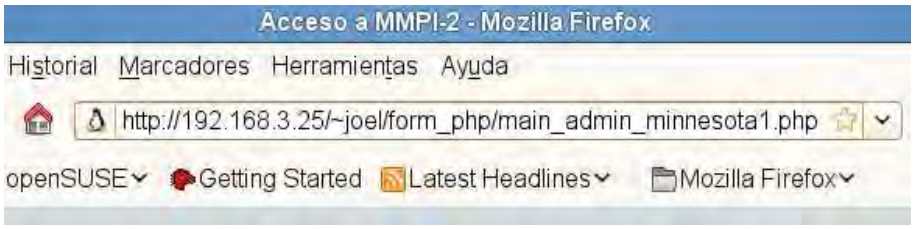


Figura 7.1: Ejemplo de ruta de acceso a ADC4.5Web.

7.3. Antecedentes del algoritmo generador de Árboles de Decisión C4.5

El algoritmo C4.5 es una extensión del algoritmo generador de árboles de decisión básico ID3 diseñado por el Dr. Ross Quinlan, para resolver algunos problemas no tratados por el algoritmo ID3, como son [Quinlan, 1993]:

- Evitar sobreadaptación (overfitting) en los datos
- Reducción de error en la poda (pruning)
- Regla de post-poda (post-pruning)
- Manipulación de atributos continuos

- Selección apropiada de atributos
- Manipulación de datos de entrenamiento con valores de atributo faltantes
- Mejoramiento de la eficacia de cómputo

El código fuente del algoritmo C4.5 así como el manual de ususario están disponibles para el público en varios sitios web, algunos de ellos son el sitio web del libro [C4.5] *C4.5: programs for machine learning [Quinlan, 1993]* y el sitio web de la universidad canadiense de Regina [University Regina].

7.4. Componentes del Ambiente de Descubrimiento C4.5Web

Con el fin de facilitar la explicación de la herramienta de minería ADC4.5Web y verificar su desempeño, se recurre a un ejemplo clásico propuesto por el Dr. Quinlan: el juego de golf (con un conjunto pequeño de ejemplos), aunque el algoritmo está diseñado para trabajar con un volumen considerable de datos.

Al igual que el algoritmo C4.5, ADC4.5Web está compuesto por tres partes principales: el preparador de datos, el generador de Árboles de Decisión y el intérprete y consultor de Árboles de Decisión . Estos componentes se describen a continuación.

7.4.1. Preparador de datos

Esta parte permite preparar fácilmente los datos que formarán a los conjuntos de datos de entrenamiento/prueba y al de evaluación, seleccionando aleatoriamente a sus elementos, a partir de la *vista minable* (tabla o tablas requeridas) de la base de datos seleccionada al accesar al sistema ADC4.5Web. Los dominios de atributo que soporta el sistema son: *discreto* y *continuo*, por lo tanto, como el sistema manejador de bases de datos empleado es MySQL 3.23.48, los tipos de atributos previamente deben fijarse como *enum* para los dominios discretos, para los dominios continuos el tipo de atributo puede ser *int, bigint, float, double o decimal* [MySQL]. En la sección 7.5.5 se muestra un ejemplo de su forma de uso.

La aplicación phpMyAdmin facilita al AI Humano con perfil en computación, las tareas de definición de datos (Data Definition Language) y manipulación de datos (Data Manipulation Language) empleando al sistema manejador de bases de datos MySQL 3.23.48.

7.4.2. Generador de Árboles de Decisión

El algoritmo C4.5 es una aplicación para inducir reglas de clasificación en forma de Árboles de Decisión a partir de un conjunto de datos. Todos los archivos leídos y escritos por el algoritmo C4.5 son de la forma **nombrearchivo.ext**, donde *nombrearchivo* es el nombre de archivo que identifica la tarea de inducción y *ext* es una extensión que define el tipo de archivo. El algoritmo C4.5 requiere dos archivos para operar: un archivo de nombres **nombrearchivo.names** que define la clase, atributos

y sus dominios, así como un archivo de datos **nombrearchivo.data** que contiene un conjunto de ejemplos, donde cada uno de ellos es descrito por los valores de atributo y clase respectivos.

7.4.2.1. Archivo de nombres

El archivo de nombres ***nombrearchivo.names*** es una serie de entradas que definen los nombres de atributos, dominios de atributos y clases. El archivo es de texto sin formato, los comentarios se preceden con "|". Cada entrada se termina con un punto (.) o con un retorno de línea (return).

El archivo comienza con los nombres de las clases, separados por comas y un punto al final como se puede apreciar en la figura 7.2. Cada nombre consiste en una cadena de caracteres que no incluya coma, signo de interrogación o dos puntos (a menos que sea precedido por una diagonal inversa "\"). Se puede incluir un punto en un nombre si éste no es seguido por un espacio. También los espacios en blanco se permiten y cuando se encuentran varios espacios consecutivos en un nombre se sustituyen por un sólo espacio.

Enseguida se incluye una línea en blanco para separar los nombres de clases de los de atributos.

El resto del archivo contiene los nombres de atributos y sus dominios. Una entrada de atributo comienza con el nombre de atributo seguido por dos puntos y enseguida los nombres de los elementos del dominio, separados por comas y un punto al final como se puede observar en la figura 7.2.

Bajo,Medio,Moderado,Alto,Muy alto.
r16: V,F.
r29: V,F.
r41: V,F.
r51: V,F.
r77: V,F.
r93: V,F.
r102: V,F.
r107: V,F.
r123: V,F.
r139: V,F.
r153: V,F.
r183: V,F.
r203: V,F.
r232: V,F.
r260: V,F.

Figura 7.2: Archivo *L.names*.

Se pueden utilizar las siguientes palabras como dominio: la palabra *"ignore"* indica que este atributo no debe ser utilizado, la palabra *"continuous"* indica que el atributo contiene valores reales (números

reales), la palabra *"discrete"* seguida por un número entero n indica que el programa debe montar una lista de n posibles valores o bien, debe escribirse una lista de todos los posibles valores discretos separados por comas (se recomienda emplear esta forma).

7.4.2.2. Archivo de datos

El archivo de datos ***nombrearchivo.data*** contiene una línea por cada ejemplo. Una línea contiene los valores de los atributos ordenados adecuadamente seguidos por la clase del ejemplo, con todas las entradas separadas por comas, como se puede observar en la figura 7.3.

F,V,V,V,V,V,V,F,F,V,F,V,V,F,V,Medio
F,V,V,V,V,V,V,F,F,V,F,V,V,F,F,Bajo
V,V,V,V,F,V,V,F,V,V,F,V,V,F,F, Bajo
V,V,V,V,V,V,V,F,F,V,F,V,F,F,V,Medio
F,V,V,V,F,V,V,F,F,V,F,V,F,F,F,Medio
.
.
.
V,V,V,V,V,V,V,F,F,V,F,V,F,F,F,Bajo
V,V,V,V,V,V,V,F,F,V,F,V,F,F,F,Medio
F,V,V,V,V,V,V,F,F,V,F,V,V,F,F,Medio

Figura 7.3: Archivo *L.data*.

Las reglas para los nombres en el archivo de nombres también son válidas para los nombres en el archivo de datos. Un valor desconocido de atributo se indica con un signo de interrogación "?". Si se utiliza un archivo de evaluación ***nombrearchivo.test*** como el de la figura 7.4, tendrá el mismo formato que el archivo de datos.

F,F,V,F,V,F,V,V,F,V,F,V,V,F,V,Medio
F,F,F,V,F,V,V,F,V,V,F,V,F,F,F,Moderado
V,F,F,V,V,V,V,F,V,V,F,V,F,V,F,Medio
.
.
.
F,V,V,V,V,V,V,F,F,V,F,V,F,F,F,Medio
V,V,V,V,V,V,V,F,F,V,F,V,F,F,F,Bajo

Figura 7.4: Archivo *L.test*.

7.4.2.3. Opciones del generador de Árboles de Decisión

El algoritmo C4.5 puede generar árboles de dos maneras:

- En el modo por lote (batch) el algoritmo genera sólamente un árbol usando todos los datos disponibles.
- En el modo iterativo (número de pruebas), el algoritmo inicia con un subconjunto de datos seleccionados al azar, genera un Árbol de Decisión de prueba, lo evalúa y repite el proceso hasta que genera un Árbol de Decisión que clasifica correctamente a todos los ejemplos.

Todos los árboles generados en el proceso se guardan en el archivo ***nombrearchivo.unpruned***. Después de que cada árbol es generado, se poda para intentar simplificarlo. Todos los árboles producidos, tanto pre como post-simplificación, se verifican con un conjunto de datos de prueba obtenido del conjunto de datos de entrenamiento. El "mejor" árbol podado (seleccionado por el algoritmo si hay más de una prueba) se guarda en el archivo ***nombrearchivo.tree***. En esta etapa ADC4.5Web permite elegir el uso del conjunto de datos no vistos contenidos en el archivo ***nombrearchivo.test***, con la finalidad de evaluar la eficiencia del Árbol de Decisión obtenido.

En la figura 7.5 se muestra el formulario de datos de entrada para el algoritmo C4.5 , con las opciones para generar el Árbol de Decisión.

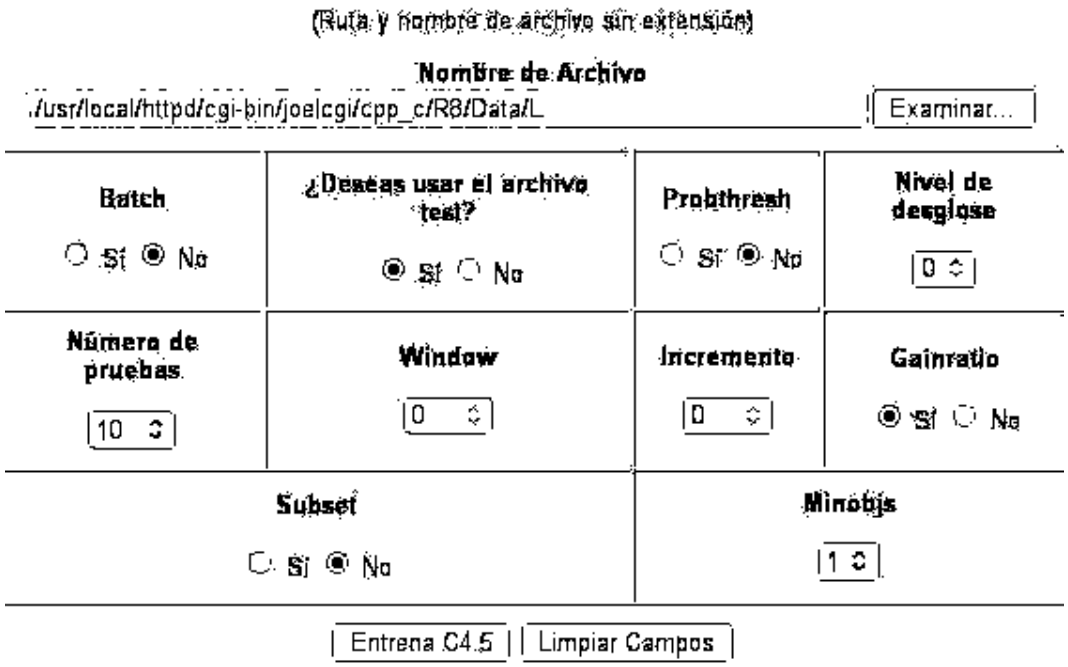


Figura 7.5: Formulario de opciones para generar un Árbol de Decisión.

- **Nombre de archivo**. En este campo se escribe el nombre del archivo sin extensión y su ruta respectiva.

- **Batch**. Esta opción le indica al algoritmo C4.5 si se desea que genere sólamente un árbol usando todos los datos disponibles o no.
- **Test**. Permite elegir si el árbol generado debe ser evaluado o no, con casos no vistos contenidos en el archivo *nombrearchivo.test*.
- **Probthresh**. Opción que sirve para indicar al algoritmo C4.5 que estime probabilísticamente los rangos para discretizar el dominio de atributos continuos.
- **Nivel de detalle para información adicional del Aŕbol de Decisión generado**. Esta opción permite fijar el nivel de detalle del resultado de salida entre (0-3), el valor por defecto es 0. Con esta opción se puede indicar al algoritmo C4.5 que genere un Árbol de Decisión con información adicional que pueda ayudar a explicar como se obtuvo.
- **Pruebas**. Permite establecer el número de árboles de prueba que el algoritmo C4.5 debe generar hasta obtener el árbol definitivo.

Las opciones siguientes también están disponibles, pero no necesitan ser utilizadas a menos que se desee experimentar en la construcción de árboles:

- **Window**. Fija el tamaño de la muestra inicial (por defecto el valor máximo es el 20 % del conjunto de datos).
- **Incremento**. Fija el número máximo de ejemplos que pueden ser agregados a la muestra con cada iteración (por defecto es el 20 % del tamaño inicial de la muestra).
- **Gainratio**. Permite fijar o no, el criterio de proporción de ganancia para la partición de los datos. Por defecto se usa este criterio.
- **Subset**. Indica al algoritmo C4.5 que construya un árbol de prueba con un subconjunto de valores asociado con cada rama.
- **Minobjs**. En todos los árboles de prueba, por lo menos dos ramas deben contener un número mínimo de objetos (por defecto 2). Esta opción permite cambiar el número mínimo.

7.4.3. Intérprete y consultor de Árboles de Decisión

El AI de Software (algoritmo C4.5) interpreta al ***archivo.tree*** que contiene al Árbol de Decisión generado, permitiendo que los AI Humanos puedan visualizarlo gráficamente o consultarlo de forma interactiva por medio de la herramienta de minería ADC4.5Web, ya sea para evaluar su eficiencia como modelo predictivo o para realizar una tarea de clasificación.

7.5. Ejemplo de la escala de validación L del MMPI-2

Se trata de estimar la tendencia del paciente a mentir al contestar el test MMPI-2. La solución consiste en que el algoritmo C4.5 aprenda una definición del **predicado de la meta** *Bajo/Medio/Moderado/Alto/Muy Alto*, expresada mediante un Arbol de Decisión.

7.5.1. Conjunto de datos de entrenamiento

En la tabla 7.1 se presenta una muestra del conjunto de ejemplos de entrenamiento para la escala de validación L.

	Atributos															**Meta o Clase**
Ej.	*r16*	*r29*	*r41*	*r51*	*r77*	*r93*	*r102*	*r107*	*r123*	*r139*	*r153*	*r183*	*r203*	*r232*	*r260*	*Bajo/Medio/Moderado/Alto/Muy Alto*
X1	F	V	V	V	V	V	V	F	F	V	F	V	V	F	V	Medio
X2	F	V	V	V	V	V	V	F	F	V	F	V	V	F	F	Bajo
X3	V	V	V	V	F	V	V	F	V	V	F	V	V	F	F	Bajo
X4	V	V	V	V	V	V	V	F	F	V	F	V	F	F	V	Medio
X5	F	V	V	V	F	V	V	F	F	V	F	V	F	F	F	Medio
X6	F	V	V	V	V	V	V	F	F	V	F	V	V	F	V	Medio
X7	V	V	F	V	V	V	V	F	V	V	F	V	F	F	V	Bajo
X8	V	V	V	V	F	V	V	F	V	V	F	V	V	F	V	Bajo
X9	F	F	V	F	V	F	V	V	F	V	F	V	V	F	V	Medio
X10	F	F	F	V	F	V	V	F	V	V	F	V	F	F	F	Moderado
X11	V	F	F	V	V	V	V	F	V	V	F	V	F	V	F	Medio
.....																
X1392	F	V	V	V	V	V	V	F	F	V	F	V	F	F	F	Medio
x1393	V	V	V	V	V	V	V	F	F	V	F	V	F	F	F	Bajo
x1394	V	V	V	V	V	V	V	F	F	V	F	V	F	F	F	Medio
x1395	F	V	V	V	V	V	V	F	F	V	F	V	V	F	F	Medio

Tabla 7.1: Datos de entrenamiento para el ejemplo escala de validación L del MMPI-2.

Los nombres de columna (nombre de atributos) se transforman en parte del archivo de nombres ***L.names***. El 80 % de las filas subsecuentes de estos ejemplos de entrenamiento serán introducidos en el archivo de datos ***L.data*** y el restante 20 % al archivo de prueba ***L.test***.

7.5.2. Menú de acceso al Ambiente de Descubrimiento con C4.5Web

Como la herramienta de minería ADC4.5Web está desarrollada para un ambiente web, para poder usarla se necesita un navegador web en el que se proporciona la dirección IP o nombre de dominio del servidor que hospeda a ADC4.5Web, como se muestra en la figura 7.6 y enseguida se oprime el botón ***Minería de datos*** para que se muestre el formulario de acceso.

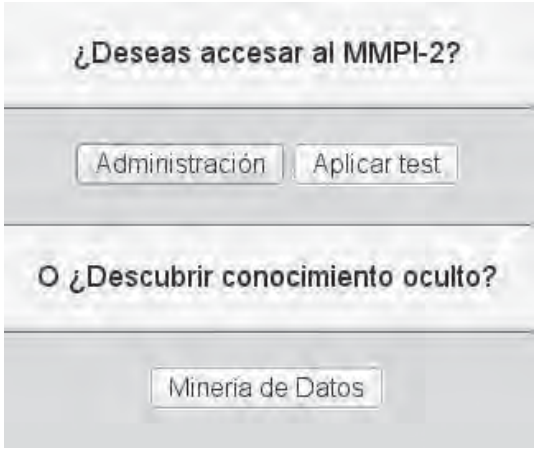


Figura 7.6: Menú de acceso a ADC4.5Web.

7.5.3. Formulario de acceso a ADC4.5Web

En este formulario se proporciona un nombre de usuario válido, la clave y el nombre de la base de datos sobre la cual se desea realizar el proceso de descubrimiento de conocimiento como se muestra en la figura 7.7 y después se oprime el botón ***Deseo entrar a Minería de Datos***.

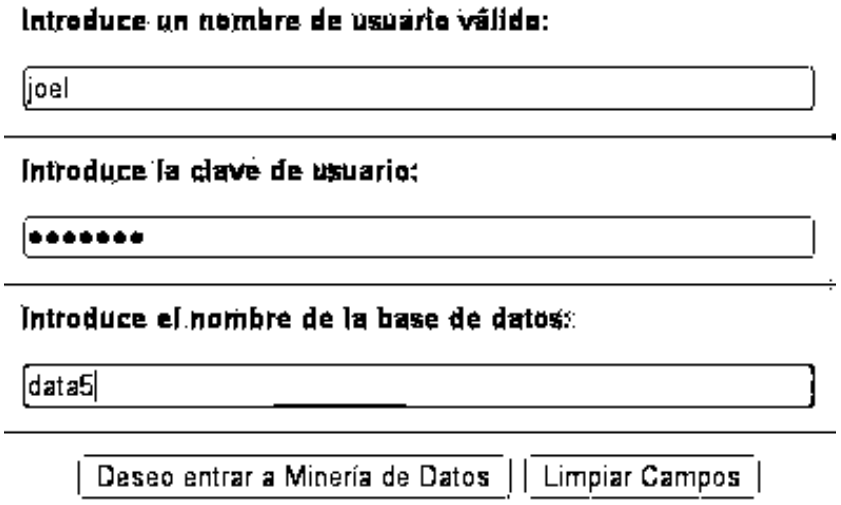


Figura 7.7: Formulario de acceso a ADC4.5Web.

7.5.4. Menú del sistema ADC4.5Web

El menú del sistema ADC4.5Web proporciona tres opciones: preparar datos, entrenar C4.5 y consultar árbol como se puede ver en la figura 7.8.

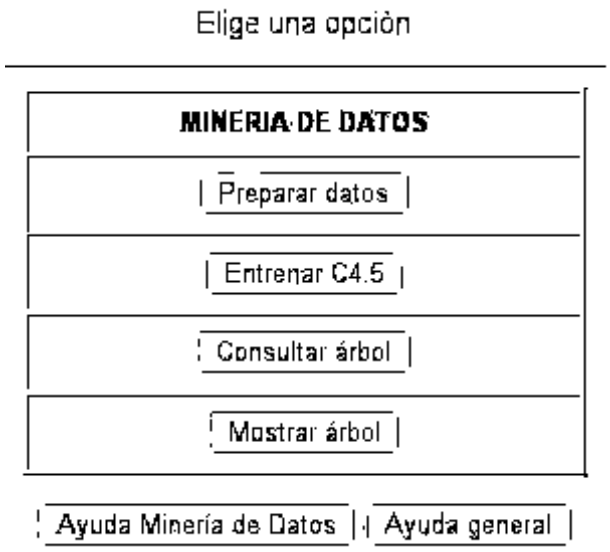


Figura 7.8: Menú del sistema ADC4.5Web.

7.5.5. Preparación de los datos

Esta opción permite preparar fácilmente los datos que formarán a los conjuntos de entrenamiento/prueba y evaluación, seleccionando aleatoriamente a sus elementos, a partir de las tablas requeridas de la base de datos seleccionada al accesar al sistema. Enseguida se muestran los pasos a seguir en esta etapa:

1. En el menú del sistema ADC4.5Web, oprimir el botón ***Preparar datos***. Después se proporciona el nombre del modelo de predicción que se generará posteriormente como se indica en la figura 7.9 y se oprime el botón ***Continuar***.

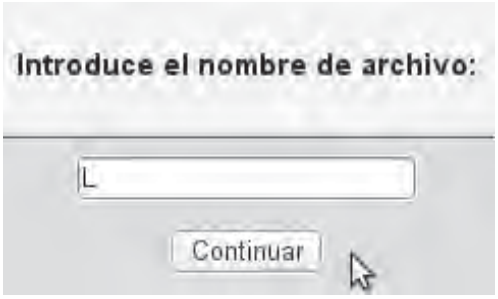


Figura 7.9: Nombre del modelo que se generará.

2. Se elige una tabla como se muestra en la figura 7.10.

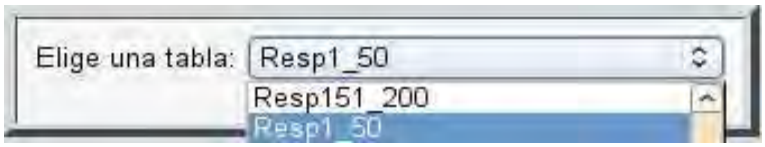


Figura 7.10: Elección de una tabla.

3. Se selecciona un campo como se indica en la figura 7.11

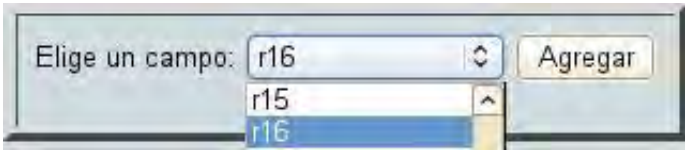


Figura 7.11: Selección de un campo.

y en seguida se agrega para formar parte de la **vista minable** (tabla temporal) a partir de la cual se van generar los conjuntos de datos de entrenamiento/prueba y evaluación como se indica en la figura 7.12.

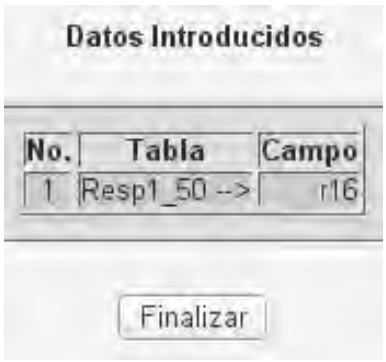


Figura 7.12: Adición de un campo a la **vista minable**.

4. Repetir los pasos 2 y 3 las veces que sean necesarias, seleccionando campos de una o varias tablas, hasta que se agreguen los atributos requeridos. Al hacerlo se debe cuidar que su cardinalidad sea idéntica de lo contrario ADC4.5Web indicará un error. Cuando ya se hayan incluido todos los atributos, al final se debe agregar un campo más, que será la clase o meta. Una vez que todo esto se realiza se obtiene la **vista minable**, como se muestra en la figura 7.13

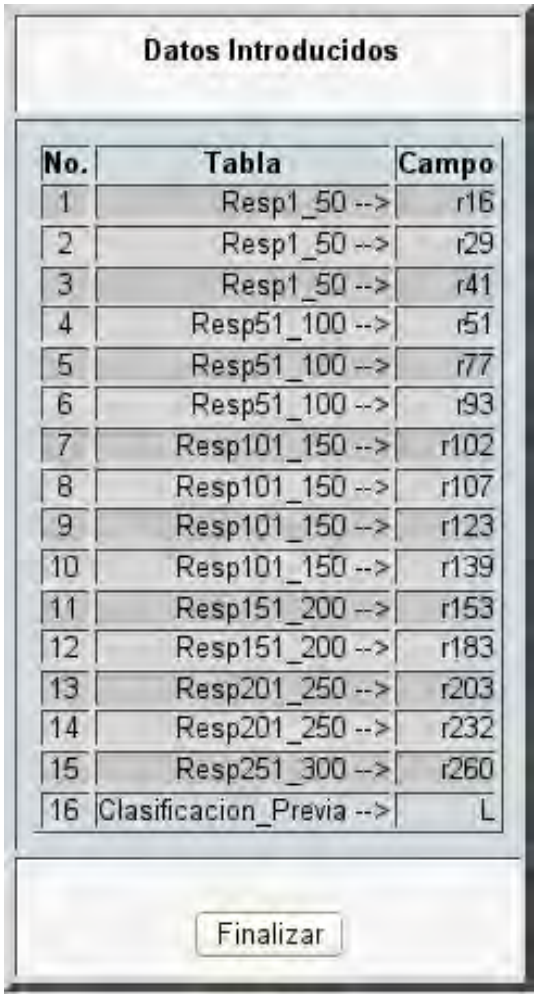


Figura 7.13: Vista minable **L**.

enseguida se oprime el botón ***Finalizar*** para terminar y deberá aparecer el texto de la figura 7.14.

Preparación de datos procesada.

Figura 7.14: Vista minable procesada.

Entonces serán creados los archivos ***L*** (nombre del modelo), ***L.names*** (nombres de clases, dominios y atributos), ***L.data*** (datos de entrenamiento/prueba) y ***L.test*** (datos de evaluación) que sirven para entrenar y evaluar al algoritmo C4.5.

7.5.6. Entrenamiento del algoritmo C4.5 (Agente Inteligente de Software)

Continuando con la secuencia del ejemplo, después de observar el texto "Preparación de datos procesada" se puede regresar una ventana en el navegador y enseguida oprimir el botón ***Menú Minería de Datos*** para volver al menú del sistema de inducción C4.5.

En general el entrenamiento se puede llevar a cabo con los archivos generados en la etapa anterior del sistema o bien, con archivos incluidos como ejemplos o archivos obtenidos de otra fuente (repositorios, etc) que cumplan con el formato adecuado. Enseguida se muestran los pasos para el entrenamiento:

1. En el menú del sistema ADC4.5Web mostrado en la figura 7.8, se oprime el botón ***Entrenar C4.5*** para ver el formulario de entrada de datos del algoritmo C4.5 mostrado en la figura 7.15. En este formulario se proporciona la ruta y nombre del modelo a entrenar, así como las opciones de entrenamiento. Estas opciones fueron discutidas en la sección 7.4.2.3.

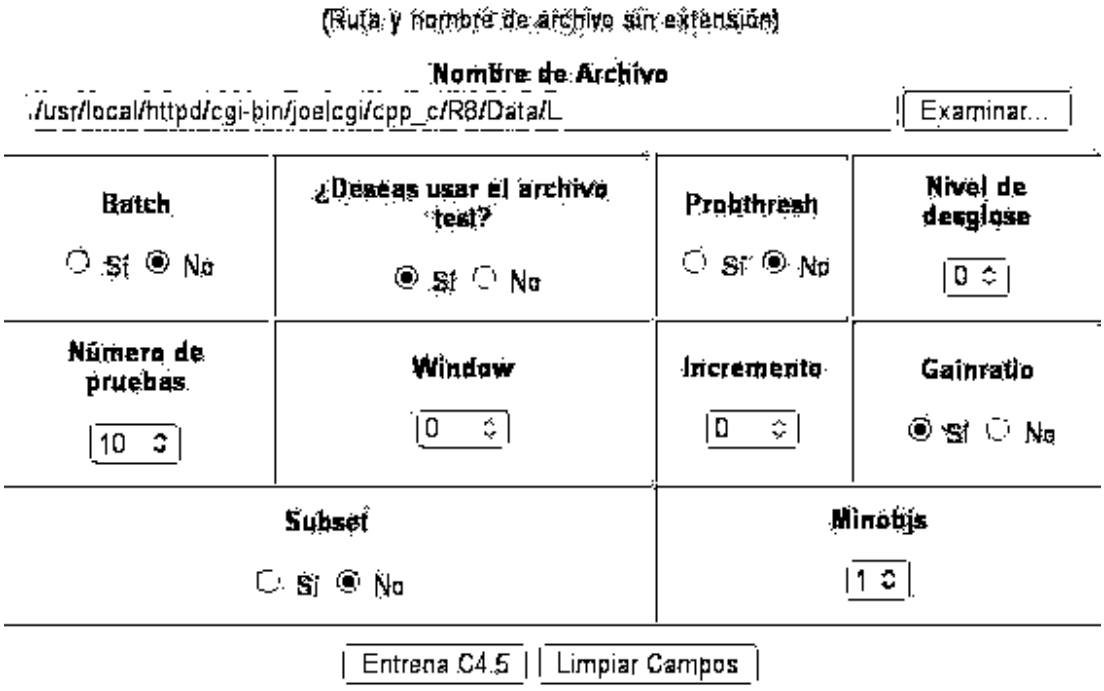


Figura 7.15: Formulario de entrada de datos para el algoritmo C4.5.

2. Enseguida se oprime el botón ***Entrena C4.5*** para generar el árbol de decisión correspondiente el cual será mostrado inmediatamente, como se puede ver en la figura 7.16.

3. Después se deben observar y analizar los resultados mostrados, si el modelo generado se considera adecuado, entonces está listo para ser consultado. En caso contrario repetir el paso 1, es decir, preparar los datos nuevamente y enseguida repetir el paso 2 (entrenar C4.5), nuevamente se deben observar y analizar los resultados mostrados, esto debe realizarse hasta que el sistema genere un modelo adecuado, si no es posible conseguir esto, entonces puede ser necesario revisar el planteamiento del problema o aumentar el volumen de datos.

El entrenamiento del algoritmo se realiza en cuestión de segundos, además como ya se ha mencionado anteriormente, no requiere de una gran capacidad de cómputo, lo cual permite su implantación en cualquier computadora de uso común.

7.5.7. Árbol de Decisión generado por el algoritmo C4.5

En la figura 7.16 se muestra una porción del Árbol de Decisión generado por el algoritmo C4.5 para el ejemplo de la escala de validación **L** del MMPI-2.

```
C4.5 [versión 8] Generador de árboles de decisión       Wed Jul 15 14:57:48 2009
-----------------------------------------------

1116 casos leídos (15 atributos) desde L.data

Arbol de decisión:

r51 = V:
|   r16 = V:
|   |   r203 = V:
|   |   |   r41 = V:
|   |   |   |   r77 = V:
|   |   |   |   |   r29 = V: Baja (281.0)
|   |   |   |   |   r29 = F:
|   |   |   |   |   |   r123 = V: Baja (15.0)
|   |   |   |   |   |   r123 = F:
|   |   |   |   |   |   |   r232 = V: Baja (3.0)
|   |   |   |   |   |   |   r232 = F:
|   |   |   |   |   |   |   |   r260 = V: Baja (1.0)
|   |   |   |   |   |   |   |   r260 = F: Medio (2.0)
|   |   |   |   r77 = F:
|   |   |   |   |   r232 = V: Baja (20.0/2.0)
|   |   |   |   |   r232 = F:
|   |   |   |   |   |   r29 = F: Medio (2.0)
|   |   |   |   |   |   r29 = V:
|   |   |   |   |   |   |   r123 = V: Baja (3.0)
|   |   |   |   |   |   |   r123 = F: Medio (4.0/1.0)
|   |   |   r41 = F:
|   |   |   |   r29 = V:
|   |   |   |   |   r123 = V: Baja (15.0)
|   |   |   |   |   r123 = F:
|   |   |   |   |   |   r77 = V:
|   |   |   |   |   |   |   r260 = F: Baja (7.0)
|   |   |   |   |   |   |   r260 = V:
|   |   |   |   |   |   |   |   r232 = V: Baja (1.0)
|   |   |   |   |   |   |   |   r232 = F: Medio (1.0)
|   |   |   |   |   |   r77 = F:
|   |   |   |   |   |   |   r232 = V: Baja (3.0/1.0)
|   |   |   |   |   |   |   r232 = F: Medio (3.0)
|   |   |   |   r29 = F:
|   |   |   |   |   r232 = F: Medio (7.0/1.0)
|   |   |   |   |   r232 = V:
|   |   |   |   |   |   r123 = V: Baja (4.0)
|   |   |   |   |   |   r123 = F: Medio (5.0/1.0)
|   |   r203 = F:
|   |   |   r41 = V:
|   |   |   |   r232 = V:
```

Figura 7.16: Porción del Árbol de Decisión generado para la escala de validación **L**.

Esta salida generada por el algoritmo C4.5 se interpreta de la forma siguiente:

Primero, la ***cabecera*** (7.16) indica:

- El número de atributos: ***15***
- El número total de ejemplos de entrenamiento, o casos, que son leídos en el archivo de datos *nombrearchivo.data* por C4.5: ***1116***
- El nombe del archivo que fue usado: ***L.data***

Segundo, se genera una o más traducciones ASCII de un Árbol de Decisión, como el árbol simplificado que se muestra en la figura 7.17.

- El árbol consiste en una estructura de valores de atributo que descienden desde una raíz de prueba de atributo.
- En este ejemplo, la raíz es el atributo de prueba *r51* (reactivo 51, "No leo diariamente todos los artículos editoriales del periodico"). Este tiene dos valores de atributo: *V* (verdadero) y *F* (falso).
- Ocurren dos subárboles: un subárbol *r16* (reactivo 16, "De vez en cuando pienso en cosas demasiado malas como para hablar de ellas") bajo *V*, y un subárbol *r102* (reactivo 102, "Algunas veces me enojo") bajo *F* y así sucesivamente.
- El número entre paréntesis que sigue a cada hoja es igual al número de ejemplos de entrenamiento, clasificados correctamente.
- Este número puede ser seguido por un segundo número (ej. 331.0/11.8), en tal caso el segundo valor (11.8) es igual al número de errores de clasificación.
- La suma de la primera serie de números es igual al número total de casos leídos por el algoritmo C4.5 desde el archivo de datos ej_golf.data. (331.0+15.0+9.0+3.0+3.0+7.0+4.0+5.0+35.0+203.0+21.0+5.
- La suma de la segunda serie de números es igual al número total de errores (239 para este ejemplo).
- Se crean dos archivos binarios durante la ejecución:

  nombrearchivo.unpruned: el árbol de decisión sin podar generado y usado por el algoritmo C4.5.

  nombrearchivo.tree: el árbol de decisión podado generado y usado por el algoritmo C4.5.

```
Simplificación del Arbol de decisión:

r51 = V:
|    r16 = V:
|    |    r203 = V:
|    |    |    r41 = V: Bajo (331.0/11.8)
|    |    |    r41 = F:
|    |    |    |    r29 = V:
|    |    |    |    |    r123 = V: Bajo (15.0/1.3)
|    |    |    |    |    r123 = F:
|    |    |    |    |    |    r77 = V: Bajo (9.0/2.4)
|    |    |    |    |    |    r77 = F:
|    |    |    |    |    |    |    r232 = V: Bajo (3.0/2.1)
|    |    |    |    |    |    |    r232 = F: Medio (3.0/1.1)
|    |    |    |    r29 = F:
|    |    |    |    |    r232 = F: Medio (7.0/2.4)
|    |    |    |    |    r232 = V:
|    |    |    |    |    |    r123 = V: Bajo (4.0/1.2)
|    |    |    |    |    |    r123 = F: Medio (5.0/2.3)
|    |    r203 = F:
|    |    |    r41 = F: Medio (35.0/8.2)
|    |    |    r41 = V:
|    |    |    |    r232 = F: Medio (203.0/87.4)
|    |    |    |    r232 = V:
|    |    |    |    |    r123 = V: Bajo (21.0/3.7)
|    |    |    |    |    r123 = F:
|    |    |    |    |    |    r77 = V: Bajo (5.0/2.3)
|    |    |    |    |    |    r77 = F: Medio (5.0/1.2)
|    r16 = F:
|    |    r203 = F: Medio (220.0/19.4)
|    |    r203 = V:
|    |    |    r29 = V:
|    |    |    |    r123 = V:
|    |    |    |    |    r41 = V: Bajo (14.0/1.3)
|    |    |    |    |    r41 = F:
|    |    |    |    |    |    r77 = V: Bajo (6.0/2.3)
|    |    |    |    |    |    r77 = F: Medio (4.0/1.2)
|    |    |    |    r123 = F:
|    |    |    |    |    r77 = F: Medio (8.0/1.3)
|    |    |    |    |    r77 = V:
|    |    |    |    |    |    r107 = V: Bajo (1.0/0.8)
|    |    |    |    |    |    r107 = F:
|    |    |    |    |    |    |    r232 = F: Medio (165.0/71.9)
|    |    |    |    |    |    |    r232 = V:
|    |    |    |    |    |    |    |    r41 = V: Bajo (6.0/1.2)
|    |    |    |    |    |    |    |    r41 = F: Medio (4.0/2.2)
|    |    |    r29 = F:
|    |    |    |    r260 = V: Medio (22.0/1.3)
|    |    |    |    r260 = F:
|    |    |    |    |    r232 = F: Medio (4.0/1.2)
|    |    |    |    |    r232 = V:
|    |    |    |    |    |    r123 = V: Bajo (7.0/3.4)
|    |    |    |    |    |    r123 = F: Medio (2.0/1.0)
r51 = F:
|    r102 = V: Moderado (6.0/2.3)
|    r102 = F: Alto (1.0/0.8)
```

Figura 7.17: Árbol de Decisión simplificado generado para la escala de validación **L**.

Tercero, el Árbol de Decisión sin podar y el Árbol de Decisión podado son evaluados con los datos de entrenamiento para verificar la eficiencia de cada uno, como se aprecia en la figura 7.18.

- La primera tabla ilustra la eficiencia de un árbol sin podar. Esta tiene dos columnas:

  1. **Tamaño**: el tamaño del árbol sin podar. Que es el número de nodos por los cuales está compuesto, **147** para este ejemplo.
  2. **Errores**: el número de errores de clasificación y su correspondiente porcentaje de error a partir del número total de casos, **165(14.8 %)** para este ejemplo.

- La segunda tabla ilustra la eficiencia del árbol podado. Esta tiene tres columnas:

  1. **Tamaño**: el tamaño del árbol podado. Es menor o igual que el árbol sin podar dependiendo de la magnitud de la poda ejecutada por el algoritmo C4.5, **55** para este ejemplo.
  2. **Errores**: el número de errores de clasificación y su correspondiente porcentaje de error actual después de la poda, **192(17.2 %)** para este ejemplo.
  3. **Estimación**: el porcentaje de error estimado después de la poda del árbol, conveniente cuando se compara con el porcentaje actual, **21.4 %** para este ejemplo.

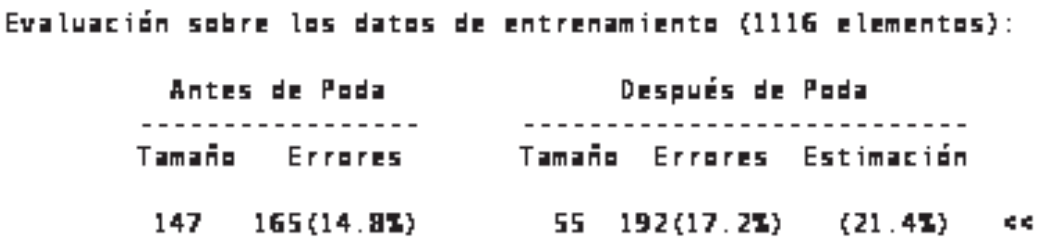


Figura 7.18: Evaluación del Árbol de Decisión con los datos de entrenamiento.

Cuarto, evaluación del modelo generado con una matriz de confusión. Una matriz de confusión (confusion matrix) contiene información sobre la realidad actual y clasificaciones predichas hechas por un sistema de clasificación. El funcionamiento de tales sistemas es evaluado comúnmente usando los datos en la matriz. La tabla 7.2 muestra la composición de la matriz para un clasificador de dos clases.

		Predicciones	
		Negativo	**Positivo**
Actual	**Negativo**	a	b
Actual	**Positivo**	c	d

Tabla 7.2: Matriz de confusión.

Las entradas en la matriz de confusión tienen los siguientes significados en el contexto de estudio: **a** es el número de predicciones correctas de ejemplos negativos clasificados como negativos, **b** es el número de predicciones incorrectas de ejemplos negativos clasificados como positivos, **c** es el número de predicciones incorrectas de ejemplos positivos clasificados como negativos, y **d** es el número de predicciones correctas de ejemplos positivos clasificados como positivos. Un modelo se considera bueno si no se presentan errores en las predicciones.

En la figura 7.19 se muestra la matriz de confusión para el ejemplo de la escala de validación L del MMPI-2, en la que se indica la cantidad de datos (número de ejemplos contenidos en el archivo ***L.test***) empleados en la evaluación del modelo, enseguida está la matriz de confusión, en la que se observa que el modelo ha clasificado correctamente 92 ejemplos con la clase *Bajo*, 136 ejemplos clasificados correctamente con la clase *Medio* y 51 ejemplos clasificados incorrectamente con la clase *Bajo*. Por lo tanto, como se presentan errores en las predicciones, esto demuestra que el fundamento empírico del instrumento MMPI-2 debería ser revisado, ya que se detecta un problema de sobreadaptación en los datos, lo que significa que algunos ejemplos están descritos de manera distinta y clasificados de forma similar.

```
Evaluación sobre los datos de prueba (279 elementos):

          Antes de Poda              Después de Poda
       ------------------      --------------------------
       Tamaño   Errores        Tamaño  Errores  Estimación

        147    55(19.7%)          55   51(18.3%)   (21.4%)   <<


         (a)   (b)   (c)   (d)   (e)        <-clasificado como
        ----  ----  ----  ----  ----
          92    51                         (a): class Bajo
               136                         (b): class Medio
                                           (c): class Moderado
                                           (d): class Alto
                                           (e): class Muy alto
```

Figura 7.19: Evaluación del Árbol de Decisión con los datos de prueba.

7.5.8. Consulta interactiva del Árbol de Decisión generado por el algoritmo C4.5

Continuando con el ejemplo de la escala **L**, una vez que se ha generado un modelo adecuado, se pueden regresar las ventanas necesarias en el navegador o bien, reingresar al sistema para volver al menú de ADC4.5Web. Enseguida se muestran los pasos para consultar un árbol:

1. En el menú del sistema ADC4.5Web mostrado en la figura 7.8, se oprime el botón ***Consultar árbol***.

2. Enseguida se selecciona el nombre del Árbol de Decisión (modelo) que se desea consultar, como se indica en la figura 7.20 y se oprime el botón ***Consulta árbol***.

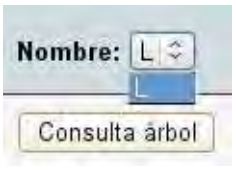


Figura 7.20: Selección del Árbol de Decisión que se desea consultar.

3. Enseguida ya se puede realizar la consulta del Árbol de Decisión, seleccionando la respuesta para el nodo de evaluación que se considere apropiada en la lista desplegable, para este ejemplo el nodo correspondiente al reactivo 51 ("No leo diariamente todos los artículos editoriales del período"), la respuesta seleccionada es **V** y enseguida se oprime el botón ***Enviar***„ como se muestra en la figura 7.21.

Figura 7.21: Inicio de la consulta del Árbol de Decisión L con el reactivo 51.

Después aparece otro nodo de evaluación, el nodo correspondiente al reactivo 16 ("De vez en cuando pienso en cosas demasiado malas como para hablar de ellas"), la respuesta seleccionada es **F** y enseguida se oprime el botón ***Enviar***, como se muestra en la figura 7.22.

Figura 7.22: Selección de respuesta para el reactivo 16.

A continuación aparece otro nodo de evaluación, el nodo correspondiente al reactivo 203 ("En ocasiones me gusta el chisme"), la respuesta seleccionada es **F** y enseguida se oprime el botón ***Enviar***, como se muestra en la figura 7.22.

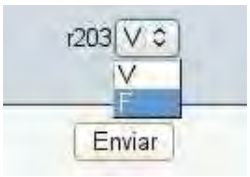


Figura 7.23: Selección de respuesta para el reactivo 203.

Finalmente el Árbol de Decisión muestra la clasificación *Medio* de este ejemplo como se aprecia en la figura 7.24.

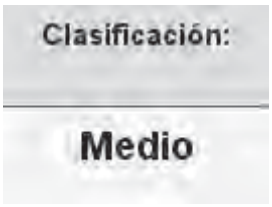


Figura 7.24: Clasificación del ejemplo de consulta.

7.5.9. Consulta gráfica del Árbol de Decisión generado por el algoritmo C4.5

Antes de consultar gráficamente un Árbol de Decisión, primero se debe consultar de forma interactiva (para que se genere la imagen correspodiente). Para el caso de este ejemplo, se pueden regresar las ventanas necesarias en el navegador o bien, reingresar al sistema para volver al menú de ADC4.5Web. Enseguida se muestran los pasos para consultar un árbol:

1. En el menú del sistema ADC4.5Web mostrado en la figura 7.8, se oprime el botón ***Mostrar árbol***.

2. Después aparece una tabla con los nombres de imágenes correspondientes a los Árboles de Decisión generados, se elige el elemento deseado y se oprime el botón ***Muestra árbol***, como se observa en la figura 7.25.

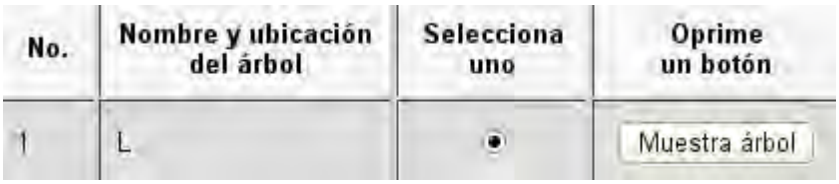


Figura 7.25: Selección de la imagen de un Árbol de Decisión.

A continuación se muestra la imagen correspondiente al Árbol de Decisión seleccionado, como se aprecia en la figura 7.26.

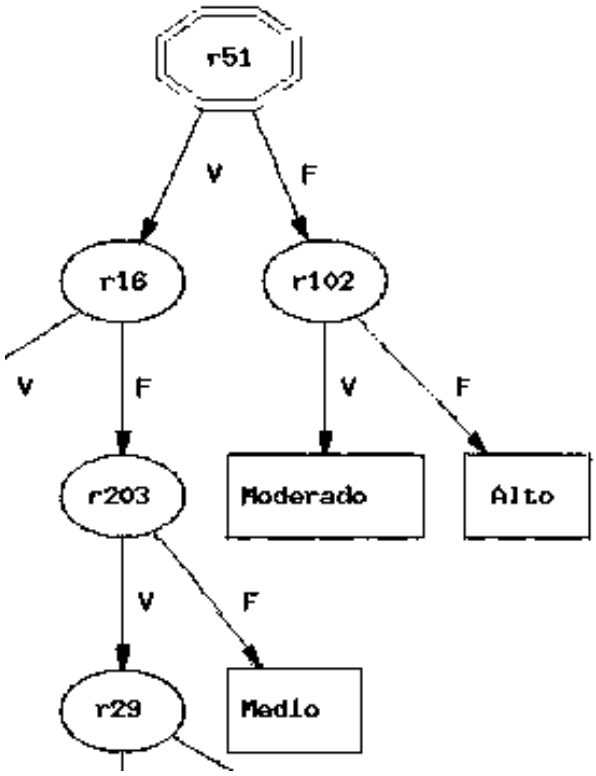


Figura 7.26: Porción de la imagen correspondiente al Árbol de Decisión de la escala **L**.

7.6. Conclusión del capítulo

La herramienta de minería Ambiente de Descubrimiento con C4.5Web (ADC4.5Web) ha sido implementada con aplicaciones de código abierto y libre distribución proporcionando una alternativa de bajo costo accesible a los usuarios. ADC4.5Web es una aplicación que integra una adecuación del algoritmo generador de árboles de decision C4.5[Quinlan, 1993] para el ambiente web, con acceso a la información minable administrada mediante el sistema manejador de bases de datos MySQL 3.23.48.

Como se menciona en las secciones 2.6, 3.2 y 4.4, el Agente Inteligente de Software se implementa con el algoritmo C4.5, el que por medio de la herramienta de minería ADC4.5Web puede interactuar con los Agentes Inteligentes Humanos involucrados en el proceso de descubrimiento de conocimiento, tomando como guía al algoritmo de referencia para minería propuesto en el capítulo 5, conformando de esta manera al Multiagente Inteligente Descubreconocimiento capaz de llevar a cabo el proceso completo de descubrimiento de conocimiento en bases de datos (KDD), como se ha mencionado en repetidas ocasiones.

El sistema ADC4.5Web, es una herramienta de manejo simple y de bajo costo que permite a los investigadores, a las personas dedicadas a los negocios o usuarios en general, llevar a cabo el proceso KDD.

Capítulo 8

Conclusiones

En este capítulo se presentan las conclusiones bajo las siguientes perspectivas:

- Agente Inteligente Descubreconocimiento
- Técnica de Minería de Datos
- Algoritmo de referencia de minería
- Conocimiento obtenido
- Aportaciones
- Resultados
- Trabajos futuros

8.1. Agente Inteligente Descubreconocimiento

Como se discute en el capítulo 2, el Agente Inteligente Descubreconocimiento capaz de llevar a cabo exitosamente el proceso de descubrimiento en bases de datos, es en realidad un Multiagente, conformado por varios Agentes Inteligentes Humanos y un Agente Inteligente de Software. En este caso los AI Humanos con perfil en computación son el autor de esta tesis y su asesor el Dr. Juan José Flores Romero. Los AI Humanos con perfil en el área de psicometría fueron en un principio, los psicólogos Ma. Inés Castro Curiel y David Urbina Zamora, del plantel del Instituto Tecnológico Superior de Uruapan donde se inició el caso de estudio y la psicóloga Luz María Elias Amezcua del CAPEP de Uruapan, que fue consultada como apoyo. Finalmente se contó con la colaboración bastante valiosa de la Dra. Emilia Lucio Gómez-Maqueo, quien ha trabajado ampliamente con el inventario MMPI-2, publicando en este país, la versión en español del Inventario Multifásico de la Personalidad Minnesota 2 (MMPI-2) junto con el Manual de la prueba, constituyendo una de las tres versiones en español, autorizadas por la Editorial de la Universidad de Minnesota [Hathaway y McKinley, 1995]. No se contó con la colaboración de AI Humanos con perfil en estadística, lo cual hubiese sido deseable, por

lo que el personal con los otros perfiles se vió obligado a realizar la parte básica de este perfil. El Agente Inteligente de Software es el algoritmo generador de Árboles de Decisión C4.5 [Quinlan, 1993] únicamente.

Con el AI Descubreconocimiento se hace énfasis en la conformación de un equipo de personal interdisciplinario que colabore en las tareas de minería, resaltando la importancia imprescindible de los seres humanos en tal proceso, lo que proporciona una mayor posibilidad de realizar con éxito el proceso de KDD.

8.2. Técnica de Minería de Datos

La técnica de Minería de Datos que se empleó fue **Clasificación**. Esta sirve para resolver problemas predictivos, como lo es la aplicación, evaluación e interpretación del inventario MMPI-2. El algoritmo que se utilizó fue el C4.5 para la generación de Árboles de Decisión. Este algoritmo se entrenó con un conjunto de 1395 ejemplos para crear el modelo de la versión experimental reducida del instrumento MMPI-2.

Cabe resaltar que la aplicación del algoritmo C4.5, presenta las siguientes limitantes: el tipo de problema que resuelve la herramienta de minería ADC4.5Web es únicamente la clasificación, sólo se emplea el algoritmo de minería C4.5 para generar Árboles de Decisión. La ventaja que proporciona ADC4.5Web es en la transformación de los datos a la granularidad correcta, puesto que permite generar **vistas minables** a partir de las bases de datos almacenadas con MySQL 3.23.48.

8.3. Algoritmo de referencia de minería

La metodología empleada para el proceso KDD fue el algoritmo de referencia de minería propuesto en el capítulo 5. Este es para uso general, como pueden ser tareas de investigación, de negocios, etc. Consta de las siguientes etapas:

1. Identificación del problema
2. Obtención de datos
3. Selección de datos
4. Preprocesamiento de los datos
5. Transformación de los datos a la granularidad correcta
6. Definición del tipo de problema de minería
7. Selección del algoritmo de minería
8. Proceso de Minería de Datos

9. Evaluación del modelo e interpretación de resultados
10. Selección del mejor modelo
11. Conocimiento descubierto

8.4. Conocimiento obtenido

Al obtener los Árboles de Decisión (modelos) para cada uno de los indicadores de validez del MMPI-2, así como para cada una de las escalas clínicas, se detectó un problema de "inconsistencia" en los datos, lo que significa que algunos son incorrectos, es decir: existen más de un ejemplos descritos de manera diferente(en función de sus atributos), pero sus puntuaciones y clasificación son idénticas, entonces el algoritmo C4.5 resulta incapaz de encontrar un Árbol de Decisión congruente para todos los ejemplos. Esto confirma el fundamento empírico del instrumento MMPI-2 y sugiere que se realice en un trabajo posterior un análisis con datos recopilados por profesionales con un alto grado de especialización en el MMPI-2 para comprobar si el problema de "inconsistencia" persiste, si es así, la recomendación es revisar al instrumento para corregir el aspecto empírico de su fundamento, en caso contrario se considerará aceptable el uso del instrumento.

Cuando se inició el caso de estudio con el departamento de psicología del Instituto Tecnológico Superior de Uruapan, el inventario se aplicaba sólamente con 370 reactivos del total de 567, los que corresponden a las escalas clínicas del inventario MMPI-2 y aunque no se aplicaba el inventario completo, aún así se consideraba útil obtener una versión todavía más reducida de este instrumento. Esta versión se obtuvo, pero cuando el autor de esta tesis tuvo contacto con la Dra. Emilia Lucio, ésta le informó que el inventario rigurosamente se debe aplicar completo a los pacientes con los 567 reactivos que lo conforman, puesto que dicho inventario posee un fundamento empírico y depende de su aplicación en forma completa para arrojar resultados adecuados.

Por tal razón la versión reducida correspondiente a las escalas clínicas no se consideró recomendable para su uso en psicometría. Ocurrió todo lo contrario con la versión estandarizada del inventario MMPI-2Web, desarrollada para comparar los resultados obtenidos con ésta y los de la versión reducida. Esta versión estándar del inventario MMPI-2 es bastante funcional y fácil de usar, hace posible la administración del instrumento MMPI-2 permitiendo la manipulación básica (consulta, inserción, eliminación y actualización) de los datos de pacientes, reactivos, respuestas del test y diagnósticos, almacenados en las bases datos del MMPI-2 contenidas en MySQL 3.23.48. Esta herramienta también permite la aplicación del test tanto en forma individual o colectiva y su respectiva calificación automatizada de acuerdo con las normas del MMPI-2, incluyendo la graficación de los perfiles. Esta versión estándar del instrumento MMPI-2 si contó con la aprobación de la Dra. Emilia Lucio, recomendó su utilización aunque por el momento no sea de manera abierta al público en general, puesto que el inventario está registrado y posee derechos de autor, por lo que se requeriría la autorización de los propietarios del MMPI-2 para poder usar sin limitaciones la versión estándar del inventario MMPI-2 desarrollada en esta tesis.

8.5. Aportaciones

Una de las aportaciones más importantes de este trabajo es el desarrollo de la herramienta de minería Ambiente de Descubrimiento con C4.5Web implementada con aplicaciones de código abierto y libre distribución proporcionando una alternativa de bajo costo accesible a los usuarios, esta aplicación integra una adecuación del algoritmo de minería C4.5 [Quinlan, 1993] para el ambiente web, con acceso a la información minable administrada mediante el sistema manejador de bases de datos MySQL 3.23.48. ADC4.5Web hace posible que el AI Descubreconocimiento guiado por el algoritmo de referencia de minería propuesto, sea capaz de realizar el proceso completo de KDD, conformando en realidad una alternativa de KDD a bajo costo para los investigadores, científicos o expertos en algún área de actividad humana

Aunque esta aplicación es de uso general, seguramente resultará limitada en algunos casos de aplicación, ya que no existe un algoritmo de minería que sea eficiente para todos los casos, sin embargo, el algoritmo empleado (C4.5) ha demostrado ser bastante eficiente.

Otra aportación importante es la aplicación MMPI-2Web, desarrollada a partir de la versión estándar del inventario MMPI-2 como se menciona en la sección anterior.

8.6. Resultados

Los archivos correspondientes a los resultados obtenidos al realizar el proceso de descubrimiento de conocimiento (KDD) sobre el instrumento MMPI-2 como caso de estudio, con el fin de obtener una alternativa simplificada de dicho instrumento, se encuentran en el CD que acompaña a este documento.

8.7. Trabajos futuros

Se propone como trabajo futuro lo siguiente:

- Entrenar la herramienta ADC4.5Web con un conjunto de ejemplos mucho mayor, incluyendo a los grupos de pacientes criterio de los hospitales psiquiátricos.
- Extender ADC4.5Web para adicionarle más algoritmos de minería y conectividad con otros sistemas manejadores de bases de datos.
- Aplicar ADC4.5Web para resolver problemas de otras áreas como la biotecnología, bioinformática, etc.

Bibliografía

[Branchman y Anand, 1996] Branchman, R.J., Anand, T. *"The Process Of Knowledge Discovery In Databases: A Human-Centered Aproach"*. En *Advances in Knowledge Discovery and Data Mining*. AAAI / MIT Press. Cambridge, Mass. USA. 1996.

[C4.5] C4.5. http://www.mkp.com.c45

[Clark y Boswell, 2000] Clark, P.; Boswell, R. Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers. USA. 2000.

[CRISP-DM] CRISP-DM. http://www.crisp-dm.org

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro G., Smyth P. *"From data mining to knowledge discovery: An overview". Advances in Knowledge Discovery and Data Mining.* AAAI / MIT Press. Cambridge, Mass. USA.1996.

[Hernández et al., 2004] Hernández Orallo, J.; Ramírez Quintana, M. J.; Ferri Ramírez, C. Introducción a la Minería de Datos. Pearson Educación, S.A. España. 2004.

[Mitchell, 1997] Mitchell, T. Machine Learning. MCB/McGraw-Hill. Carnegie Mellon University, USA. 1997.

[Butcher, 2001] Butcher, J.N. *MMPI-2. Guía para principiantes*. Manual Moderno. México. 2001.

[Hathaway y McKinley, 1995] Hathaway, S.R., McKinley, J.C. I*nventario Multifásico de la Personalidad Minnesota-2 (MMPI-2)*. Manual Moderno. México. 1995.

[Lucio y León, 2003] Lucio, E., León, M.I. *Uso e interpretación del MMPI-2 en español*. Manual Moderno. México. 2003.

[MySQL] MySQL. http://www.mysql.org

[Quinlan, 1990] Quinlan, J.R. Induction of Decision Trees. Machine Learning. Morgan Kaufmann. USA. 1990.

[Quinlan, 1988] Quinlan, J.R. Decision trees and multi-valued attributes. Machine Intelligence. Oxford University Press. Oxford, UK. 1988.

[Quinlan, 1993] Quinlan, J.R. C4.5: programs for machine learning. Morgan Kaufmann Publishers, Inc. USA. 1993.

[Quinlan, 1993b] Quinlan, J.R. Learning Efficient Classification Procedures and Their Application to Chess Games. Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann. USA. 1993.

[Russell y Norvig, 1996] Russell, S., Norvig, P. *Inteligencia Artificial: un enfoque moderno.* Prentice Hall Hispanoamericana S.A. México. 1996.

[Russell y Norvig, 2004] Russell, S., Norvig, P. *Inteligencia Artificial: un enfoque moderno.* 2da. Ed. Pearson Educación S.A. España. 2004.

[Triola, 2004] Triola, Mario F. Estadística. 9na. Ed. Pearson Educación. México. 2004.

[University Regina] University Regina. http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html