Predicción de pacientes diabéticos. Preprocesado para Minería de Datos

Prediction of diabetic patients. Preprocessing for Data Mining

Ing. Rolando Acosta Sánchez, Facultad de Ingeniería Informática, Ciudad Universitaria José Antonio Echeverría (CUJAE), Cuba. Email: rosete@ceis.cujae.edu.cu

Dr. Alejandro Rosete Suárez, Facultad de Ingeniería Informática, Ciudad Universitaria José Antonio Echeverría (CUJAE), Cuba

Lic. Alfredo Rodríguez Díaz, Centro para el Desarrollo Informático en la Salud, MINSAP, Cuba

Resumen:

La presente investigación expone el desarrollo de las fases de comprensión y preparación de los datos, dentro de la metodología para desarrollar procesos de Minería de Datos, CRISP-DM 1.0. Se refleja el caso práctico del trabajo con los datos asociados a encuestas realizadas con el objetivo de determinar factores influyentes en el padecimiento de diabetes. Como herramienta de apoyo para la descripción y comprensión de los datos se emplearon Microsoft Excel 2007 y WEKA 3.5.8.

Palabras Claves:

Minería de Datos, KDD, Comprensión de Datos, CRISP-DM, Microsoft Excel 2007, WEKA, Diabetes.

Abstract:

This research exposes the development stage of understanding and data preparation, within the CRISP-DM 1.0 methodology for conducting data mining. It reflects a practical study on data associated to a survey that aimed to determine factors that influence the appearance of diabetes. Microsoft Excel 2007 and WEKA 3.5.8 where used as support tools for describing and understanding the data.

Keywords:

Data Mining, KDD, data compression, CRISP-DM, Microsoft Excel 2007, WEKA, Diabetes.

Introducción

La diabetes constituye un problema de salud delicado que afecta a un gran número de personas a nivel mundial [1], [2], [3]. A continuación se exponen algunos datos tomados del "Atlas of Diabetes Mellitus" (en su versión de 2008) que revelan la magnitud del problema [4]:

- La Diabetes afecta a 246 millones de personas; para el 2025 el número de afectados ascenderá a 380 millones.
- La Diabetes es la cuarta causa de muerte a nivel mundial.
- Al menos el 50% de las personas diabéticas ignoran que lo son.
- El 80% de la Diabetes tipo 2 es prevenible mediante la adopción de una dieta saludable y el incremento de la actividad física.

En nuestro país (Cuba) la diabetes es un problema creciente. La Figura 1 muestra el incremento que se ha producido en la cantidad de personas identificadas como diabéticas entre los años 1991 y 2006 [5].

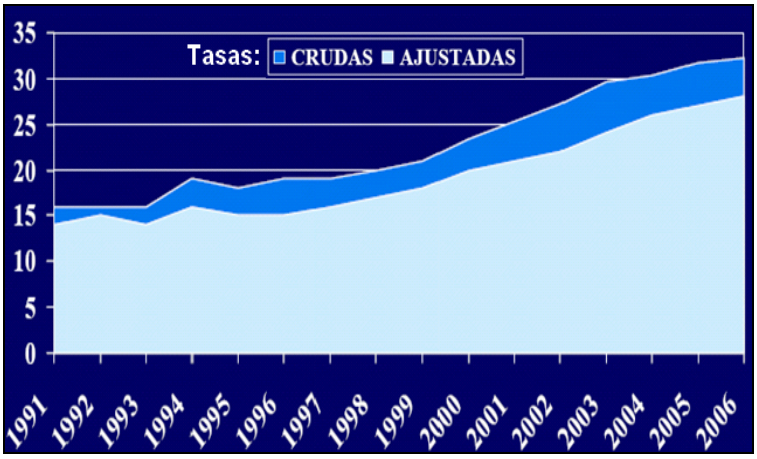

Figura. 1: Prevalencia de Diabetes Mellitus en Cuba. Tasas Crudas y Ajustadas por edad x 1000 habitantes [5].

Desde épocas remotas, el estudio de los casos particulares de determinados padecimientos y características de los pacientes ha sido fundamental para el desarrollo de la medicina como ciencia. El estudio de casos particulares permite la generalización; así como encontrar síntomas o causas de enfermedades u otro tipo de padecimiento.

Actualmente, con el surgimiento de la denominada era digital y el desarrollo de las nuevas técnicas de almacenamiento y recuperación de la información, el registro y análisis de los procesos que tienen lugar en una empresa o sector se ha visto particularmente potenciado. Una forma muy valiosa de análisis de información es la Minería de Datos. A pesar de la popularidad del término, "la Minería de Datos es sólo una etapa, si bien la más importante, de lo que se ha venido llamando el proceso de extracción de conocimiento a partir de datos. Este proceso consta de varias fases e incorpora muy diferentes técnicas de los campos del aprendizaje automático, la estadística, las bases de datos, los sistemas de toma de decisión, la inteligencia artificial y otras áreas de la informática y de la gestión de información." [6]

El proceso de extracción de conocimiento a partir de datos, en inglés Knowledge Discovery in Databases, KDD (o simplemente Minería de Datos, como comúnmente se le llama) ha sido definido de diferentes maneras por diversos autores. Sin embargo, todos refieren las mismas ideas. Por citar una de las definiciones, se puede entender la Minería de Datos como: "el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y

estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, Data-Warehouses, o cualquier otro medio de almacenamiento de información." [7]

Uno de los campos en los que la Minería de Datos se está viendo cada día más utilizada es precisamente en la medicina. En una encuesta realizada por el portal para el análisis de datos KDnuggets, en diciembre de 2008, sobre las diversas áreas en las que se emplea la Minería de Datos, aparecen las aplicaciones en la medicina en los lugares 15, con un 9.3% de empleo (cuando se refiere al cuidado da la salud) y en el lugar 20, con un 7.5% de empleo (cuando se refiere a procesos farmacéuticos) [8]. La Figura 2 muestra los resultados de la encuesta referida. Un análisis comparativo entre los resultados de la encuesta realizada en ese año (2008) y el anterior (2007), reflejó que en tanto las aplicaciones referidas a procesos farmacéuticos bajaron 1.9 puntos porcentuales, las aplicaciones al cuidado de la salud aumentaron 2.1 puntos porcentuales. Los resultados para el año 2007 pueden ser consultados en [9]. Estos análisis evidencian el amplio uso que está teniendo la Minería de Datos en estas áreas, y su estabilidad de empleo.

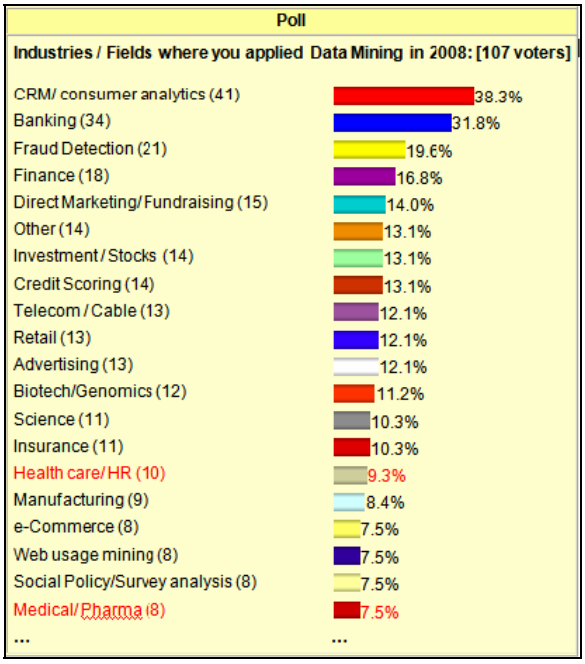

Figura 2 - Campos en los que se emplea la Minería de Datos. Encuesta realizada por KDnuggets, en diciembre de 2008[8].

Aprovechando esta tendencia, y las bondades brindadas por las nuevas tecnologías, surge la idea de aplicar técnicas de Minería de Datos a los resultados de una encuesta realizada en la localidad de Jaruco, Provincia Habana, Cuba; a pacientes identificados por el personal médico como diabéticos o con riesgo de padecer la enfermedad.

El objetivo de la encuesta fue recopilar la mayor cantidad de información posible, asociada a las características de cada paciente (síntomas, análisis médicos, etc.). El objetivo del análisis mediante técnicas de Minería de Datos es: Obtener modelos que permitan clasificar a un paciente, según los indicadores registrados por la encuesta, en una de las siguientes categorías: Diabético conocido (DC), Diabético detectado (DD), Grupo de no riesgo (GNR), Grupo de riesgo (GR), Tolerancia a la glucosa alterada (TGA), o Alteración de la Glucosa en Ayunas (AGA).

Como se planteó anteriormente, el KDD engloba varias fases. Una de las fases concebidas dentro de este proceso, comúnmente denominada "Pre-procesado de datos", se centra en garantizar la mayor fidelidad y corrección de los datos que se van a emplear como materia prima para los análisis. Múltiples autores coinciden en que la fase de "Pre-procesado de los datos" es la más engorrosa y costosa, en todo proceso de análisis de datos. En el caso particular de la Minería

de Datos esta etapa ocupa cerca del 70% del esfuerzo [10]. En la Figura 3 se refleja el por ciento de esfuerzo promedio necesario para desarrollar cada una de las fases del KDD.

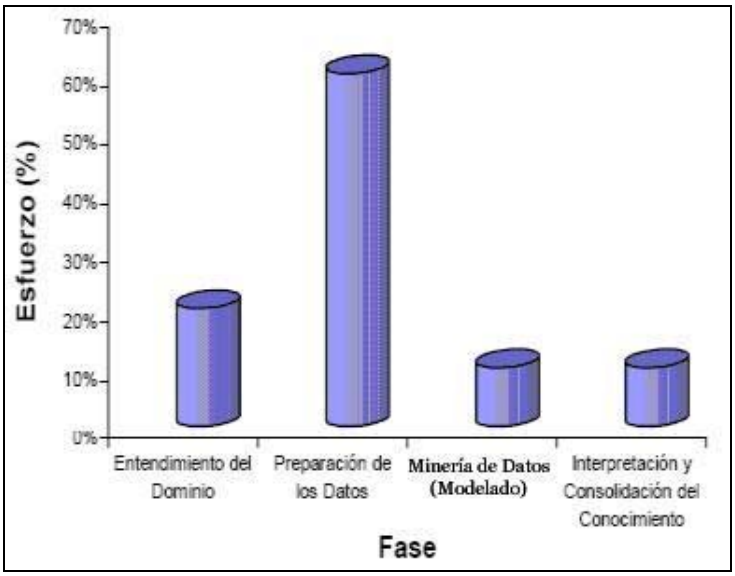

Figura 3 - Esfuerzo requerido para el desarrollo de cada una de las fases en un proceso de KDD [10].

Por la complejidad y lo abarcadora que resulta la etapa de pre-procesado algunos autores la dividen en dos: la "Comprensión de los Datos" y la "Preparación de los datos" [11]. Durante la "Comprensión de los datos" se hace una recolección y exploración inicial de los datos para familiarizarse con ellos e identificar problemas de calidad. Además, se trata de descubrir o estimar las relaciones más evidentes para formular las primeras hipótesis sobre información oculta en ellos. La fase de "Preparación de los datos" cubre todas las actividades necesarias para construir la colección de datos que finalmente será minada a partir de la colección inicial de datos. Las tareas asociadas a la preparación de los datos se desarrollan en diferentes ocasiones y no necesariamente siguen un orden prescrito.

En el presente trabajo se reflejan los métodos empleados para el desarrollo de la fase de "Pre-procesado de los datos"; y se exponen los principales resultados obtenidos. Dentro de la fase de pre-procesado se hace particular énfasis en la "Compresión de los datos".

Materiales y Métodos

Metodología

Como se ha planteado, el proceso de descubrimiento de conocimientos a partir de datos (y en particular la etapa de procesado de los datos) suele ser engorroso y difícil. Con el objetivo de guiar y organizar el trabajo se han desarrollado metodologías y estrategias de trabajo.

Dentro de las metodologías que podemos encontrar en la actualidad, la más seguida y referenciada, se nombra CRISP-DM (CRoss Industry Standard Process for Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos) [11]. Esta metodología fue concebida de forma tal que resulte independiente de la herramienta que se utilice para el desarrollo del proyecto; y es de distribución libre, por lo que se encuentra en constante desarrollo por la comunidad internacional. En la Figura 4 se muestran los resultados de una encuesta realizada por el portal para el análisis de datos KDnuggets, en agosto del 2007, sobre las diferentes metodologías empleadas para afrontar procesos de Minería de Datos.

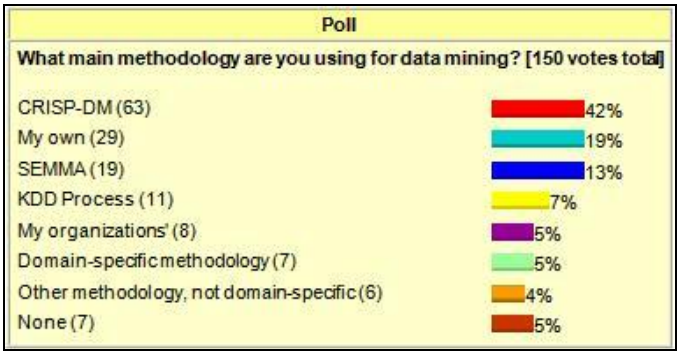

Figura 4 - Principales metodologías empleadas para realizar procesos de KDD según una encuesta realizada por KDnuggets en agosto de 2007. [12]

En esta metodología, que consta de seis fases generales: Análisis del problema, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación y Despliegue; el "Pre-procesado de los Datos" está reflejado en la segunda y tercera fases ("Comprensión de los datos" y "Preparación de los datos"). La Figura 5 muestra las principales relaciones que se establecen entre cada una de las fases de la metodología.

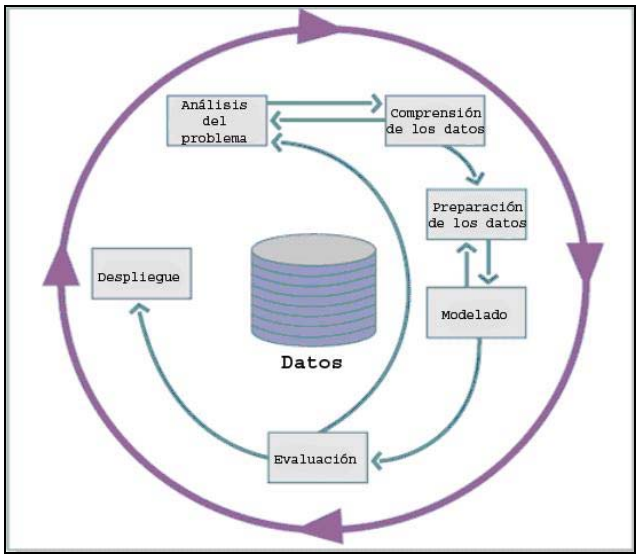

Figura 5 - Fases del modelo de referencia CRISP-DM 1.0 y sus principales relaciones [11].

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

Las ventajas encontradas en esta metodología y el hecho de que sea la más empleada en la actualidad para afrontar procesos de Minería de Datos, unido a que muchas de las restantes se basan en esta (Two Crows [13] y Metodología SQL Server- 2005 [14], entre otras); hizo que fuera la elegida para guiar el análisis de las encuestas realizadas con el fin de predecir padecimientos relacionados con la diabetes.

Herramienta

Para realizar el pre-procesado, los que deseen extraer conocimientos a partir de datos deben, además de contar con una metodología adecuada: apoyarse en herramientas Software que les faciliten la tarea. Para ello se puede emplear todo un arsenal de diversas herramientas.

Situémonos en la recopilación inicial de los datos: estos datos pueden provenir de diversas fuentes (digitales o no); entonces entra en acción un conjunto diverso de gestores de bases de

datos, o de información digital no necesariamente almacenada en bases de datos, así como de medios para la digitalización de la información. Para las tareas de exploración de datos nos podemos apoyar en Software estadísticos o incluso en el conocido Microsoft Excel (una herramienta muy potente para el análisis de datos, fundamentalmente para la exploración y transformación de estos).

Otros Software más específicos para el análisis de datos como Clementine [15], SAS Enterprise Miner [16], o WEKA (Waikato Enviroment for Knowledge Analysis) [17], entre otros; incluyen una serie de prestaciones para importar datos de diversas fuentes digitales o gestores de bases de datos reconocidos. Estas herramientas incorporan además, un conjunto de filtros que permiten realizar las operaciones necesarias para la transformación y limpieza de datos. Muchas permiten incluso la exploración y el filtrado visual de estos. Estas herramientas están designadas específicamente para el análisis de datos mediante el empleo de técnicas de Inteligencia Artificial y algoritmos matemáticos o estadísticos. Durante la fase de Modelado de Datos se vuelve esencial el empleo de alguna de estas.

En el caso práctico que se presenta, los datos fueron reflejados por los encuestadores en una hoja de cálculo de Microsoft Excel 2003 que luego se llevó a la versión 2007 [18] para utilizar las facilidades para filtrar y transformar datos, así como para la elaborar gráficos que brinda este Software. Este será empleado de conjunto con la Suite WEKA (en su versión 3.5.8). WEKA, desarrollada por un equipo de investigadores de la Universidad de Waikato (Nueva Zelanda), es un entorno de experimentación de análisis de datos desarrollado en Java, constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, asociación y visualización. Es en la actualidad una de las herramientas libres para la Minería de Datos de mayor empleo a nivel mundial.

A modo de comparación, la Figura 6 presenta el resultado de una encuesta realizada por el portal para el análisis de datos KDnuggets, en mayo del 2009 [19] donde se reflejan las herramientas más empleadas.

Data Mining Tools Used Poll

What data mining tools have you used for a real project (not just for evaluation) in the past 6 months? (364 voters)	
SPSS PASW Modeler (formerly Clementine) (68 alone, 52 with other tools, 120 total)	
RapidMiner (36 alone, 41 w. other tools, 77 total)	
SAS (39 alone or with SAS EM; 36 with other tools, 75 total)	
Excel (1 alone, 68 total)	
SAS Enterprise Miner (39 alone or with SAS; 28 w/ other tools; 67 total)	
R (2 alone, 51 total)	
Your own code (3 alone, 44 total)	
KXEN (25 alone, 31 total)	
Weka (now Pentaho) (0 alone, 31 total)	
...	...

Figura 6 - Principales herramientas empleadas para el desarrollo de proyectos de Minería de Datos [19].

Resultados alcanzados

En este epígrafe se reflejan los principales resultados alcanzados producto del desarrollo de la fase “Comprensión de los datos” propuesta por la metodología CRISP-DM 1.0. Esta macro-fase está subdividida por la metodología en las siguientes tareas generales: “Recopilación inicial de los datos”, “Descripción de los datos”, “Exploración de los datos” y “Verificación de la calidad de los datos”.

Recopilación inicial de los datos

Como parte del desarrollo de la fase “Recopilación inicial de los datos” se obtuvo un “Reporte de la colección inicial de los datos”. En dicho reporte quedaron reflejadas las fuentes de datos que se emplearán para los futuros análisis, entre las que están las siguientes: los datos colectados reflejan una serie de características presentadas por un grupo de pacientes de la localidad de Jaruco, Provincia Habana, Cuba; los datos se encuentran en una única fuente: un hoja de cálculo de Microsoft Excel 2003; y se tienen registrados los datos correspondientes a 9314 pacientes, de cada paciente se reflejan 63 características.

Descripción de los datos

El desarrollo de la fase de “Descripción de los datos” permitió la obtención del “Reporte de la descripción de los datos”. El objetivo es familiarizarse con la forma en que se encuentran almacenados los datos y sus características. La descripción refleja aspectos como el formato de los datos, la cantidad, y el tipo de cada uno de los campos. La Tabla 1 muestra un fragmento de la tabla original de 63 campos La tabla puede ser consultada en forma integra en el Anexo1.

Tabla I- Fragmento de la tabla “Descripción de campos”.

#	Identificador del Campo	Característica reflejada	Tipo del campo	Importancia para la investigación
…				
15	Ha padecido de presión alta o hipertensión	Si el paciente ha padecido de presión alta o hipertensión.	Booleano (Verdadero o Falso)	Relevante
…				
49	Área Geográfica	Refiere si el paciente es de la zona Urbana o Rural	Nominal	Interesante
…				
61	Nombre y Apellidos	Nombre y apellidos del paciente	Nominal	Sin Importancia
…				

Para cada campo la tabla expone:

- Identificador del campo.
- Breve descripción. Conocer semánticamente qué refleja dicho campo para poder interpretar el resultado mostrado por una herramienta de análisis.
- Tipo de dato que almacena el campo (Numérico, Nominal, Booleano). Conocer esto es importante pues todas las herramientas de análisis no tratan de igual forma cada tipo de datos. Incluso existen algoritmos que implementan técnicas de Minería de Datos que no pueden tratar con determinados tipos.

[1] Se refiere a la importancia que a primera vista, solo con la opinión de los expertos del área de estudio se considera posee el campo. Como se podrá ver posteriormente esta relevancia se va refinando a medida que se profundiza en los análisis.

- Importancia ***A Priori*** [1] del campo para la investigación. Fueron identificados 31 atributos relevantes, 25 que pueden resultar de interés para investigaciones futuras y avalar los resultados de la actual; y los restantes 7 se consideraron sin importancia, por reflejar características triviales que no tienen relación con el estudio emprendido.

[1] Se refiere a la importancia que a primera vista, solo con la opinión de los expertos del área de estudio se considera posee el campo. Como se podrá ver posteriormente esta relevancia se va refinando a medida que se profundiza en los análisis.

Exploración de los datos

La fase de “Exploración de los datos” aborda las interrogantes del análisis de datos que se pueden solucionar usando consultas, visualización y reportes. Estos análisis pueden responder directamente a los objetivos concretos de investigación o contribuir a una mejor descripción de estos; pueden ayudar a la detección de problemas en la calidad y a formular algunas hipótesis sobre las relaciones entre los datos. Son importantes y reflejan una vista inicial sobre el problema.

El “Reporte de la exploración de los datos” obtenido resulta de gran valor para las posteriores fases de la investigación pues en él se detallan las distribuciones que tienen los posibles valores para cada uno de los 31 atributos que se consideró relevante para la investigación. Conociendo la distribución de valores de cada campo es posible una mejor comprensión de los resultados que se obtengan posteriormente durante la fase “Modelado de los datos”. Entre los resultados de interés producto del desarrollo de esta fase podemos citar que se acuerda eliminar de los análisis posteriores al campo “Alteración de Glucosa en Ayuna” pues se pudo observar que el 99.9% de los pacientes dio negativo a este indicador (una distribución tan uniforme de valores no aporta información). La Tabla 2 muestra la distribución de valores para ese campo.

Tabla II Distribución de valores para el campo Alteración de Glucosa en Ayunas (AGA)

Count of AGA	
AGA	**Total**
FALSE	9313
TRUE	1
(blank)	
Grand Total	**9314**

Como ejemplo de las 31 gráficas obtenidas durante el “Reporte de la exploración de los datos” se muestra la Figura 7 que representa la distribución de valores de campo Índice de Masa Corporal (IMC) agrupado según las normas establecidas por la Organización Mundial de la Salud (OMS) [20].

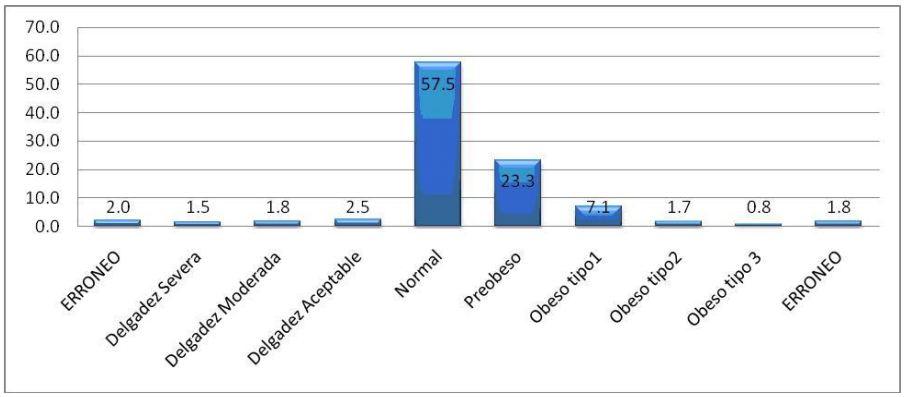

Figura 7 - Distribución de valores para el campo IMC. El valor que se observa en cada una de las barras representa el por ciento de entrevistados que está en esa categoría.

Los análisis exploratorios, permiten además, identificar relaciones entre los datos, que formulan las primeras hipótesis sobre relaciones entre ellos y posible conocimiento a obtenerse en la fase de Modelado de Datos. A modo de ejemplo se exponen algunas relaciones que se aprecian mediante la Suite para realizar procesos de KDD: WEKA 3.5.8.

Una de las preguntas que las personas frecuentemente se hacen es: ¿fumar puede provocar diabetes? Es por lo tanto uno de los principales atributos a tener en cuenta en nuestros análisis. La Figura 8 muestra una gráfica donde se aprecia que (hasta el punto que en está la investigación), no hay relación entre haber fumado y padecer de diabetes. Puede observarse que no hay predominio de ningún color para los posibles valores del atributo Fumar (true, false).

Tampoco se observa diferencia significativa para el atributo Ha Fumado; esta conclusión pudiera parecer trivial luego del análisis previo, sin embargo es interesante contrastar posibles diferencias entre los fumadores activos y aquellos que han dejado de serlo: Figura 9.

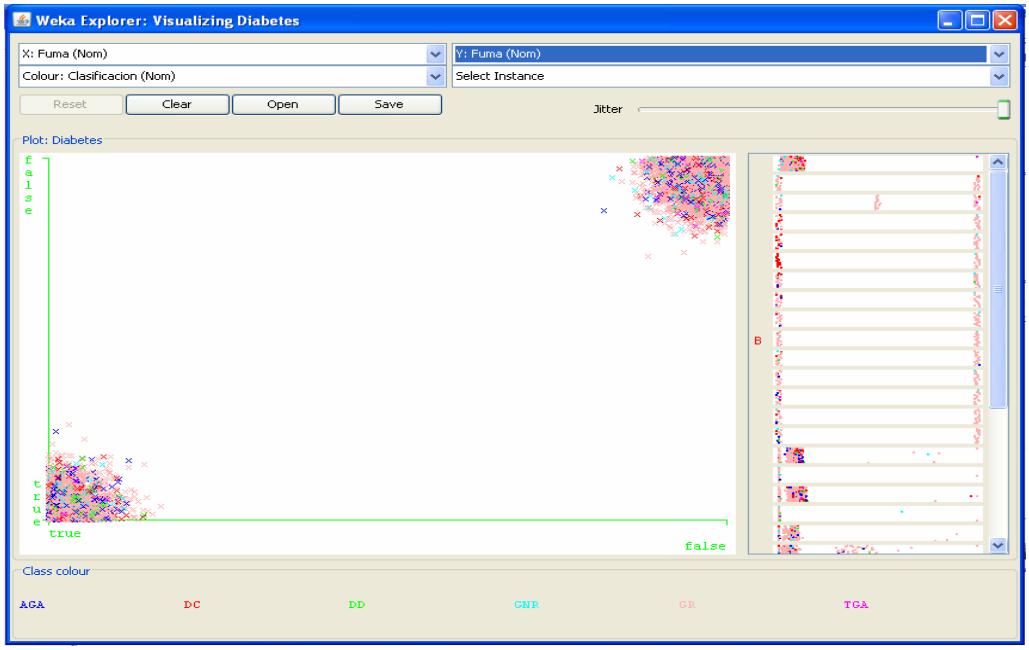

Figura 8 - Relación entre ser fumador activo y ser clasificado por los especialistas médicos en alguna de las siguientes categorías: Diabético conocido (DC), Diabético detectado (DD), Grupo de no riesgo (GNR), Grupo de riesgo (GR), Tolerancia a la glucosa alterada (TGA), o Alteración de la Glucosa en Ayunas (AGA).

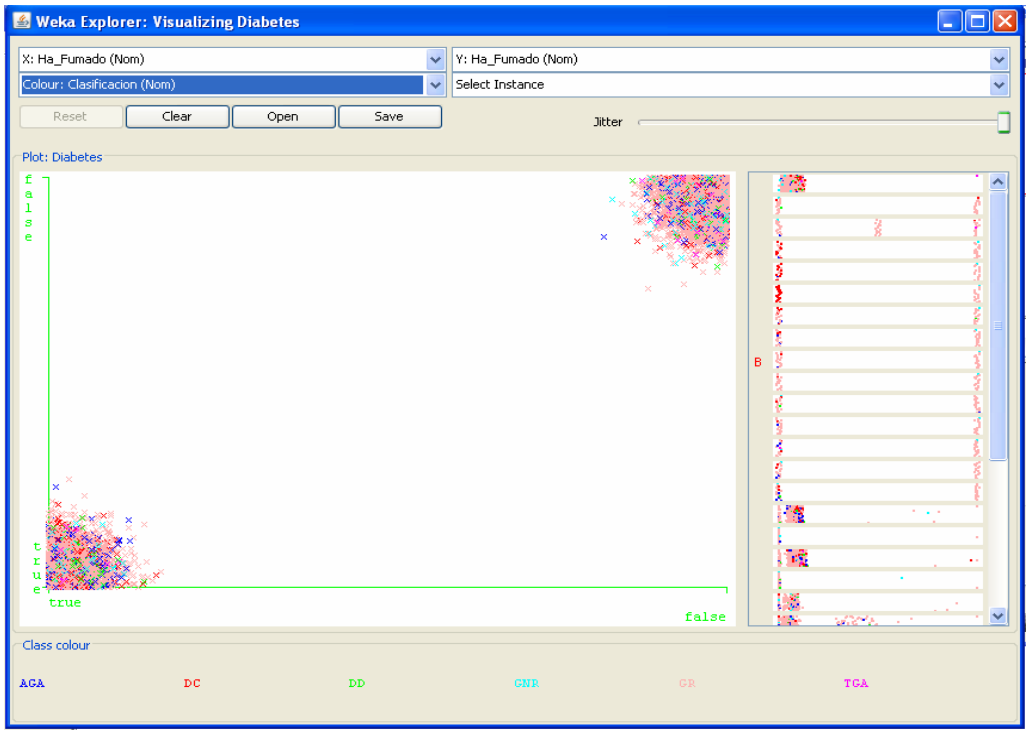

Figura 9 - Relación entre haber sido fumador y ser clasificado por los especialistas médicos en alguna de las siguientes categorías: Diabético conocido (DC), Diabético detectado (DD), Grupo de no riesgo (GNR), Grupo de riesgo (GR), Tolerancia a la glucosa alterada (TGA), o Alteración de la Glucosa en Ayunas (AGA).

Otra hipótesis producto de los análisis exploratorios (que tiene amplio aval por parte de los especialistas) es que padecer diabetes está estrechamente relacionado con la cantidad de azúcar en sangre y en orina que posea el paciente. Las figuras 10 y 11 reflejan cómo se comporta la variable objetivo para los atributos citados (los posibles valores para los atributos azúcar en sangre y azúcar en orina son true o false, o sea: si se le detectó o no).

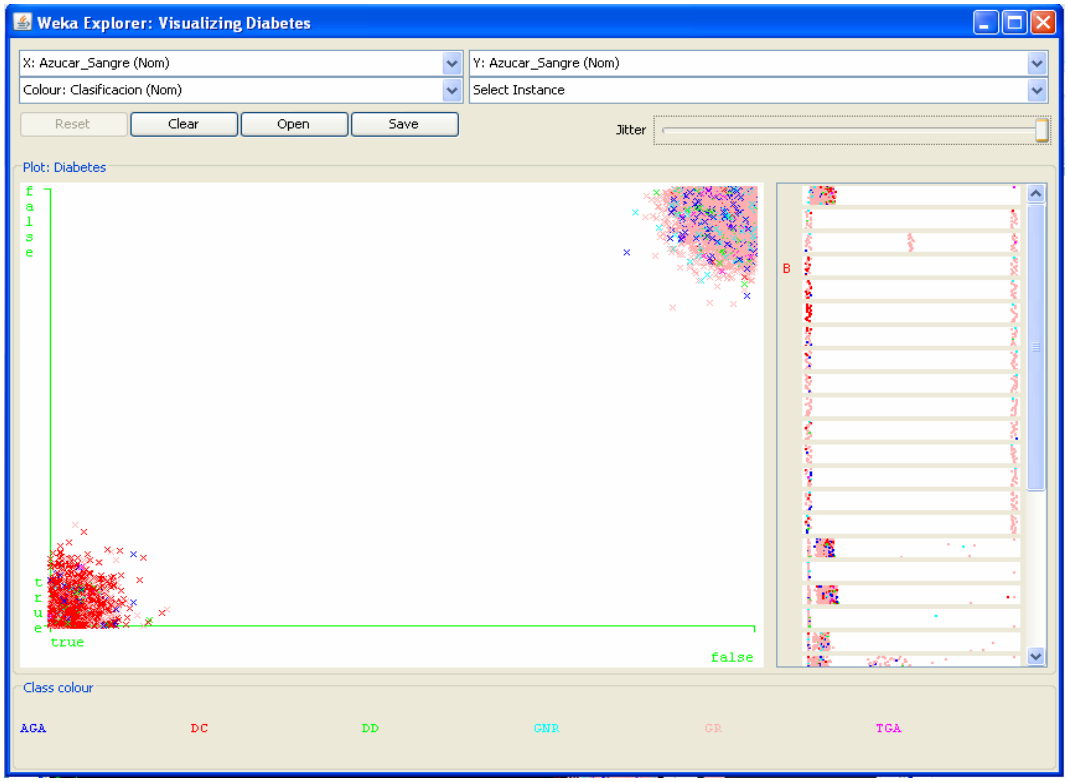

Figura 10 - Relación entre poseer niveles elevados de azúcar en sangre y ser clasificado por los especialistas médicos en alguna de las siguientes categorías: Diabético conocido (DC), Diabético detectado (DD), Grupo de no riesgo (GNR), Grupo de riesgo (GR), Tolerancia a la glucosa alterada (TGA), o Alteración de la Glucosa en Ayunas (AGA).

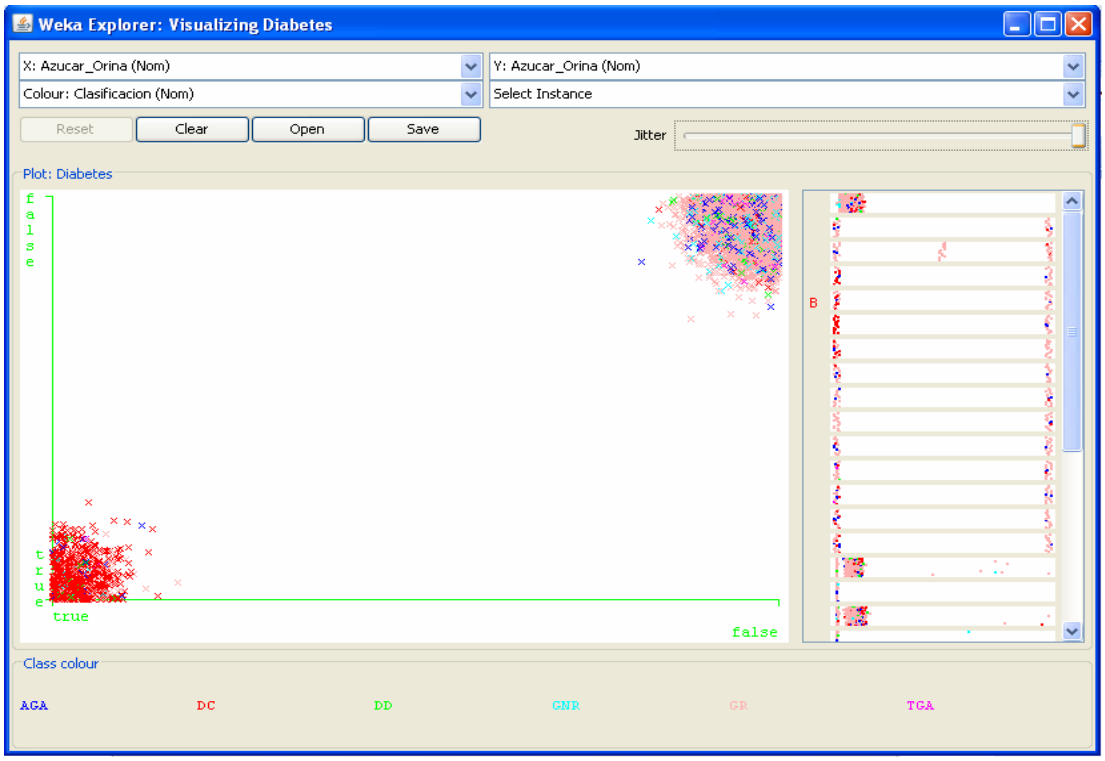

Figura 11 - Relación entre poseer niveles elevados de azúcar en orina y ser clasificado por los especialistas médicos en alguna de las siguientes categorías: Diabético conocido (DC), Diabético detectado (DD), Grupo de no riesgo (GNR), Grupo de riesgo (GR), Tolerancia a la glucosa alterada (TGA), o Alteración de la Glucosa en Ayunas (AGA).

Puede observarse que si se detecta azúcar en orina o en sangre existe una alta probabilidad de ser clasificado como Diabético Conocido. Puede incluso inferirse (producto de la representación gráfica) que el poseer el indicador como verdadero para el atributo Azúcar en Orina es más riesgoso que el poseerlo para el atributo Azúcar en Sangre.

Verificación de la calidad de los datos

El objetivo fundamental de esta tarea es identificar los problemas en la calidad de los datos. Como resultado de esta etapa se obtuvo el "Reporte de la calidad de los datos", donde se reflejaron los problemas de calidad identificados y las soluciones o medidas a tomar en el caso de los campos e incluso tuplas (entrevistados específicos) que presentasen problemas. Como resultado, se decidió eliminar el campo "Familiares diabéticos" que A Priori resultaba de gran interés pero que reflejó insalvables problemas de calidad. Siempre se ha dicho que la herencia juega un papel fundamental para determinar ciertos tipos de padecimiento, en la diabetes en particular, la experiencia de muchos expertos apunta en esta dirección. La Figura 12 muestra la distribución de valores para el citado atributo. Puede observarse la ocurrencia de valores erróneos como: *n* o *nb* por ejemplo; pero este tipo de problema no es el más grave pues dado el poco porciento que representan se pueden eliminar sin mayores consecuencias. El principal problema se encuentra en el solapamiento entre valores, y en la poca ocurrencia de algunas instancias que resulta sospechosa. ¿Cómo es posible que sólo uno de los entrevistados tenga precedencia de abuelo diabético?; el sentido común indica que esto no es lógico. No queda claro tampoco que sean excluyentes las categorías, por ejemplo: MADRE + HERMANA y, HERMANOS o MADRE.

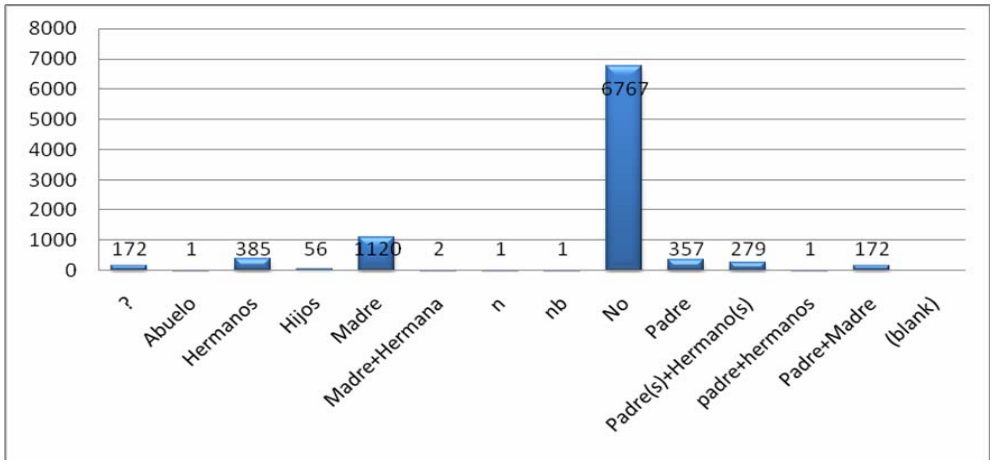

Figura 12- Distribución de valores para el atributo Familiares Diabéticos. En las barras se observa la cantidad neta de individuos pertenecientes a cada posible valor.

Otro resultado relevante fue la constatación de que el resto de los atributos presentaban una calidad excelente (ninguno sobrepasa el 5 % de valores erróneos entre valores nulos y fuera de rango). La medida adoptada fue por lo tanto eliminar las tuplas que presentasen problemas en la calidad.

La Figura 13 muestra la distribución de valores para el campo edad (en grupos de cinco años), se puede observar que a pesar de la existencia de valores anómalos (se consideran estrictamente anómalos los valores por encima de 100 y por debajo de 5), los valores se mueven en un rango lógico. Sin embargo, se puede observar además que un ínfimo porciento de personas tiene edades por debajo de 15 años y por encima de 90. Para la investigación actual se propone eliminar todos aquellos que estén fuera del citado rango (15-90) pues los menores de 15 son niños; que no poseen iguales características que los adultos, y es sabido que la Diabetes Mellitus es una enfermedad cuyo riesgo de padecimiento aumenta con la edad. Por otra parte, las personas mayores de 90 años de edad son muy pocas afortunadas y los modelos que se obtengan para ellos no tienen valor para la mayoría de la población. Con estos cambios se pierde sólo el 2.27% de los datos. Obsérvese como dato curioso la perfecta distribución normal para los valores y *la cola* más larga a la derecha, reflejo de que incluso durante las entrevistas no se le dio peso a la población infantil (pues es de suponer que exista un mayor número de personas entre los cero y 15 años de edad que mayores de 80 años).

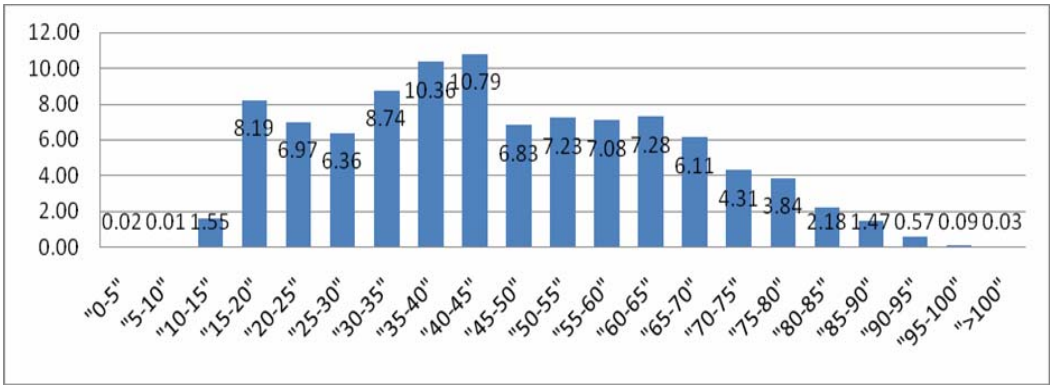

Figura 13 - Distribución del atributo edad del paciente (en grupos de cinco años). El valor representado en las barras representa el por-ciento de individuos pertenecientes a cada posible valor.

Discusión

Al concluir esta iteración del pre-procesado de datos, se tiene una *vista minable* para enfrentar las próximas etapas de CRISP-DM con 9314 registros y 29 campos: Edad, Sexo, Piel,

Estatura, Peso, Índice masa corporal, Circunferencia de Cintura, Azúcar en Sangre, Azúcar en Orina, Diagnosticado Diabético, Síntomas de Diabetes, Antecedentes Cardiopatías y/o Cerebro-Vasculares, Ha fumado, Fuma, Ha consumido Bebidas Alcohólicas, Alcohólico de riesgo, Ha tenido colesterol o triglicéridos altos, Ha padecido de presión alta o hipertensión, Tiene tratamiento con hipotensores, Tensión Arterial Sistólica, Tensión Arterial Diastólica, Tensión Arterial Promedio Sistólica, Tensión Arterial Promedio Diastólica, HDL-Colesterol, Triglicéridos, Glicemia Capilar Ayuna, PTG Ayuna, PTG 2 Horas, Clasificación.

Se considera que los resultados alcanzados son alentadores pues la calidad de los datos es buena y se tiene un número considerable de atributos para el análisis. Se pudo establecer las primeras hipótesis sobre relaciones en los datos y tener una vista detallada de los posibles valores por cada campo y sus distribuciones.

Con esta *vista minable* se debe seguir a la próxima fase de la metodología "Modelado de datos" en aras de concretar resultados y obtener las primeras reglas y relaciones explicitas entre estos; producto de la aplicación de técnicas y algoritmos matemáticos. El primer objetivo a ser abordado es determinar la relación entre cada uno de los 28 atributos "independientes" y el atributo objetivo "Clasificación del paciente", con el fin de encontrar modelos que puedan ser valorados y posteriormente utilizados por médicos y especialistas en este tipo de padecimiento. En un futuro, y en dependencia de los resultados que se obtengan, deberá valorarse la aplicación de las técnicas expuestas a otros conjuntos de datos (relativos a otros grupos de pacientes) y la inclusión de nuevos atributos que permitan una predicción más exacta.

La fase que se concluye no es para nada definitiva, pues como bien se pudo observar en la Figura 5, se trata de un proceso iterativo. Este proceso incluye la periódica evaluación de los resultados y la constante *vuelta atrás* en el desarrollo de cada fase.

Conclusiones

El trabajo permitió identificar y describir las características principales de los datos a emplear para clasificar a los pacientes de la localidad de Jaruco. Se logró identificar los atributos relevantes para la investigación, así como los que pueden servir de aval o apoyo a esta. El análisis exploratorio de los datos permitió conocer sus características (distribución, media, valores más frecuentes) lo que resulta de gran valor para comprender el significado de los modelos que se obtengan posteriormente. Se identificaron los problemas de calidad de los datos y se tomaron medidas para tratarlos. Se puede dar por concluida la "Comprensión de los datos" y se puede seguir con las otras fases de la metodología (CRISP-DM 1.0).

Agradecimientos

A los que creyeron en el proyecto y en las nuevas tecnologías. A los encuestadores que hicieron el, muchas veces injustamente subestimado, trabajo de base.

Referencias

[1] American Center for Chronic Disease Prevention and Health Promotion, "Diabetes Statistics and Research". [En línea]. 2008 Disponible en: URL: http://www.cdc.gov/diabetes/faq/research.htm. [Consultado: 5 de septiembre del 2008].

[2] International Diabetes Federation. "Diabetes Statistics". [En línea]. 2008. Disponible en: URL: http://www.idf.org/home/index.cfm?node=37. [Consultado: 5 de septiembre del 2008].

[3] Oramas, J. " La diabetes no anda sola". [En línea]. 2005 Disponible en: URL: http://cubahora.co.cu/index.php?tpl=columnas/ojosvistas/share-tpls/ver-ot.tpl.html&newsid_obj_id=1010158. [Consultado: 5 de septiembre del 2008].

[4] Scobie, I.N., Atlas of Diabetes Mellitus, 3ed. UK: Informa Healthcare, 2008. ISBN-10: 0-415-37649-1.

[5] Díaz Díaz, O. (Marzo 2008). El problema de la diabetes en Cuba. [En línea]. 2008. Disponible en: http://www.sld.cu/galerias/pdf/sitios/diabetes/epidemiologia_de_la_diabetes_en_cuba.pdf. [Consultado: 10 de septiembre del 2008].

[6] Hernández Orallo J, Ramírez Quintana M J, Ferri Ramírez C. Introducción a la minería de datos. Madrid, Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación: Ed. PEARSON EDUCACIÓN, S.A, 2004. ISBN: 84-205-4091-9.

[7] Servente M. Algoritmos TDIDT aplicados a la Minería de Datos Inteligente. Prof. Dr. Ramón García Martínez (Dir.). Universidad de Buenos Aires, Facultad de Ingeniería. Tesis de Grado en Ingeniería Informática, 2002.

[8] KDnuggets.Polls.Data Mining Applications in 2008. [En línea]. 2008. Disponible en: URL: http://www.kdnuggets.com/polls/2008/data-mining-applications.htm. [Consultado: 24 de septiembre de 2009].

[9] KDnuggets Polls. Data Mining Applications - Industries. [En línea]. 2007. Disponible en: URL: http://www.kdnuggets.com/polls/2007/data_mining_applications.htm. [Consultado: 10 de junio del 2007].

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

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

[12] KDnuggets Polls. Data Mining Methodology. [En línea]. 2007. Disponible en: URL: [Consultado: 10 de agosto del 2007].

[13] Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. 3ra Edición. USA: Two Cows Corporation, 2005. ISBN: 1-892095-02-5.

[14] Tang Zhao Hui, MacLennan Jamie. Data Mining with SQL Server 2005. Wiley, 2005. ISBN-13: 978-0471462613.

[15] SPSS, Statistical Product and Service Solutions. SPSS Clementine. [Software]. SPSS, 2007. Disponible en: URL: http://www.spss.com/clementine/index.htm. [Consultado: 5 de marzo del 2008].

[16] SAS Institute (Statistical Analysis Systems). SAS Enterprise Miner. [Software]. SAS, 2007. Disponible en: URL: http://www.sas.com/technologies/analytics/datamining/miner/. [Consultado: 5 de marzo del 2008].

[17] WEKA, Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java. [Software]. Weka 3.4.7, 2005. Disponible en: URL: http://www.cs.waikato.ac.nz/ml/weka/. [Consultado: 5 de marzo del 2008].

[18] Walkenbach John. Excel® 2007 Bible. Indiana, Indianapolis, Wiley, 2007. ISBN-13: 978-0-470-04403-2

[19] KDnuggets Polls. Data Mining Tools Used (May 2009). [En línea]. 2009 Disponible en: URL: http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm. [Consultado: 15 de febrero del 2007].

[20] World Health Organization. Global Database on Body Mass Index. [En línea]. 2008 Disponible en: URL: http://www.who.int/bmi/index.jsp?introPage=intro_3.html.[Consultado: 10 de junio del 2008].

Dirección para correspondencia

Dr. Alejandro Rosete Suárez.
Email: rosete@ceis.cujae.edu.cu
Postal: Calle Lugareño 52, e/ Luaces y Almendares, Plaza de la Revolución, Ciudad de La Habana, Cuba.

Anexo1: Tabla Resumen de la Descripción de los datos.

#	Identificador del Campo	Característica reflejada	Tipo del campo	Importancia para la investigación
1	Edad	Edad del paciente	Numérico	Relevante
2	Sexo	Sexo del paciente	Nominal	Relevante
3	Piel	Color de la piel del paciente	Nominal	Relevante

4	Azúcar en Sangre	Refleja si al paciente se le detectó azúcar en sangre.	Booleano	Relevante
5	Azúcar en Orina	Refleja si al paciente se le detectó azúcar en orina.	Booleano	Relevante
6	Diagnosticado Diabético	Refleja si el paciente, previamente, ha sido diagnosticado como diabético	Booleano	Relevante
7	Familiares Diabéticos	Si tiene o no, familiares diabéticos. En caso de tenerlos se refleja el parentesco de estos con el paciente.	Nominal	Relevante
8	Síntomas de Diabetes	Refleja si el paciente presenta síntomas evidentes de diabetes.	Booleano	Relevante
9	Antecedentes Cardiopatías y/o Cerebro-Vasculares	Refleja si el paciente presenta antecedentes de (Cardiopatías y/o Cerebro-Vasculares).	Booleano	Relevante
10	Ha fumado	Refleja si el paciente ha fumado	Booleano	Relevante
11	Fuma	Refleja si el paciente fuma en la actualidad	Booleano	Relevante
12	Ha consumido Bebidas Alcohólicas	Refleja si el paciente ha consumido bebidas alcohólicas	Booleano	Relevante
13	Alcohólico de riesgo	Refleja si el paciente es o no, un alcohólico de riesgo.	Booleano	Relevante
14	Ha tenido colesterol o triglicéridos altos	Refleja si el paciente ha presentado colesterol o triglicéridos altos	Booleano	Relevante
15	Ha padecido de presión alta o hipertensión	Refleja si el paciente ha padecido de presión alta o hipertensión	Booleano	Relevante
16	Tiene tratamiento con hipotensores	Refleja si el paciente está bajo tratamiento con hipotensores	Booleano	Relevante
17	Lectura Tensión Arterial 1 Sistólica	Valor de la lectura para tensión arterial sistólica.	Numérico	Relevante
18	Lectura Tensión Arterial 1 Diastólica	Valor de la lectura para tensión arterial diastólica	Numérico	Relevante
19	Lectura Tensión Arterial Promedio Sistólica	Valor promedio de la lectura para tensión arterial sistólica	Numérico	Relevante
20	Lectura Tensión Arterial Promedio Diastólica	Valor promedio de la lectura para tensión arterial diastólica	Numérico	Relevante
21	Estatura	Refleja la estatura del paciente	Numérico	Relevante
22	Peso	Refleja el peso del paciente	Numérico	Relevante
23	Índice masa corporal	El índice de masa corporal del paciente se calcula como la razón entre el peso de la persona (en **Kg**) y el cuadrado de la estatura (en $\mathbf{m^2}$). Da una medida de la cantidad de grasa corporal de una persona	Numérico	Relevante
24	Circunferencia de Cintura	Refleja la medida de la circunferencia de la cintura del paciente (en **cm**)	Numérico	Relevante
25	HDL-Colesterol	Refleja la medida de la cantidad de colesterol del tipo HDL del paciente.	Numérico	Relevante
26	Triglicéridos	Refleja la medida de los triglicéridos del paciente.	Numérico	Relevante
27	Glicemia Capilar Ayuna	Refleja la medida de la glicemia capilar del paciente en ayunas.	Numérico	Relevante
28	AGA	Refleja si el paciente presienta o no, Alteración de Glucosa en Ayunas.	Booleano	Relevante
29	PTG Ayuna	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa en Ayunas.	Numérico	Relevante
30	PTG 2 Horas	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa 2 horas después de ser alimentado.	Numérico	Relevante
31	Clasificación	Refleja la clasificación asignada al paciente producto del análisis médico.	Nominal	Relevante

32	Área Geográfica	Refiere si el paciente es de la zona Urbana o Rural	Nominal	Interesante
33	Consejo Popular	Refiere el Consejo Popular al que pertenece el paciente.	Nominal	Interesante
34	Circunscripción	Refiere la Circunscripción a la que pertenece el paciente.	Numérico	Interesante
35	Área-Salud	Refiere el Área de Salud para la atención al paciente.	Numérico	Interesante
36	Consultorio	Refiere el Consultorio de Salud que atiende al paciente	Numérico	Interesante
37	Entrevistador	Refiere el nombre y apellidos de la persona que realizó la entrevista	Nominal	Interesante
38	Nivel Escolar	Refiere el nivel escolar del paciente del paciente	Nominal	Interesante
39	Ocupación actual	Refiere la ocupación actual del paciente	Nominal	Interesante
40	Estado Conyugal	Refiere el estado conyugal del paciente.	Nominal	Interesante
41	Antecedentes Obstétricos	Refiere si el paciente presenta antecedentes obstétricos.	Nominal	Interesante
42	Ha tenido Diabetes durante algún embarazo	Se refleja si el paciente, en caso de ser femenino, ha tenido diabetes durante el embarazo.	Booleano	Interesante
43	Fecha Ultima Menstruación	Refiere la fecha de la última menstruación de los pacientes femeninos	DATE	Interesante
44	Antecedentes Bajo Peso al Nacer	Refiere si el paciente presentó bajo peso al nacer.	Booleano	Interesante
45	Ha sentido que debe beber menos	Refleja si el paciente ha sentido la necesidad de beber menos	Booleano	Interesante
46	Le ha molestado que le critiquen su forma de beber	Refleja si al paciente le ha molestado que le critiquen su forma de beber	Booleano	Interesante
47	Se ha sentido molesto o culpable por su forma de beber	Refleja si el paciente se ha sentido molesto o culpable por su forma de beber	Booleano	Interesante
48	Ha tomado algún trago por la mañana para la resaca	Refleja si el paciente ha tomado algún trago en la mañana para la resaca.	Booleano	Interesante
49	En la última semana cuantos días realizó actividades vigorosas	Refleja cuantos días de la última semana antes de la encuesta el paciente realizó actividades vigorosas.	Numérico	Interesante
50	Cuanto tiempo en total empleo en actividades vigorosas	Refleja el tiempo total empleado por el paciente en actividades vigorosas.	Numérico	Interesante
51	En la última semana cuantos días realizó actividades moderadas	Refleja cuantos días de la última semana antes de la encuesta el paciente realizó actividades moderadas.	Numérico	Interesante
52	Cuanto tiempo en total empleó en actividades moderadas	Refleja el tiempo total empleado por el paciente en actividades moderadas.	Numérico	Interesante
53	En la última semana cuantos días caminó al menos 10 min seguidos	Refleja cuantos días de la última semana antes de la encuesta el paciente caminó al menos 10 minutos seguidos.	Numérico	Interesante
54	Cuanto tiempo en total caminó en esos días	Refleja el tiempo total que caminó el paciente en la semana anterior a la encuesta.	Numérico	Interesante
55	En la última semana cuanto tiempo paso sentado en días hábiles	Refleja cuanto tiempo de la última semana el paciente pasó sentado durante días hábiles.	Numérico	Interesante
56	Realizó alguna actividad física durante 10 min seguidos	Refleja si el paciente realizó actividad física durante 10 minutos seguidos.	Booleano	Interesante
57	Id	Refiere un identificador de cada paciente	Numérico	Sin Importancia

58	Fecha	Refiere la fecha en que se registra el dato del paciente	DATE	Sin Importancia
59	Provincia	Refiere la provincia de residencia del paciente	Nominal	Sin Importancia
60	Municipio	Refiere el municipio de residencia del paciente	Nominal	Sin Importancia
61	Nombre y Apellido	Refiere el nombre y los apellidos del paciente	Nominal	Sin Importancia
62	CI	Refiere el carnet de identidad del paciente	Nominal	Sin Importancia
63	Dirección	Refiere la dirección de residencia del paciente	Nominal	Sin Importancia