Universidad de Cienfuegos “Carlos Rafael Rodríguez”

Facultad de Ingeniería

Trabajo de Diploma para optar por el título de Ingeniero Informático

“Conjunto de entrenamiento útil a sistema experto para la prevención y diagnóstico de diabetes tipo 2.”

AUTORA:

Thays Burke Mariño

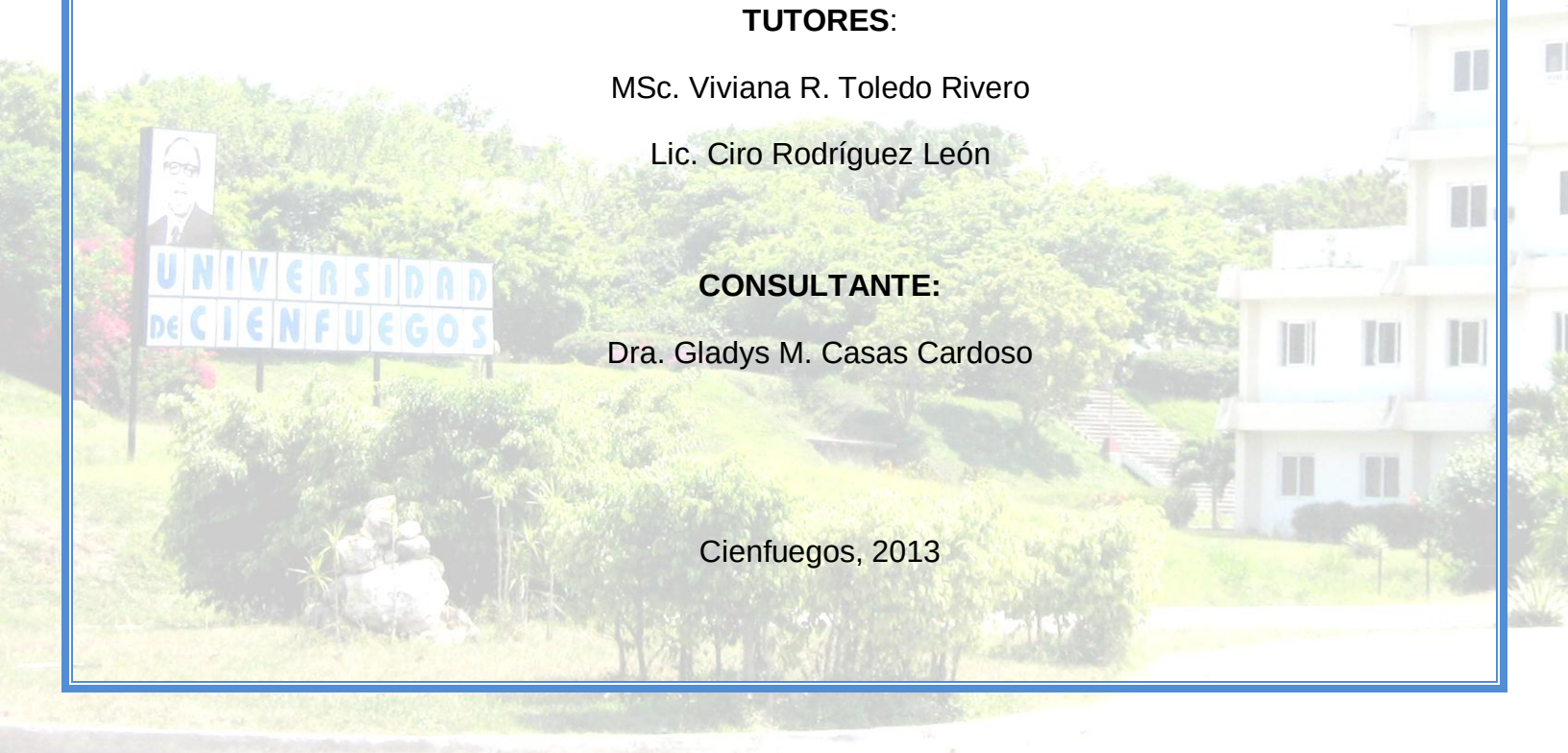
TUTORES:

MSc. Viviana R. Toledo Rivero

Lic. Ciro Rodríguez León

CONSULTANTE:

Dra. Gladys M. Casas Cardoso



Cienfuegos, 2013

Declaración de autoría

Yo, Thays Burke Mariño, declaro que soy la única autora del trabajo de diploma titulado "Conjunto de entrenamiento útil a sistema experto para la prevención y diagnóstico de diabetes tipo 2." y autorizo al Departamento de Informática de la Facultad de Ingeniería en la Universidad de Cienfuegos "Carlos Rafael Rodríguez", para que haga el uso que estime pertinente con el mismo.
Para que así conste firmo la presente a los 10 días del mes de junio del 2013.

_____

Firma de la Autora

Thays Burke Mariño

Los firmantes abajo certificamos que el presente trabajo ha sido revisado según acuerdo de la dirección de nuestro centro y el mismo cumple los requisitos que debe tener un trabajo de esta envergadura referente a la temática señalada.

_____	_____
Firma de la Tutora	Firma del Tutor
MSc. Viviana R. Toledo Rivero	Lic. Ciro Rodríguez León
_____	_____
Firma ICT	Firma Vicedecano

‘‘No se alabe el sabio en su sabiduría, ni en su valentía se alabe el valiente, ni el rico se alabe en sus riquezas, mas alábese en esto el que se hubiere de alabar: en entenderme y conocerme, que yo soy Jehová, que hago misericordia, juicio y justicia en la tierra.’’

La Biblia, Jeremías 9.23

Agradecimientos

- *A Tí, que eres Dios y Rey del mundo creado. Mi Señor y Amigo, por tener este tiempo planeado; por acompañarme y guiarme; por fortalecerme y proveer: te agradezco.*
- *A tí, mi amado esposo, mi colaborador, por tu paciencia y esfuerzo; por tu amor y orientación siempre acertada: te agradezco.*
- *A ustedes, mis padres, por apoyarme y hacerme saber la incondicionalidad de su amor. A tí mamá por tu ayuda incansable. A tí papá porque sabes animarme: les agradezco.*
- *A ustedes Ciro y Viviana, mis tutores, por dedicarme su tiempo, intelecto y esfuerzo; por acomodar el "pastel" y ayudarme a salir adelante: les agradezco.*
- *A ustedes mis hermanos y amigos, por su apoyo en oración y por su preocupación; por dejarse usar por Dios para animarme en momentos claves: les agradezco.*
- *A mi familia, y la de Alain que ya es mía, porque fue suficiente una palabra de auxilio para demostrarme su disposición: les agradezco.*
- *A Yailem, por tu disposición y por tus buenas observaciones de los detalles: te agradezco.*
- *A Gladys, por tu oportuna aparición, por dedicar tiempo a un proyecto ajeno: te agradezco.*
- *A mis compañeros de aula, a ustedes que me permitieron conocerlos más de cerca: les agradezco.*
- *A mis profesores, a muchos de ustedes les llevo en mi corazón, por su amor y dedicación; por cómo nos consideraron y ayudaron: les agradezco.*
- *Al personal del CAED, por su ayuda; por hacer espacio en su ocupada agenda para atenderme; por su paciencia con mi ignorancia médica: les agradezco.*
- *A ustedes, que se han preocupado por la tesis y que a veces indirectamente, han puesto su granito de arena: les agradezco.*

Resumen

En la investigación “Conjunto de entrenamiento útil a sistema experto para la prevención y diagnóstico de diabetes tipo 2” se desarrolla minería de datos, utilizando CRISP-DM como metodología, en el grupo de pacientes que ingresaron en el Centro de Atención y Educación en Diabetes de Cienfuegos del 2005 al 2012. Se aplican técnicas de clustering para identificar clases o grupos de diabéticos que conformen el conjunto de entrenamiento para un sistema experto útil a la prevención y el diagnóstico de la diabetes mellitus tipo 2.

La investigación resulta relevante dado que la diabetes mellitus tipo 2 es un problema de salud tanto para el mundo desarrollado como para el subdesarrollado y en Cuba es creciente el número de pacientes con la enfermedad cada año, lo que justifica los esfuerzos a nivel del Ministerio de Salud Pública para su tratamiento y prevención a niveles primarios de atención.

Los resultados, comprobados por los índices de validación interna y el criterio experto, demuestran que los algoritmos de clustering “Expectation Maximization” y “Conglomerado en dos fases” son apropiados a los propósitos del trabajo. Se obtiene un conjunto de datos con tres clases para los hombres y otro con cuatro clases para las mujeres, que interpretados como niveles de riesgo de complicación, asistirán la construcción de un sistema basado en el conocimiento para elevar la eficiencia del diagnóstico, al tiempo que medidas preventivas en pacientes con riesgo de padecer diabetes mellitus tipo 2.

Índice

Índice de figuras

Índice de tablas

Introducción

Durante la segunda mitad del siglo XX, gran parte de los avances de la ciencia han sido consecuencia del desarrollo de los conocimientos y tecnologías que permiten el estudio y la manipulación de las moléculas de la vida (Bioquímica y Biología Molecular), y de aquellas que facilitan la interpretación e integración de datos, y la ejecución de procesos a velocidad mucho mayor que la propia acción humana (Ciencias de la Computación) [1].

Ambas áreas de la ciencia se han desarrollado simultáneamente. Así en 1956, Arthur Kornberg sintetizó por primera vez ADN in vitro a partir de nucleótidos, mientras John Backus inventaba el primer lenguaje de programación de computadoras (FORTRAN). Tres años más tarde mientras Severo Ochoa y Kornberg recibían el premio Nobel por la biosíntesis de ácidos nucleicos Grace Murray Hopper inventaba el lenguaje COBOL. En 1965, sólo seis años más tarde Kemeny y Kurtz desarrollaban el BASIC y se hablaba ya de sistemas expertos, los cuales proceden de la inteligencia artificial.

Los primeros pasos en Inteligencia Artificial (IA) se dieron en los años 50. A comienzos de los años 50 el conocido A.M.Turing publicó su "Computing Machinery and Intelligence" y a partir de entonces, aparecen varias definiciones de lo que significaba la inteligencia en una máquina. La inteligencia artificial es una subdivisión de las ciencias de la computación dedicada a crear software y hardware para computadoras que imitan la mente humana [2]. Su principal objetivo es hacer las computadoras más inteligentes, creando softwares que permitan a una computadora imitar algunas de las funciones del cerebro en áreas de aplicación seleccionada. La idea no es reemplazar a los seres humanos sino proveerlos de una poderosa herramienta para asistirlos en su trabajo.

Un área en la que se evidencia la limitación humana y la efectividad de las técnicas inteligentes es la deducción de nuevo conocimiento a partir de grandes volúmenes de información. Es obvia la incapacidad del hombre de procesar y extraer nueva información de grandes cantidades de datos, mientras que una importante aplicación de la inteligencia artificial es la minería de datos. La Minería de Datos (MD) es definida como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [3]. Esta surge por el reconocimiento de un nuevo potencial: el valor de la gran cantidad de datos

almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares [4].

Un área en la que actualmente se almacena gran cantidad de información sobre los pacientes y donde ha demostrado su utilidad la minería de datos es la medicina.

La Medicina, tan antigua como el hombre mismo, se ha visto beneficiada y enriquecida por el surgimiento de otras ciencias como la Cibernética. Si vieja es una, joven es la otra, que al decir de Wiener, la cibernética es, "la ciencia sobre los rasgos generales de los procesos y sistemas de mando en los dispositivos técnicos, los organismos vivos y las organizaciones humanas" [5]. La cibernética constituye en esencia una de las principales bases de la revolución tecnológica que vive el mundo de nuestros días, representa un logro de la ciencia actual, es una rica fuente de ideas nuevas que han ayudado a la cosmovisión científica actual.

Nuestro tiempo se caracteriza por la diferenciación e integración de las ciencias exactas, las ciencias naturales y las ciencias sociales. La medicina actual dentro de su desarrollo no ha escapado a este fenómeno y junto a la cibernética, y en particular la Computación, y como condición necesaria, su vínculo estrecho con la Ciencia de la Información (Informática), han dado lugar a la Informática Médica, que agrupa los campos del software y el hardware para su uso en la medicina. Han devenido en ciencias integradas, vinculadas muy estrechamente por lazos que cada día son más fuertes.

Las condiciones actuales, el desarrollo científico-tecnológico, la interrelación con otras ciencias y sus métodos, el modo de vida de una sociedad altamente desarrollada y muchos más factores, han cambiado cualitativamente la problemática de la medicina teórica. Han surgido nuevos fenómenos, nuevos problemas, como la actitud de la medicina ante otras ciencias (la matemática, la cibernética), el proceso de integración del conocimiento médico, la informatización de la sociedad, etc., que posibilitan un análisis más dialéctico del desarrollo de la ciencia de la medicina en el mundo actual. Las nuevas tecnologías surgidas dentro de la computación rápidamente han sido aplicadas a la medicina.

La inteligencia artificial cobra cada día más fuerza en el mundo, con el desarrollo de la Robótica, los Sistemas Expertos (SE) y más recientemente la Realidad Virtual y la

Minería de Datos. El hombre ve la posibilidad de dotar a las computadoras de cierta "inteligencia" para incorporarlas a disímiles, agotadoras y complejas tareas.

La minería de datos permite que los datos pasen de ser un "producto" a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento [6] Dado que la minería de datos excede la capacidad humana para el análisis de grandes volúmenes de datos, la utilización plena de los datos almacenados depende del uso de técnicas del descubrimiento del conocimiento [7], entre las que se encuentran las de IA.

Probablemente en pocos años, el uso de la MD se haya extendido a todas las actividades humanas complejas en las que interviene gran cantidad de datos y variables. Se podrán analizar y comprimir datos para nosotros, tomar decisiones de poca importancia y servir como medio de apoyo para decisiones complejas o de gran trascendencia.

Como un lógico proceso de desarrollo la Medicina ha ido asimilando la introducción de las computadoras para agilizar y mejorar los procesos de apoyo médico, teniendo una gran influencia, que aumenta cada día, la introducción de la inteligencia artificial en la vigilancia del paciente con complejos equipos biomédicos, realización de procesamiento voluminoso de información para la toma de decisiones y muchas otras aplicaciones.

Otros usos de las computadoras en este campo son las pruebas para detectar e identificar alteraciones, como por ejemplo, la Tomografía Axial Computarizada (TAC), la resonancia magnética, el ultrasonido, los análisis de electrocardiogramas por computadoras, análisis de imágenes y muchos más.

La integración ha permitido extender la aplicación de las computadoras a los servicios administrativos y de apoyo, la dirección, la investigación, el diagnóstico y el tratamiento, sin dejar de mencionar la educación. [8], [9], [10], [11], [12]

Es el diagnóstico, quizás, el más controvertido de los sectores de aplicación de las computadoras en la medicina, por las implicaciones éticas que puede traer. Se sabe que el diagnóstico médico es el arte de identificar una enfermedad por sus signos y síntomas. La introducción de computadoras para apoyar el diagnóstico ha planteado la interrogante: ¿Sustituirá la computadora al médico algún día? Según Ávila, uno de los problemas cuando se usan para el diagnóstico es que no toman en cuenta que una persona puede tener más de una enfermedad, que los síntomas pueden ser

independientes, o que el paciente puede estar fingiendo [13]. Él mismo responde a la interrogante planteando “Si bien es cierto que la computadora tiene gran capacidad de cálculo, velocidad y exactitud, está claro que una computadora no puede sustituir al médico. Sólo éste es capaz de razonar lógicamente y mezclar la razón con la intención, la ética, lo afectivo y la experiencia, algo que una máquina no puede hacer. No puede mantener el aspecto más importante: la relación médico-paciente”.

Desde el triunfo revolucionario de 1959 en Cuba, la salud pública constituye un objetivo primordial de este proceso. Desde entonces, nuestro estado se esfuerza por mantener una atención sanitaria a la altura de países desarrollados. La modernización del Sistema Nacional de Salud, y la construcción de modernos hospitales, han permitido la introducción de tecnologías de punta para servir de apoyo a la asistencia médica.

En los últimos años se han introducido el ultrasonido, la Tomografía Axial Computarizada y más recientemente la Resonancia Magnética Nuclear, todas, tecnologías de un elevadísimo costo, pero utilizadas en la salud de nuestro pueblo de forma gratuita. Varios centros de investigación dedican parte de su trabajo a crear equipos computarizados de apoyo a la actividad médica. Un ejemplo fehaciente de esto es el Instituto Central de Investigaciones Digitales (I.C.I.D), creador de un número importante de equipos de la más alta tecnología, utilizando para ello las computadoras: el CardioCid, el NeuroCid, el SUMA (Sistema Ultra Micro Analítico) utilizado en la detección del SIDA, por mencionar algunos, constituyen aportes significativos al Sistema Nacional de Salud.

No solo en el campo del hardware se han alcanzado avances, también vale mencionar productos de software para la investigación, como el Sistema Morfo-Estereológico Asistido por Computadoras con Digitalización de Imágenes (COMSDI-Plus), desarrollado en la Facultad de Ciencias Médicas de Holguín, con el cual se han realizado investigaciones histológicas y patológicas, hoy introducido en muchos centros del país. Dentro de los trabajos más relevantes realizados en la rama de la inteligencia artificial encontramos DIAG un SE para el diagnóstico de un grupo de anomalías craneofaciales, encontradas en la clínica, este sistema es una herramienta de diagnóstico para odontólogos, residentes y estomatólogos dedicados a la ortodoncia y puede también ser empleado como sistema tutor inteligente para el estudio de la ortodoncia, desarrollado en la Universidad de Ciego de Ávila en 1997. O el SISI

(Sistema Inteligente de Selección de Información), desarrollado por un grupo de investigadores de la Universidad de Las Villas. Este programa es una variante del Shell, diseñado por Stanfill y Waltz en 1986, y se ejecuta sobre el sistema operativo Windows 95. Su efectividad ha sido ampliamente reconocida, por presentar una interface amigable para el usuario, quien sólo necesita conocimientos mínimos para su utilización, y ya existen diferentes expertos que lo utilizan [14]. Varios trabajos en esta temática se han desarrollado en los últimos años en la UCLV [15], [16] y más recientemente se han desarrollado sistemas para el diagnóstico y tratamiento de enfermedades, como el embarazo eptópico [17], las infecciones de transmisión sexual [18] y el fibroma uterino [19] en los que han participado como expertos profesionales de nuestra provincia.

Por otra parte un gran número de centros asistenciales cuentan con actividades económicas y administrativas automatizadas y se trabaja intensamente para lograr un mayor nivel de automatización sin pasar por alto la labor docente, donde futuros especialistas en medicina y alumnos de postgrado reciben los conocimientos básicos para poder explotar sistemas de apoyo a su trabajo. Otro aspecto importante es la creación desde 1996 de la Red Telemática de Salud en Cuba (INFOMED), que permite la comunicación entre los centros de investigación, hospitales, policlínicos, y centros de información tanto dentro de nuestro país como con el resto del mundo, permitiendo el intercambio de información, y abriendo posibilidades a la realización de proyectos hasta hace poco inimaginables, propiciando el intercambio constante, la realización de trabajos colaborativos y la gestión de proyectos investigativos sin frontera.

No obstante todo lo logrado existen áreas de aplicación como la prevención y el diagnóstico temprano de la diabetes mellitus tipo 2, donde es posible resolver un problema real creando una nueva oportunidad aún insuficientemente explorada.

La diabetes mellitus tipo 2 es una enfermedad que se caracteriza por el aumento de los niveles de glucosa en la sangre. Las personas con diabetes tienen una esperanza de vida reducida y una mortalidad dos veces mayor que la población general. En el mundo, la diabetes afecta a 366 millones de personas, y se estima que para el 2030 el número de afectados ascenderá a 552 millones. Esta enfermedad es la cuarta causa de muerte a nivel mundial y se estima que al menos el 50% de las personas diabéticas ignoran que lo son.

Son varias las complicaciones de esta enfermedad. Entre ellas se encuentran: daños de los pequeños vasos sanguíneos, de los nervios periféricos y de la piel. El pie diabético, complicación que consiste en heridas difícilmente curables y la mala irrigación sanguínea de los pies, puede conducir a laceraciones y a la amputación de las extremidades inferiores. Daños de la retina, problemas renales y afectaciones de los vasos sanguíneos grandes son dificultades mayores producidas por la diabetes. Esta última conduce a infartos, apoplejías y trastornos de la circulación sanguínea en las piernas. El hígado graso, la hipertensión arterial, las cardiopatías y el coma diabético también afectan a estas personas, pudiendo este último llegar a ocasionar la muerte.

La Federación Internacional de Diabetes (IDF) reportó un gasto en 2011 de 465,000 millones de dólares [20]. Esto da la medida de lo costosa que resulta y la importancia de trabajar en la prevención para evitar el alto costo que tiene para la vida y la estabilidad de la familia, que se ve afectada en sus estadios más avanzados.

Cuba no es ajena a este problema de salud mundial. La prevalencia de esta enfermedad en la población cubana se ha acrecentado en los últimos 20 años y en el año 2000 existían 263 808 diabéticos dispensariados y se estimaba una cantidad similar sin diagnosticar para un total de 40 por cada 1000 habitantes.

Como parte de la respuesta a este problema, desde el año 1975 el Instituto Nacional de Endocrinología elaboró un Programa Nacional de Atención Integral al Diabético. Este Programa recibió una mayor prioridad en 1992 cuando se definieron los Objetivos, Prioridades y Directivas para el año 2000. [21]

Los esfuerzos se han dirigido desde entonces a la disminución de la mortalidad por diabetes, a reducir la frecuencia y severidad de las complicaciones agudas y crónicas y a mejorar la calidad de vida de los diabéticos. Además, a mejorar el conocimiento de la magnitud del problema en Cuba, desarrollar metodologías educativas para la población en general, disminuir los costos de esta enfermedad a la sociedad y apoyar investigaciones destinadas a la prevención y control de la diabetes mellitus. [21]

Los objetivos anteriores se cumplen mediante el desarrollo de determinadas actividades en los tres niveles de atención establecidos por el Ministerio de Salud Pública. Específicamente, el nivel primario es el encargado de gran parte de estas actividades por el gran peso de este en las acciones de promoción y prevención de salud que preconiza estilos de vida saludables, de prevención primaria así como de acciones de

detección de la enfermedad y de sus potenciales complicaciones agudas y crónicas. [21]

Una de las dificultades que se presentan en el proceso de prevención es que la diabetes mellitus tipo 2 en la mayoría de los casos (90%) se diagnostica entre 4 y 7 años después de la existencia de hiperglicemias en el paciente no diagnosticadas [21], lo que dificulta el diagnóstico de estos individuos en esta etapa "silente" de modo que se puedan ejecutar medidas de control.

En nuestro país, el conjunto de acciones deberá ir dirigido a lograr con un enfoque multisectorial la disminución de la vulnerabilidad de las personas o grupos aprovechando las indiscutibles ventajas de nuestra organización político social.

El grupo de trabajo gubernamental para el enfrentamiento y prevención de la diabetes mellitus tipo 2 viene trabajando activamente con un enfoque intersectorial.

Es obvio que es el sector salud el líder en el enfrentamiento a los problemas que genera la diabetes mellitus tipo 2 y para ello se requiere una cada vez mejor preparación científica de todos los profesionales y técnicos, pues sin dudas sólo mediante una preparación de excelencia es que se podrá enfrentar con éxito este colosal reto.

Es por ello que, teniendo en cuenta todo lo anterior, se identifica como **problema científico**: ¿Cómo crear un conjunto de entrenamiento capaz de contribuir a la construcción de un sistema experto útil a la prevención y diagnóstico de la diabetes mellitus tipo 2 en la atención primaria de salud de Cienfuegos?

En consecuencia el **objeto** de la presente investigación es: las técnicas de inteligencia artificial en la prevención de la diabetes mellitus tipo 2, y el **campo de acción** el agrupamiento usando técnicas de clustering en apoyo a la prevención de la diabetes mellitus tipo 2 en la provincia Cienfuegos

Como **objetivo** se plantea: Construir un conjunto de entrenamiento, utilizando técnicas de agrupamiento, que contribuya a la creación de un sistema experto de apoyo a la prevención y diagnóstico temprano de la diabetes mellitus tipo 2 en la provincia Cienfuegos.

Como **idea a defender** se plantea que con el uso de técnicas de agrupamiento puede crearse un conjunto de entrenamiento, capaz de contribuir a la creación de un sistema

experto que apoye la prevención y detección temprana de la diabetes mellitus tipo 2 en la población cienfueguera.

Para lograr esta investigación se realizaron las siguientes **tareas científicas**:

1. Revisión y análisis de la bibliografía contemporánea para caracterizar el estado actual de la problemática planteada tanto en Cuba como en el mundo.
2. Selección de las técnicas apropiadas.
3. Preprocesamiento de los datos.
4. Experimentación con las técnicas seleccionadas.
5. Validación de la conformación de los grupos.
6. Conclusiones y Recomendaciones.

Se aplicaron los siguientes **métodos**:

1. Inducción - deducción, con el objetivo de estructurar el conocimiento científico a partir de la revisión bibliográfica.
2. Histórico - comparativo, para conocer el problema estudiado en su origen y desarrollo; desde el punto de vista de la informática y de la medicina, así como del empleo de nuevas tecnologías como recurso válido para la prevención y diagnóstico temprano.
3. Análisis y síntesis, para poder establecer nexos, comparar resultados, determinar enfoques comunes y aspectos distintivos de los diferentes enfoques estudiados, lo que permite arribar a conclusiones.

La **novedad científica** de la investigación está dada en que propone un conjunto de entrenamiento útil a la construcción de un sistema experto que apoye la prevención y detección temprana de la diabetes mellitus tipo 2.

La tesis está **estructurada** en introducción, 3 capítulos, conclusiones y recomendaciones.

El primer capítulo, titulado *"Técnicas de la Inteligencia Artificial útiles a la prevención y diagnóstico precoz de enfermedades."*, nos introduce en las distintas técnicas de IA empleadas en el diagnóstico médico y en algunos aspectos teóricos de las empleadas en la investigación.

En el capítulo dos, *"CRISPDM: metodología para identificar clases útiles al diagnóstico de diabetes tipo 2"* se expone el desarrollo de las tres primeras fases de la metodología.

El proceso de obtención de la solución propuesta se explica en el capítulo III, titulado *"Experimentación y análisis de los resultados"*

1. *Capítulo I: "Técnicas de la Inteligencia Artificial útiles a la prevención y diagnóstico precoz de enfermedades."*

1.1. Introducción

A grandes rasgos, el problema de la prevención y diagnóstico en Inteligencia Artificial consiste en determinar a partir del conocimiento de las leyes que rigen el comportamiento de un sistema y de un conjunto de medidas, observaciones o síntomas, cuáles son las causas, o los componentes del sistema responsables en última instancia de un posible comportamiento anómalo. Las distintas técnicas de IA empleadas para el diagnóstico son abordadas en este capítulo, referenciándose ejemplos fundamentales en esta área y en particular en el sector de la salud pública, objeto de nuestro trabajo.

1.2. Principales objetivos de la Salud en Cuba.

El Ministerio de Salud Pública (MINSAP) es el Organismo rector del Sistema Nacional de Salud (SNS), encargado de dirigir, ejecutar y controlar la aplicación de la política del Estado y del Gobierno en cuanto a la Salud Pública, el desarrollo de las Ciencias Médicas y la Industria Médico Farmacéutica.

El SNS se estructura en tres niveles que se corresponden con la estructura político-administrativa del país. El nivel nacional está representado por el Ministerio de Salud Pública que es el órgano rector con funciones metodológicas, normativas y de coordinación y control, al cual se le subordinan directamente los centros universitarios, institutos de investigaciones, centros hospitalarios de asistencia médica altamente especializados, centros de distribución y comercializadoras de suministros y tecnologías médicas, así como otros centros y entidades nacionales destinados a actividades técnicas y de apoyo. [22]

Los otros dos niveles están representados por las direcciones provinciales y municipales de salud que agrupan a las instituciones de salud en su respectivo nivel y que, al igual que en el nivel central, se subordinan desde el punto de vista administrativo a las estructuras de Gobierno en los distintos niveles organizativos, representando sus intereses ante ellos y dando respuesta a las demandas y necesidades de la población.

El Sistema Nacional de Salud se organiza en 3 niveles de atención:

Atención Primaria: Se brinda a nivel de los policlínicos y/o hospitales rurales a través del Programa de Medicina Familiar y abarca a todos los Equipos Básicos de Salud (EBS).

Constituye el primer contacto del paciente sano o enfermo con el sistema de salud, que puede brindarse en locales adaptados para consultas o en el domicilio de los pacientes, a cualquier instancia del sistema de salud, aunque generalmente se realiza en el Consultorio Médico.

La función principal de la atención primaria es la promoción-prevención de salud en las diferentes comunidades, además se realizan procedimientos diagnósticos y terapéuticos que no requieren técnicas complejas, que aplicadas con calidad pueden resolver la mayor parte de los padecimientos que afectan a las poblaciones. Se diagnostican enfermedades graves que pueden ser derivadas a niveles de atención superiores, realizan seguimiento a personas con padecimientos crónicos y pueden otorgar bienestar a pacientes con patologías incurables. En general tiene carácter ambulatorio y comprende tanto a personas aparentemente sanas como a enfermas y/o discapacitadas. [9]

La atención primaria de salud es un nivel cualitativamente superior de atención médica, cuya esencia radica en la participación activa de la comunidad; donde las poblaciones de objetos pasivos, en espera de que se le ofrezcan soluciones, pasan a ser sujetos protagónicos activos ante sus propios problemas de salud. Decir participación comunitaria, es decir liderazgo, comunicación, cambio de hábitos y de estilos de vida, autorresponsabilidad y acción creadora.

Atención Secundaria: Se brinda a nivel de las instituciones hospitalarias, por lo general son de carácter provincial, o sea atienden a toda la población de una provincia determinada. Se proporciona en un segundo escalón, al cual el paciente tiene acceso a través de una remisión del personal médico de la atención primaria o sin ella, acudiendo directamente la persona necesitada de atención médica.

Atención Terciaria: Es aquella que por su condición muy especializada, sólo se brinda en determinados centros, ejemplo: Instituto de Neurocirugía, Instituto de Cirugía Cardiovascular, Instituto de Nefrología, Instituto de Gastroenterología, entre otros o en centros hospitalarios y/o de investigación categorizados como centros de referencia nacional y en algunos casos de referencia internacional.

Podemos destacar entre los Principios Rectores del MINSAP el carácter estatal y social de la medicina, accesibilidad y gratuidad de los servicios, orientación profiláctica, aplicación adecuada de los adelantos de la ciencia y la técnica, participación de la comunidad e intersectorialidad, colaboración internacional y la centralización normativa y descentralización ejecutiva.

Tiene como Funciones Rectoras ejercer el control y la vigilancia epidemiológica de las enfermedades y sus factores de riesgo, la vigilancia sanitaria de todos los productos que pueden tener influencia sobre la salud humana, regulación y control de las investigaciones biomédicas, normar las condiciones higiénicas y el saneamiento del medio ambiente, regular el ejercicio de la medicina y de las actividades que le son afines y ejercer la evaluación, el registro, la regulación y el control de los medicamentos de producción nacional y de importación, equipos médicos y material gastable y otros de uso médico. [22]

En el actual proceso de perfeccionamiento, el MINSAP se ha trazado como estrategias de desarrollo el perfeccionamiento de la atención primaria, la revitalización hospitalaria, el desarrollo del programa nacional de medicamentos y medicina natural y tradicional, el desarrollo de la tecnología de punta e investigación, así como contar con sistemas para urgencia, óptica, estomatología, asistencia social, control económico, atención al hombre y los cuadros. [22]

En respuesta a estas políticas trazadas por el MINSAP y para apoyar la futura creación de un sistema experto que determine de forma rápida y eficiente si un paciente que presenta ciertos síntomas y un cuadro clínico concreto, presenta riesgo de padecer o padece de forma “silente” diabetes mellitus tipo 2 y hacerlo desde el nivel de atención primaria para evitar en lo posible la llegada del paciente a los niveles superiores, esta investigación se propone:

La confección de un conjunto de entrenamiento, a partir de determinar clases existentes en la población diabética tipo 2, para la prevención y diagnóstico precoz de la DM tipo 2

1.3. La Diabetes Mellitus tipo 2 en Cuba. Importancia de su prevención y diagnóstico precoz.

La Diabetes Mellitus (DM) es una enfermedad que se caracteriza por el aumento de los niveles de glucosa en la sangre. Esta tiene múltiples afectaciones para el organismo humano y es hoy, junto con las enfermedades cardiovasculares, el cáncer y las

enfermedades respiratorias crónicas, una de las cuatro Enfermedades No Transmisibles (ENT) prioritarias según la definición de la Organización Mundial de la Salud (OMS) [23]

En Cuba dado el aumento de la expectativa de vida de la población, producto a la disminución de la mortalidad infantil, las enfermedades trasmisibles y de la tasa de mortalidad en general, se ha originado un envejecimiento de la población que provoca un continuo aumento de la prevalencia de la diabetes.

La DM, por tanto ha devenido en un problema creciente para la población cubana. En la figura 1 se muestra el comportamiento de la cantidad de diabéticos identificados por cada 1000 habitantes desde el año 1991 hasta el 2006. Para el año 2000 existían 40X1000 diabéticos en Cuba entre diabéticos dispensariados y no dispensariados.

En el año 1968 se hablaba de la razón de 3 diabéticos desconocidos por cada caso conocido, actualmente se habla de 1 diabético desconocido por cada caso conocido en el país y se considera baja la cantidad de diabéticos dispensariados con respecto a la cantidad de diabéticos existentes. [21]

Ilustración 1-1 Prevalencia de Diabetes Mellitus en Cuba. Tasas Crudas y Ajustadas por edad x 1000 habitantes. [20]

A la situación del aumento de la prevalencia de la enfermedad se unen las complicaciones que esta genera y las implicaciones que tienen para la calidad de vida de los pacientes.

De acuerdo al Programa Nacional de Diabetes [21] las complicaciones fundamentales de esta enfermedad son:

- Las agudas, en las que se encuentran comprendidas el coma diabético y la cetoacidosis diabética. En el año 1994 la mortalidad por esta causa era de 64 x 100 000 diabéticos, (en USA era de 50 x 100 000 habitantes)

- Las macroangiopatías, que son los daños de los vasos sanguíneos grandes y que provocan infartos, apoplejías y trastornos de la circulación sanguínea en las piernas, lo que produce el conocido pie diabético.
- Las microangiopatías como la retinopatía y nefropatía diabéticas, que son daños en los ojos y en los riñones que se producen a largo plazo en los pacientes.

  En estudios poblacionales realizados por el Instituto Nacional de Nefrología en Cienfuegos, el 25% de todos los casos de insuficiencia renal eran diabéticos (similar a países desarrollados). En el caso de la retinopatía, estudios del Instituto Nacional de Endocrinología y el Nivel de Atención Primaria, realizados en la década del 90 muestran que el 26% de los diabéticos tienen algún grado de Retinopatía (lo que significa alrededor de 40 000 casos en el país), pero el 4% padece de Retinopatía Proliferativa (8 000 casos), es decir con riesgo de pérdida de visión o ya ciegos.
- Las neuropatías diabéticas, que son los daños de los nervios que produce dolor, pérdida de la sensibilidad e incapacidad para controlar los músculos

Todas estas, sumado a complicaciones de menor envergadura como lo son la hipertensión arterial (HTA), el hígado graso, daños en la piel y cardiopatías, afectan no solo la calidad de vida de los pacientes, sino también la familia y generan grandes gastos a la sociedad en general

Ante esta difícil realidad y el hecho de que el 80% de la diabetes tipo 2 es prevenible mediante la adopción de una dieta saludable y el incremento de la actividad física [21] se hace cada vez más necesario aunar esfuerzos en pro de la prevención y el diagnóstico temprano de esta enfermedad.

1.4. Flujo actual de los procesos.

Para el diagnóstico de la DM tipo 2, en el nivel primario de atención de la provincia Cienfuegos, un paciente aquejado debe transitar por el siguiente proceso [21]:

- El paciente acude al médico general integral de su área de atención primaria.
- El médico realiza la entrevista que le proporciona los datos generales del paciente y el conjunto primario de los síntomas.
- El médico realiza un examen físico que arroja un conjunto más completo de síntomas que no son identificables por el paciente.

En este momento el médico con todos los datos obtenidos debe poder sospechar un diagnóstico certero, y para corroborarlo entonces:

- El médico le indica una Alteración de la Glicemia en Ayunas (AGA) al paciente y espera por los resultados para llegar a un diagnóstico.
- Si este análisis tiene cifras mayores que 7 mmol/L el médico arriba a un diagnóstico certero.
- Si las cifras están entre 5.6 y 6.9 mmol/L entonces el análisis se considera dudoso y se le orienta una Tolerancia a la Glucosa Alterada (TGA).
- En caso de obtenerse en este nuevo análisis cifras dudosas al paciente se le da seguimiento cada seis meses repitiendo el mismo procedimiento.

En este punto, de no tener aún suficientes elementos entonces:

- El médico decide remitir al paciente a un especialista, generalmente un clínico, por no tener elementos suficientes para emitir un diagnóstico certero.

Una vez en manos del especialista, el médico especialista puede:

- Emitir un diagnóstico certero.
- Considerar repetir parte o todo el proceso anterior para llegar a emitir un diagnóstico certero.

En todos los casos cuando el médico arriba a un diagnóstico certero se entiende que inmediatamente indica el tratamiento que a su elección alivia el padecimiento del paciente.

Actualmente, este proceso no cuenta en ninguna de sus etapas con niveles de automatización que faciliten el arribo al diagnóstico precoz de esta enfermedad en la población.

1.5. Análisis de la ejecución de los procesos.

Cuando se analiza el flujo actual de los procesos es posible determinar que en el caso del diagnóstico de diabetes mellitus tipo 2, un paciente acude a su médico cuando se le presenta algún síntoma. Esto es interesante puesto que por lo general la sintomatología en esta enfermedad aparece cuando los niveles de glicemia se elevan por encima de 10 mmol/L y en este punto ya el paciente comienza a sufrir afectaciones en sus órganos y corre el riesgo de complicarse.

A esto se le suma que desde que acude a su médico de asistencia primaria hasta que obtiene un diagnóstico y con él el tratamiento que provoca alivio a su mal debe esperar como promedio unas 4 semanas, teniendo en cuenta que normalmente un médico general integral, joven, de poca experiencia, en la mayoría de los casos indica

exámenes complementarios al paciente, los cuales no se obtienen de forma inmediata y que tienen un flujo actual de proceso que no describimos aquí por no ser de nuestro interés pero que consume un tiempo considerable (a consideración de los expertos, consume como promedio dos semanas) y que tienen adicionalmente un costo asociado. Una vez llegado a este punto el estado de salud del paciente puede haber empeorado con lo que su condición puede tornarse, si no lo es ya, crítica lo que podría conducir en el peor de los casos a la muerte del paciente. De no tratarse de un caso grave, podría entonces demorarse la aplicación del tratamiento adecuado con lo cual se retrasa el período de recuperación del paciente y se encarece el costo del tratamiento.

Comúnmente sucede que la experiencia del joven médico no alcanza para una vez realizados análisis complementarios determinar, en un primer caso, el nivel de riesgo que tiene el paciente o, en un segundo caso, la posibilidad de prevenir o demorar su debut con la enfermedad. Por lo que se ve imposibilitado de emitir un diagnóstico preciso y de indicar el tratamiento adecuado al caso en cuestión. Por tanto, en el primer caso, el paciente es remitido a un especialista quien deberá comenzar por estudiar la situación con la ayuda de los análisis y síntomas ya establecidos y de ser necesario reorientar la búsqueda por otros caminos indicando nuevos análisis según su criterio. En el segundo caso, al paciente se le hace esperar por el seguimiento semestral que se le brinda hasta que debute con cifras de glicemia por encima de los 7 mmol/L. Lo que puede provocar daños irreversibles en la salud del paciente [24].

1.6. Procesos objeto de automatización.

En el proceso antes descrito es posible minimizar los tiempos de diagnóstico y lo que resulta más eficiente aún, lograr la identificación de personas con riesgo de padecer la enfermedad y de pacientes en etapas tempranas si aplicamos técnicas de IA a esta fase del proceso.

El proceso actual comienza cuando el paciente llega a la consulta y termina cuando es diagnosticado por el médico (MGI o especialista) quien emite su criterio de si el paciente padece o no la enfermedad y de padecerla ofrece su clasificación y sugiere un tratamiento adecuado.

Dentro de este proceso es importante mencionar algunas tareas que podrían ser consideradas como subprocesos dentro de un proceso general y único y que son objeto de automatización.

- Registro de los datos generales del paciente.
- Registro de los síntomas del paciente que se obtienen por entrevista y/o examen físico.
- Registro de los resultados de los exámenes complementarios.
- Emisión de resultados que contempla la obtención de un diagnóstico para el paciente y la sugerencia de el(los) posible(s) tratamientos a indicar en este caso.

1.7. Técnicas de IA útiles a los procesos de prevención y diagnóstico.

El problema del diagnóstico ha sido uno de los más estudiados en el campo de la Inteligencia Artificial. Son ya clásicos algunos de los sistemas expertos utilizados con fines de diagnóstico que ha producido esta rama del saber y en particular se encuentran un número importante de estos que se han dedicado ya bien sea con fines prácticos o de investigación al campo del diagnóstico médico en cuestión.

Así puede citarse el sistema MYCIN [25], que constituye el primer SE aplicado a un problema real y precisamente, se diseñó para dar solución a problemas de diagnóstico médico. Este sistema constituyó un proyecto cooperado entre el Dpto. de Ciencias de la Computación y la Escuela de Medicina de la Universidad de Stanford, y su objetivo era diagnosticar y a la vez, remitir recomendaciones de tratamientos de la meningitis y la bacteremia, ambas infecciones de la sangre.

Posteriormente, se desarrollan múltiples ejemplos de SE entre los que podemos citar: CADUCEUS desarrollado por Pople, Myers y Miller en 1975 para diferentes diagnósticos de medicina interna; PUFF para el diagnóstico de enfermedades pulmonares de Feigenbaum en 1977; y CASNET un SE para diagnóstico y tratamiento del glaucoma que ha favorecido el diseño de otros muchos en el área de la oftalmología, la endocrinología y la reumatología. Constituyen ejemplos también: INTERNIST [26], un sistema experto de consulta en el dominio de la medicina interna, CASNET [27], para diagnóstico médico del glaucoma, o sistemas como MDX [28], también utilizado para diagnóstico médico.

Los primeros SE desarrollados en el área de diagnóstico utilizaron como formalismo de representación del conocimiento las reglas de producción y como Método de Solución de Problemas (MSP) el algoritmo primero en profundidad con búsqueda dirigida por objetivos.

Dado que una de las ideas centrales de MYCIN era la separación entre el conocimiento del dominio y el conocimiento de inferencia, esto dio lugar a su generalización en el sistema EMYCIN, un "shell " o concha que incluye los procedimientos de inferencia de MYCIN y al que pueden incorporarse bases de conocimiento externas construidas sobre distintos dominios, de manera que a partir del núcleo del sistema MYCIN pudiesen desarrollarse otros sistemas expertos. Con ello se pretendía que la labor del "ingeniero del conocimiento" se limitara a "extraer" del experto una experiencia o un conocimiento que era plasmado en reglas de producción que podían ser procesadas por el sistema-concha siguiendo la misma estrategia de MYCIN.

Las críticas que siguieron a MYCIN, en particular la realizada por Clancey, [29], dan origen a la construcción de un nuevo sistema denominado NEOMYCIN en el que se separa el conocimiento médico de la estrategia de diagnóstico, permitiendo que ambos sean incorporados en el sistema de forma independiente mediante sus correspondientes reglas y metarreglas.

La generalización del sistema NEOMYCIN dio origen al sistema-concha HERACLES (Heuristic Classification Shell). En él aparece ya la idea de que parte del proceso de diagnóstico llevado a cabo por NEOMYCIN corresponde a una tarea genérica de clasificación heurística dentro de una taxonomía de conceptos correspondientes a las enfermedades o en general a las disfunciones que se intentan diagnosticar. [29]

Haciendo una revisión de los principales sistemas y métodos empleados en inteligencia artificial para resolver el problema del diagnóstico, comenzando por los primeros sistemas expertos, encontramos además del clásico sistema MYCIN, y sus sucesivos desarrollos, la línea seguida por el grupo de McDermott, en la que se destaca por su aplicación al problema del diagnóstico el sistema MOLE [30], una de cuyas principales características es la inclusión de un método automático de adquisición del conocimiento. Años más tarde, a partir de las experiencias de este sistema y del sistema SALT [31] se construye el entorno de desarrollo de sistemas inteligentes PROTÉGÉ [32]. Este entorno propone una metodología propia para el desarrollo de los sistemas inteligentes, que se desarrolló en el sistema PROTÉGÉ-II, [33]. Otra línea de sistemas expertos de diagnóstico que da lugar a propuestas metodológicas es la seguida por el grupo de la Universidad de Ohio, que se inicia con la aparición del sistema experto MDX [28] en donde claramente se pone de manifiesto el problema del

"conocimiento profundo" de los sistemas expertos y que posteriormente evoluciona dando lugar al concepto de Tarea Genérica [34]. En esta línea existen otros trabajos [35], que también están encaminados hacia propuestas metodológicas centradas en la idea de Tareas Genéricas y Métodos.

A partir de estos se desencadenó el diseño de múltiples SE de ayuda al diagnóstico. En IFIPIMIA International Working Conference on Computer Aided Medical Decision-Making se presentan distintos trabajos en tal sentido [36], uno de ellos es un prototipo para el diagnóstico y tratamiento de la epilepsia usando reglas para representar el conocimiento de los expertos.

NEUREX constituye un SE tutorial que imita el proceso de diagnóstico de un neurólogo, ayuda al usuario en la planificación de pruebas y la interpretación de los resultados y asegura que se alcancen los diagnósticos más adecuados.

Adicionalmente, los Sistemas de Apoyo a las Decisiones Clínicas (DSS) [37] son actualmente un área de intensa investigación y desarrollo, que comprende a diversas técnicas –redes probabilísticas Bayesianas, redes neuronales, sistemas basados en reglas—y ha logrado producir aplicaciones que compiten con la pericia diagnóstica clínica de un médico. Los DSS tienen un gran potencial como herramientas para entrenar a estudiantes de medicina y médicos jóvenes en razonamiento diagnóstico. Dxplain[1], un sistema experto desarrollado por la Universidad de Harvard es una de tales aplicaciones, con éxito probado.

En Cuba se han dado algunos pasos al respecto. Baste recordar, algunos de los productos de software ya mencionados en la introducción de este trabajo.

1.8. Estructura de los Sistemas Expertos.

Un SE o Sistema Basado en el Conocimiento (SBC) es definido como: "*Un sistema computarizado que usa conocimiento sobre un dominio para arribar a una solución de un problema de ese dominio. Esta solución es esencialmente la misma que la obtenida por una persona experimentada en el dominio del problema cuando se enfrenta al mismo problema"* [38]

Estos se diferencian de otros programas en:

[1] Version Date: April 6, 2003 en http://www.lcs.mgh.harvard.edu/dxplain.htm

- La separación del conocimiento de cómo este es usado (distinción entre conocimiento y estrategia de control).
- El uso de conocimiento muy específico del dominio.
- La naturaleza heurística, en lugar de algorítmica, del conocimiento empleado. [38]

Estos sistemas son codificados por un ingeniero de software experto en IA, a este se le llama “ingeniero del conocimiento” y es quien establece las reglas y los caminos a seguir por el sistema.

La estructura básica de los SBC es vista desde tres puntos diferentes: el usuario final, el constructor de herramientas y el ingeniero del conocimiento. Para este último estos sistemas se componen de un programa inteligente y el shell o concha de desarrollo.

El programa inteligente es el producto desarrollado para el usuario final y está compuesto por una base de conocimientos (BC) y una máquina de inferencia (MI). Por otra parte el shell es quien asiste al ingeniero de conocimiento en estructurar, depurar (debuging), modificar y expandir el conocimiento extraído desde el experto [38].

La BC es la componente más importante del SBC. Toda BC tiene asociada un formato el cual indica como el conocimiento se representa internamente. A este formato se denomina Forma de Representación del Conocimiento (FRC).La MI es el interprete del conocimiento almacenado en la BC. La MI implementa algún método de solución de problemas con una dirección (forward o backward) de búsqueda dada. [38]

Para la conformación de la BC de un sistema experto el ingeniero adquiere información de diversas fuentes. Estas fuentes de conocimiento varían en dependencia de la FRC del sistema. En la gran mayoría de los SBC una de estas fuentes son los ejemplos, que no es más que un conjunto de casos tomados del dominio de aplicación del SE. Para la conformación de las BC no solo se tienen en cuenta las fuentes, sino que en muchos casos, como los Sistemas Basados en Reglas (SBR) y Redes Expertas, se utilizan técnicas de IA supervisadas, las cuales necesitan el conjunto de ejemplos para su entrenamiento. En estos casos, los ejemplos se componen de un conjunto de variables que los caracteriza y una clasificación.

Existen dominios de aplicación de los SE donde se cuenta con los ejemplos, pero estos carecen de una clasificación o clase. Esto reduce las posibilidades de aplicar para la conformación de la BC técnicas de IA supervisadas, pues como su nombre lo indica requieren clases en su conjunto de entrenamiento para su aprendizaje.

No obstante se han desarrollado técnicas de Inteligencia Artificial que permiten identificar clases en conjuntos de datos que carecen de las mismas. Entre ellas se encuentran las técnicas de clustering o agrupamiento, que han demostrado su utilidad en la clasificación de grandes cantidades de casos.

Específicamente en esta investigación, se aplican estas técnicas para desarrollar minería de datos en la información de los pacientes diabéticos tipo 2, con el propósito de identificar clases que sean útiles para la conformación de un SE que contribuya a la prevención y el diagnóstico temprano de la DM.

1.9. Técnicas de clustering en la Minería de Datos.

1.9.1. Minería de Datos: Proceso para convertir datos en información.

La minería de datos (Knowledge Discovery in Databases, KDD) es definida como:

“Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.” [39]

La minería de datos surge especialmente, por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares. Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento; un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos. [6]

Lo anterior permite concluir que la minería de datos se encarga de dos retos fundamentales, el trabajo con grandes volúmenes de datos y el uso de técnicas adecuadas para analizarlos de modo que se pueda extraer conocimiento novedoso y útil de los mismos.

La minería de datos excede la capacidad humana para el análisis de grandes volúmenes de datos. Por consiguiente, la utilización plena de los datos almacenados depende del uso de técnicas del descubrimiento del conocimiento. [7]

Estas técnicas son utilizadas en KDD para la inducción de reglas, el reconocimiento de patrones, el modelado predictivo, la detección de dependencia y los problemas de clasificación o agrupamiento.

Las técnicas de clustering encuentran su utilidad en MD en el cumplimiento de estas tareas para la solución de problemáticas reales de la ciencia, por sus características y efectividad en la determinación de clases o grupos en los conjuntos de datos.

Son muchas las investigaciones que vinculan estas dos ramas de la IA para aprovechar sus ventajas en la solución de diversas problemáticas, que van desde la determinación de patrones característicos de la población carcelaria en Argentina [40] hasta la búsqueda y caracterización de subgrupos de pobreza en datos de encuestas hechas en Nicaragua [41]. Otros se han dedicado a utilizar las bondades de estas técnicas para descubrir estilos de aprendizaje en los estudiantes [42], [43], siendo la segunda una investigación realizada en la Universidad de Cienfuegos.

Son estas investigaciones evidencia de la utilidad de la MD. Una rama en la que se almacena gran cantidad de información es la medicina, donde el descubrimiento del conocimiento se puede emplear para:

- La identificación de terapias médicas satisfactorias para diferentes enfermedades.
- La asociación de síntomas y clasificación diferencial de patologías.
- El estudio de factores (genéticos, precedentes, hábitos alimenticios,...) de riesgo para la salud en distintas patologías.
- La segmentación de pacientes para una atención más inteligente según su grupo.
- Los estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.
- La identificación de terapias médicas y tratamientos erróneos para determinadas enfermedades. [7]

Investigaciones enfocadas en la reducción de los rasgos que actualmente se registran de los pacientes en las bases de datos médicas [44]; en evaluar la forma en la que se consumen medicamentos en un hospital peruano [45]; en el descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino [46]; y en la determinación de la eficacia de la braquiterapia en el tratamiento de cáncer [47], son muestra de los avances de la aplicación de la MD y el clustering en la solución de problemas de la salud en América Latina. Específicamente en Cuba esta relación se ha aplicado a la predicción de pacientes diabéticos [48] y a la contribución del diagnóstico de entidades clínicas [49].

Los ejemplos antes citados demuestran las múltiples aplicaciones de la MD en las ciencias médicas y estos han necesitado para su concreción de la aplicación del análisis de cluster.

1.9.2. El análisis de cluster.

El análisis de cluster es parte de la vida cotidiana. Constantemente estamos distinguiendo entre una cosa de un tipo y otra, a la que consideramos de otra clase, por ser muy diferente a la primera. La verdad es que como Han y Kamber [50] expresan, el análisis de cluster es una importante actividad humana. Desde temprano en la niñez, uno aprende a distinguir entre perros y gatos, o entre animales y plantas.

El Clustering (Agrupamiento) es definido como el proceso de agrupar un conjunto de objetos físicos o abstractos en clases (cluster) de objetos similares. Permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. [50]

Expresado en términos de variabilidad se hablaría de minimizarla dentro de los grupos para al mismo tiempo maximizarla entre los distintos grupos. En el proceso de clustering, no hay clases predefinidas que permitan conocer las relaciones existentes entre los datos, esto se puede ver como un proceso no supervisado. [51]

Básicamente, el agrupamiento consiste en clasificar objetos o individuos en clusters donde los elementos de un grupo son más similares entre sí, que respecto a elementos de otros grupos.

Una vez conformada la clasificación, si esta es acertada, los elementos de un grupo estarán gráficamente muy próximos entre sí, a la vez que se mostrarán distantes con elementos de otros clusters.

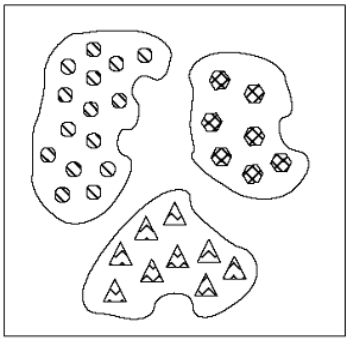


Ilustración 1-2 Ejemplo de clustering. [43]

1.9.3. Técnicas de clustering.

Para realizar el análisis de cluster existe gran variedad de técnicas o algoritmos.

Estas se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. [7]

Como característica fundamental de las técnicas de agrupamiento se tiene que estas utilizan una medida de semejanza, la cual se basa en las características de los objetos y frecuentemente se define por la proximidad en un espacio multidimensional.

Existen varias medidas de distancia. La más popular es la distancia Euclídea que se define como:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Donde “x” y “y” son vectores de dimensión “n”.

Otra medida, bien conocida, es la distancia Manhattan o distancia por cuadras (city-block) que hace referencia a recorrer un camino zigzagueando por el camino más corto, como se haría en Manhattan. Su fórmula de cálculo es:

$$d(x,y) = \sum_{i=1}^{n}\left|x_i - y_i\right|$$

Una medida que generaliza las dos anteriores es la distancia de Minkowsky y se define como:

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^q\right)^{1/q}$$

Siendo “q” un entero positivo.

Las técnicas de clusters se clasifican principalmente en particionales o jerárquicas. Aunque en las últimas décadas, debido al desarrollo de la Inteligencia Artificial se han presentado métodos de clustering basados en otras teorías o técnicas [52], [53].

Las técnicas particionales generan una división única de los datos en un intento por recuperar grupos naturales presentes en los datos [54]. Dentro de estos algoritmos están:

- Clustering Numérico (k-medias)

Es un algoritmo muy utilizado ya que es muy sencillo. En primer lugar se debe especificar por adelantado cuántos clusters se van a crear, este es el parámetro k, para lo que se pueden seleccionar k elementos aleatoriamente, que representarán el centro o media de cada cluster. A continuación cada una de las instancias, ejemplos, es asignada al centro del cluster más cercano de acuerdo con una medida de distancia que le separa de él. Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos clusters. Finalmente se repite el proceso completo con los nuevos centros de los clusters. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración. [7]
En la Figura 3 se muestra un ejemplo de lo que se explicó anteriormente.

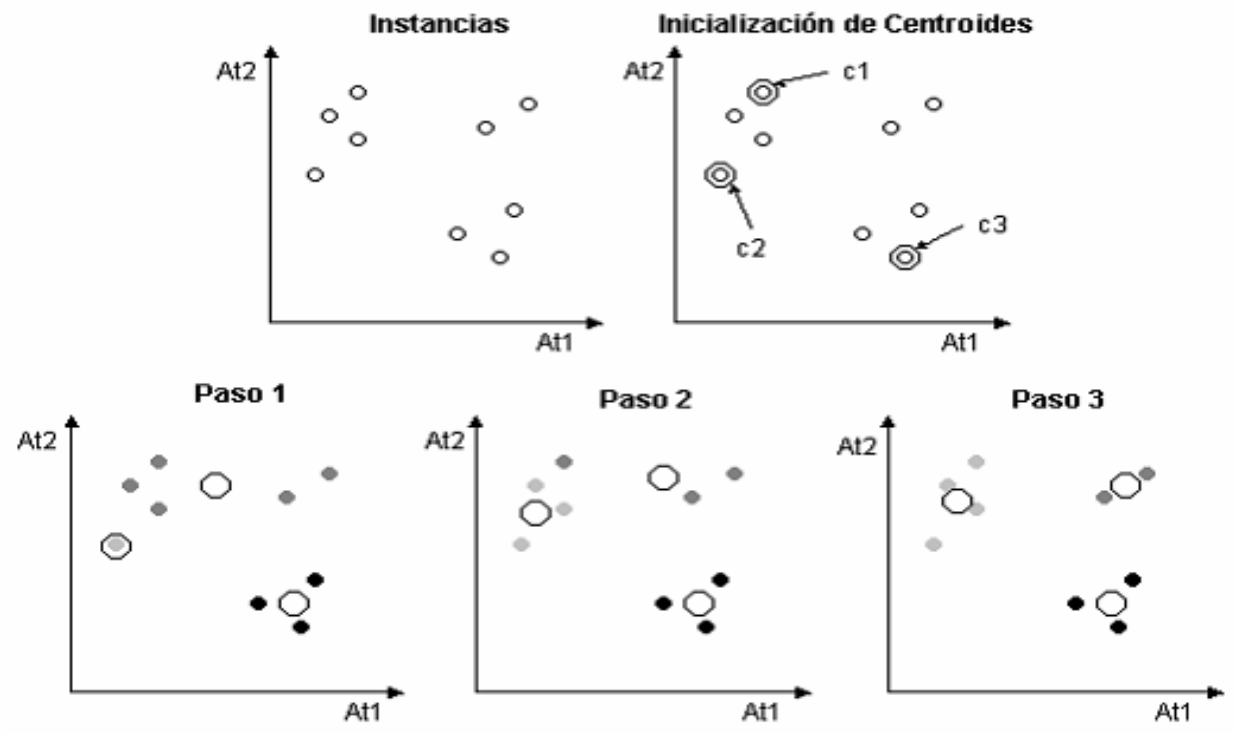


Ilustración 1-3 Ejemplo de clustering con K medias.

- Clustering Conceptual (COBWEB)

Forma los conceptos por agrupación de ejemplos con atributos similares. Representa los clusters como una distribución de probabilidad sobre el espacio de los valores de los atributos, generando un árbol de clasificación jerárquica en el que los nodos intermedios definen subconceptos. El objetivo de COBWEB es hallar un conjunto de

clases o clusters (subconjuntos de ejemplos) que maximice la utilidad de la categoría (partición del conjunto de ejemplos cuyos miembros son clases). [7]

En la Figura 4 se muestra un ejemplo de árbol generado por el algoritmo. Vale destacar que este es sensible al orden en que se le introducen los datos, por tanto una variación en este sentido provocaría otra salida.

Datos de Entrada

Nombre	Recubierto de	Cavidades corazón	temperatura del cuerpo	Fertilización
Mamifero	Pelo	Cuatro	Regulada	Interna
Pez	Escamas	Dos	Sin regulación	Externa
Anfibio	Piel húmeda	Tres	Sin regulación	Externa
Ave	Plumas	Cuatro	Regulada	Interna
Reptil	Piel dura	Cuatro Imperfectas	Sin regulación	Interna

Árbol Generado

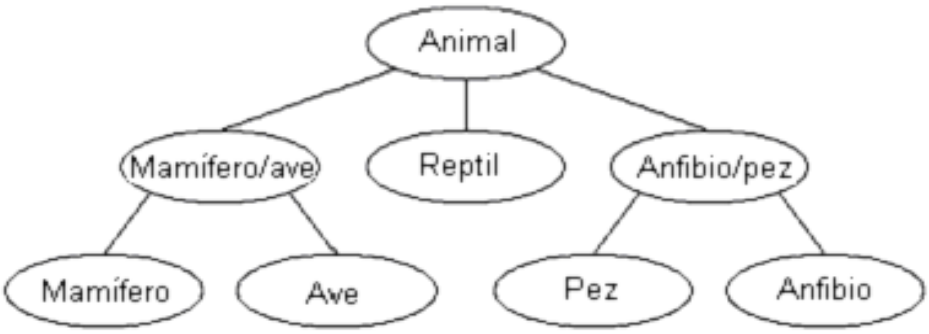


Ilustración 1-4 Ejemplo de árbol generado por COBWEB.

- Clustering Probabilístico (Expectation Maximization)

Busca el grupo de clusters más probable en dependencia a los datos, solucionando así los defectos de los algoritmos anteriores al establecer dependencia entre los resultados y el orden en el que se le introducen los datos y la tendencia al sobreajuste.

En este algoritmo los ejemplos tienen ciertas probabilidades de pertenecer a un cluster. La base de este tipo de clustering se encuentra en un modelo estadístico llamado mezcla de distribuciones. Cada distribución representa la probabilidad de que un objeto tenga un conjunto particular de pares atributo-valor, si se supiera que es miembro de ese cluster. Se tienen k distribuciones de probabilidad que representan los k clusters. [7]

El problema básico en el clustering es determinar la cantidad óptima de clusters. La teoría expuesta anteriormente es útil para conformar grupos en un determinado conjunto de datos. Una vez conformados estos grupos, un problema es que en algunas

ocasiones los grupos obtenidos después de aplicar algún algoritmo de conglomerados, no representan la estructura real que la base de datos posee. Por esta razón se necesitan medidas cuantitativas para evaluar el resultado del algoritmo de conglomerados. Esta tarea es llamada Validación de Conglomerados. [55]

Numerosas aproximaciones se han hecho a lo largo de los años en este sentido. En el próximo sub-epígrafe se abordan algunas de estas medidas desarrolladas en la literatura estadística.

1.9.4. Índices de validación de los clusters.

Son muchas las técnicas desarrolladas para la validación de los conglomerados. Estas estrategias son denominadas índices y se pueden clasificar en dependientes del algoritmo utilizado para el agrupamiento o independientes.

Para validar los modelos en esta investigación se utilizarán los índices de validación independientes de los algoritmos de conglomeración.

Estos procedimientos, a su vez, pueden ser clasificados en externos, los que utilizan una partición de referencia obtenida de manera independiente, y en internos, los que utilizan información que se obtiene a partir del mismo proceso de clasificación. [55]

Por la ausencia de clases en los datos de los pacientes diabéticos tipo 2 en esta investigación se utilizan índices internos.

Los criterios de validación independientes internos permiten verificar si la estructura del cluster producido por un algoritmo de agrupamiento coloca adecuadamente los datos, pero usando solamente información inherente a la base de datos. [55]

Algunos de los índices que están dentro de esta clasificación son:

- Dunn: corresponde al radio de la distancia más pequeña entre las observaciones de diferentes clusters y la distancia inter-cluster más grande. Tiene valores entre 0 e infinito y debe maximizarse para que la agrupación sea la óptima. Mide cuan compactos y separados están los clusters entre ellos. [56]

  Es definido como:

$$D = \min_{i=1,\ldots,m} \left\{ \min_{j=i+1,\ldots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} diam(C_k)} \right) \right\}$$

donde la función de diferencia entre dos clusters C_i y C_j es $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$ y el diámetro de un cluster C es definido como $diam(C) = \max_{x,y \in C} d(x, y)$

- Calinski y Harabasz: está basado en la suma cuadrática interna y entre clusters. Utiliza las matrices de dispersión y el valor que maximice esta función será el candidato para especificar el número de clusters que se usará para clasificar los datos [56], [57]

  Si se tiene una partición con k clusters el índice de Calinski y Harabasz (CH) se calcula de la siguiente manera:

$$CH(k) = \frac{tr(S_B)/(k-1)}{tr(S_W)/(n-k)}$$

  S_B y S_W son las matrices de dispersión externa e interna respectivamente.

$$S_W = \sum_{k=1}^{k} \sum_{i \in P_k} (x_i - \bar{x}_k)^T (x_i - \bar{x}_k)$$

$$S_B = \sum_{k=1}^{k} n_k (x_i - \bar{x}_k)^T (x_i - \bar{x}_k)$$

- Ball and Hall: se basa en las matrices de dispersión interna y externa. Su fórmula de cálculo es: SSW/k donde k es el número de clusters y SSW es la suma de cuadrados dentro de los clusters. El máximo valor de las segundas diferencias respecto al de la izquierda es tomado como el número de clusters apropiado. [58]
- RMSSDT: Del inglés Root Mean Square Standard Deviation (Raíz Media Desviación Estándar Cuadrada) es la varianza que mide la homogeneidad de los clusters, formalmente definida como:

$$\sqrt{\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} (x_k - \overline{x_j})^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} (n_{ij} - 1)}}$$

  donde nc es el número de clusters, d es el número de dimensión, nij es el número de elemento en el cluster i y dimensión j y $\overline{x_j}$ es el valor esperado en la dimensión j.

Como el objetivo del proceso de clustering es identificar grupos homogéneos el valor de RMSSTD más pequeño indica el mejor agrupamiento. [57]

- RS: Del inglés R Squared (R Cuadrado) es la medida de la diferencia de los clusters. Formalmente mide el grado de homogeneidad entre grupos. Los valores de RS están en un rango de 0 a 1, donde 0 significa que no hay diferencias entre los clusters y 1 que existen diferencias significativas entre estos. [57]

La aplicación de estos índices permite resolver el problema de identificar la cantidad óptima de clusters y su correcta distribución, encontrando en los datos grupos bien conformados que respondan a los objetivos de la MD. Por esta razón se propone su uso para el cumplimiento de estos fines en la presente investigación.

1.10. Conclusiones parciales.

La Inteligencia Artificial y sus técnicas contribuyen de manera activa a la prevención y el diagnóstico de enfermedades, entre ellas la diabetes mellitus tipo 2. Específicamente, las técnicas de clustering permiten realizar minería de datos para determinar clases en los pacientes diabéticos, para conformar la base de conocimientos de un SE para la prevención y el diagnóstico de esta enfermedad.

2. Capítulo II: "CRISP-DM: metodología para identificar clases útiles al diagnóstico de diabetes tipo 2"

2.1. Introducción.

La minería de datos, encargada de extraer conocimiento de grandes bases de datos, es un proceso que para su correcta ejecución se estructura en varias fases o etapas de desarrollo. Es claro que al tratarse de trabajo con datos almacenados en diferentes instituciones y en diferentes formatos, estos requieren ser adecuados para las técnicas de descubrimiento del conocimiento. Es este proceso el que demanda el 70% del esfuerzo en KDD [7]. En el presente capítulo se expone cómo se desarrollan las primeras fases de la MD en la información de los pacientes diabéticos, siguiendo la metodología CRISP-DM. Además de explicar aspectos fundamentales de esta metodología y la justificación de su uso.

2.2. Metodología para el proceso de minería de datos.

La minería de datos como proceso que involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario ha sido estructurada en la literatura en las siguientes etapas [59]:

- Comprensión del dominio de la aplicación, del conocimiento relevante y de los objetivos del usuario final.
- Creación del conjunto de datos: consiste en la selección del conjunto de datos, o del subconjunto de variables o muestra de datos, sobre los cuales se va a realizar el descubrimiento.
- Limpieza y preprocesamiento de los datos: Se compone de las operaciones, tales como: recolección de la información necesaria sobre la cual se va a realizar el proceso, decidir las estrategias sobre la forma en que se van a manejar los campos de los datos no disponibles, estimación del tiempo de la información y sus posibles cambios.
- Reducción de los datos y proyección: Encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.

- Elegir la tarea de Minería de Datos: Decidir si el objetivo del proceso de KDD es: Regresión, Clasificación, Agrupamiento, etc.
- Elección del algoritmo(s) de Minería de Datos: Selección del método(s) a ser utilizado para buscar los patrones en los datos. Incluye además la decisión sobre que modelos y parámetros pueden ser los más apropiados.
- Minería de Datos: Consiste en la búsqueda de los patrones de interés en una determinada forma de representación o sobre un conjunto de representaciones, utilizando para ello métodos de clasificación, reglas o árboles, regresión, agrupación, etc.
- Interpretación de los patrones encontrados. Dependiendo de los resultados, a veces se hace necesario regresar a uno de los pasos anteriores.
- Consolidación del conocimiento descubierto: consiste en la incorporación de este conocimiento al funcionamiento del sistema, o simplemente documentación e información a las partes interesadas.

Todas estas etapas son englobadas básicamente en cuatro fases: entendimiento del dominio, preparación de los datos, minería de datos (modelación) e interpretación y consolidación del conocimiento.

No cabe dudas que el proceso de extraer información, a partir de un amplio conjunto de datos, en muchas ocasiones se torna engorroso y difícil. Por esta razón se han desarrollado metodologías y estrategias que permiten organizar y guiar el trabajo.

Actualmente se puede hablar de metodologías como, el proceso KDD, CRISP-DM o Catalyst. Cada una de ellas responde a los diversos intereses y necesidades de las áreas de aplicación de KDD.

De estas la más seguida y referenciada es CRISP-DM (CRoss Industry Standard Process for Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos). Esto es corroborado por una encuesta realizada por el portal para análisis de datos KDnuggets en agosto del 2007 [61], sobre las metodologías empleadas para afrontar procesos de minería de datos. Los resultados se muestran en la Figura 2.1.

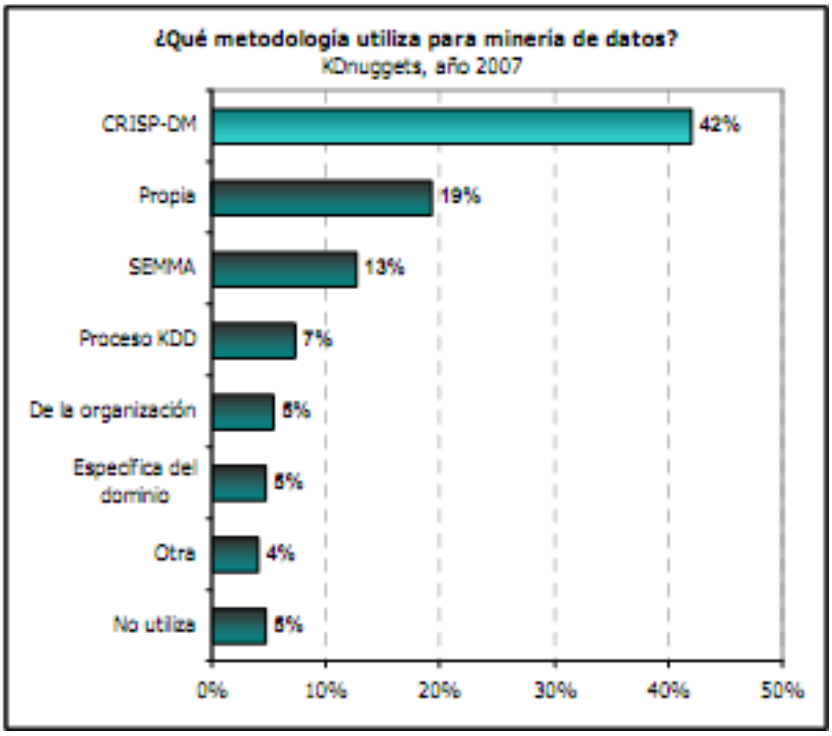


Ilustración 2-1 Principales metodologías empleadas para realizar procesos de KDD según una encuesta realizada por KDnuggets en agosto de 2007. [62]

CRISP-DM creada en el 2000 por el grupo de empresas SPSS, NCR y Daimer Chrysler es una metodología de distribución libre lo que permite que esté en constante desarrollo por la comunidad internacional. Una de sus ventajas es que resulta independiente de la herramienta que se utilice para llevar a cabo el proceso de MD. Esta metodología estructura el proceso en seis fases generales: análisis del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. [63]. Estas fases se relacionan conformando un proceso iterativo de desarrollo. (Figura 2.2).

Las fases de CRISP-DM se descomponen en tareas generales de segundo nivel y estas a su vez en tareas específicas, aunque en la literatura no se especifica cómo realizar estas. La sucesión de estas fases y tareas no es rígida.

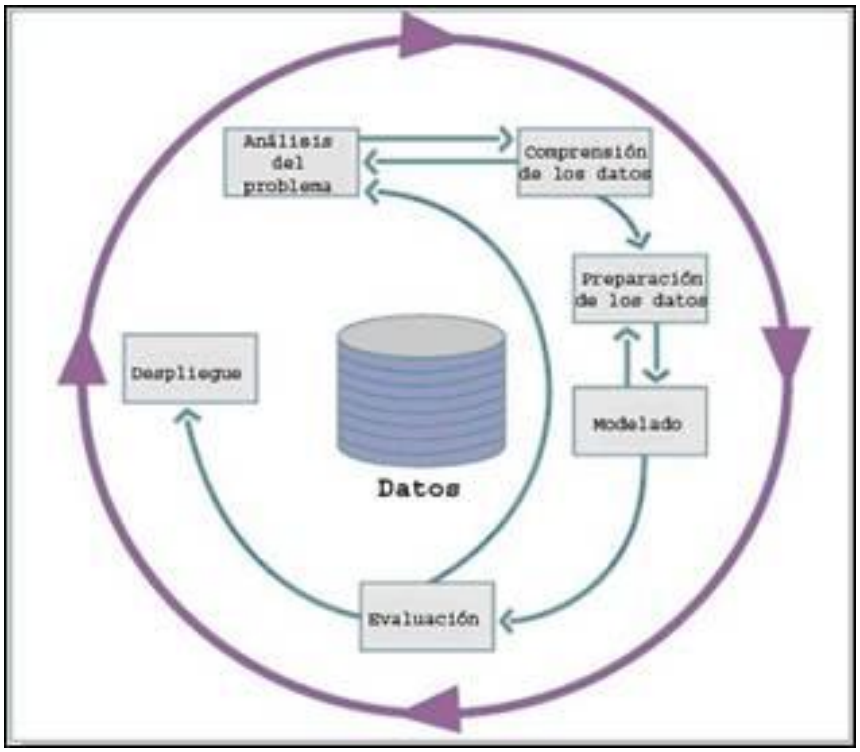


Ilustración 2-2 Fases del modelo de referencia CRISP-DM 1.0 y sus principales relaciones. [63]

2.2.1. Justificación de la elección de CRISP-DM.

Las ventajas que CRISP-DM ofrece como metodología para desarrollar MD, dígase su libre distribución, su independencia del tipo de herramienta (a diferencia de SEMMA, desarrollada para softwares de SAS) etc., la sugieren como candidata para realizar KDD en los datos de los pacientes con diabetes tipo 2.

Por otra parte se tiene que la experiencia del ingeniero de conocimientos en este caso particular, es pobre en el desarrollo de MD, proveyendo CRISP-DM con su desglose en tareas generales y específicas una guía clara, lo que no hacen otras metodologías como el proceso KDD, por solo citar una.

Sumado a todo lo anterior razones como su evidente preferencia a nivel internacional para llevar a cabo la minería de datos y su aplicación en procesos de descubrimiento del conocimiento en la medicina cubana con resultados satisfactorios [48], [49], justifican la elección de CRISP-DM para rectorar la MD en la información de los diabéticos tipo 2, con el objetivo de identificar clases presentes en los datos que contribuyan al diagnóstico temprano y la prevención de la diabetes mellitus.

2.3. CRISP-DM: análisis del problema y preprocesamiento de los datos.

Los primeros pasos en MD consisten en definir lo que se espera del proceso y preparar los datos para ser explorados en busca del conocimiento que solucione la problemática identificada. CRISP-DM propone para el cumplimiento de estas funciones el desarrollo de tres fases, una para el análisis y dos para la adecuación de los datos.

2.3.1. Fase I: Análisis del problema.

Es en la fase Análisis del problema, también llamada Comprensión del negocio, donde se identifica la expectativa del cliente con el proceso KDD. Durante esta etapa CRISP-DM propone el cumplimiento de tareas como la determinación de los objetivos del negocio, evaluación de la situación, determinación de los objetivos de minería de datos y producción del plan del proyecto [64], [65]. En el caso que nos ocupa esta fase se desarrolla como se describe en los sub-epígrafes siguientes.

2.3.1.1. Determinación del objetivo del negocio.

Saber cómo se puede contribuir a la problemática de la DM tipo 2 en Cienfuegos es parte del reto de esta tarea. Determinar el objetivo del negocio permitió comprender lo que los médicos quieren lograr.

El objetivo es mejorar el proceso de diagnóstico de la DM tipo 2, permitiendo que este se haga en etapas tempranas de la enfermedad o aún antes de su debut, en las áreas de atención primaria de Cienfuegos.

2.3.1.2. Determinar el objetivo de minería de datos.

Una vez determinado el objetivo del negocio es preciso expresar la meta que regirá la MD en los diabéticos tipo 2. Esta consiste en identificar grupos existentes dentro de los datos, que conformen las clases de un conjunto de entrenamiento útil a la construcción de un SE de apoyo a la prevención y el diagnóstico temprano de la diabetes tipo 2.

2.3.1.3. Evaluación de la situación.

Con los objetivos anteriores en mente se hace necesario identificar los recursos de los que se dispone para el proceso de minería de datos. Evaluar la situación ayudó a tener conciencia de la realidad en la que se llevaría a cabo el proceso.

Las salidas de esta tarea son los recursos disponibles, tanto humanos como materiales y digitales, para realizar el proceso.

Un recurso importante en todo proceso donde se usen técnicas de aprendizaje automático es el conjunto de casos o base de casos. Los datos para este proyecto se

obtienen de una hoja de cálculo Excel disponible en el Centro de Atención y Educación en Diabetes (CAED), donde se encuentra registrada la información de las historias clínicas de los pacientes que han ingresado en él.

Otro aspecto a tener en cuenta es las herramientas de software a utilizar en KDD. Son varios los softwares que se han implementado para realizar minería de datos y muchos de estos permiten hacerlo mediante técnicas de Inteligencia Artificial. En la figura 2-3 se muestra el resultado de una encuesta realizada por KDNuggets en mayo del 2012 sobre las herramientas de minería de datos utilizadas en los años 2011 y 2012[66].

Las herramientas seleccionadas para el proceso de minería en este proyecto fueron:

- Microsoft Excel 2007, una de las razones es porque el formato en el que se encuentran los datos del centro (.xls) es interpretado por esta herramienta. Otras son, las facilidades que brinda para el análisis y la transformación de los datos y sus múltiples ventajas en las opciones de cálculo, filtrado y para graficar información.

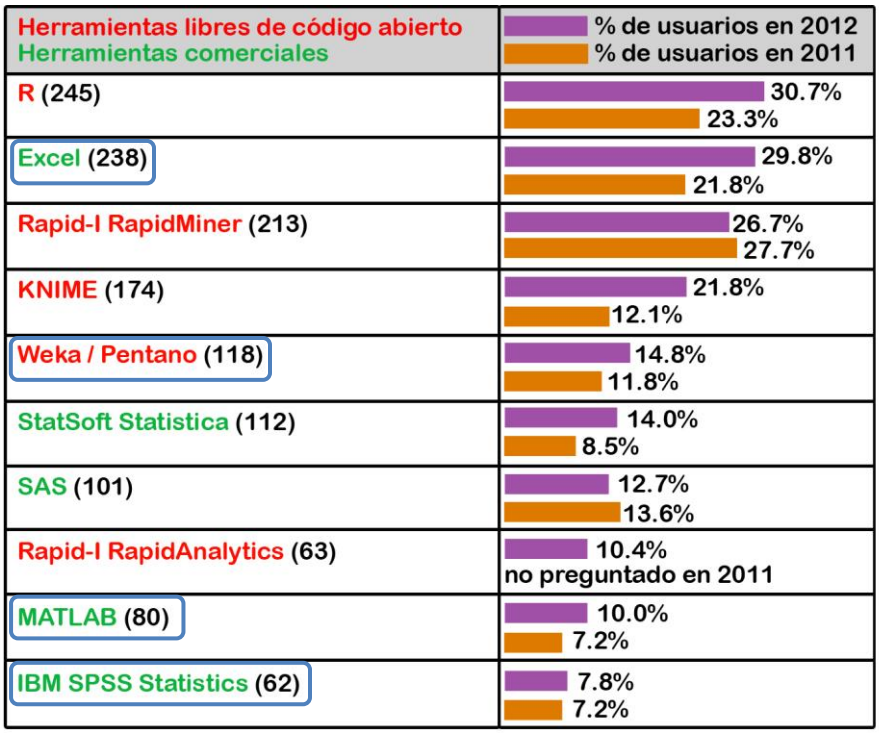


Ilustración 2-3 Herramientas de Software de Minería de Datos utilizadas en los años 2011 y 2012. Encuesta realizada por KDNuggets en mayo del 2012.

- SPSS para Windows en sus versiones 15.0 y 20. El SPSS, acrónimo de Statistical Package for the Social Sciences (Paquete Estadístico para las Ciencias Sociales), es un instrumento de análisis multivariante de datos cuantitativos que está diseñado para el manejo de datos estadísticos. [67]

En el caso del análisis en los diabéticos tipo 2 son aprovechables las ventajas que brinda al cumplir con todas las fases que implica un análisis de datos (planificación, elaboración de una base de datos, preparación de estos, análisis de los mismos y elaboración de un informe), permitiendo así un análisis integral de los datos [67]. Otra ventaja es su capacidad de trabajar con grandes volúmenes de datos, como es el caso de esta investigación. Otros motivos para su elección son su utilidad para realizar análisis de clusters y la implementación del Cluster Bietápico, técnica de agrupamiento ausente en otras herramientas.

- WEKA (Waikato Environment for Knowledge Analysis) en su versión 3.7.5, fue desarrollado en la Universidad de Waikato, Nueva Zelanda para la exploración de datos a partir de distintas funcionalidades [68].

  El ser de libre distribución, lo que permite actualizar su código fuente para incorporar nuevas utilidades o modificar las ya existentes; el ser una de las suites más utilizadas en la MD durante los últimos años; el tener implementaciones de varios algoritmos de clustering de interés para la investigación, entre ellos “K medias” y “EM”; unido a las facilidades que brinda para experimentar y realizar análisis la convierten en una herramienta útil para realizar KDD en esta investigación.

- MATLAB (abreviatura de MATrix LABoratory, "laboratorio de matrices") en su versión R2011b, es un software matemático que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio (lenguaje M). Está disponible para las plataformas Unix, Windows y Apple Mac OS X.

  Entre sus prestaciones básicas se hallan: la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario (GUI) y la comunicación con programas en otros lenguajes y con otros dispositivos hardware. Este se utiliza para la validación de los modelos de clusters resultantes de los algoritmos utilizados, ya que se cuenta con los índices de validación implementados en el lenguaje M.

Al finalizar esta fase se cuenta con ideas claras acerca de hacia dónde dirigir la minería de datos y de aquellas herramientas que son útiles en el proceso. Dando paso a la etapa de preprocesamiento de los datos.

2.3.2. Preprocesamiento de los datos.

Esta etapa 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 "Preprocesado de los datos" es la más engorrosa y costosa, en todo proceso de análisis de datos [48]. Por ser esta tan abarcadora y compleja CRISP-DM la divide en dos fases más simples Comprensión de datos y Preparación de datos.

2.3.2.1. Fase II: Comprensión de datos.

Durante la Comprensión de los datos se obtiene una visión más realista del conjunto de datos del que se extrae el conocimiento. Esto es posible mediante la ejecución de tareas como: la recolección de los datos iniciales, describir y explorar los datos y verificar su calidad. [69], [70], [71]

2.3.2.1.1. Recolección de datos iniciales.

Una vez identificadas las líneas a seguir el primer paso es obtener los datos de los pacientes. Esto se hace a partir de una hoja de cálculo de Microsoft Office Excel 2003 donde están registradas 77 características, correspondientes a 1951 pacientes, recogidas de las historias clínicas (anexo 1) de los diabéticos que han ingresado en el CAED.

2.3.2.1.2. Descripción de los datos.

El segundo paso es describir estos de modo que se pueda identificar de manera general las características del conjunto de pacientes.

La descripción de las variables se hizo de acuerdo a una tabla, propuesta por [48], que se muestra en el anexo 2. De cada campo se especifica el identificador, una breve descripción de la información que aporta a los médicos, el tipo de datos (Booleano, Nominal, Numérico) y la importancia del atributo para la investigación (Relevante y Sin importancia). En la tabla 2-1 se muestran algunas variables a modo de ejemplo.

La información referente al tipo de datos es necesaria para la adecuación de los datos a las herramientas que se utilizan en la MD. En cuanto a la relevancia de la variable para la investigación, este análisis se hizo necesario porque los médicos registran en la base de casos variables útiles al seguimiento y a la atención especializada que brindan a los pacientes. Como el objetivo de minería de datos es identificar clases útiles al diagnóstico temprano de la DM tipo 2, solo se seleccionan aquellas variables que contribuyen a este objetivo, para lo cual se tuvo en cuenta el criterio de los médicos.

Identificador	Descripción	Tipo	Relevancia
Nombres y Apellidos	Nombre y Apellidos del paciente	Nominal	Sin importancia
Edad	Edad en años del paciente al momento de ingresar	Numérico	Relevante
HDL-c	Refleja la medida de la cantidad de colesterol del tipo HDL del paciente.	Numérico	Sin importancia

Tabla 2-1 Ejemplo de la descripción de las variables

Del total de las variables se identificaron 37 atributos que tienen relevancia para la investigación, básicamente son aquellos factores que los médicos identifican como de riesgo, algunos análisis que son comunes en la atención primaria y las complicaciones que presentan los pacientes. Las 40 restantes no tienen importancia para los objetivos del proceso de MD. Entre estas se encuentran variables que no aportan información sobre el debut y aquellas que son segundos valores de una misma variable (peso final, glicemia final, ppd final).El caso específico de la variable HDL-c es considerada sin importancia, aunque es relevante para los expertos por los problemas de calidad que serán posteriormente explicados.

2.3.2.1.3. Exploración de los datos.

La descripción de los datos permite tener nociones globales sobre la muestra. Un paso que profundiza en el conocimiento de los datos es la exploración. Durante esta tarea se utilizan interrogantes de minería, que son solucionadas mediante la gráfica y visualización de los datos, para realizar una mejor descripción de los ellos.

Para el desarrollo de la exploración se realizan análisis de visualización, a través de plotear y graficar algunas variables utilizando las herramientas seleccionadas y las opciones que brindan para esto.

En el Excel se aprovechan sus opciones para graficar. La opción del Weka 3.7.5 *"Plot"* permite plotear las variables y hacer análisis preliminares sobre sus relaciones.

Con el objetivo de comprobar la concordancia de las características del conjunto de datos con lo que se expone en la teoría y la práctica médicas se analiza el comportamiento de algunas variables en los datos. Para ello se plotean los valores del atributo "tipo_diabetes", seleccionándolo tanto por el eje "X" como por el eje "Y". Para ver mejor la distribución de los puntos obscuros en la gráfica se desplaza al máximo la opción "Jitter", la cual ejecuta un desplazamiento aleatorio para cada uno de los puntos de la gráfica de modo que se puedan determinar mejor los colores predominantes. Por último, se varía el parámetro "color" (colour) con cada una de las variables de interés para analizar su influencia en el tipo de diabetes.

El primer análisis se hizo con la variable "Edad". En las figuras 2-4 y 2-5 se muestran las gráficas de este análisis, donde se observa un predominio de un rango de edad entre los 40 y los 69 años. Esto concuerda con el criterio de experto que plantea que la mayoría de las personas que padecen diabetes tipo 2 son mayores de 40 años [20], [21].

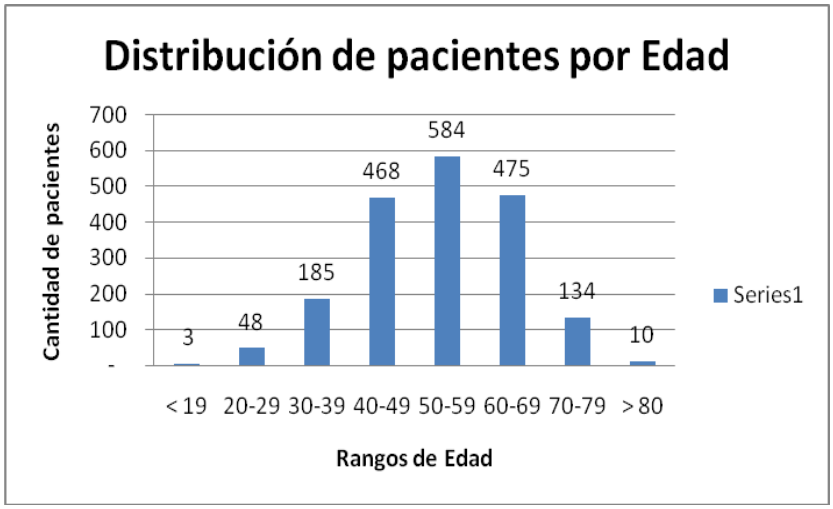


Ilustración 2-4 Distribución de los pacientes diabéticos por rangos utilizando Excel.

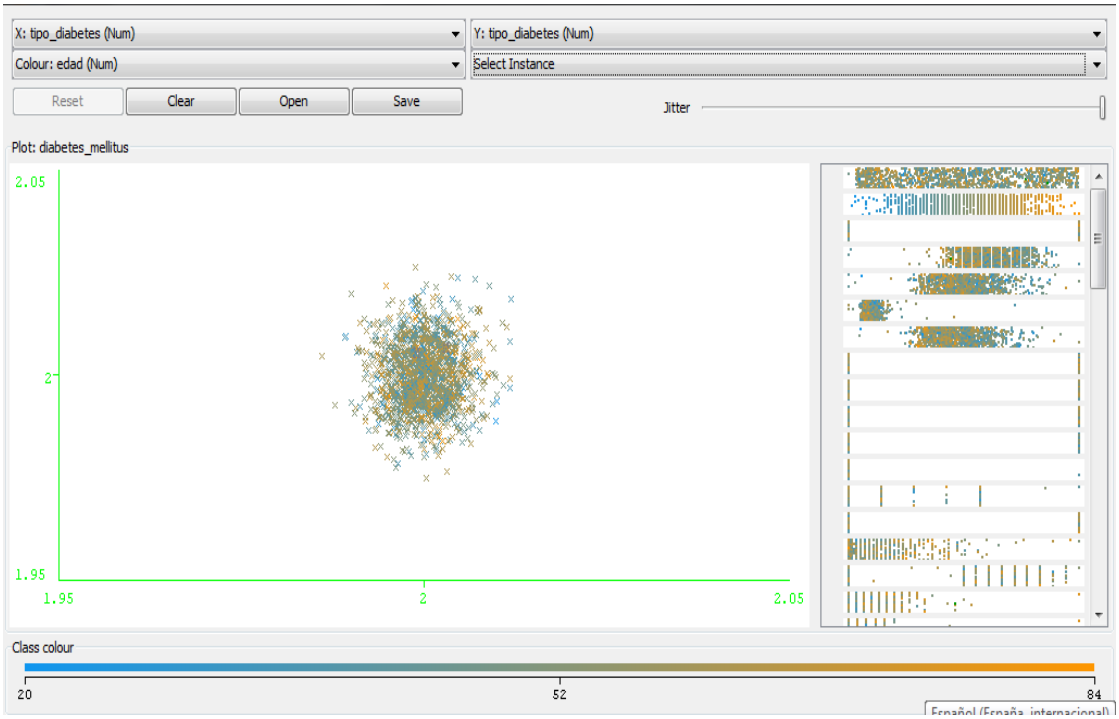


Ilustración 2-5- Relación de edades en los pacientes con DM tipo 2 utilizando Weka.

Otros análisis, se hicieron con las variables referentes al: sexo, hábito de fumar, la obesidad, los antecedentes familiares de DM y el padecimiento de hipertensión arterial (HTA) pues son factores de riesgo de padecer esta enfermedad. Para su estudio se generaron gráficas como las anteriores que se muestran en los anexos del 3 al 7.

Los análisis resultantes arrojan que existe un predominio del sexo femenino lo que es corroborado por la teoría médica, puesto que existen situaciones obstétricas desfavorables que constituye factores de riesgo en las mujeres [21] y aumenta su propensión a padecer la enfermedad.

En cuanto al hábito tóxico de fumar no se observa en los datos que abunden los fumadores, aunque los malos hábitos son un aspecto preocupante para los médicos.

La obesidad, los antecedentes familiares de diabetes mellitus y el padecimiento de HTA son rasgos que predominan en el conjunto de pacientes. Esto concuerda con el criterio médico de que estos son factores de riesgo de padecer diabetes mellitus tipo 2. [21]

Al concluir esta tarea se evidencia una concordancia entre el comportamiento de las variables en los datos y lo que la teoría médica expresa.

2.3.2.1.4. Verificación de la calidad de los datos.

Concluir las tareas de caracterización y exploración de los datos y demostrar la correspondencia del conjunto de datos con la teoría de las ciencias médicas es un

resultado alentador en el proceso de minería. No obstante, se hace necesario verificar la calidad de los mismos para aplicarles técnicas inteligentes en próximas fases y proponer en la medida de lo posible soluciones que mejoren la misma.

Pese a los resultados antes obtenidos, al analizar la calidad de los datos se identifican varios problemas. De primera instancia se nota que existen valores perdidos en la muestra. En algunos casos la cantidad es considerable como en los años 2005 y 2006 que variables como “Modo de debut”, “Obesidad al debut”, “Antecedentes obstétricos”, “APF DM”, “APP”, “APF Otras”, “Complicaciones” y “Pie con riesgo” están completamente en blanco. Por esta razón se omiten los 65 pacientes que ingresaron en estos años. Quedando 1887 pacientes restantes.

En cuanto a las variables, “HDL-c” solo es tomada en el año 2010 en algunos pacientes, por ser un análisis cuyo reactivo escasea con frecuencia, esto representa el 2% de la muestra y por tanto se descarta para la investigación. En una situación similar se encuentra “Reingreso” ya que solo tienen valores en esta variable 5 pacientes de los años 2005 y 2006 para un 0.26% del total.

En la variable “Sexo” se identificaron pacientes con nombres de mujer y con valores en los antecedentes obstétricos, que tenían especificado sexo masculino. Las historias clínicas de estos pacientes son 680, 925,981, 1282 y se les sustituye el valor del sexo por femenino.

Uno de los problemas que se detecta en las variables continuas, por los errores propios que genera la manera de registrar los datos, es la forma de delimitar las cifras decimales, ya que estas se delimitan en algunas ocasiones con una coma y otras con un punto. Esto, sumado a dejar espacios innecesarios entre la unidad y las cifras decimales. Para solucionar los problemas de calidad anteriores se corrigen sustrayendo los espacios y estableciendo como carácter delimitador de números decimales: el punto.

En el caso específico de la talla se encuentran 23 celdas con algunas imprecisiones, las cuales fueron solucionadas sustituyendo los valores por aquellos que se consideraron lógicos. En la tabla 2-2 se ejemplifica este proceso, todas las sustituciones se muestran en el anexo 8.

Historia Clínica	Talla	Talla sustituida
469	1.64.5	1.64
565	!,70	1.70
2213	177	1.77

Tabla 2-2- Ejemplo de la sustitución de valores imprecisos en la variable talla.

En las variables relacionadas con el peso también se encontraron imprecisiones. En peso inicial y peso final se observan valores incoherentes, los cuales fueron sustituidos como se muestra en las tablas 2-3 y 2-4, teniendo como referencia la variable "IMC". Todos los valores sustituidos se exponen en los anexos 9 y 10.

Historia Clínica	Peso inicial	Peso inicial sustituido	IMC
1329	10.1	101	33.7
1703	1.63	78.9	29.7

Tabla 2-3 Ejemplo de la sustitución de valores imprecisos en la variable peso inicial.

Historia Clínica	Peso final	Peso final Sustituido
1329	10.1	101
1513	94 1/2	94.5

Tabla 2-4 Ejemplo de la sustitución de valores imprecisos en la variable peso final.

En el caso de la variable "IMC" (Índice de Masa Corporal) existen 11 pacientes que tienen valores fuera del rango lógico, lo que representa un 0.56% del total de los datos. Para corregir esto se recalcula el IMC (anexo 11) para dichos casos mediante la fórmula:

$$IMC = \frac{masa(Kg)}{estatura^2(m)}$$

Otro problema en los datos, específicamente en las variables nominales es que se utilizan varios términos para describir un mismo valor. La solución a este problema se describe en la próxima fase.

Por último, se evidencia una confusión en las columnas asociadas a la "Microalbuminuria", puesto que en la primera columna se ubica el valor del resultado de

dicho análisis y en la segunda si este es positivo o negativo, no siendo así para todos los pacientes.

Al concluir la fase “Comprensión de los datos” se cuenta con una idea clara de la información contenida en los datos y de las dificultades que estos presentan. Haciéndose posible la corrección de algunas de estas dificultades, con el propósito de mejorar la calidad de los datos. No obstante a estos problemas, es relevante la correspondencia entre el comportamiento en los datos y la teoría de las ciencias médicas, de las variables consideradas significativas en estos pacientes.

2.3.2.2. Fase III: Preparación de los datos.

Una vez conscientes de la condición de los datos se requiere realizar acciones concretas para alistarlos para la etapa de minería. Estas funciones se cumplen durante la etapa “Preparación de los datos” a través de realizar tareas como seleccionar, limpiar, construir, integrar y formatear los de datos. [72], [73], [74]

2.3.2.2.1. Selección de los datos.

En este punto del preprocesamiento donde ya se han identificado problemas de calidad y se conocen las características de las variables, se está en condiciones de decidir qué datos serán usados para el análisis. Los criterios para esta selección incluyen: la importancia para el cumplimiento de los objetivos de la minería de datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Este proceso incluye no solo la selección de las variables sino también de los casos.

Las variables que se decide usar para la investigación son aquellas que en la descripción de los datos se clasificaron como relevantes. Estas serán utilizadas en la etapa de modelado para determinar posibles clases dentro de los pacientes diabéticos tipo 2.

Por tanto las variables que se usarán en la próxima fase son: “Historia Clínica” (como identificador), “Edad”, “Sexo”, “Talla”, “Peso inicial”, “IMC”, “Hábitos tóxicos(Fuma)”, “Hábitos tóxicos(Café)”, “Hábitos tóxicos (Bebidas Alcohólicas)”, “Hábitos tóxicos(Otros)”, “Obesidad al debut”, “Antecedentes obstétricos (Menarca)”, “Antecedentes obstétricos (Embarazos)”, “Antecedentes obstétricos (Abortos)”, “Antecedentes obstétricos (Malformaciones)”, “Antecedentes obstétricos (Macrofetos)”, “Antecedentes obstétricos (Muerte perinatal)”, “Antecedentes obstétricos (Menopausia)”, “APF DM”, “APP”, “APF otros”, “TGP”, “Glicemia (Inicio)”, “Glicemia (Postpandrial)”,

"Creatinina", "Microalbuminuria (1)", "Eritro", "Hemoglobina", "Triglicérido", "Acido úrico", "Colesterol", "Complicación micro (RD)", "Complicación micro (ND)", "Complicación micro (Neuro.D)", "Complicación macro (CI)", "Complicación macro (AVE)", "Complicación macro (IAP)", "Tipo de diabético".

De los 1887 casos que quedan de los años 2007 al 2012 se descartan aquellos pacientes que presentan un tipo de diabetes diferente a la tipo 2, puesto que es objetivo de la investigación trabajar con los pacientes que padecen este tipo de diabetes. Se eliminaron 159 filas lo que representa un 8,42% del total de los datos.

Algunas Técnicas de IA son sensibles a los valores perdidos por esta razón se analizó la cantidad de valores perdidos que presentaba cada paciente. En la muestra se identifican 43 filas que presentaban más de 10 valores perdidos. Por no ser esta una cantidad considerable con respecto al total de casos que se tiene se descartan, quedando un total de 1685 casos en este punto del proceso KDD.

2.3.2.2.2. Limpieza de los datos.

Durante la fase "Limpieza de los datos" se eleva la calidad de los datos seleccionados al nivel requerido por las técnicas de análisis. Para esto se solucionan los problemas de calidad descritos en la fase anterior.

Un problema de calidad importante es la cantidad de valores perdidos en la muestra, esto se evidencia en que solo 511 casos no presentan valores perdidos. Esto representa un 30,33% de los casos seleccionados. Es necesario por tanto, mejorar la calidad de los datos restantes para no descartarlos, teniendo en cuenta la necesidad de las técnicas inteligentes de tener una cantidad considerable de ejemplos para su entrenamiento y la sensibilidad de algunas de ellas a los datos perdidos.

Algunas herramientas tienen la opción de remplazar los espacios en blanco por la media o la moda de la variable. El SPSS tienen una opción llamada "*Analizar valores perdidos*" (Figura 2-6) que permite hacer una estimación de estos valores por varios métodos, lo que resulta en muchas ocasiones una mejor opción. Para este caso específico se selecciona el método "EM" (Maximización Esperada). En la figuras 2-7 se muestran las opciones seleccionadas para este método.

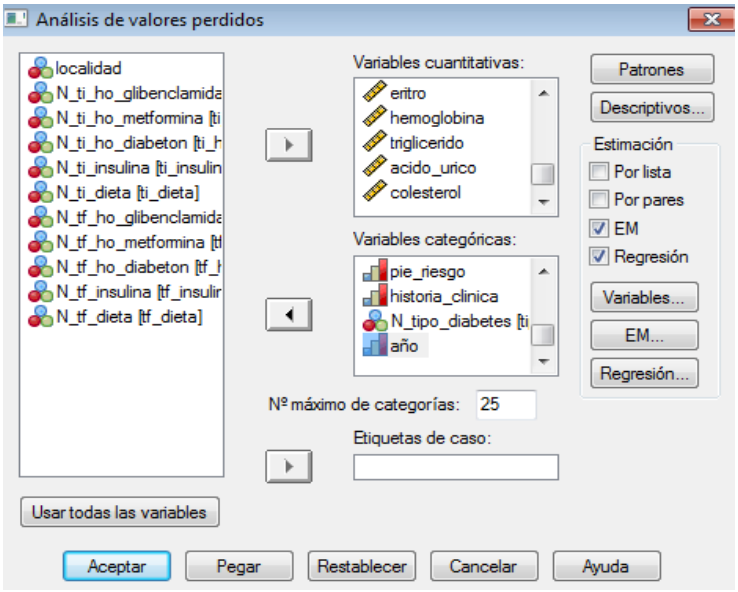


Ilustración 2-6 Ventana "Análisis de valores perdidos" del SPSS.

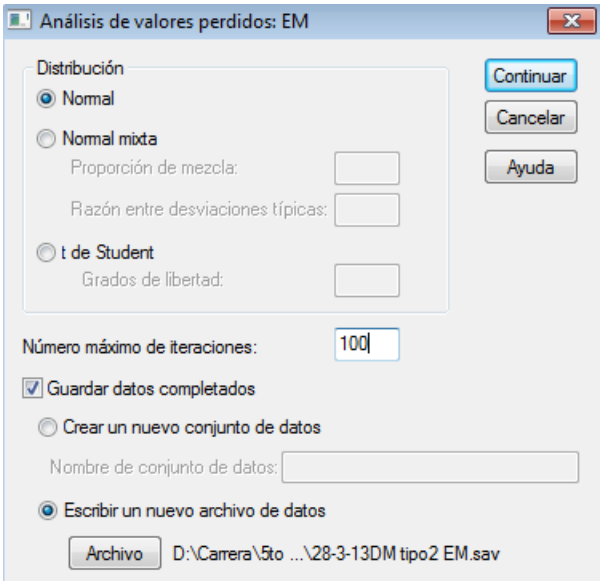


Ilustración 2-7 Ventana "Análisis de valores perdidos: EM" del SPSS.

Otro elemento que puede afectar el buen desempeño de las técnicas es la presencia en los datos de valores atípicos. Como se pretende hacer un análisis teniendo en cuenta el comportamiento típico de los diabéticos tipo 2, estos valores muy diferentes del comportamiento predominante de los datos se consideran como ruido y se descartan del análisis. Para estos se utiliza la opción "Identificar casos atípicos" del SPSS, como se muestra en la figura 2-8. Los resultados arrojan un total de 18 casos anómalos, los que se descartan.

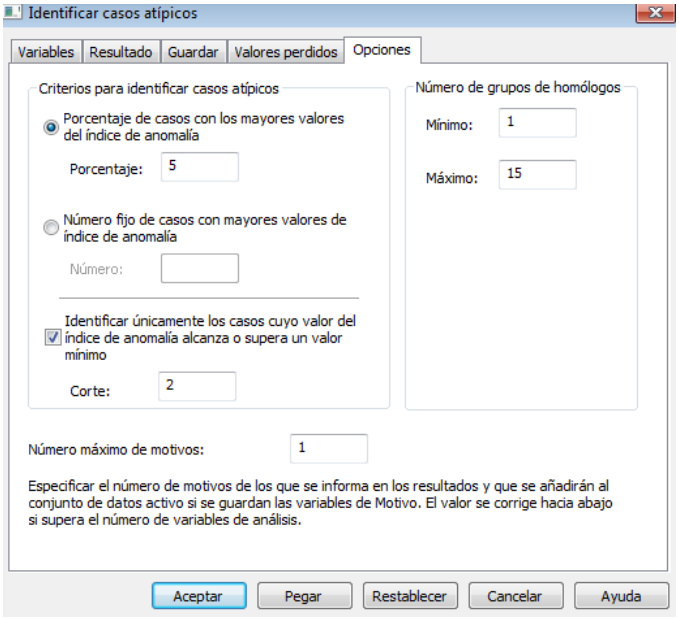


Ilustración 2-8 Ventana de opciones de la funcionalidad "Identificar valores atípicos del SPSS".

Finalmente, para nivelar el comportamiento de las variables se estandarizan los datos. En el SPSS se selecciona en el menú "*Analizar*" la opción "*Estadísticos descriptivos*", "*Descriptivos*", que se muestra en la figura 2-9. Donde se eligen las variables continuas y la opción "*Guardar valores tipificados como variables*", como resultado se crean nuevas variables estandarizadas.

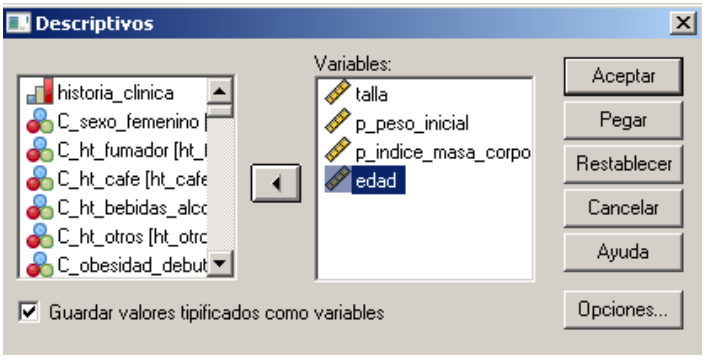


Ilustración 2-9 Ventana de la opción "Estadísticos Descriptivos", "Descriptivos".

En el Weka se hace mediante la opción del "*Explorer*", "*Filter*" (Filtro) donde se elige de los filtros no supervisados de atributos "*Standardize*" (Estandarizar) y luego se aplica al conjunto de datos. En ambos casos el resultado es, variables continuas con media 0 y desviación estándar 1.

En este punto se cuenta con un conjunto de datos con una mejor calidad para aplicarles las técnicas de clustering durante la fase de modelado.

2.3.2.2.3. Construcción de los datos.

Después de todo lo que se ha hecho aún se encuentran en los datos dificultades para la aplicación de técnicas de IA. Tal es el caso de variables nominales cuyos valores tienen múltiples combinaciones. Modificar esta situación codificando valores y creando nuevas variables son funciones que se llevan a cabo en "Construir los datos."

Una de esas variables es "APF DM" que contiene aquellos familiares del paciente que padecen diabetes mellitus, lo que supone existan disímiles combinaciones de valores dentro de la muestra que dificultan el análisis. El factor de riesgo consiste en que el paciente presente antecedentes familiares de DM [21], no en quien específicamente la padece. Por eso para la investigación es suficiente conocer si el paciente tiene o no familiares que padecen la enfermedad. Una transformación que soluciona el problema y brinda la información necesaria es codificar la variable para que tome solo dos valores:

1: El paciente tiene al menos un familiar que padece DM.

0: El paciente no tiene ningún familiar que padezca esta enfermedad.

Otras variables que presentan dificultades similares son las relacionadas con las patologías que padece el paciente ("APP") y las que padecen su familia ("APF"), a parte de la DM. Estas contienen para cada paciente un listado de estas enfermedades, que como en el caso anterior dificulta la correcta interpretación de la información por parte de los algoritmos inteligentes. Crear nuevas variables dicotómicas, para cada una de las enfermedades que propone la historia clínica (anexo 1), y agrupar en una variable el resto es la solución determinada para este problema.

Las variables que se crean son "APP HTA" (Hipertensión Arterial), "APP hiperlipoproteinemia", "APP cardiopatía isquémica", "APP clauditación intermitente" y "APP otros" y las mismas enfermedades pero para los Antecedentes Patológicos Familiares (APF). Cada una de estas variables toma valor 1 si el paciente o su familiar la padece, sino toma valor 0. La variable otros es igual 1 si el paciente o su familia padecen otra enfermedad diferente a las que se definieron anteriormente. La descripción de estas variables y la justificación de su inclusión se pueden ver en el anexo 12.

2.3.2.2.4. Formateo de los datos.

Un último paso en el preprocesamiento es la adecuación de los datos a los requerimientos sintácticos de las herramientas a utilizar en la etapa de modelado, SPSS y Weka. Esto se lleva a cabo en la tarea "Formateo de los datos"

Un primer paso consistió en codificar las variables booleanas y nominales, asignando un número natural a cada uno de los posibles valores. A las booleanas se les asigna 1 a los valores positivos y 0 a los negativos. En cuanto a la variable "Sexo" se codifican sus valores como sigue:

Masculino: 0

Femenino: 1

Para el SPSS los datos se introducen directamente desde el fichero xls con las opciones de Copiar y Pegar, teniendo en cuenta que las columnas especifican variables y las filas casos. Dos requerimientos de esta herramienta son que los números decimales deben estar separados por coma y los valores perdidos se especifican dejando las celdas en blanco.

La herramienta Weka carga los datos desde un fichero con formato ARFF (Attribute-Relation File Format.) Para crear este archivo los datos se guardan en formato csv (delimitado por coma). Es necesario tener en cuenta que cada línea representa un caso y la coma separa los valores de las variables. La estructura básica de este tipo de archivos es la siguiente:

```
% Comentarios
@RELATION Nombre de la base de datos
@ATTRIBUTE r1 REAL
@ ATTRIBUTE r2 REAL
...
@ ATTRIBUTE i1 INTEGER
@ ATTRIBUTE i2 INTEGER
…
@ ATTRIBUTE s1 {v1_s1, v2_s1,…vn_s1}
@ ATTRIBUTE s2 {v1_s1, v2_s1,…vn_s1}
…
@DATA
Datos
```

Algunos de los requerimientos del Weka son: los números decimales deben estar separados por punto, los valores perdidos se indican con un signo de interrogación "?", no deben existir espacios vacíos entre un valor y otro y el orden de los valores debe coincidir con el orden especificado en la definición de los atributos.

Teniendo en cuenta las especificaciones anteriores se generan dos ficheros: uno con extensión arff para el Weka y otro con extensión csv para el SPSS.

2.4. Conclusiones parciales.

El desarrollo de las tres primeras fases de CRISP-DM, que a su vez corresponden con las dos primeras de la MD, provee un conjunto de datos con la calidad y el formato requeridos por las herramientas de software, para ser explorados por las técnicas de agrupamiento.

El preprocesamiento, ofrece, no solo el conjunto de datos para el aprendizaje de los algoritmos de clustering, sino también el conjunto de ejemplos que, añadidas las clases, contribuirá al entrenamiento de las técnicas supervisadas en la creación de un SE útil a la prevención y el diagnóstico temprano de la DM tipo 2.

El conjunto de datos, conformado por 45 atributos y 1667 pacientes diabéticos tipo 2, presenta en sus variables un comportamiento que concuerda con lo que la medicina expresa al respecto.

3. Capítulo III: "Experimentación y análisis de los resultados."

3.1. Introducción.

Todo el esfuerzo realizado en comprender los datos, corregirlos y formatearlos ha sido teniendo como meta alistarlos para la fase "Minería de datos". Es la etapa de modelado, la cúspide en el proceso de descubrir el conocimiento. Es durante esta fase que los algoritmos "cavarán" y se adentrarán en los casos de los pacientes y extraerán el preciado producto: el conocimiento. Dígase en este caso, grupos de diabéticos tipo 2, que serán el completamiento de un conjunto de entrenamiento útil a la formación de un sistema experto, cuyo fin será contribuir al diagnóstico temprano y la prevención de esta enfermedad. En este capítulo se aborda la ejecución de los experimentos y se evalúan los resultados de la misma, desarrollando las últimas fases de CRISP-DM, llegando a conclusiones respecto al cumplimiento del objetivo de MD.

3.2. CRISP-DM. Fase IV: Modelado.

Hablar de MD tiene sentido cuando existe un punto en el proceso en el que se realiza la extracción del conocimiento de los datos. Es precisamente en la fase "Modelado" o "Minería de datos" donde se cumple esta función realizando tareas como: seleccionar las técnicas de modelado, generar la prueba de diseño, construir y evaluar el modelo. [75], [76], [77]

3.2.1. Selección de las técnicas de modelado.

"Selección de las técnicas de modelado" es la tarea que permite definir qué algoritmos del Weka y del SPSS se utilizarán para determinar clases en los diabéticos tipo 2.

Estas herramientas tienen implementaciones de varios algoritmos de clustering. Los utilizados en el agrupamiento de los pacientes con DM son:

- Simple EM (Expectation Maximization). Implementado en el Weka, el algoritmo "Maximización Esperada" es una de las opciones para el agrupamiento que brinda el "Explorer" en la pestaña "Cluster". Es una implementación del clásico "EM" que puede decidir cuántos clusters crear por el método "cross validation".

  Los parámetros que requiere para su ejecución son [78]:

  - Debug: Si es verdadero se muestra información adicional en la consola.
  - displayModelInOldFormat: Usa el antiguo formato para la salida del modelo. El formato antiguo es mejor cuando hay muchos grupos. El nuevo formato es mejor cuando hay pocos clusters y muchos atributos.

- maxIterations: Número máximo de iteraciones.
- minStdDev: Fija la mínima desviación estándar permitida.
- numClusters: Cantidad de clusters. El -1 indica seleccionar el número de clusters automáticamente por cross validation.
- Seed. La semilla a usar para la generación de números aleatorios.

- Simple KMeans. El "K Medias" del Weka es un algoritmo de clustering que puede ser usado con dos medidas de distancia la Euclidiana y la Manhattan. Es un algoritmo no supervisado que admite cualquier tipo de datos y tiene los siguientes parámetros [78]:
  - displayStdDevs. Muestra la desviación estándar de los atributos numéricos y cuenta los atributos nominales.
  - distanceFunction. La función de distancia a usar en la comparación de las instancias.
  - dontReplaceMissingValues. Remplazar los valores perdidos globalmente con la media/moda.
  - fastDistanceCalc. Usa los valores límites para acelerar el cálculo de las distancias.
  - initializeUsingKMeansPlusPlusMethod. Inicializa los centroides de los clusters usando el método probabilístico farthest first (el último primero) del algoritmo K medias++.
  - maxIterations. Fija el máximo número de iteraciones
  - numClusters. Cantidad de clusters.
  - preserveInstancesOrder. Preservar el orden de las instancias.
  - Seed. La semilla a usar para la generación de números aleatorios.
- Conglomerado Bietápico. Implementado por el SPSS el "Conglomerado en dos fases" es un algoritmo de análisis de clusters con escalabilidad diseñado para manejar conjuntos de datos muy grandes. Es capaz de manipular ambos tipos de variables o atributos, continuas y categóricas. En la primera fase del procedimiento, pre-agrupa los casos en muchos pequeños sub-clusters. Luego, agrupa los sub-clusters del paso anterior en el número de clusters deseado. Si el número de clusters deseado es desconocido, el "Conglomerado en dos fases" del SPSS

encontrará el número apropiado de clusters automáticamente utilizando dos criterios de conglomeración. Permite modificar las siguientes opciones [79]:

- Medida de distancia. Esta opción determina cómo se calcula la semejanza entre dos conglomerados. (Euclidiana, Log-verosimilitud).
- Número de conglomerados. Esta opción permite especificar cómo se va a determinar el número de conglomerados. (automáticamente o especificar un número fijo).
- Recuento de variables continuas. Este grupo proporciona un resumen de las especificaciones acerca de la tipificación de variables continuas realizadas en el cuadro de diálogo "*Opciones"*.
- Criterio de conglomeración. Esta opción determina cómo el algoritmo de conglomeración determina el número de conglomerados. Se puede especificar tanto el criterio de información bayesiano (BIC) como el criterio de información de Akaike (AIC).

- Conglomerado de K medias. La implementación en el SPSS de este algoritmo es útil cuando se dispone de un gran número de casos. Necesita que se le especifique la cantidad de clusters y utiliza para la aglomeración la distancia Euclidiana. Los parámetros que se le especifican son [79]:
  - Número de conglomerados. Este número no debe ser inferior a 2 ni superior al número de casos del archivo de datos.
  - Método. Puede ser iterar y clasificar o solo iterar.
  - Cantidad de iteraciones y criterio de convergencia.

De las técnicas anteriores, el "EM" y el "Conglomerado en dos fases" se utilizan para determinar la cantidad óptima de clusters en los diabéticos. Las dos implementaciones del "K medias" se usan por su utilidad para agrupar grandes cantidades de datos y porque las especificaciones de cada una pueden variar, modificando los resultados.

3.2.2. Generación de la prueba de diseño.

El resultado de las técnicas seleccionas son grupos de pacientes. Algunos de los cuales estarán mejor conformados que otros. Una manera de determinarlo es aplicando a los grupos formados los índices de validación explicados en el epígrafe 1.9.4.

Los índices se tienen implementados en ficheros de tipo M, que son interpretados por la herramienta MatLab. La función que ejecuta los índices internos tiene como parámetros

principales: una matriz de datos, donde cada columna representa una variable y un vector con las etiquetas de las clases conformadas por los algoritmos.

Una vez calculados los índices, se considera que un algoritmo es el mejor, si tiene la mayor cantidad de índices con mejores valores.

3.3. Experimentación con las técnicas seleccionadas y los datos.

Con las técnicas seleccionadas y el mecanismo de prueba establecido, el próximo paso en CRISP-DM es ejecutar las técnicas, para luego determinar la calidad de los resultados desde una mirada técnica. Este proceso se lleva a cabo en las tareas "Construcción y Evaluación del modelo", que forman parte de la fase "Modelado".

En el proceso de determinar clases en los pacientes diabéticos tipo 2 estas tareas se desarrollan a la par durante la experimentación, que consta de un primer análisis general seguido de otros análisis específicos.

3.3.1. Experimentación con el total de datos.

En el primer experimento se aplican los algoritmos de clustering a todo el conjunto de datos resultante del preprocesamiento.

El primer algoritmo es el "Conglomerado en dos fases", para su ejecución es necesario distinguir entre las variables categóricas y las continuas. Una vez seleccionadas las variables, la medida de distancia a utilizar es "Log-verosimilitud", pues la distancia Euclidiana no permite el uso de variables categóricas. Se elige como forma de determinar el número de conglomerados, la automática para un máximo de 15 clusters y como criterio de conglomeración el "BIC" (al usar el "AIC" como criterio de conglomeración los resultados son los mismos). Estos parámetros se ajustan en la ventana "Análisis de conglomerados en dos fases" del SPSS que se muestra en la figura 3-1. Otras opciones que se modificaron se encuentran en la figura 3-2, donde se establece a un 10% el tratamiento de ruido.

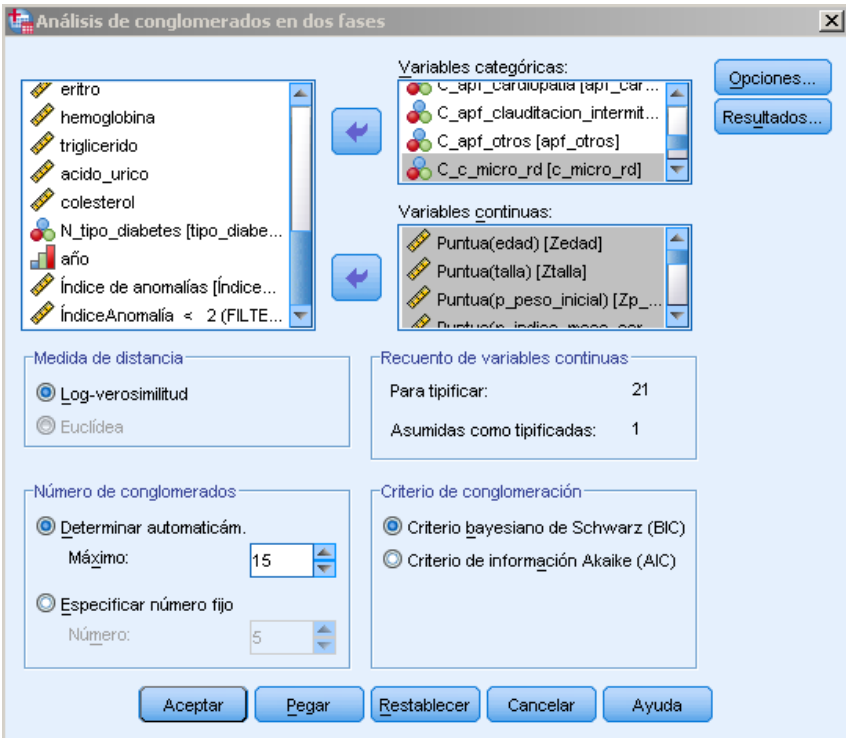


Ilustración 3-1 Ventana "Análisis del conglomerado en dos fases" del SPSS.

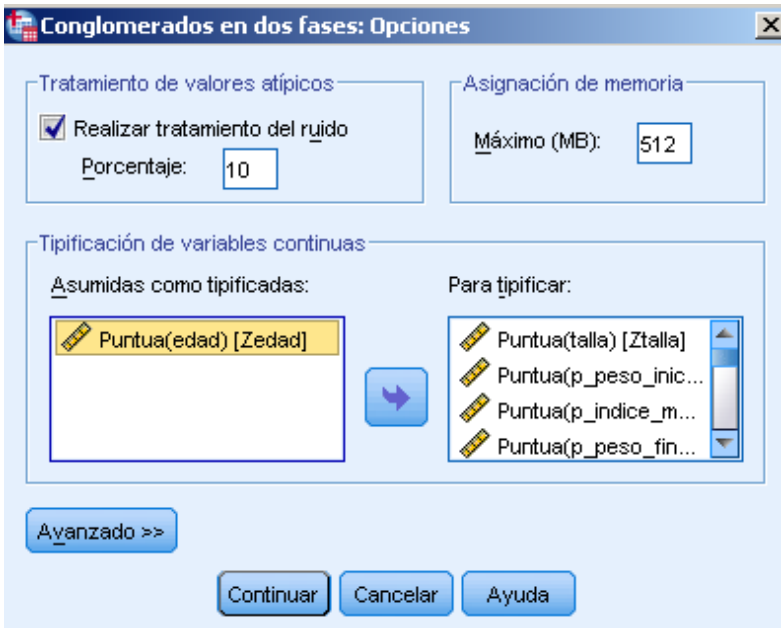


Ilustración 3-2 Ventana de opciones del "Conglomerado en dos fases".

El resultado de aplicar este algoritmo a los datos es la conformación de dos conglomerados con una distribución de 669 casos en el cluster 1 y 645 en el 2 (figura 3-3), el resto de los casos el algoritmo los excluye por presentar valores perdidos. El proceso de selección automática de la cantidad de clusters se muestra en el anexo 13.

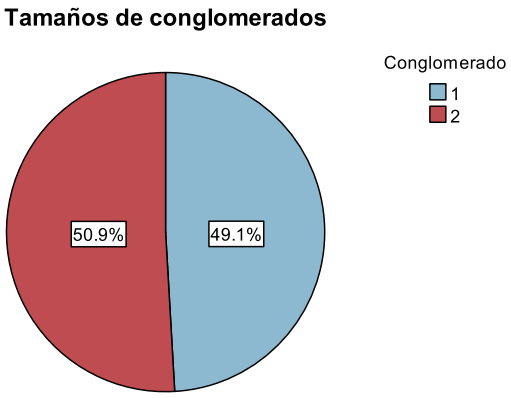


Ilustración 3-3 Tamaño de los conglomerados formados por el algoritmo de "Conglomerado en dos fases".

Dos grupos de pacientes fue el resultado de aplicar el "EM" en el Weka (ver anexo 14) con los parámetros de la figura 3-4. Los grupos quedaron conformados por 936 casos, el primero y 731, el segundo.

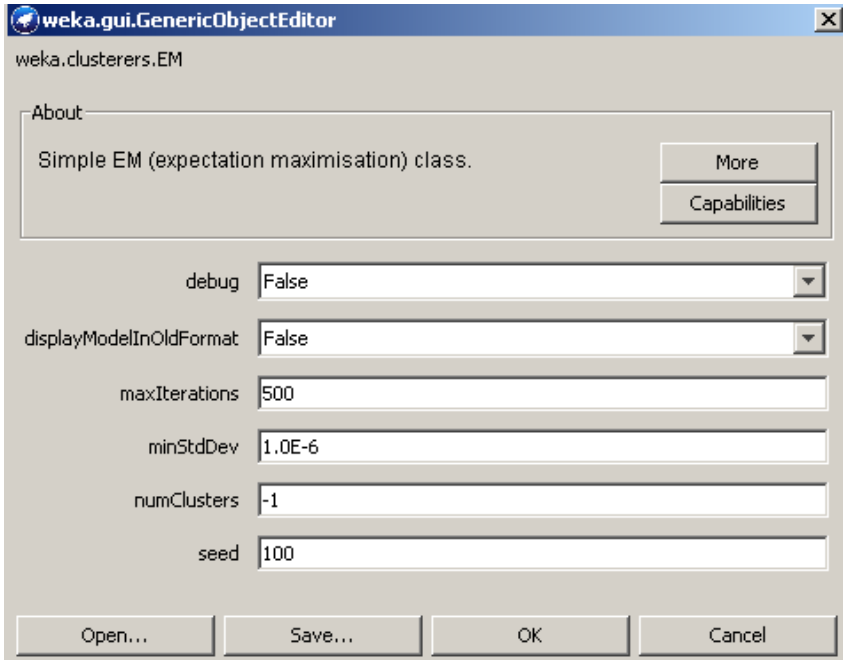


Ilustración 3-4 Editor de los parámetros del "Simple EM".

Al aplicarle a los datos el algoritmo "K medias" del Weka se ajustan los parámetros como se muestra en la figuras 3-5. Se aplica este algoritmo para valores de k entre 2 y 4 y se utiliza además la función de distancia Manhattan. Los resultados se muestran en los anexos del 15 al 20.

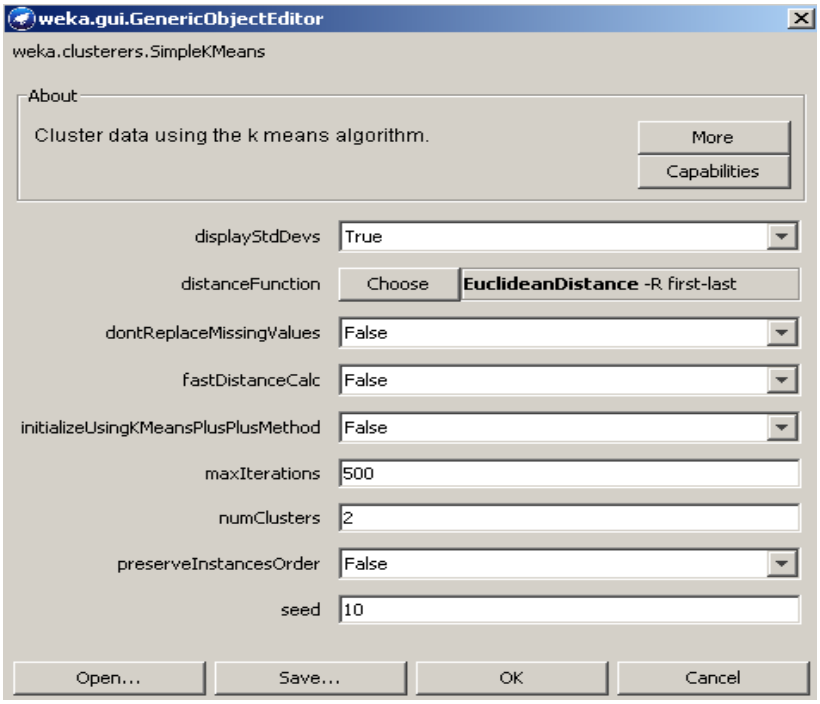


Ilustración 3-5 Editor de los parámetros del "Simple K means".

Al ejecutar el K medias del SPSS seleccionando las opciones de la figura 3-6 y variando k entre 2 y 4, en la distribución de los casos se observa que existen 6 pacientes que se asignan a un grupo en cada una de las variantes (ver anexo 21). Esto provoca que la distribución de casos esté desbalanceada. Para solucionar esto se retiran los pacientes del análisis. Los resultados se muestran en los anexos del 22 al 24.

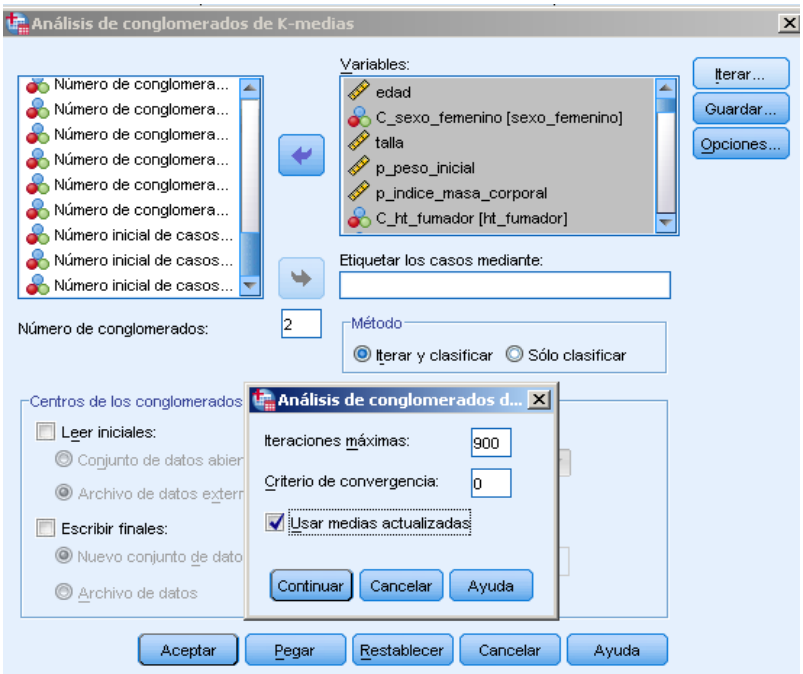


Ilustración 3-6 Ventana "Análisis del conglomerado K medias" del SPSS.

Un resultado importante en la conformación de los grupos son los centroides o prototipos de los clusters. Estos caracterizan el comportamiento promedio de cada grupo. Los centroides de los grupos de diabéticos, conformados por los algoritmos, se exponen en el anexo 25.

Finalmente, en este primer experimento se le aplican los índices de validación a los clusters resultantes de cada algoritmo y se comparan para la selección del mejor. (Anexo 26)

En cada índice se hace una comparación para determinar qué algoritmo tiene el mejor valor de acuerdo a sus especificaciones. Esta comparación se hace entre los algoritmos de cada herramienta por separado.

Los algoritmo que maximizan el índice Ball son el “EM” del Weka y el “K medias”, para k=2, del SPSS. El “Conglomerado en dos fases” y el “K means” del Weka, para k=2, maximizan el Calinski Harabasz y minimizan el Dunn. En los índices RMSSTD y RS se destacan el “Conglomerado en dos fases” y el “K means” del Weka para k=3. Los K medias mencionados utilizan la distancia Euclidiana. De lo anterior se concluye que es el “Conglomerado en dos fases” el que tiene la mayor cantidad de índices con mejores valores, dígase CH; Dunn, RMSSTD, y RS, por lo que es seleccionado como el conjunto de grupos apropiado para proseguir la minería de datos.

El conglomerado en dos fases dividió los diabéticos en dos grupos. Al analizar los centroides, resalta que en un grupo se encuentran las mujeres y en otro los hombres. También en la gráfica que genera este algoritmo sobre la importancia de las variables en la conglomeración (Figura 3-7) se observa que dentro de las variables de mayor importancia se encuentran aquellas que son propias de las mujeres como menarca, embarazos y menopausia, siendo la segunda en importancia el sexo. Debido a que este agrupamiento no aporta información útil para el diagnóstico temprano de la diabetes y que existen diferencias marcadas entre los sexos, se divide el conjunto de datos en hombres y mujeres. A estos nuevos conjuntos se le aplican los algoritmos de clustering para obtener clases útiles para el diagnóstico temprano y la prevención de la DM.

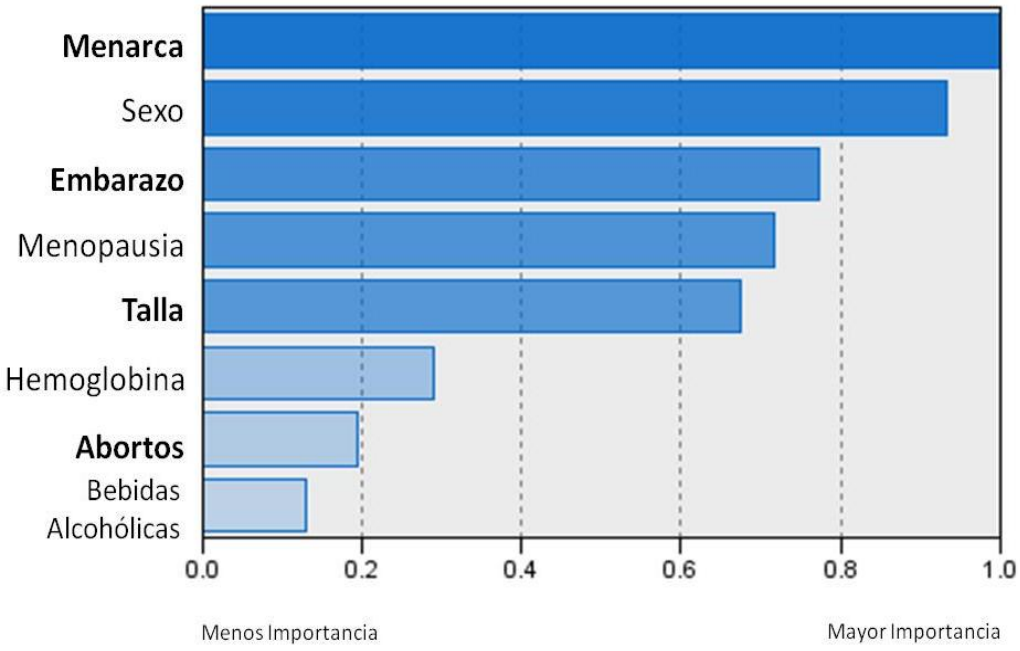


Ilustración 3-7 Gráfico "Importancia del predictor".

3.3.2. Experimentación con los datos de los hombres.

Para realizar el análisis de los hombres en el Weka se crea un nuevo fichero arff con los casos masculinos y se retiran las variables correspondientes a los antecedentes obstétricos. En el SPSS se utilizan las opciones de selección de casos para descartar las mujeres. Quedando un total de 712 hombres para el proceso de agrupamiento.

Se aplican a los datos los algoritmos "EM" y "Conglomerado bietápico" con el propósito de determinar la cantidad óptima de clusters. El "EM" conformó tres grupos y el "Conglomerado en dos fases" cuatro, los resultados se muestran en los anexos 27 y 28 respectivamente. Posteriormente, se aplican los "K medias" variando los valores de k entre 2 y 4 y utilizando la distancia Euclidiana, pues la Manhattan no tuvo buenos resultados en análisis anteriores. (Ver anexos 29-31).

Para validar la calidad de conglomeración se aplican los índices de igual forma que en el análisis general (anexo 32). Resultando los grupos formados por el "EM" y el "K medias" para k=4 los de mejor calidad, puesto que tienen los índices CH, Dunn, RMSSTD, RS con los mejores valores. Los centroides de los clusters formados por estos algoritmos se encuentran en el anexo 33.

3.3.3. Experimentación con los datos de las mujeres.

El experimento en las mujeres se hace con los 955 casos restantes y con el total de variables. De igual forma se ejecutan los algoritmos "EM" y "Conglomerado bietápico"

con los datos para identificar la cantidad de clusters. El resultado fue 8 grupos con el "EM" y 4 con el bietápico (anexos 34 y 35). Estos algoritmos también se ejecutan para fijándoles el valor de k a 3. En la ejecución de los "K medias" se varía el valor de k entre 2 y 4, y en el SPSS se corre este algoritmo para k=4 y k=8 (anexos 36-42).

El proceso de validación de los conglomerados resultantes se hace con los índices, como anteriormente. El anexo 43 muestra los valores de los índices para cada técnica. Resultando ser el "K means" del Weka para k=2 y el "K medias" del SPSS para k=8, los algoritmos con mayor cantidad de índices con mejores valores. El primero con CH, Dunn, RS y el segundo con CH, RMSSTD y RS. En el anexo 44 se pueden ver los prototipos de estos algoritmos.

3.4. Análisis de los resultados.

Finalizada la etapa de experimentación y la evaluación de los clusters resultantes, desde una mirada técnica, la minería de datos propone la fase "Consolidación del conocimiento". CRISP-DM desarrolla esta en dos etapas: "Evaluación" [81], [82], [83] y "Desarrollo" [84], [85], [86]. El reto de estas fases es comprobar cuán útiles son los modelos resultantes para la prevención y el diagnóstico temprano de la DM tipo 2 verificando el cumplimiento de los objetivos de la minería de datos. Este proceso se lleva a cabo durante el análisis de los resultados.

Durante la experimentación se dividió el conjunto de datos para aplicar las técnicas a los pacientes masculinos y a los femeninos, obteniéndose conglomerados en ambos grupos. Por esta razón los resultados se analizan de manera independiente en los conjuntos.

3.4.1. Análisis en los hombres.

De los mejores algoritmos seleccionados por los índices de validación para el conjunto de hombres el "EM" es el que propone la conglomeración más adecuada para el diagnóstico de la DM tipo 2, de acuerdo al criterio de expertos.

El "EM" conformó tres grupos distribuidos como se muestra en la figura 3-7. Obesidad, 50 años de edad como promedio, antecedentes familiares de diabetes mellitus e HTA son características comunes de los hombres de estos grupos. No obstante existen en los pacientes de cada uno de estos clusters diferencias significativas.

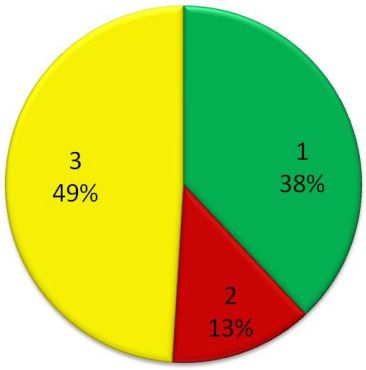


Ilustración 3-7 Distribución de los hombres en los grupos del "EM".

En el primer grupo se encuentran los hombres con un peso promedio de 86,27 Kg y un IMC de 30.17, lo que indica un alto nivel de obesidad. Además tienen el hábito tóxico de tomar café. Niveles de glicemia bajos y resultados de análisis médicos en niveles normales son características distintivas de los hombres de este cluster.

Un segundo grupo está conformado por pacientes tomadores de café, que con un peso promedio de 80 Kg y un IMC de 28.79, tienen glicemias por encima de los 10mmol/L y niveles de colesterol, triglicéridos y microalbuminuria muy altos.

El tercer grupo dentro de los hombres se compone de aquellos que con un IMC de 27,32, que implica niveles de obesidad bajos, y sin hábitos tóxicos presentan niveles de glicemia altos pero por debajo de los 10mmol/L. Además de tener el colesterol en niveles de riesgo y los triglicéridos un poco altos.

En los pacientes del grupo 1 los médicos observan que aunque son los hombres de mayor obesidad, el tener la glicemia y los análisis médicos en niveles normales son factores favorables para la situación clínica de los pacientes. Por el contrario en el grupo 2 se encuentran aquellos, que a consideración de los médicos, debutan tóxicos. Ya que niveles de glicemia por encima de los 10mmol/L implican la posibilidad de complicarse y de presentar afectaciones en los órganos, lo que se evidencia en el resto de los análisis. El grupo 3 caracteriza pacientes que presentan niveles de colesterol, triglicéridos y glicemias propios de una persona al debut, que aunque no están en los niveles saludables, tampoco lo están en niveles críticos. Por lo anterior se observa en la conglomeración diferentes niveles de riesgo en los pacientes, de complicarse o de tener afectaciones en los órganos.

Por tanto se proponen como clases en los hombres, para el conjunto de entrenamiento, tres niveles de riesgo: bajo, medio y alto. Que coinciden con los grupos 1, 3 y 2 respectivamente, propuestos por el "EM". (Anexo 45)

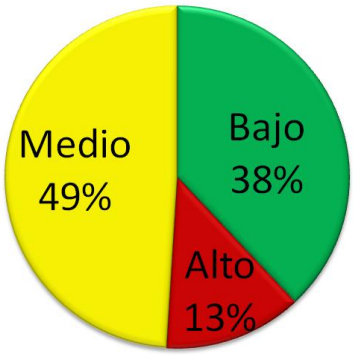


Ilustración 3-8 Clases en el conjunto de hombres.

Una vez identificadas las clases cabe preguntarse qué seguimiento proponen los expertos otorgar a las personas en cada uno de estos grupos.

3.4.1.1. Procedimiento recomendado por los médicos para cada grupo de hombres.

Debido a que el seguimiento de un paciente depende de su situación clínica los médicos proponen actuar de las siguientes formas para cada uno de los grupos:

- Grupo de riesgo bajo. Estas personas se encuentran con un cuadro cínico favorable, aunque factores agravantes son su obesidad, la HTA y los antecedentes familiares de DM. El primer paso consiste en realizar una prueba de tolerancia a la glucosa. El segundo paso es modificar los estilos de vida del paciente, para lo que se le asigna una dieta y un plan de ejercicios físicos. Estas personas deben tener seguimiento con el fin de observar el comportamiento de los factores de riesgo.
- Grupo de riesgo medio. Los hombres de este grupo ya debutaron con diabetes, pero al tener niveles de glucosa por debajo de los 10mmol/L los médicos proponen modificar su estilo de vida y, en caso de ser necesario, indicar un hipoglicemiante oral.
- Grupo de riesgo alto. A estos pacientes con un cuadro clínico tóxico los médicos no solo le modifican su estilo de vida, mediante una dieta y ejercicios, sino que le indican tratamiento medicamentoso. Este consiste, en muchos casos, en la indicación de insulina, durante una primera etapa, para lograr la estabilidad del

paciente. Posterior a esto la persona puede mantenerse con hipoglicemiantes orales.

3.4.2. Análisis en las mujeres.

En los experimentos realizados en las féminas la mejor calidad, de acuerdo a los índices de validación, la tuvieron el “K means” del Weka para k=2 y el “K medias” de SPSS para k=8.

Al analizar las conglomeraciones de conjunto con los expertos, la conformación de 8 grupos resulta inadecuada por ser muchas clases. Esto sumado a que la distribución de casos, producto a la cantidad de grupos, queda desequilibrada conlleva a descartar el agrupamiento.

El agrupamiento del “K means”, pese a su buena calidad técnica, no refleja características distintivas relevantes para los médicos entre los grupos, y por esta razón se desecha. Por su parte el “Conglomerado en dos fases” es el próximo algoritmo que presenta mejores resultados en los índices (el “EM” también se descarta por proponer 8 grupos). Al analizar los clusters conformados por esta técnica (anexo 46), desde el punto de vista de los expertos, se encuentran características interesantes en los diferentes grupos de mujeres.

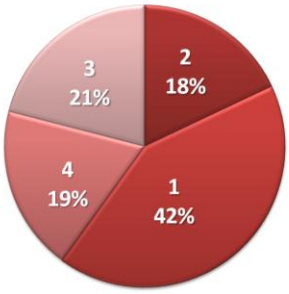


Ilustración 3-9 Distribución de mujeres por el “Conglomerado bietápico”.

Este algoritmo conforma cuatro grupos de mujeres distribuidos de acuerdo a la figura 3-9. Es interesante que en ellos las glicemias se encuentren entre 7 y 10mmol/L, lo que es considerado por los médicos como glicemias altas, aunque no en niveles críticos. Otra característica común es que las pacientes tienen antecedentes familiares de diabetes.

En el grupo 1 se encuentran las mujeres que con 74,52Kg de peso tienen un IMC de 30,52, lo que indica niveles de obesidad elevados. Estas mujeres, muy obesas, toman café, son hipertensas y tienen los triglicéridos y el colesterol un poco altos.

El segundo grupo, identificado por la técnica, está formado por las mujeres de 47 años de edad aproximadamente, con 82,4 Kg de peso promedio y un IMC de 33,48. Lo que indica mujeres jóvenes muy obesas. Otras características interesantes son que padecen HTA y tienen los triglicéridos y la microalbuminuria altos.

El tercer grupo lo constituyen mujeres de 56 años, 64Kg y un IMC de 26,52. Estas pacientes no son obesas, no tienen hábitos tóxicos, no padecen HTA y sus análisis están normales.

El último grupo se compone de mujeres de 63 años, 64,24 Kg y 27,69 de IMC. Estas personas con niveles de obesidad bajos, son tomadoras de café, hipertensas y con los triglicéridos y el colesterol un poco elevados.

Al analizar las características de estos grupos se observa que las diferencias notables no están en los niveles de glicemia, como en los grupos de hombres, sino en la edad y el peso, específicamente los niveles de obesidad.

Según los médicos la obesidad es un factor agravante del riesgo que tienen los diabéticos de complicarse, por las consecuencias que esta tiene para la salud humana. La edad es otro agente que aumenta la posibilidad de riesgo ya que varias de las complicaciones de la diabetes mellitus aparecen después de padecer la enfermedad por algún tiempo. Además cuando se debuta en edades tempranas es porque se tienen los factores de riesgo en niveles altos. Por lo anterior se pueden identificar en los grupos formados por la herramienta niveles de riesgo de complicación, ya no determinados por la glicemia, sino por estos factores igual de importantes en la prevención de la DM.

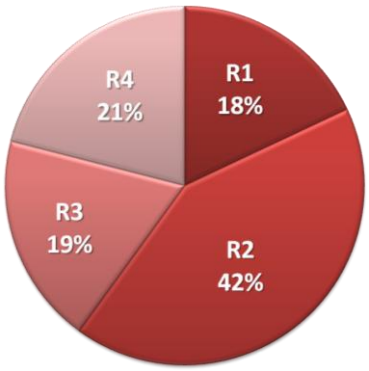


Ilustración 3.10 Clases en el conjunto de mujeres.

Por tanto se proponen como clases para el conjunto de datos de mujeres, que contribuirá al entrenamiento de las técnicas supervisadas en la creación de un SE para la prevención y el diagnóstico de la DM, cuatro niveles de riesgo donde 1 indica mayor riesgo y 4 menor riesgo.

La clasificación R1 corresponde al grupo 2 conformado por la técnica, ya que estas son las mujeres más jóvenes y con más obesidad. R2 son las mujeres muy obesas pero menos jóvenes, agrupadas en el cluster 1. El grupo de R3 coincide con el cuarto conglomerado, donde las pacientes son obesas en menor medida. El cluster 3 se considera como R4 ya que estas mujeres no son obesas, ni hipertensas y sus análisis estás en niveles saludables. (Anexo 47)

3.4.2.1. Procedimiento recomendado por los médicos para los grupos de mujeres.

Debido a que en todos los grupos las mujeres tienen glicemias entre 7 y 10mmol/L el procedimiento a seguir es el mismo para los cuatro. Este consiste en modificar el estilo de vida de la paciente, mediante la dieta y el plan de ejercicios, e indicarle un hipoglicemiante oral, si es necesario.

Para la indicación de la dieta los médicos tienen en cuenta el peso, la talla, el IMC, la edad, el sexo y la actividad física. Algunos de estos factores influyen también en el tipo de hipoglicemiante oral que le prescriben al paciente. Por lo que, al existir diferencias notables en los grupos de mujeres en estos factores se presupone que los habrá también en el tipo de dieta y el tratamiento de cada una de estas pacientes.

3.5. Conclusiones parciales.

Las técnicas de clustering que brindan las herramientas Weka y SPSS permitieron la experimentación con los datos y ofrecieron conjuntos de grupos en los diabéticos tipo 2.

Dividir el conjunto de datos en hombres y mujeres resultó del análisis con el “Conglomerado en dos fases”.

Los mejores agrupamientos para la propuesta de las clases en los diabéticos tipo dos, de acuerdo a los índices de validación y al criterio de expertos, es la de los algoritmos “EM” y “Conglomerado bietápico”.

El uso de las herramientas de la IA, de conjunto con el criterio médico, contribuyó al desarrollo exitoso de la MD, proporcionando tres clases en el grupo de hombres y

cuatro en el grupo de mujeres, que serán las clases de un conjunto de entrenamiento de apoyo a la construcción de un SE, para la prevención y el diagnóstico de la DM.

Conclusiones

Llegado el final de esta investigación se arriban a las siguientes conclusiones:

1. El proceso de minería de datos desarrollado con la metodología CRISP-DM y utilizando técnicas de agrupamiento brindó dos conjuntos de entrenamiento, uno por cada sexo, que contribuirán a la creación de un sistema experto de apoyo a la prevención y diagnóstico temprano de la diabetes mellitus tipo 2 en Cienfuegos.
2. En el análisis del conjunto de hombres el mejor resultado fue alcanzado con el algoritmo “Expectation Maximization” implementado en Weka, obteniéndose 3 clases.
3. En el análisis del conjunto de mujeres el mejor resultado fue alcanzado por el algoritmo “Conglomerado en dos fases” implementado en el SPSS, obteniéndose 4 clases.
4. Las clases resultantes de la experimentación, comprobadas mediante los índices de validación y el criterio de experto, fueron interpretadas como niveles de riesgo de complicación de la diabetes mellitus tipo 2.

Recomendaciones

Concluida la investigación se recomienda utilizar los conjuntos de entrenamiento obtenidos en la minería de datos para construir un sistema experto de apoyo al diagnóstico de la diabetes mellitus tipo 2 en las áreas de atención primaria de Cienfuegos.

Referencias Bibliográficas

[1] A Hellemans y B Bunch, *The Timetables of Science*. Nueva York: Simon and Schuster, 1988.

[2] Louis E. Jr Frenzel, *Crash Course in Artificial Intelligence and Experts Systems*. Addison Wesley, 1985.

[3] José Luis Rodríguez Sotelo, «Análisis de bioseñales en la identificación de arritmias cardíacas mediante técnicas no supervisadas», Trabajo de grado para optar al título de Doctor en Ingeniería, Universidad Nacional de Colombia, Manizales, 2010.

[4] Andrea Villagra, Ana Guzmán, Daniel Pandolfi, y Guillermo Leguizamón, «Análisis de medidas no-supervisadas de calidad en clusters obtenidos por K-means y Particle Swarm Optimization», Universidad Nacional de la Patagonia Austral, Argentina.

[5] Germinal Álvarez Batard, *Pensamiento Médico y Cibernética*.

[6] J Hernández Orallo, M J Ramírez Quintana, y C Ferri Ramírez, *Introducción a la minería de datos*. Madrid: PEARSON EDUCACIÓN, S.A, 2004.

[7] José Manuel Molina López y Jesús García Herrero, «Técnicas de análisis de datos aplicaciones prácticas utilizando Microsoft Excel y Weka», Carlos III, Madrid, 2006.

[8] J Yetano, AB Montero, y R Saracho, «Disminución de errores archivados en un archivo hospitalario», *Rev Calidad Asistencial*, vol. 3, pp. 118–120, 1995.

[9] A Esteban, E Cerda, MA de la Cal, y JA Laronte, «Control de calidad del archivo de datos computarizado de una unidad de cuidados intensivos.», *Rev Calidad Asistencial*, vol. 1, n^{o}. 23, p. 6, 1995.

[10] TP Clemmer y RM Gardem, *Informática Médica en la unidad de cuidados intensivos : estado de la cuestión 1995*. Rev Calidad Asistencial, 1996.

[11] S Fojon, JG Pardo, y JD Fernández, «Sistema de información en medicina intensiva», *Rev Calidad Asistencial*, 1996.

[12] JE George, «Standarization in health care informatic and telematic in Europe: CEN TC 251 activities», *Med inform*, vol. 17, n^{o}. 3, 1992.

[13] Víctor R Ávila, «Medicina y Computación: Una integración necesaria».

[14] Mercedes Medina Pagola, «Utilización del aprendizaje basado en problemas bajo la óptica de la Inteligencia Artificial».

[15] María M García, «Diagnóstico presuntivo de cardiopatía isquémica», UCLV, 1998.

[16] Yanet Rodríguez, «Diagnóstico de enfermedades infectorespiratorias en niños menores de dos años», UCLV, 1998.

[17] Karina L. Fernández Sánchez y Daniel Gálvez, «Sistema Experto para el diagnóstico y tratamiento de embarazos eptópicos», Trabajo de Diploma, UCLV, Villa Clara, 2004.

[18] Maribel García García y Viviana Toledo, «Sistema Experto para el diagnóstico y tratamiento de las ITS», Trabajo de Diploma, UCf, Cienfuegos, 2005.

[19] Yariel Ramos Negrín, Karina L. Fernández Sánchez, y Viviana Toledo, «Sistema Experto para el diagnóstico y tratamiento de fibroma uterino», Trabajo de Diploma, UCf, Cienfuegos, 2006.

[20] «La carga mundial | International Diabetes Federation», *IDF Diabetes Atlas Fifth Edition*. [Online]. Available: http://www.idf.org/diabetesatlas/... [Accessed: 08-may-2013].

[21] Instituto Nacional de Endocrinología, «Programa Nacional de diabetes», Cuba.

[22] Abelardo Ramírez Marquez, *El Sistema Nacional de Salud en Cuba*. La Habana: ENSAP, 2003.

[23] Belkis M. Vicente Sánchez, José Fermín Sánchez Pedraza, Guillermo Alexander Llaguno Pérez, y Miriam Costa Cruz, «Policlínico Docente Universitario Área V “Manuel Piti Fajardo”. Municipio Cienfuegos Efecto del ejercicio físico en pacientes con diabetes mellitus tipo 2».

[24] «El cardiólogo tratando la hiperglicemia de la Diabetes Mellitus tipo 2», *Rev. costarric. cardiol.*, vol. 5, n^{o}. 2, pp. 27–34, ago. 2003.

[25] E H Shortliffe, *Computer based medical consultation: MYCIN*. Nueva York: Elsevier, 1976.

[26] H Popple, *The formation of composite hypotheses in diagnostic problem solving - An exercise in synthetic reasoning*. Proceeding IJCAI’77., 1987.

[27] S M Weiss, C A Kulikowski, S Amarel, y A Safir, «A model-based method for computer aided medical decision-making», in *Artificial Inteligence*, vol. 11, 1978.

[28] B Chandrasekaran, S Mittal, F Gomez, y J Smith, *An Aproach to Medical Diagnosis Based on Conceptual Structures*. Proceedings, 1979.

[29] W J Clancey, *Knowledge Based Problem Solving.*, Segunda. New Jersey: Prentice-Hall, 1986.

[30] L Eshelman, D Eiret, J McDermott,, y M Tan, «International Journal of Man-Machine Studies», *MOLE: a tenacious knowledge-acquisition tool*, 1987.

[31] S Marcus y J McDermott, *SALT: A knowledge Acquisition Language for Propose-and-Revise Systems.* 1989.

[32] M A Musen, *Automated Suport for Building and Extending Expert Models.* 1989.

[33] A R Puerta, S W Tu, y M A Musen, «Modeling task with mechanisms», *International Journal of Intelligent Systems*, 1993.

[34] B Chandrasekaran, *Generic Task in Knowledge Based Reasoning: High level building blocks for expert systems design*. 1986.

[35] B Chandrasekaran, Johnson, R Todd, Smith, y W Jack, *Task-structure analysis for knowledge modeling*. 1992.

[36] R Bello, «Aplicaciones de la Inteligencia Artificial», Universidad de Guadalajara, Jalisco, 2002.

[37] F Calvo, «Capítulo XVIII: e-Learning para Educación Médica Continua», in *TELEMEDICINA E INFORMATICA MEDICA*, C F Rubio, J Aparcana, V M Changa, y F Calvo, Eds. Lima, 2002.

[38] Dr. Daniel Gálvez Lio, *Curso de Sistemas Basados en el Conocimiento*. Grupo de Investigación en Inteligencia Artificial Departamento de Ciencia de la Computación Facultad de Matemática, Física y Computación Universidad Central «Martha Abreu» de Las Villas: 1998.

[39] P Clark y R Boswell, *PracticalMachineLearning Toolsand Techniques with Java Implementation*. Morgan Kaufmann Publishers, 2000.

[40] P Gutiérrez Rüegg, H Merlino, C Rancan, C Procopio, D Rodríguez, P Britos, y R García Martínez, «Identificación de patrones característicos de la población carcelaria mediante minería de datos», Universidad de Buenos Aires.

[41] Marta Sananes, Elizabeth Torres, Surendra P. Sinha, y Luis Nava Puente, «Búsqueda y caracterización de subgrupos de pobreza mediante la aplicación de algunas técnicas de Minería de datos», Universidad de Los Andes, Mérida, Venezuela.

[42] Elena Durán y Rosanna Costaguta, «Minería de datos para descubrir estilos de aprendizaje», *Revista Iberoamericana de Educación*, vol. 42, n°. 2, mar. 2007.

[43] Yessica Milagros Quiñones Alvárez, «Aplicación de técnicas de Análisis de Cluster a la exploración de los Estilos de Aprendizaje en estudiantes de Ingeniería Informática», Universidad de Cienfuegos «Carlos Rafael Rodríguez», Cienfuegos, 2011.

[44] «ScienceDirect.com - Artificial Intelligence in Medicine - Medical data mining by fuzzy modeling with selected features», 10-dic-2012. [Online]. Available: http://hinari-gw.who.int/whalecomwww.sciencedirect.com/whalecom0/science/article/pii/S0933655708000523. [Accessed: 10-dic-2012].

[45] Iván Gildo Tapia Rivas, «Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando Datamart y Datamining en un Hospital Nacional», Para optar el título profesional de: Ingeniero de Sistemas, Universidad Nacional Mayor de San Marcos, Lima, Perú, 2006.

[46] Ricardo Timarán Pereira y María Clara Yépez Chamorro, «La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino», *Revista Universidad y Salud*, vol. 14, n°. 2, pp. 117–129, 2012.

[47] D Reparaz, H Merlino, C Rancan, D Rodríguez, P Britos, y R García Martínez, «Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basada en minería de datos», Universidad de Buenos Aires.

[48] Rolando Acosta Sánchez., Alejandro Rosete Suárez, y Alfredo Rodríguez Díaz, «Predicción de pacientes diabéticos. Preprocesado para Minería de Datos», *Revista Cubana de Informática Médica*.

[49] Frank Dávila Hernández y Yovannys Sánchez Corales, «Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas», *Revista Cubana de Informática Médica*, vol. 4, n°. 2, dic. 2012.

[50] Jiawei Han y Micheline Kamber, *Data Mining: Concepts and Techniques*. Simon Fraser University: Morgan Kaufmann Publishers., 2000.

[51] P Honey y A Munford, *Using your learning styles*.

[52] U Fayyad, G Piatetsky-Shapiro, y P Smith, *From data mining to knowledge discovery in database*. American Association for Artificial Intelligence, 1996.

[53] U E Backer, *Computer-assisted reasoning in cluster analysis*. Prentice-Hall, 1995.

[54] Anil K. Jain y Richard C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[55] Marggie D. González Toledo, «Una comparación de índices de validación de conglomerados», Grado de Maestría en Ciencias, Universidad de Puerto Rico, Puerto Rico, 2005.

[56] José Fernández Márquez, «Interfaz web para estudiar el efecto de diferentes condiciones sobre la expresión de los genes», Tesis de Grado, Universidad Autónoma de Barcelona, 2011.

[57] Ferenc Kovács, Csaba Legány, y Attila Babos, «Cluster Validity Measurement Techniques», Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary.

[58] Andreas Weingessel, Evgenia Dimitriadou, y Sara Dolnicar, «An examination of indexes for determining the number of clusters in binary data sets», Vienna University of Economics and Business Administration, Austria, 1999.

[59] María N. Moreno García, «Guía Docente de Introducción a la Minería de Datos (1,5 ECTS)», Departamento de Informática y Automática Facultad de Ciencias – Universidad de Salamanca, España.

[60] U Fayyad, «Advanced in Knowledge Discovery and Data Mining». MIT Press, 1996.

[61] «Data Mining Applications in 2008.», *KDnuggets Polls.*, 2008. [Online]. Available: http://www.kdnuggets.com/polls/2008/data-mining-applications.htm. [Accessed: 24-sep-2009].

[62] Juan Miguel Moine, Ana Silvia Haedo, y Silvia Gordillo, «Estudio comparativo de metodologías para minería de datos», Universidad Nacional de La Plata, Argentina.

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

[64] «Comprendiendo el negocio| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprendiendo-el-negocio. [Accessed: 12-feb-2013].

[65] «Comprensión del negocio| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-del-negocio-salidas. [Accessed: 12-feb-2013].

[66] «Poll: What Analytics, Data mining, Big Data software you used in the past 12 months for a real project?», *KDnuggets*, may-2012.

[67] Jackson Rondón, «Qué es el SPSS y sus ventajas», 11-may-2011.

[68] J. Keefe, *Profiling and Utilizing Learning Style.* Reston Virginia: National Association of Secondary School Principals, 1988.

[69] «Comprensión de datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprensi%C3%B3n-de-datos. [Accessed: 12-feb-2013].

[70] «Comprensión de datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-de-datos-guia. [Accessed: 12-feb-2013].

[71] «Comprensión de datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-datos. [Accessed: 12-feb-2013].

[72] «Preparación de datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparaci%C3%B3n-de-datos. [Accessed: 12-feb-2013].

[73] «Preparación de datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparacion-de-los-datos-guia. [Accessed: 12-feb-2013].

[74] «Preparación de los datos | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparacion-de-los-datos-salidashttp://www.dataprix.com/preparacion-de-los-datos-salidas. [Accessed: 12-feb-2013].

[75] «Modelado | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado. [Accessed: 12-feb-2013].

[76] «Modelado | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado-guia. [Accessed: 12-feb-2013].

[77] «Modelado | Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado-salidas. [Accessed: 12-feb-2013].

[78] *Waikato Environment for Knowledge Analysis*. Nueva Zelanda: Universidad de Waikato y Hamilton, 1999.

[79] IBM Corporation, «IBM SPSS Modeler CRISP-DM Guide». 2011.

[80] «The SPSS TwoStep Cluster Component».

[81] «Evaluación de las salidas de CRISP DM| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluacion-salidas. [Accessed: 12-feb-2013].

[82] «Evaluación| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluacion. [Accessed: 12-feb-2013].

[83] «Evaluación| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluaci%C3%B3n-guia. [Accessed: 12-feb-2013].

[84] «Desarrollo| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo. [Accessed: 12-feb-2013].

[85] «Desarrollo| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo-guia. [Accessed: 12-feb-2013].

[86] «Desarrollo| Manual IT online», *Administrador de Dataprix*, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo-salidas. [Accessed: 12-feb-2013].

Bibliografía

[1] E. Fowlkes y C. Mallows, A method for comparing two hierarchical clusterings. Journal of the American Statistical Asociation, 1983.

[2] S M Weiss, C A Kulikowski, S Amarel, y A Safir, «A model-based method for computer aided medical decision-making», in Artificial Inteligence, vol. 11, 1978.

[3] U Fayyad, «Advanced in Knowledge Discovery and Data Mining». MIT Press, 1996.

[4] Anil K. Jain y Richard C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[5] B Chandrasekaran, S Mittal, F Gomez, y J Smith, An Aproach to Medical Diagnosis Based on Conceptual Structures. Proceedings, 1979.

[6] Andreas Weingessel, Evgenia Dimitriadou, y Sara Dolnicar, «An examination of indexes for determining the number of clusters in binary data sets», Vienna University of Economics and Business Administration, Austria, 1999.

[7] G. W Milligan y M.C. Cooper, An examination of procedures for determining the number of clusters in a data set. Psychometrika, 1985.

[8] R. E Bellman, An Introduction to Artificial Intelligence: Can Computers Think? San Francisco: Boyd & Fraser Publishing Company, 1978.

[9] Carmen López y Isidoro Cortés, «Análisis avanzados de grandes volúmenes de datos en el sector seguros (2a parte)», ACTUARIOS, no. 25, dic. 2006.

[10] José Luis Rodríguez Sotelo, «Análisis de bioseñales en la identificación de arritmias cardíacas mediante técnicas no supervisadas», Trabajo de grado para optar al título de Doctor en Ingeniería, Universidad Nacional de Colombia, Manizales, 2010.

[11] Ingrid Wilford Rivera, Alejandro Rosete Suárez, y Alfredo Rodríguez Díaz, «Análisis de Información Clínica mediante técnicas de Minería de Datos», RevistaeSalud.com, vol. 5, no. 20.

[12] Andrea Villagra, Ana Guzmán, Daniel Pandolfi, y Guillermo Leguizamón, «Análisis de medidas no-supervisadas de calidad en clusters obtenidos por K-means y Particle Swarm Optimization», Universidad Nacional de la Patagonia Austral, Argentina.

[13] Yessica Milagros Quiñones Alvárez, «Aplicación de técnicas de Análisis de Cluster a la exploración de los Estilos de Aprendizaje en estudiantes de Ingeniería

Informática», Universidad de Cienfuegos «Carlos Rafael Rodríguez», Cienfuegos, 2011.

[14] Osvaldo M. Spositto, Martín E. Etcheverry, Hugo L. Ryckeboer, y Julio Bossero, «Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil», Universidad Nacional de La Matanza, San Justo, Provincia de Buenos Aires, Argentina.

[15] R Bello, «Aplicaciones de la Inteligencia Artificial», Universidad de Guadalajara, Jalisco, 2002.

[16] E Rich y K Knight, Artificial Intelligence, segunda ed. New York: McGraw-Hill, 1991.

[17] P H Winston, Artificial Intelligence, tercera ed. Massachusetts: Addison-Wesley, 1992.

[18] R I Schalkoff, Artificial Intelligence: An Engineering Approach. New York: McGraw-Hill, 1990.

[19] G F Luger y W A Stubblefield, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, segunda ed. Redwood City, California: Benjamin/Cummings, 1993.

[20] J Haugeland, Artificial Intelligence: The Very Idea. Cambridge, Massachusetts: MIT Press, 1985.

[21] I N Scobie, Atlas of Diabetes Mellitus, 3ra ed. UK: Informa Healthcare, 2008.

[22] M A Musen, Automated Suport for Building and Extending Expert Models. 1989.

[23] Marta Sananes, Elizabeth Torres, Surendra P. Sinha, y Luis Nava Puente, «Búsqueda y caracterización de subgrupos de pobreza mediante la aplicación de algunas técnicas de Minería de datos», Universidad de Los Andes, Mérida, Venezuela.

[24] Harrison, «Capítulo 338. Diabetes mellitus», in Principios de Medicina Interna, 16a(2006) ed., McGraw-Hill.

[25] F Calvo, «Capítulo XVIII: e-Learning para Educación Médica Continua», in Telemedicina e Informatica Medica, C F Rubio, J Aparcana, V M Changa, y F Calvo, Eds. Lima, 2002.

[26] Ferenc Kovács, Csaba Legány, y Attila Babos, «Cluster Validity Measurement Techniques», Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary.

[27] Y. Batistakis, M. Halkidi, y M. Vazirgiannis, Cluster validity methods: Part i. Sigmod Record, 2002.

[28] «Comprendiendo el negocio| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprendiendo-el-negocio. [Accessed: 12-feb-2013].

[29] «Comprensión de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprensi%C3%B3n-de-datos. [Accessed: 12-feb-2013].

[30] «Comprensión de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-de-datos-guia. [Accessed: 12-feb-2013].

[31] «Comprensión de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-datos. [Accessed: 12-feb-2013].

[32] «Comprensión del negocio| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/comprension-del-negocio-salidas. [Accessed: 12-feb-2013].

[33] E H Shortliffe, Computer based medical consultation: MYCIN. Nueva York: Elsevier, 1976.

[34] U E Backer, Computer-assisted reasoning in cluster analysis. Prentice-Hall, 1995.

[35] A Esteban, E Cerda, MA de la Cal, y JA Laronte, «Control de calidad del archivo de datos computarizado de una unidad de cuidados intensivos.», Rev Calidad Asistencial, vol. 1, no. 23, p. 6, 1995.

[36] Warren Hart y Manuel Collazo Herrera, «Costos del diagnóstico y tratamiento de la diabetes mellitus en diferentes países del mundo», Revista Cubana de Endocrinología, vol. 9, no. 3, pp. 212–220, 1998.

[37] Louis E. Jr Frenzel, Crash Course in Artificial Intelligence and Experts Systems. Addison Wesley, 1985.

[38] «Crear particiones de los datos en conjuntos de entrenamiento y de pruebas (Analysis Services - Minería de datos)», 12-feb-2013. [Online]. Available: http://msdn.microsoft.com/es-es/library/bb895173(v=sql.100).aspx. [Accessed:12-feb-2013].

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

[40] L MTierney Tierney, S J McPhee, y M A Papadakis, «Current medical Diagnosis & Treatment», New York: Lange Medical Books/McGraw-Hill, 2002, pp. 1203–1215.

[41] Dr. Daniel Gálvez Lio, Curso de Sistemas Basados en el Conocimiento. Grupo de Investigación en Inteligencia Artificial Departamento de Ciencia de la Computación Facultad de Matemática, Física y Computación Universidad Central «Martha Abreu» de Las Villas: 1998.

[42] «Data Mining Applications in 2008.», KDnuggets Polls., 2008. [Online]. Available: http://www.kdnuggets.com/polls/2008/data-mining-applications.htm. [Accessed:24-sep-2009].

[43] «Data Mining Tools Used Poll», KDnuggets, may-2009. .

[44] Jiawei Han y Micheline Kamber, Data Mining: Concepts and Techniques. Simon Fraser University: Morgan Kaufmann Publishers., 2000.

[45] «Desarrollo| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo. [Accessed:12-feb-2013].

[46] «Desarrollo| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo-guia. [Accessed:12-feb-2013].

[47] «Desarrollo| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/desarrollo-salidas. [Accessed:12-feb-2013].

[48] «Descripción de partes | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/descripci%C3%B3n-de-partes. [Accessed: 12-feb-2013].

[49] D Reparaz, H Merlino, C Rancan, D Rodríguez, P Britos, y R García Martínez, «Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basada en minería de datos», Universidad de Buenos Aires.

[50] Mark Ming-Tso Chiang y Boris Mirkin, «Determining the number of clusters in the Straight K-means: Experimental comparison of eight options», Birkbeck College, University of London, School of Computer Science & Information Systems.

[51] Yanet Rodríguez, «Diagnóstico de enfermedades infectorespiratorias en niños menores de dos años», UCLV, 1998.

[52] María M García, «Diagnóstico presuntivo de cardiopatía isquémica», UCLV, 1998.

[53] J Yetano, AB Montero, y R Saracho, «Disminución de errores archivados en un archivo hospitalario», Rev Calidad Asistencial, vol. 3, pp. 118–120, 1995.

[54] «El cardiólogo tratando la hiperglicemia de la Diabetes Mellitus tipo 2», Rev. costarric. cardiol., vol. 5, no. 2, pp. 27–34, ago. 2003.

[55] O Díaz Díaz, «El problema de la diabetes en Cuba.», mar. 2008.

[56] Abelardo Ramírez Marquez, El Sistema Nacional de Salud en Cuba. La Habana: ENSAP, 2003.

[57] Juan Miguel Moine, Ana Silvia Haedo, y Silvia Gordillo, «Estudio comparativo de metodologías para minería de datos», Universidad Nacional de La Plata, Argentina.

[58] «Evaluación de las salidas de CRISP DM| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluacion-salidas. [Accessed:12-feb-2013].

[59] «Evaluación| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluacion. [Accessed:12-feb-2013].

[60] «Evaluación| Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/evaluaci%C3%B3n-guia. [Accessed:12-feb-2013].

[61] U Fayyad, G Piatetsky-Shapiro, y P Smith, From data mining to knowledge discovery in database. American Association for Artificial Intelligence, 1996.

[62] B Chandrasekaran, Generic Task in Knowledge Based Reasoning: High level building blocks for expert systems design. 1986.

[63] Wordl Health Organization, «Global Database on Body Mass Index», 2008. [Online]. Available: http://www.who.int/bmi/index.jsp?introPage=intro_3.html. [Accessed:10-jun-2008].

[64] «Glosario/Terminología | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/glosario-terminologia-crisp-dm. [Accessed:12-feb-2013].

[65] María N. Moreno García, «Guía Docente de Introducción a la Minería de Datos (1,5 ECTS)», Departamento de Informática y Automática Facultad de Ciencias – Universidad de Salamanca, España.

[66] Roberto Espinosa, Jose Zubcoff, Marta Zorrilla, y Jose Norberto Mazón, «Hacia la consideración de aspectos de calidad de datos en procesos de minería: el caso de las técnicas de clasificación», 2010.

[67] Yuniet Rodríguez Suárez y Anolandy Díaz Amador, «Herramientas de Minería de Datos», in RCCI, vol. 3, 2009, pp. 73–80.

[68] IBM Corporation, «IBM SPSS Modeler CRISP-DM Guide». 2011.

[69] P Gutiérrez Rüegg, H Merlino, C Rancan, C Procopio, D Rodríguez, P Britos, y R García Martínez, «Identificación de patrones característicos de la población carcelaria mediante minería de datos», Universidad de Buenos Aires.

[70] TP Clemmer y RM Gardem, Informática Médica en la unidad de cuidados intensivos : estado de la cuestión 1995. Rev Calidad Asistencial, 1996.

[71] José Fernández Márquez, «Interfaz web para estudiar el efecto de diferentes condiciones sobre la expresión de los genes», Tesis de Grado, Universidad Autónoma de Barcelona, 2011.

[72] L Eshelman, D Eiret, J McDermott,, y M Tan, «International Journal of Man-Machine Studies», MOLE: a tenacious knowledge-acquisition tool, 1987.

[73] J Hernández Orallo, M J Ramírez Quintana, y C Ferri Ramírez, Introducción a la minería de datos. Madrid: PEARSON EDUCACIÓN, S.A, 2004.

[74] E Charniak y D McDermott, Introduction to Artificial Intelligence. Massachusetts: Addison-Wesley, 1985.

[75] W J Clancey, Knowledge Based Problem Solving., Segunda. New Jersey: Prentice-Hall, 1986.

[76] «La carga mundial | International Diabetes Federation», IDF Diabetes Atlas Fifth Edition. [Online]. Available: http://www.idf.org/diabetesatlas/... [Accessed:08-may-2013].

[77] «La metodología CRISP-DM | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/la-metodolog%C3%AD-crisp-dm.....la metodologi-crisp-dm. [Accessed: 12-feb-2013].

[78] Ricardo Timarán Pereira y María Clara Yépez Chamorro, «La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino», Revista Universidad y Salud, vol. 14, no. 2, pp. 117–129, 2012.

[79] Víctor R Ávila, «Medicina y Computación: Una integración necesaria».

[80] Sofia J. Vallejos, «Minería de Datos», Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste, Corrientes - Argentina, 2006.

[81] Elena Durán y Rosanna Costaguta, «Minería de datos para descubrir estilos de aprendizaje», Revista Iberoamericana de Educación, vol. 42, no. 2, mar. 2007.

[82] R. Agrawal, T. Imielinski, y A. Swami, «Mining association rules between sets of items in large databases.», Washington, DC, may-1993.

[83] «Modelado | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado. [Accessed:12-feb-2013].

[84] «Modelado | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado-guia. [Accessed:12-feb-2013].

[85] «Modelado | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/modelado-salidas. [Accessed:12-feb-2013].

[86] A R Puerta, S W Tu, y M A Musen, «Modeling task with mechanisms», International Journal of Intelligent Systems, 1993.

[87] Noé Ruiz García, Félix V. González Cossío, Alberto Castillo Morales, y Fernando Castillo González, «Optimización y validación del análisis de conglomerados aplicado a la clasificación de razas mexicanas de maíz». .

[88] «Pasaje de modelos genéricos a modelos especializados | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/pasaje-modelos-crisp-dm. [Accessed: 12-feb-2013].

[89] S. Theodoridis y K. Koutroumbas, Pattern Recognition. Academic Press, 1999.

[90] Germinal Álvarez Batard, Pensamiento Médico y Cibernética.

[91] Mónica Arnold Rodríguez, Yuri Arnold Domínguez, Yanira Alfonso Hernández, Clara Villar Guerra, y Teresa Margarita González Calero, «Pesquisaje y prevención

de la diabetes mellitus tipo 2 en población de riesgo», Revista Cubana de Higiene y Epidemiología, vol. 50, no. 3, dic. 2012.

[92] Belkis M. Vicente Sánchez, José Fermín Sánchez Pedraza, Guillermo Alexander Llaguno Pérez, y Miriam Costa Cruz, «Policlínico Docente Universitario Área V "Manuel Piti Fajardo". Municipio Cienfuegos Efecto del ejercicio físico en pacientes con diabetes mellitus tipo 2».

[93] «Poll: What Analytics, Data mining, Big Data software you used in the past 12 months for a real project?», KDnuggets, may-2012.

[94] «Poll: Where did you apply Analytics / Data Mining in 2012?», KDnuggets, dic-2012. .

[95] P Clark y R Boswell, PracticalMachineLearning Toolsand Techniques with Java Implementation. Morgan Kaufmann Publishers, 2000.

[96] Rolando Acosta Sánchez., Alejandro Rosete Suárez, y Alfredo Rodríguez Díaz, «Predicción de pacientes diabéticos. Preprocesado para Minería de Datos», Revista Cubana de Informática Médica.

[97] «Preparación de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparaci%C3%B3n-de-datos. [Accessed:12-feb-2013].

[98] «Preparación de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparacion-de-los-datos-guia. [Accessed: 12-feb-2013].

[99] «Preparación de los datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/preparacion-de-los-datos-salidashttp://www.dataprix.com/preparacion-de-los-datos-salidas. [Accessed: 12-feb-2013].

[100] Paola Britos, «Procesos de explotación de información basados en sistemas inteligentes», Universidad Nacional de La Plata, Argentina., 2008.

[101] J. Keefe, Profiling and Utilizing Learning Style. Reston Virginia: National Association of Secondary School Principals, 1988.

[102] Instituto Nacional de Endocrinología, «Programa Nacional de diabetes», Cuba.

[103] Jackson Rondón, «Qué es el SPSS y sus ventajas», 11-may-2011.

[104] «Resumen de dependencias | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/resumen-de-dependencias. [Accessed: 12-feb-2013].

[105] «Revista Cubana de Endocrinología - Trastornos metabólicos asociados con la evolución hacia la diabetes mellitus tipo 2 en una población en riesgo», 13-nov-2012. [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-29532010000200001&nrm=iso. [Accessed:13-nov-2012].

[106] «Revista Cubana de Farmacia - Aplicación de la minería de datos al Sistema Cubano de Farmacovigilancia». .

[107] «Revista Cubana de Medicina General Integral - Diabetes mellitus: Diagnóstico positivo», 13-nov-2012. [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21252006000100012&nrm=iso. [Accessed:13-nov-2012].

[108] S Marcus y J McDermott, SALT: A knowledge Acquisition Language for Propose-and-Revise Systems. 1989.

[109] «ScienceDirect.com - Artificial Intelligence in Medicine - Medical data mining by fuzzy modeling with selected features», 10-dic-2012. [Online]. Available: http://hinari-gw.who.int/whalecomwww.sciencedirect.com/whalecom0/science/article/pii/S0933365708000523. [Accessed: 10-dic-2012].

[110] S Fojon, JG Pardo, y JD Fernández, «Sistema de información en medicina intensiva», Rev Calidad Asistencial, 1996.

[111] Karina L. Fernández Sánchez y Daniel Gálvez, «Sistema Experto para el diagnóstico y tratamiento de embarazos eptópicos», Trabajo de Diploma, UCLV, Villa Clara, 2004.

[112] Yariel Ramos Negrín, Karina L. Fernández Sánchez, y Viviana Toledo, «Sistema Experto para el diagnóstico y tratamiento de fibroma uterino», Trabajo de Diploma, UCf, Cienfuegos, 2006.

[113] Maribel García García y Viviana Toledo, «Sistema Experto para el diagnóstico y tratamiento de las ITS», Trabajo de Diploma, UCf, Cienfuegos, 2005.

[114] JE George, «Standarization in health care informatic and telematic in Europe: CEN TC 251 activities», Med inform, vol. 17, no. 3, 1992.

[115] B Chandrasekaran, Johnson, R Todd, Smith, y W , Jack, Task-structure analysis for knowledge modeling. 1992.

[116] José Manuel Molina López y Jesús García Herrero, «Técnicas de análisis de datos aplicaciones prácticas utilizando Microsoft Excel y Weka», Carlos III, Madrid, 2006.

[117] Gladys M. Casas Cardoso, Gladys Cardoso Romero, Vivian Guerra Morales, y Luis Felipe Herrera Jiménez, «Técnicas de detección de clusters aplicadas a la investigación psicológica», Revista Cubana de Psicología, vol. 19, no. 1, 2002.

[118] Frank Dávila Hernández y Yovannys Sánchez Corales, «Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas», Revista Cubana de Informática Médica, vol. 4, no. 2, dic. 2012.

[119] R Kurzweil, The Age of Intelligent Machines. Cambridge, Massachusetts: MIT Press, 1990.

[120] H Popple, The formation of composite hypotheses in diagnostic problem solving - An exercise in synthetic reasoning. Proceeding IJCAI'77., 1987.

[121] A Hellemans y B Bunch, The Timetables of Science. Nueva York: Simon and Schuster, 1988.

[122] «Tipos de problemas de minería de datos | Manual IT online», Administrador de Dataprix, 12-feb-2013. [Online]. Available: http://www.dataprix.com/tipos-de-problemas-de-mineria-de-datos. [Accessed: 12-feb-2013].

[123] Marggie D. González Toledo, «Una comparación de índices de validación de conglomerados», Grado de Maestría en Ciencias, Universidad de Puerto Rico, Puerto Rico, 2005.

[124] Iván Gildo Tapia Rivas, «Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando Datamart y Datamining en un Hospital Nacional», Para optar el título profesional de: INGENIERO DE SISTEMAS, UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS, LIMA, PERÚ, 2006.

[125] P Honey y A Munford, Using your learning styles. .

[126] María N. Moreno García y Vivian F. López Batista, «Uso de técnicas no supervisadas en la construcción de modelos de clasificación en Ingeniería del Software», Universidad de Salamanca.

[127] Mercedes Medina Pagola, «Utilización del aprendizaje basado en problemas bajo la óptica de la Inteligencia Artificial».

[128] Waikato Environment for Knowledge Analysis. Nueva Zelanda: Universidad de Waikato y Hamilton, 1999.

[129] «World Health Organisation Department of Noncommunicable Disease Surveillance». 2006.

Anexos

Anexo 1: “Historia clínica de los pacientes del CAED”.

CENTRO DE ATENCIÓN Y EDUCACIÓN EN DIABETES CIENFUEGOS	HISTORIA CLINICA	EXPEDIENTE CLINICO: _____ CARNET DE IDENTIDAD: _____

1 ER APELLIDO	2DO APELLIDO	NOMBRE (S)	FECHA DE CONFECCION ____ ____ ____ DIA MES AÑO

DIRECCION:

_____ _____ _____ _____

Calle No. Entrecalles Municipio.

_____ _____ _____ _____

Ciudado Publico Provincia Teléfono (T) Teléfono (C)

OCUPACION:	ESCOLARIDAD	SEXO: F:____ M: ____	EDAD:	EDAD AL DEBUT.

TIEMPO DE EVOLUCION	MODO DE DEBUT:	OBESIDAD AL DEBUT
	___ SINTOMAS CLINICOS. ___ CHEQUEO SIN SINTOMAS. ___ CETOACIDOSIS (PROBADA). ___ DURANTE EL EMBARAZO. ___ OTROS. ___ NO SABE.	___ SI. ___ NO ___ NO PRECISADO.

ANTECEDENTES FAMILIARES DE DIABETES		HISTORIAOBSTETRICA MENARCA FM : G___P____A____ MACROFETOS MALFORMACIONES MUERTES ANTICONCEPCION MENOPAUSIA EDAD
VIA MATERNA. ___ NADIE. ___ MADRE ___ ABUELO (A). ___ MADRE + ABUELO (A). ___ NO SABE.	VIA PATERNA. ___ NADIE. ___ PADRE ___ ABUELO (A). ___ MADRE + ABUELO (A). ___ NO SABE.	

HERMANO:	HIJO.
___ SI. ___ NO. ___ NO SABE	___ SI. ___ NO. ___ NO SABE

ANTECEDENTES PATOLOGICOS	PERSONALES			FAMILIARES			HABITOS TOXICOS.
	SI	NO	NO PREC.	SI	NO	NO PREC.	___ NO FUMADOR. ___ EX FUMADOR (6 MESES) ___ FUMADOR. SI FUMADOR (# POR DIAS) ___ CIGARROS_____ ___ TABACO _____ ALCOHOL ___ NO ___ SI (GR/DIAS): _____ CAFÉ -3T ____ *3 T ____
HIPERLIPOPROTEINEMIA							
HIPERTENSION ARTERIAL							
CARDIOPATIA ISQUEMICA.							
CLAUDITACION INTERMITENTE.							
AVE							
OTROS (ESPECIFICAR).							

TRATAMIENTO AL INICIO	TRATAMIENTO ACTUAL	OTROS TTOS	SINTOMAS ACTUALES
____ SOLO DIETA: ____ COH. ____ INSULINA. ____ INSULINA + COH. ____ NO PRECISADO AÑOS DE COMIENZO _____	____ SOLO DIETA: ____ COH. ____ INSULINA. ____ INSULINA + COH. ____ NO PRECISADO AÑOS DE COMIENZO _____		

EXAMEN FISICO (EXAMENES POR APARATOS).	.
MUCOSAS	TALLA _____M. PESO _____ Kg. IMC: _____ PI_____ DIETA_____ CA_____ 0

TCS

ARESP

FR
ACV

FC TA

ABDOMEN

EX CUELLO

EX ORAL

ESCALA NSS
-SENSACION.
QUEMAZON,ENTUMECIMIENTO,HORMIGUEO EN PIE (2 PTOS)
FATIGA,CALAMBRES ODOLORIMIENTO (1 PTO)
-LOCALIZACIÓN
PIES (2 PTOS)
PANTORRILLAS (1 PTO)
EN OTRO LUGAR (0PTO)
-TIENE SIEMPRE SINTOMAS AL DESPERTAR POR LA NOCHE
SI (1 PTO)
-RELACION HORARIA
EMPEORA POR NOCHE (2 PTOS)
PRESENTE DIA Y NOCHE (1 PTO)
PRESENTE DURANTE EL DIA (0 PTO)

-COMO MEJORAN LOS SINTOMAS
PASEANDO (2 PTOS)
DE PIE (1 PTO)
SENTADO O ACOSTADO O NO MEJORAN (O PTO)

EXISTE NEUROPATIA
NDS ES ≥6 O SI NDS ESTA ENTRE 3 Y 5 Y NSS ≥ 5

RESULTADO DE EDUCACION DIABETOLOGICA.

INICIO	FINAL
SUFICIENTE	SUFICIENTE
NO SUFICIENTE	NO SUFICIENTE

EXAMEN DE MIEMBROS INFERIORES.

ABDOMEN (AORTA) : NORMAL____ DILATADA____

SISTEMA VENOSO: VARICES____ MICROVARICES____

SISTEMA LINFÁTICO: LINFANGITIS____ LINFEDEMA____

DEFORMIDADES PODÁLICAS: PIE PLANO____ METATARSO CAIDO____

HALLUX VALGUS____ DEDOS EN GARRA____ PIE DE CHARCOT____

AMP. MENORES____ OTROS____

LESIONES DERMATOLÓGICAS: EPIDERMOFITOSIS____ FISURAS____ PÌEL SECA

Y CALLOSA____ HIPERPIGMENTACION____ DERMATITIS____

CELULITIS____ ÚLCERAS____ PALIDEZ____ CIANOSIS____ RUBICUNDEZ____

DEDOS CON NECROSIS____ OTRAS____

LESIONES A NIVEL DE LAS UÑAS: ONICOMICOSIS____ PARONIQUIA____

TIPO DE ANTEPIÉ: EGIPCIO____ GRIEGO____ 1 Y 2 IGUAL____ CUADRADO____

ESTANDAR____

EXAMEN ARTERIAL:

PULSOS	DERECHO	IZQUIERDO
FEMORAL:	_____	_____
POPLÍTEO:	_____	_____
TIBIAL POSTERIOR:	_____	_____
PEDEO:	_____	_____

SOPLO FEMORAL DERECHO_____ IZQUIERDO_____

IMPRESIÓN DIAGNOSTICA:

PIE DE RIESGO GRADO O ____,___

PIE DE RIESGO GRADO I____,____

PIE DE RIESGO GRADO II____,____

PIE DE RIESGO GRADOIII____,____

	DERECHO		IZQUIERDO
	___	PRESENTE	___
REFLEJOS	___	PRESENTE CON REFUERZO	___
AQUILEANO	___	AUSENTE	___

	DERECHO		IZQUIERDO
	___	NORMAL	___
	___	REDUCIDA	___
PALESTESIA (PUNTA DEL DEDO GORDO)			

	DERECHO		IZQUIERDO
TEMPERATURA	___	NORMAL	___
(DORSO DEL PIE)	____	REDUCIDA	___

	DERECHO		IZQUIERDO
	___	NORMAL	___
	___	REDUCIDA	___
PERCEPCION DE PUNTA ROMA			

PUNTUACION ESCALA NDS
0 A2 NO NEUROPATIA
3-5 LIG
6-8 MODERADA
9-10 GRAVE

VISION BORROSA: SI ___ NO ___

AV: OD: _____ OI: _____

RD: OD: _____ OI: _____

OD: A: _____ SA: _____ M; _____ FO: _____

OI: A: _____ SA: _____ M; _____ FO: _____

RETINOPATIA DIABETICA. NO ___ NO PROLIFERATIVA ___ PROLIFERATIVA ___ NO PRECISADA ____

RETINOPATIA HIPERTENSIVA NO ___ GRADO 1 ____ GRADO 2 ___ GRADO 3 ____ GRADO 4 ____ NO PRECISADA ____

RETINOPATIA ARTEREO ESCLEROSIS NO ___ GRADO 1 ____ GRADO 2 ___ GRADO 3 ____ GRADO 4 ____ NO PRECISADA ____

HEMORRAGIA VITREA OD: NO ___ SI ____ NO PRECISADA ____

HEMORRAGIA VITREA OI: NO ___ SI ____ NO PRECISADA ____

MACULOPATIA OD NO ____ SI ____ NO PRECISADA

MACULOPATIA OI NO ____ SI ____ NO PRECISADA

CATARATA NO ____ METABOLICA ____ SENIL ____ OTRAS _____ NO PRECISADA _____

GLAUCOMA: NO ____ ANGULO ABIERTO _____ SECUNDARIO ____ ANGULO ESTRECHO ____ NO PRECISADO ____.

COMPLICACIONES: SI NO NO PRECISADO.

NEUROPATIAS REVERSIBLES:

MONONEUROPATIA

PARALISIS DE N CRANEALES

NEUROPATIAS POR COMPRESION

DSE

SEPSIS ESPECIFICAR: _____

DERMOPATIA

HIPERLIPOPROTEINEMIA ESPECIFICAR: _____

NEUROPATIA AUTONOMICA ESPECIFICAR: _____

OTRAS COMPLICACIONES ESPECIFICAR: _____

NEFROLOGIA

FGT_____
CREATININA____
UREA____
AC URICO_____
MICROALBUMINURIA
1____
2_____
3_____
ID:

DIAGNOSTICO DEFINITIVO

PACIENTE:

1ER APELLIDO	2DO APELLIDO	NOMBRE	H.C

Anexo 2: “Descripción de las variables”.

Identificador	Descripción	Tipo	Relevancia
No.(Número)	Identificador del paciente, coincide con la historia clínica	Numérico	Sin importancia
Grupo	Número del grupo al que el paciente perteneció al ingresar	Numérico	Sin importancia
Historia Clínica	Número de la historia clínica del paciente, esta queda archivada en formato físico en la clínica.	Numérico	Sin importancia
Nombres y Apellidos	Nombre y Apellidos del paciente	Nominal	Sin importancia
Edad	Edad en años del paciente al momento de ingresar	Numérico	Relevante
Sexo	Sexo del paciente	Nominal	Relevante
Dirección	Dirección de la residencia del paciente	Nominal	Sin importancia
Área	Área de salud del municipio Cienfuegos a la que pertenece el paciente	Nominal	Sin importancia
Municipio	Municipio en el que vive el paciente	Nominal	Sin importancia
Talla	Altura del paciente en metros	Numérico	Relevante
Peso Inicial	Peso en Kg. del paciente al ingresar	Numérico	Relevante
Índice de Masa Corporal	El índice de masa	Numérico	Relevante

	corporal del paciente se calcula como la razón entre el peso de la persona (en Kg) y el cuadrado de la estatura (en m2). Da una medida de la cantidad de grasa corporal de una persona		
Peso Final	Peso en Kg. del paciente al terminar el ingreso	Numérico	Sin importancia
Escolaridad	Nivel educacional más alto culminado por el paciente	Nominal	Sin importancia
Ocupación	Trabajo u oficio que realiza el paciente	Nominal	Sin importancia
Hábitos Tóxicos(Fuma)	Si el paciente ha fumado o fuma	Booleano	Relevante
Hábitos Tóxicos(Café)	Si el paciente toma café	Booleano	Relevante
Hábitos Tóxicos (Beb. Alcohólicas)	Si el paciente toma bebidas alcohólicas	Booleano	Relevante
Hábitos Tóxicos(Otros)	Si el paciente practica algún hábito tóxico diferente a los anteriores.	Booleano	Relevante
Modo de Debut	Describe la forma en la que el paciente debutó con la diabetes	Nominal	Sin importancia
Obesidad al Debut	Si el paciente es obeso o no cuando debuta con la	Booleano	Relevante

	enfermedad		
Tiempo de Evolución de la Enfermedad	Tiempo que ha transcurrido desde el debut del paciente hasta la fecha en que ingresó a la clínica	Numérico	Sin importancia
Antecedentes Obstétricos (Menarca)	Edad de la primera menstruación de los pacientes femeninos	Numérico	Relevante
Antecedentes Obstétricos (Embarazos)	Cantidad de embarazos hasta el momento del ingreso	Numérico	Relevante
Antecedentes Obstétricos (Abortos)	Cantidad de abortos realizados a la paciente	Numérico	Relevante
Antecedentes Obstétricos (Malformaciones)	En algunos casos se toma la cantidad de niños con malformaciones que ha tenido la paciente, en otros solo si las ha tenido o no	Numérico o Booleano	Relevante
Antecedentes Obstétricos (Macrofetos)	En algunos casos se toma la cantidad de partos donde el niño ha pesado más de 9 libras, en otros solo si los ha tenido o no	Numérico o Booleano	Relevante
Antecedentes Obstétricos (Muerte Perinatal)	Si algún niño de la paciente ha muerto antes del parto	Booleano	Relevante
Antecedentes Obstétricos (Menopausia)	Edad en la que la paciente tuvo su última menstruación	Numérico	Relevante
APF DM (Antecedentes	Qué antecesores del	Nominal	Relevante

Patológicos Familiares de Diabetes Mellitus)	paciente padecen diabetes mellitus		
APP (Antecedentes Patológicos Personales)	Qué patologías presenta el paciente a parte de la diabetes	Nominal	Relevante
APF Otras	Qué patologías diferentes de la diabetes mellitus presentan los ancestros del paciente	Nominal	Relevante
Tratamiento Inicial (Glibenclamida)	Si el tratamiento que le indican al paciente al debut es Glibenclamida	Booleano	Sin importancia
Tratamiento Inicial (Metformina)	Si el tratamiento que le indican al paciente al debut es Metformina	Booleano	Sin importancia
Tratamiento Inicial (Diabeton)	Si el tratamiento que le indican al paciente al debut es Diabeton	Booleano	Sin importancia
Tratamiento Inicial (Insulina)	Si el tratamiento que le indican al paciente al debut es Insulina	Booleano	Sin importancia
Tratamiento Inicial (Dieta)	Si el tratamiento que le indican al paciente al debut es Dieta	Booleano	Sin importancia
Tipo Dieta (Inicial)	Representa el tipo de dieta que presenta el paciente al inicio de su ingreso.	Numérico	Sin importancia
Tipo Dieta (Final)	Representa el tipo de dieta que presenta el paciente al final de su	Numérico	Sin importancia

	ingreso.		
TGP	Resultado del análisis Transaminasa Glutámico Pirática.	Numérico	Relevante
Glicemia (Inicio)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa en Ayunas al inicio del ingreso	Numérico	Relevante
Glicemia (PPD-Postprandial)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa 2 horas después de ser alimentado al inicio del ingreso.	Numérico	Relevante
Glicemia (Final)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa en Ayunas al final del ingreso	Numérico	Sin importancia
Glicemia (PPD-Postprandial)	Refleja la medida determinada en la Prueba de Tolerancia a la Glucosa 2 horas después de ser alimentado al final del ingreso.	Numérico	Sin importancia
Creatinina	Resultado del análisis que expresa la función renal	Numérico	Relevante
Micro albuminuria (1)	Medición de la	Numérico	Relevante

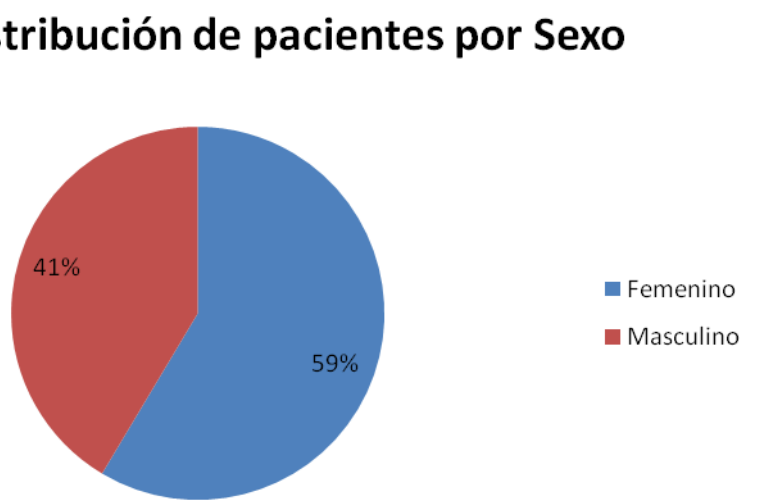
	cantidad de albumina excretada en la orina		
Micro albuminuria (2)	Si la medición de la cantidad de albumina excretada en la orina es positiva o negativa.	Nominal	Sin importancia
Eritro	Resultado del análisis que expresa la velocidad de sedimentación de los eritrocitos (glóbulos rojos)	Numérico	Relevante
Hemoglobina	Resultado del análisis realizado para determinar los valores de hemoglobina en sangre.	Numérico	Relevante
Triglicérido	Refleja los valores de triglicérido en sangre	Numérico	Relevante
Acido Úrico	Representa el nivel de ácido úrico en sangre	Numérico	Relevante
Colesterol	Resultado del análisis realizado para determinar el nivel del colesterol en la sangre	Numérico	Relevante
HDL-c	Refleja la medida de la cantidad de colesterol del tipo HDL del paciente.	Numérico	Sin importancia
Tratamiento Final (Glibenclamida)	Si el tratamiento actual del paciente es	Booleano	Sin importancia

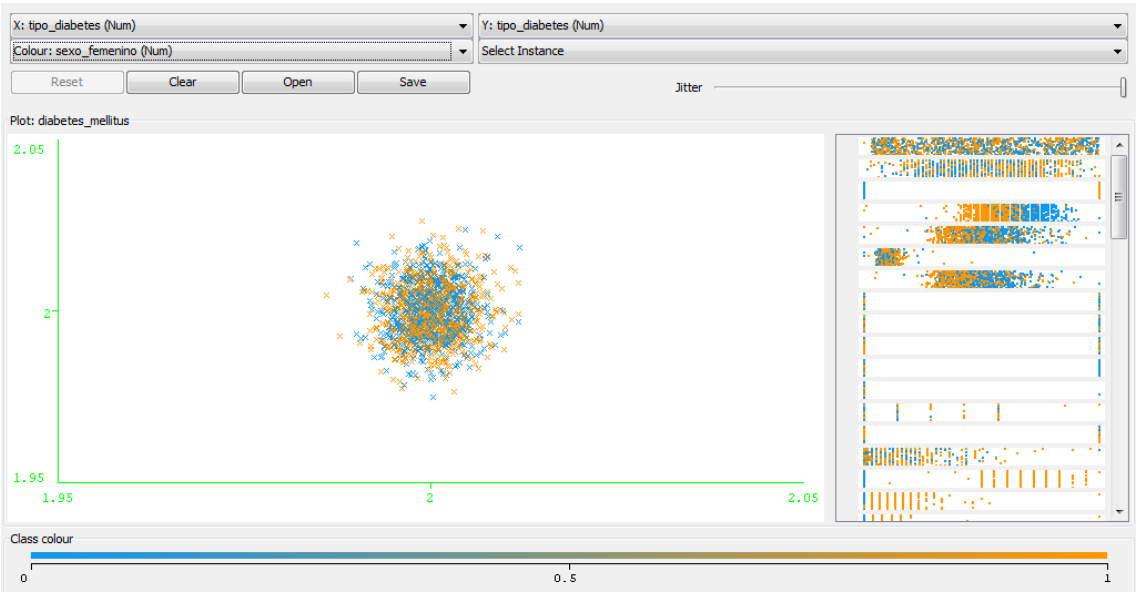
	con Glibenclamida		
Tratamiento Final (Metformina)	Si el tratamiento actual del paciente es con Metformina	Booleano	Sin importancia
Tratamiento Final (Diabeton)	Si el tratamiento actual del paciente es con Diabeton	Booleano	Sin importancia
Tratamiento Final (Insulina)	Si el tratamiento actual del paciente es con Insulina	Booleano	Sin importancia
Tratamiento Final (Dieta)	Si el tratamiento actual del paciente es con Dieta	Booleano	Sin importancia
Con Complicación Micro (RD)	El paciente presenta una retinopatía diabética	Booleano	Relevante
Con Complicación Micro (ND)	El paciente presenta una nefropatía diabética.	Booleano	Relevante
Con Complicación Micro (Neuro.D)	El paciente presenta una neuropatía diabética.	Booleano	Relevante
Con Complicación Macro (CI)	El paciente presenta una cardiopatía isquémica	Booleano	Relevante
Con Complicación Macro (AVE)	El paciente ha sufrido algún accidente vascular encefálico	Booleano	Relevante
Con Complicación Macro (IAP)	El paciente presenta alguna insuficiencia arterial periférica.	Booleano	Relevante
Pie con Riesgo	Se toma un número que representa el tipo del pie diabético que	Numérico	Sin importancia

	presenta el paciente		
Tipo de Diabético	Tipo de diabetes que presenta el paciente. Se representa con un número.	Numérico	Relevante
Reingreso	Se señalan los pacientes que ya habían ingresado antes en el centro	Booleano	Sin importancia
Test de Inicio (Suficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es suficiente	Booleano	Sin importancia
Test de Inicio (Necesario)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es necesario	Booleano	Sin importancia
Test de Inicio (Excelente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es excelente	Booleano	Sin importancia
Test de Inicio (Notable)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es notable	Booleano	Sin importancia
Test de Inicio (Insuficiente)	Si el resultado del test inicial realizado al paciente arroja que	Booleano	Sin importancia

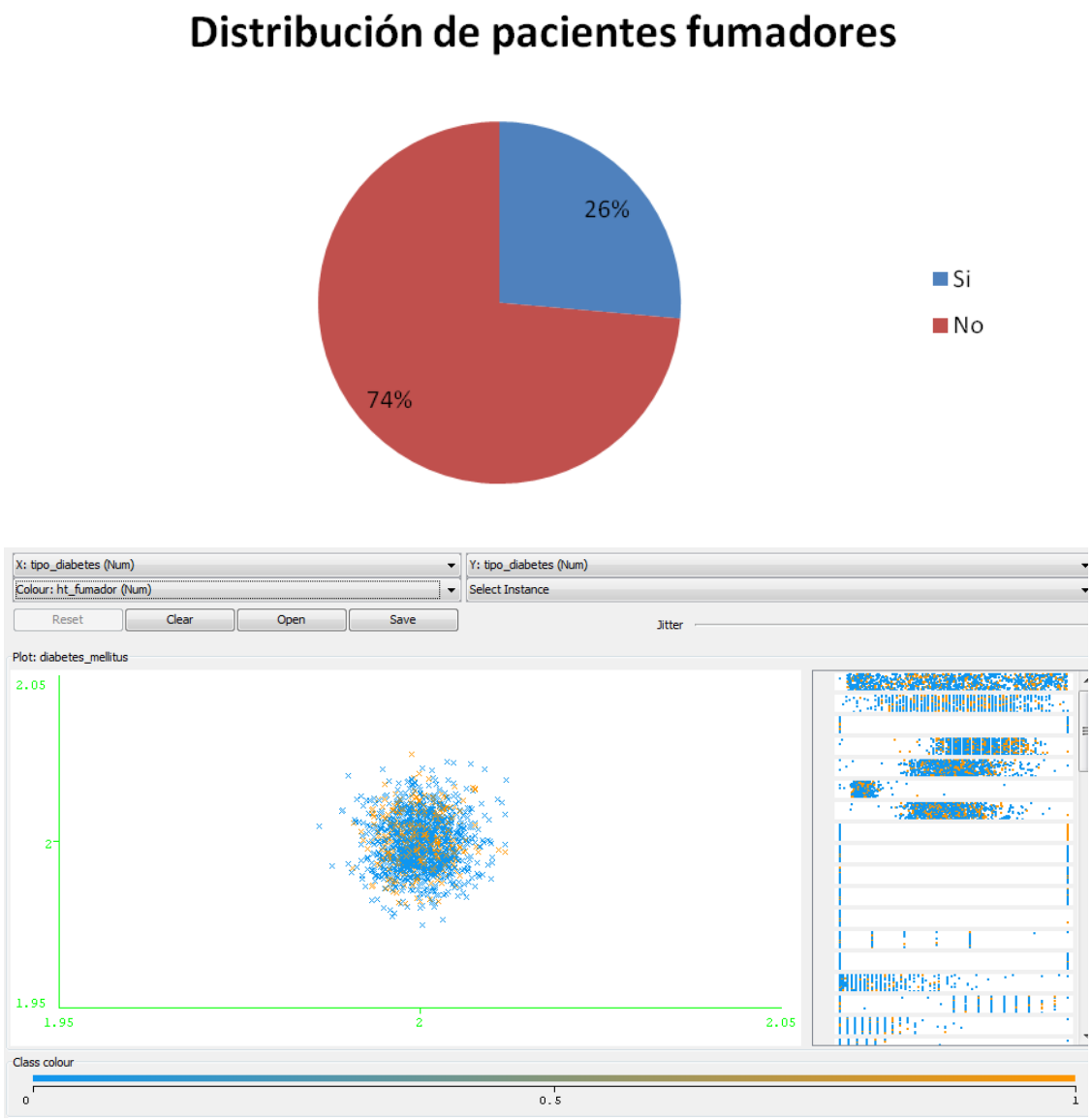
	su conocimiento sobre DM es insuficiente		
Test Final (Suficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es suficiente	Booleano	Sin importancia
Test Final (Necesario)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es necesario	Booleano	Sin importancia
Test Final (Excelente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es excelente	Booleano	Sin importancia
Test Final (Notable)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es notable	Booleano	Sin importancia
Test Final (Insuficiente)	Si el resultado del test inicial realizado al paciente arroja que su conocimiento sobre DM es insuficiente	Booleano	Sin importancia

Anexo 3: “Distribución de pacientes por Sexo (gráficas de Excel y Weka)”.

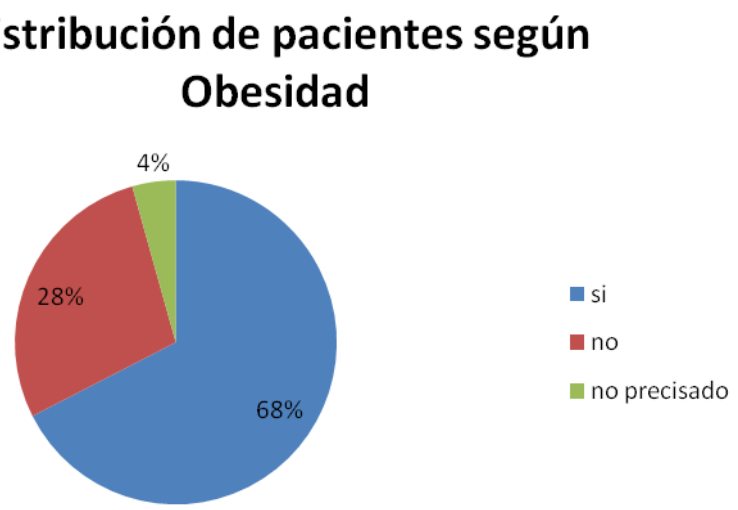


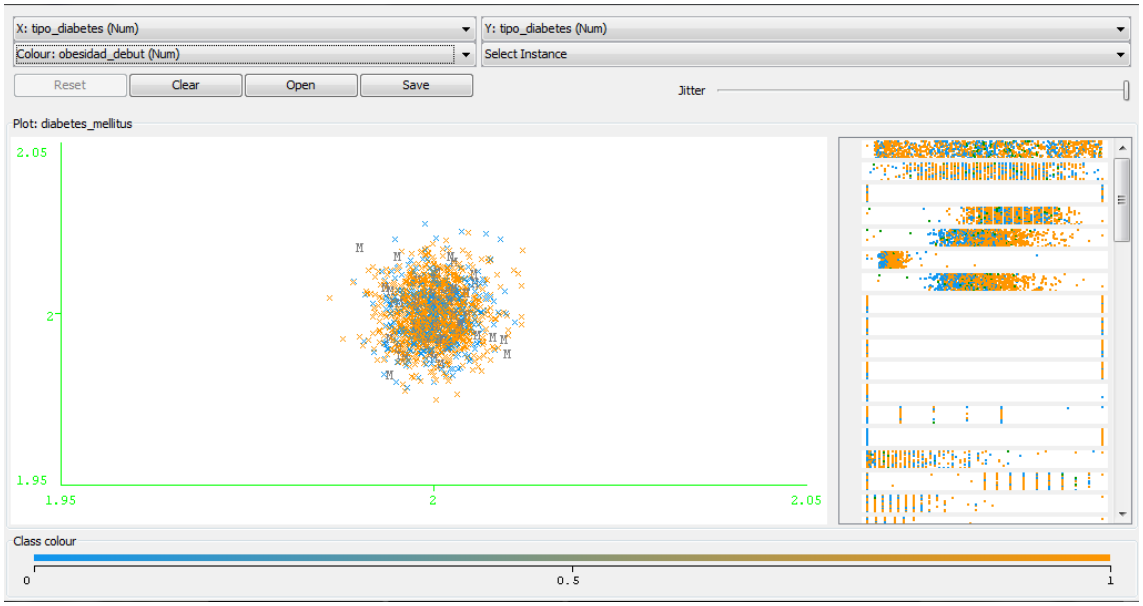


Anexo 4: “Distribución de pacientes por Hábito tóxico fumar (gráficas de Excel y Weka)”.

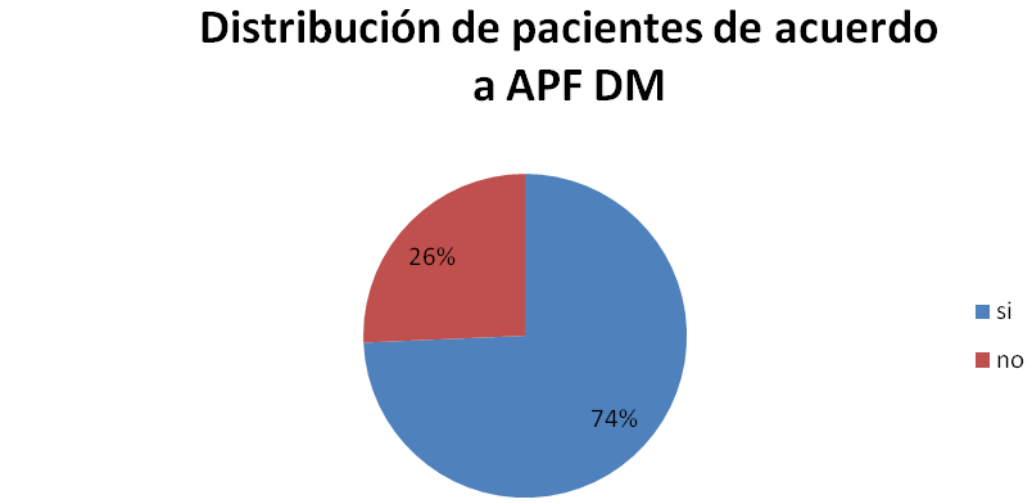


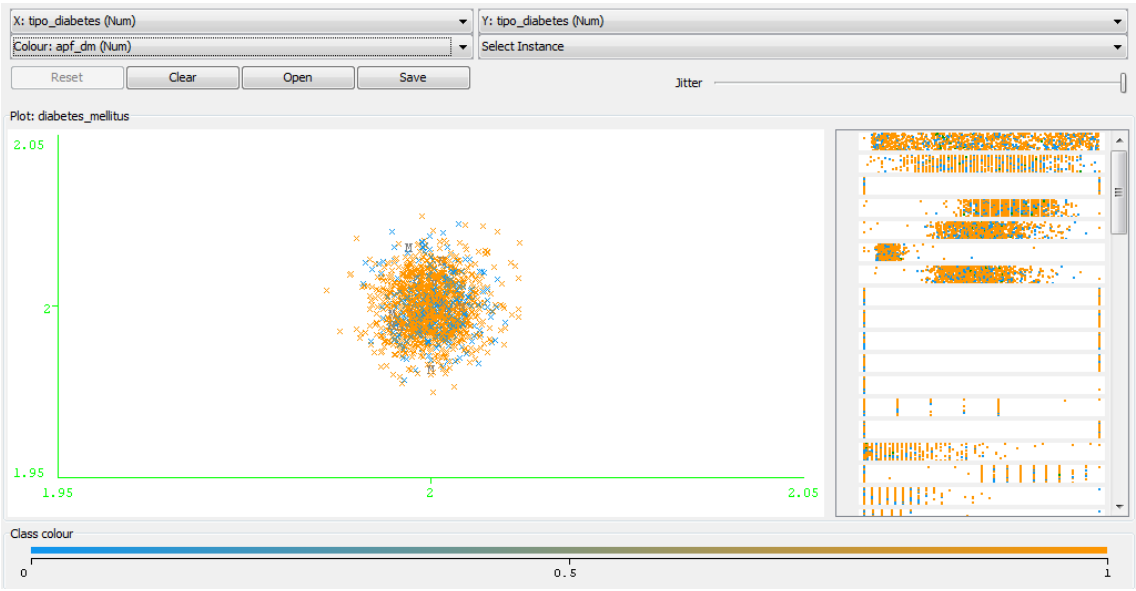
Anexo 5: “Distribución de pacientes por Obesidad (gráficas de Excel y Weka)”.



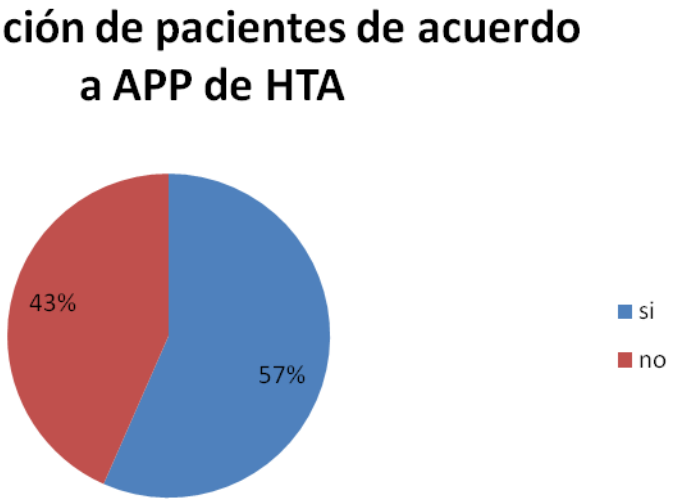


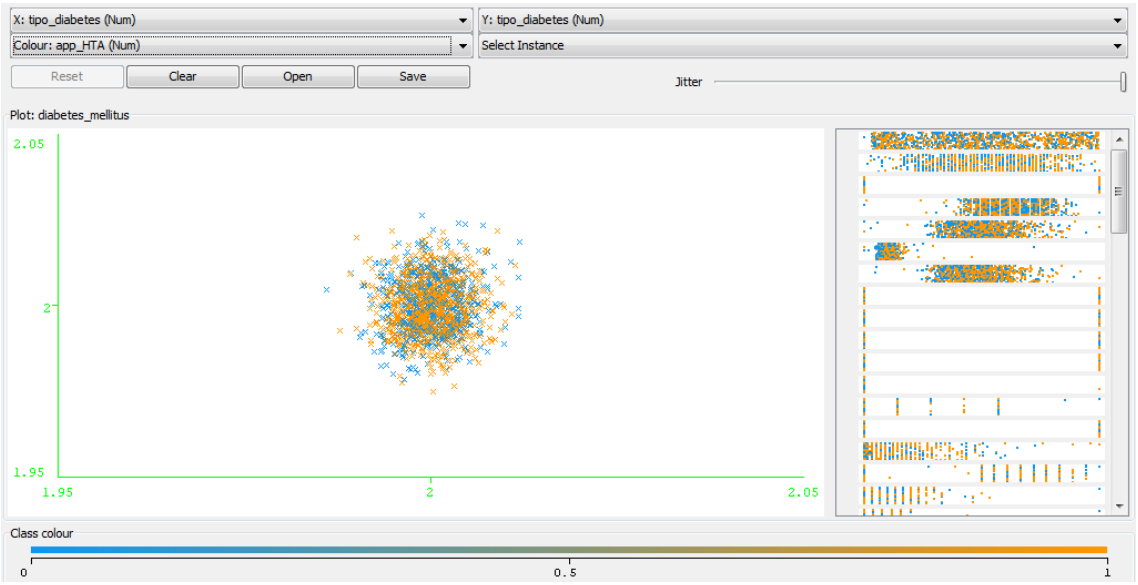
Anexo 6: “Distribución de pacientes por Antecedentes Patológicos Familiares de Diabetes Mellitus (gráficas de Excel y Weka)”.





Anexo 7: “Distribución de pacientes por Antecedentes Patológicos Personales de HTA (gráficas de Excel y Weka)”.





Anexo 8: “Sustitución de valores imprecisos en la variable Talla”.

Historia Clínica	Talla	Talla sustituida
469	1.64.5	1.64
470	166	1.66
477	1.54.5	1.54
478	164.5	1.64
485	1.51.5	1.51
487	1.67.5	1.67
488	1.63.5	1.63
492	1.71.5	1.71
565	!,70	1.70
1121	17.6	1.76
1123	18.2	1.82
1135	15.4	1.54

Historia Clínica	Talla	Talla sustituida
1141	17.3	1.73
1180	17.30	1.73
1183	17.3	1.73
1219	14.7	1.47
1364	161	1.61
1392	41.42	1.42
1414	6.9	
1415	162	1.62
1434	180	1.8
1722	179	1.79
2160	174	1.74
2213	177	1.77

Anexo 9: “Sustitución de valores imprecisos en la variable Peso inicial”.

Historia Clínica	Peso inicial	Peso inicial sustituido	IMC
1329	10.1	101	33.7
1353	0.95	95	31
1367	1.15		
1414	5.8		7.8
1547	77 1/2	77.5	26.3
1555	69 1/2	69.5	24.2
1567	82 1/2	82.5	32.9
1585	92 1/2	92.5	35.1
1596	11.5	115	39.7
1651	92 1/2	92.5	31.5
1697	90 1/2	90.5	30.7
1703	1.63	78.9	29.7
1826	66 1/2	66.5	25.1
1827	79 1/2	79.5	33.3

Anexo 10: “Sustitución de valores imprecisos en la variable Peso final”.

Historia Clínica	Peso final	Peso final Sustituido
1329	10.1	101
1414	9.4	
1513	94 1/2	94.5
1541	66 1/2	66.5
1544	73 1/2	73.5
1563	85 1/2	85.5
1565	101 1/2	101.5
1580	55 1/2	55.5
1585	91 1/2	91.5
1612	59 1/2	59.5
1638	84 1/2	84.5
1640	63 1/2	63.5
1642	92 1/2	92.5
1699	82 1/2	82.5
1827	78 1/2	78.5
1828	831/2	83.5
1830	95 1/2	95.5

Anexo 11: “Sustitución de valores imprecisos en la variable IMC”.

Historia Clínica	IMC	Cálculo IMC
632	55.30	36.11
730	283	28.23
833	72.54	25.25
848	261	26.12
868	205	20.52
1041	1.8	22.22
1047	61.4	35.56
1411	1.8	25.82
1414	7.8	
1438	2.93	29.34
1472	62.1	24.98
1494	97.6	37.7
1521	85	35.38
1610	67.1	31.77
1669	2.53	25.3
1671	67.8	24.39
1751	130.80	32.25
1801	55.5	27.36
2298	66.1	36.11

Anexo 12: “Variables añadidas referente a los padecimientos personales y familiares”.

Identificador	Valor	Justificación
APP HTA	1: El paciente padece de hipertensión arterial. 0: El paciente no padece de hipertensión arterial	Es una de las enfermedades presentes en la historia clínica y además predomina en los pacientes (52,78%)
APP hiperlipoproteinemia	1: El paciente padece de hiperlipoproteinemia. 0: El paciente no padece de hiperlipoproteinemia	Aunque no predomina en la muestra (9,9%) es una de las enfermedades de la historia clínica.
APP cardiopatía isquémica	1: El paciente padece de cardiopatía isquémica. 0: El paciente no padece de cardiopatía isquémica	Aunque no predomina en la muestra (10,3%) es una de las enfermedades de la historia clínica.
APP clauditación intermitente	1: El paciente padece de clauditación intermitente. 0: El paciente no padece de clauditación intermitente	Esta variable no predomina en la muestra, solo la padece el 1,3% de los pacientes, pero es especificada por los médicos en la historia clínica.
APP otros	1: El paciente padece una de las siguientes enfermedades: ave, asma, problemas renales,	En esta variable se añaden los pacientes que padecen ave porque aunque es una de las enfermedades explicitas

	hipercolesterolemia, neurosis, enfisema pulmonar, osteoporosis, glaucoma, hipotiroidismo, epilepsia y siclemia. 0: El paciente no padece ninguna de las enfermedades mencionadas.	en la historia clínica solo la padecen 3 personas y el resto son enfermedades que no están especificadas en la historia clínica y que no es considerable el número de pacientes que las padecen.
APF HTA	1: El paciente presenta antecedentes familiares con hipertensión arterial. 0: El paciente no presenta antecedentes familiares con hipertensión arterial.	Es una de las enfermedades presentes en la historia clínica y es considerable la cantidad de pacientes que tienen este antecedente. (16,6%)
APF hiperlipoproteinemia	1: El paciente presenta antecedentes familiares con hiperlipoproteinemia. 0: El paciente no presenta antecedentes familiares con hiperlipoproteinemia	Esta variable solo es positiva para un 4,17 % de los pacientes pero es especificada en la historia clínica
APF cardiopatía isquémica	1: El paciente presenta antecedentes familiares con cardiopatía isquémica. 0: El paciente no presenta antecedentes	El 11,99% de los pacientes presentan este antecedente y es especificado en la historia clínica.

	familiares con cardiopatía isquémica	
APF clauditación intermitente	1: El paciente presenta antecedentes familiares con cardiopatía isquémica. 0: El paciente no presenta antecedentes familiares con cardiopatía isquémica	Aunque solo presenta este antecedente el 1,85% de la muestra es una enfermedad que se especifica en la historia clínica
APF otros	1: El paciente presenta antecedentes familiares con ave, hipotiroidismo, asma, y dermatitis. 0: El paciente no presenta antecedentes familiares con estas enfermedades	La enfermedad AVE fue incluida, aunque está en la historia clínica, porque solo un paciente presenta este antecedente. El resto son enfermedades no especificadas en la historia clínica y que no es considerable su cantidad en la muestra.

Anexo 13: “Resultado del SPSS: proceso de agrupación automática”.

Agrupación automática

Número de conglomerados	Criterio bayesiano de Schwarz (BIC)	Cambio en BIC(a)	Razón de cambios en BIC(b)	Razón de medidas de distancia(c)
1	47104,503			
2	42255,270	-4849,233	1,000	4,697
3	41837,062	-418,208	,086	1,340
4	41722,780	-114,282	,024	1,108
5	41695,967	-26,813	,006	1,352
6	41879,150	183,183	-,038	1,003
7	42064,244	185,095	-,038	1,313
8	42391,159	326,914	-,067	1,027
9	42730,040	338,881	-,070	1,070
10	43097,853	367,813	-,076	1,027
11	43476,708	378,855	-,078	1,101
12	43892,518	415,810	-,086	1,170
13	44361,384	468,866	-,097	1,098
14	44858,184	496,800	-,102	1,008
15	45357,370	499,187	-,103	1,043

a Los cambios proceden del número anterior de conglomerados de la tabla.

b Las razones de los cambios están relacionadas con el cambio para la solución de los dos conglomerados.

c Las razones de las medidas de la distancia se basan en el número actual de conglomerados frente al número de conglomerados anterior.

Anexo 14: “Resultado de la ejecución del “EM” en el Weka con el total de casos”.

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 500 -N -1 -M 1.0E-6 -S 100
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
EM
==
Number of clusters selected by cross validation: 2
                                  Cluster
Attribute                              1       2
                                  (0.57)  (0.43)
=============================================
historia_clinica
  mean                           -0.0352  0.0472
  std. dev.                       1.0117  0.9814
edad
  mean                            0.1031 -0.1382
  std. dev.                       1.0061   0.974
sexo_femenino
  0                                    1     713
  1                                  956       1
  [total]                            957     714
talla
  mean                           -0.5966  0.8003
  std. dev.                       0.7371  0.7023
p_peso_inicial
  mean                           -0.2441  0.3274
  std. dev.                       0.9674  0.9474
p_indice_masa_corporal
  mean                            0.0959 -0.1286
  std. dev.                       1.0557  0.9034
ht_fumador
  0                                  787     502
  1                                  170     212
  [total]                            957     714
ht_cafe
  0                                  352     292
  1                                  605     422
  [total]                            957     714
ht_bebidas_alcoholicas
  0                                  925     512
  1                                   32     202
  [total]                            957     714
ht_otros
  0                                  956     711
  1                                    1       3
  [total]                            957     714
obesidad_debut
```

```
  0                                   253     210
  1                                   704     504
  [total]                             957     714
ao_menarca
  mean                             0.8049 -1.0796
  std. dev.                         0.477       1
ao_embarazos
  mean                             0.6294 -0.8443
  std. dev.                        0.9038       1
ao_abortos
  mean                             0.3798 -0.5094
  std. dev.                        1.1861       1
ao_malformaciones
  0                                   936     713
  1                                    21       1
  [total]                             957     714
ao_macrofetos
  0                                   801     713
  1                                   156       1
  [total]                             957     714
ao_muerte
  0                                   927     713
  1                                    30       1
  [total]                             957     714
ao_menopausia
  mean                             0.6029 -0.8087
  std. dev.                        0.9452       1
apf_dm
  0                                   179     154
  1                                   778     560
  [total]                             957     714
app_HTA
  0                                   345     316
  1                                   612     398
  [total]                             957     714
app_hiperlipoproteinemia
  0                                   850     655
  1                                   107      59
  [total]                             957     714
app_cardiopatia_isquemica
  0                                   811     627
  1                                   146      87
  [total]                             957     714
app_clauditacion_intermitente
  0                                   938     706
  1                                    19       8
  [total]                             957     714
app_otros
  0                                   814     625
  1                                   143      89
  [total]                             957     714
apf_HTA
  0                                   722     592
```

```
  1                                   235     122
  [total]                             957     714
apf_hiperlipoproteinemia
  0                                   908     698
  1                                    49      16
  [total]                             957     714
apf_cardiopatia_isquemica
  0                                   765     636
  1                                   192      78
  [total]                             957     714
apf_clauditacion_intermitente
  0                                   926     697
  1                                    31      17
  [total]                             957     714
apf_otros
  0                                   927     697
  1                                    30      17
  [total]                             957     714
tgp
  mean                             -0.1247  0.1672
  std. dev.                         0.8998  1.0978
glicemia_inicio
  mean                              0.0151 -0.0202
  std. dev.                         1.0037  0.9939
ppd_inicio
  mean                              0.0143 -0.0191
  std. dev.                         0.9985   1.001
creatinina
  mean                             -0.1964  0.2635
  std. dev.                         0.9874  0.9545
microalbuminuria
  mean                               0.024 -0.0322
  std. dev.                         1.0254  0.9633
eritro
  mean                               0.263 -0.3528
  std. dev.                         1.0306  0.8355

hemoglobina
  mean                             -0.3646   0.489
  std. dev.                         0.8555  0.9699
triglicerido
  mean                             -0.0075    0.01
  std. dev.                         0.9109  1.1076
acido_urico
  mean                             -0.1244  0.1668
  std. dev.                         0.9855  0.9943
colesterol
  mean                             -0.0065  0.0087
  std. dev.                         0.8887  1.1315
c_micro_rd
  0                                   885     649
  1                                    72      65
  [total]                             957     714
```

```
c_micro_nd
  0                                   922     681
  1                                    35      33
  [total]                             957     714
c_micro_neurod
  0                                   933     697
  1                                    24      17
  [total]                             957     714
c_macro_ci
  0                                   868     665
  1                                    89      49
  [total]                             957     714
c_macro_ave
  0                                   953     706
  1                                     4       8
  [total]                             957     714
c_macro_iap
  0                                   944     709
  1                                    13       5
  [total]                             957     714
Time taken to build model (full training data): 226.6 seconds
=== Model and evaluation on training set ===

Clustered Instances

0       936 (56%)
1       731 (44%)
Log likelihood: -32.81876
```

Anexo 15: “Resultado de la ejecución del “K medias” en el Weka con el total de casos (distancia Euclidiana y k=2)”.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 9
Within cluster sum of squared errors: 5845.842191657523
Missing values globally replaced with mean/mode
Cluster centroids:
                                          Cluster#
Attribute                    Full Data          0          1
                               (1667)     (1325)      (342)
=============================================================
historia_clinica                    0     0.2485    -0.9626
edad                                0    -0.0044     0.0171
sexo_femenino                       1          1          1
talla                               0     0.0841    -0.3259
p_peso_inicial                      0      0.074    -0.2868
p_indice_masa_corporal              0     0.0284    -0.1102
ht_fumador                          0          0          0
ht_cafe                             1          1          1
ht_bebidas_alcoholicas              0          0          0
ht_otros                            0          0          0
obesidad_debut                      1          1          1
ao_menarca                          0     -0.163     0.6314
ao_embarazos                        0    -0.1338     0.5185
ao_abortos                          0    -0.0735     0.2846
ao_malformaciones                   0          0          0
ao_macrofetos                       0          0          0
ao_muerte                           0          0          0
ao_menopausia                       0    -0.1084     0.4201
apf_dm                              1          1          1
app_HTA                             1          1          1
app_hiperlipoproteinemia            0          0          0
app_cardiopatia_isquemica           0          0          0
app_clauditacion_intermitente       0          0          0
app_otros                           0          0          0
apf_HTA                             0          0          1
apf_hiperlipoproteinemia            0          0          0
apf_cardiopatia_isquemica           0          0          1
apf_clauditacion_intermitente       0          0          0
apf_otros                           0          0          0
tgp                                 0     0.0164    -0.0636
```

```
glicemia_inicio                          0     -0.0546      0.2116
ppd_inicio                               0     -0.0432      0.1672
creatinina                               0      0.0256     -0.0991
microalbuminuria                         0      0.0133     -0.0513
eritro                                   0     -0.0109      0.0424
hemoglobina                              0      0.1096     -0.4246
triglicerido                             0      0.0144     -0.0558
acido_urico                              0      0.0548     -0.2124
colesterol                               0     -0.0087      0.0338
c_micro_rd                               0           0           0
c_micro_nd                               0           0           0
c_micro_neurod                           0           0           0
c_macro_ci                               0           0           0
c_macro_ave                              0           0           0
c_macro_iap                              0           0           0

Time taken to build model (full training data): 1.27 seconds

=== Model and evaluation on training set ===
Clustered Instances
0      1325 (79%)
1       342 (21%)
```

Anexo 16: “Resultado de la ejecución del “K medias” en el Weka con el total de casos (distancia Manhattan y k=2)”.

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.ManhattanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45

Test mode:    evaluate on training data


=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 4
Sum of within cluster distances: 8260.038719707341
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                      Full Data          0          1
                                  (1667)     (1312)      (355)
=============================================================
historia_clinica                 0.015     0.2882    -0.7943
edad                            0.0248     0.0248    -0.0624
sexo_femenino                        1          1          1
talla                          -0.0392    -0.0392    -0.1387
p_peso_inicial                 -0.1157    -0.1157    -0.2345
p_indice_masa_corporal         -0.1469    -0.1469    -0.2173
ht_fumador                           0          0          0
ht_cafe                              1          1          1
ht_bebidas_alcoholicas               0          0          0
ht_otros                             0          0          0
obesidad_debut                       1          1          1
ao_menarca                      0.2564     0.0894     0.7574
ao_embarazos                   -0.3869    -0.3869     0.0705
ao_abortos                     -0.5094    -0.5094    -0.5094
ao_malformaciones                    0          0          0
ao_macrofetos                        0          0          0
ao_muerte                            0          0          0
ao_menopausia                  -0.8087    -0.8087    -0.8087
apf_dm                               1          1          1
app_HTA                              1          1          1
app_hiperlipoproteinemia             0          0          0
app_cardiopatia_isquemica            0          0          0
app_clauditacion_intermitente        0          0          0
app_otros                            0          0          0
apf_HTA                              0          0          1
apf_hiperlipoproteinemia             0          0          0
apf_cardiopatia_isquemica            0          0          0
```

apf_clauditacion_intermitente	0	0	0
apf_otros	0	0	0
tgp	-0.1589	-0.1617	-0.141
glicemia_inicio	-0.2286	-0.2731	-0.1101
ppd_inicio	-0.2491	-0.2491	-0.165
creatinina	-0.177	-0.1433	-0.2159
microalbuminuria	-0.4822	-0.4822	-0.4822
eritro	-0.2202	-0.2202	-0.2729
hemoglobina	-0.1543	-0.1068	-0.2921
trigliceride	-0.1496	-0.1496	-0.1168
acido_urico	-0.0996	-0.0658	-0.1631
colesterol	-0.0796	-0.0838	-0.0125
c_micro_rd	0	0	0
c_micro_nd	0	0	0
c_micro_neurod	0	0	0
c_macro_ci	0	0	0
c_macro_ave	0	0	0
c_macro_iap	0	0	0

Time taken to build model (full training data): 1.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	1312	(79%)
1	355	(21%)

Anexo 17: “Resultado de la ejecución del “K medias” en el Weka con el total de casos (distancia Euclidiana y k=3)”.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 5
Within cluster sum of squared errors: 5392.779375846654
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data          0          1          2
                              (1667)      (341)      (327)      (999)
=============================================================================
historia_clinica                  0     0.0936    -0.9817     0.2894
edad                              0     0.0202     0.0168    -0.0124
sexo_femenino                     1          0          1          1
talla                             0     0.7062    -0.2879    -0.1468
p_peso_inicial                    0     0.2893    -0.2602    -0.0136
p_indice_masa_corporal            0    -0.1515    -0.1123     0.0885
ht_fumador                        0          1          0          0
ht_cafe                           1          1          1          1
ht_bebidas_alcoholicas            0          0          0          0
ht_otros                          0          0          0          0
obesidad_debut                    1          1          1          1
ao_menarca                        0    -0.8984     0.6067     0.1081
ao_embarazos                      0    -0.6887     0.5041     0.0701
ao_abortos                        0    -0.4218     0.2525     0.0613
ao_malformaciones                 0          0          0          0
ao_macrofetos                     0          0          0          0
ao_muerte                         0          0          0          0
ao_menopausia                     0    -0.6721      0.425     0.0903
apf_dm                            1          0          1          1
app_HTA                           1          1          1          1
app_hiperlipoproteinemia          0          0          0          0
app_cardiopatia_isquemica         0          0          0          0
app_clauditacion_intermitente     0          0          0          0
app_otros                         0          0          0          0
apf_HTA                           0          0          1          0
apf_hiperlipoproteinemia          0          0          0          0
apf_cardiopatia_isquemica         0          0          1          0
apf_clauditacion_intermitente     0          0          0          0
apf_otros                         0          0          0          0
tgp                               0     0.0756    -0.0503    -0.0093
glicemia_inicio                   0    -0.0267      0.181    -0.0501
ppd_inicio                        0    -0.0346     0.1401    -0.0341
creatinina                        0     0.2202     -0.111    -0.0388
microalbuminuria                  0    -0.0032    -0.0439     0.0155
eritro                            0    -0.2626     0.0046     0.0881
hemoglobina                       0     0.3597    -0.3684    -0.0022
```

trigliceridо	0	0.1047	-0.0566	-0.0172
acido_urico	0	0.1421	-0.1768	0.0094
colesterol	0	-0.0027	0.0515	-0.0159
c_micro_rd	0	0	0	0
c_micro_nd	0	0	0	0
c_micro_neurod	0	0	0	0
c_macro_ci	0	0	0	0
c_macro_ave	0	0	0	0
c_macro_iap	0	0	0	0

```
Time taken to build model (full training data): 0.89 seconds
=== Model and evaluation on training set ===
Clustered Instances
0       341 (20%)
1       327 (20%)
2       999 (60%)
```

Anexo 18: “Resultado de la ejecución del “K medias” en el Weka con el total de casos (distancia Manhattan y k=3)”.

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.ManhattanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 10
Sum of within cluster distances: 7251.370324280135
Missing values globally replaced with mean/mode
Cluster centroids:
```

Attribute	Full Data (1667)	Cluster# 0 (150)	1 (804)	2 (713)
historia_clinica	0.015	1.0377	-0.2581	0.0477
edad	0.0248	0.2865	0.112	-0.1496
sexo_femenino	1	1	1	0
talla	-0.0392	-0.7361	-0.6366	0.7573
p_peso_inicial	-0.1157	-0.2493	-0.4127	0.211
p_indice_masa_corporal	-0.1469	0.029	-0.0765	-0.2349
ht_fumador	0	1	0	0
ht_cafe	1	1	1	1
ht_bebidas_alcoholicas	0	0	0	0
ht_otros	0	0	0	0
obesidad_debut	1	1	1	1
ao_menarca	0.2564	0.9244	0.7574	-1.0796
ao_embarazos	-0.3869	0.5279	0.5279	-0.8443
ao_abortos	-0.5094	0.4238	0.4238	-0.5094
ao_malformaciones	0	0	0	0
ao_macrofetos	0	0	0	0
ao_muerte	0	0	0	0
ao_menopausia	-0.8087	0.9865	0.8484	-0.8087
apf_dm	1	0	1	1
app_HTA	1	1	1	1
app_hiperlipoproteinemia	0	0	0	0
app_cardiopatia_isquemica	0	0	0	0
app_clauditacion_intermitente	0	0	0	0
app_otros	0	0	0	0
apf_HTA	0	0	0	0
apf_hiperlipoproteinemia	0	0	0	0
apf_cardiopatia_isquemica	0	0	0	0
apf_clauditacion_intermitente	0	0	0	0
apf_otros	0	0	0	0
tgp	-0.1589	-0.242	-0.2485	0.0335
glicemia_inicio	-0.2286	-0.3027	-0.199	-0.199
ppd_inicio	-0.2491	-0.2491	-0.207	-0.2491
creatinina	-0.177	-0.4106	-0.3327	0.1513
microalbuminuria	-0.4822	-0.159	-0.4822	-0.4822
eritro	-0.2202	0.2542	0.0433	-0.5892
hemoglobina	-0.1543	-0.3918	-0.5344	0.3208
trigliceridо	-0.1496	-0.1455	-0.1209	-0.1989
acido_urico	-0.0996	-0.0171	-0.2424	0.051
colesterol	-0.0796	-0.0377	-0.0586	-0.1215

```
c_micro_rd                              0          0          0          0
c_micro_nd                              0          0          0          0
c_micro_neurod                          0          0          0          0
c_macro_ci                              0          0          0          0
c_macro_ave                             0          0          0          0
c_macro_iap                             0          0          0          0
Time taken to build model (full training data): 1.86 seconds
=== Model and evaluation on training set ===
Clustered Instances
0       150 (9%)
1       804 (48%)
2       713 (43%)
```

Anexo 19: "Resultado de la ejecución del "K medias" en el Weka con el total de casos (distancia Euclidiana y k=4)".

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 4
Within cluster sum of squared errors: 4908.6264097246485
Missing values globally replaced with mean/mode
Cluster centroids:
                                        Cluster#
Attribute                     Full Data         0          1          2          3
                                (1667)     (161)      (138)      (664)      (704)
=============================================================================================
historia_clinica                      0    0.1762    -0.8797     0.0774     0.0591
edad                                  0    0.3493     0.0419     0.0533    -0.1384
sexo_femenino                         1         1          1          1          0
talla                                 0   -0.6217    -0.4129    -0.6093     0.7978
p_peso_inicial                        0   -0.2113    -0.2978    -0.2343     0.3276
p_indice_masa_corporal                0    0.1688    -0.0799     0.1099    -0.1266
ht_fumador                            0         0          0          0          0
ht_cafe                               1         1          1          1          1
ht_bebidas_alcoholicas                0         0          0          0          0
ht_otros                              0         0          0          0          0
obesidad_debut                        1         1          1          1          1
ao_menarca                            0    0.8269     0.7719     0.7838    -1.0796
ao_embarazos                          0    0.6671     0.6472     0.5989    -0.8443
ao_abortos                            0    0.3716     0.2682     0.3942    -0.5094
ao_malformaciones                     0         0          0          0          0
ao_macrofetos                         0         0          0          0          0
ao_muerte                             0         0          0          0          0
ao_menopausia                         0     0.718     0.5265     0.5739    -0.8087
apf_dm                                1         0          1          1          1
app_HTA                               1         1          1          1          1
app_hiperlipoproteinemia              0         0          1          0          0
app_cardiopatia_isquemica             0         0          0          0          0
app_clauditacion_intermitente         0         0          0          0          0
app_otros                             0         0          0          0          0
apf_HTA                               0         0          1          0          0
apf_hiperlipoproteinemia              0         0          0          0          0
apf_cardiopatia_isquemica             0         0          1          0          0
apf_clauditacion_intermitente         0         0          0          0          0
apf_otros                             0         0          0          0          0
tgp                                   0   -0.1867    -0.1332     -0.106     0.1688
glicemia_inicio                       0   -0.0673     0.2253    -0.0104    -0.019
ppd_inicio                            0   -0.0285     0.1214     0.0008    -0.018
creatinina                            0   -0.2355    -0.1166    -0.1905     0.2563
microalbuminuria                      0   -0.0727     0.1395     0.0183    -0.028
eritro                                0    0.3468     0.0989     0.2688    -0.3522
hemoglobina                           0   -0.4022    -0.5435    -0.3112     0.4921
trigliceridо                          0   -0.0429     0.0471    -0.0081     0.0082
acido_urico                           0   -0.0266    -0.1947    -0.1323      0.169
colesterol                            0   -0.1383     0.1079      0.003     0.0077
c_micro_rd                            0         0          0          0          0
c_micro_nd                            0         0          0          0          0
c_micro_neurod                        0         0          0          0          0
c_macro_ci                            0         0          0          0          0
```

c_macro_ave	0	0	0	0	0
c_macro_iap	0	0	0	0	0

```
Time taken to build model (full training data): 0.81 seconds
=== Model and evaluation on training set ===
Clustered Instances
0       161 (10%)
1       138 (8%)
2       664 (40%)
3       704 (42%)
```

Anexo 20: “Resultado de la ejecución del “K medias” en el Weka con el total de casos (distancia Manhattan y k=4)”.

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 4 -A
"weka.core.ManhattanDistance -R first-last" -I 500 -S 10
Relation:     diabetes_mellitus
Instances:    1667
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 6
Sum of within cluster distances: 6939.252801087261
Missing values globally replaced with mean/mode
Cluster centroids:
                                        Cluster#
Attribute                  Full Data          0          1          2          3
                             (1667)     (162)      (274)      (526)      (705)
=============================================================================
historia_clinica             1392       1492      855.5     1682.5       1433
edad                           54         59         54         55         52
sexo_femenino                   1          1          1          1          0
talla                        1.62       1.56       1.58       1.56        1.7
p_peso_inicial                 75      71.75         70         70         81
p_indice_masa_corporal       28.5       29.3       28.2       29.1       27.9
ht_fumador                      0          0          0          0          0
ht_cafe                         1          1          1          1          1
ht_bebidas_alcoholicas          0          0          0          0          0
ht_otros                        0          0          0          0          0
obesidad_debut                  1          1          1          1          1
ao_menarca                      8         12         12         11          0
ao_embarazos                    1          3          3          3          0
ao_abortos                      0          1          0          1          0
ao_malformaciones               0          0          0          0          0
ao_macrofetos                   0          0          0          0          0
ao_muerte                       0          0          0          0          0
ao_menopausia                   0         39       39.5         34          0
apf_dm                          1          0          1          1          1
app_HTA                         1          1          1          1          1
app_hiperlipoproteinemia        0          0          0          0          0
app_cardiopatia_isquemica       0          0          0          0          0
app_clauditacion_intermitente   0          0          0          0          0
app_otros                       0          0          0          0          0
apf_HTA                         0          0          1          0          0
apf_hiperlipoproteinemia        0          0          0          0          0
apf_cardiopatia_isquemica       0          0          1          0          0
apf_clauditacion_intermitente   0          0          0          0          0
apf_otros                       0          0          0          0          0
tgp                         24.81      21.77     24.075         23         29
glicemia_inicio               7.4        7.1          8        7.1        7.5
ppd_inicio                    7.4       7.35       8.25        7.3        7.4
creatinina                     62         57     58.995         57      70.31
microalbuminuria                0          0          0          0          0
eritro                         22         27         25         29         15
hemoglobina                    14       13.3         13      13.46         15
trigliceridо                  1.8      1.805      1.885        1.8       1.73
acido_urico                274.53     261.55     255.63    266.765     292.25
colesterol                    5.1        5.3        5.5       5.19        4.9
c_micro_rd                      0          0          0          0          0
c_micro_nd                      0          0          0          0          0
c_micro_neurod                  0          0          0          0          0
c_macro_ci                      0          0          0          0          0
c_macro_ave                     0          0          0          0          0
c_macro_iap                     0          0          0          0          0
```

```
Time taken to build model (full training data): 1.59 seconds
=== Model and evaluation on training set ===
Clustered Instances
0       162 (10%)
1       274 (16%)
2       526 (32%)
3       705 (42%)
```

Anexo 21: “Distribución de casos al ejecutar el “K medias” en el SPSS con el total de datos”.

Número de casos en cada conglomerado (k=2)

Conglomerado	1	6.000
	2	1311.000
Válidos		1317.000
Perdidos		350.000

Número de casos en cada conglomerado (k= 3)

Conglomerado	1	830.000
	2	6.000
	3	481.000
Válidos		1317.000
Perdidos		350.000

Número de casos en cada conglomerado (k=4)

Conglomerado	1	496.000
	2	681.000
	3	6.000
	4	134.000
Válidos		1317.000
Perdidos		350.000

Anexo 22: “Resultado de la ejecución del “K medias” en el SPSS con el total de casos (k=2)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados	
	1	2
1	11.859	4.927
2	.036	3.505
3	1.971	4.406
4	.702	.837
5	.002	.001
6	4.389E-006	1.006E-006
7	1.097E-008	1.103E-009
8	2.743E-011	1.209E-012
9	6.921E-014	1.696E-015
10	6.308E-016	.000
11	1.110E-016	.000
12	1.110E-016	.000
13	1.213E-026	.000
14	3.032E-029	.000
15	7.579E-032	.000
16	1.895E-034	.000
17	4.737E-037	.000
18	1.184E-039	.000
19	2.961E-042	.000
20	7.402E-045	.000
21	1.850E-047	.000
22	4.626E-050	.000
23	1.157E-052	.000

24	2.891E-055	.000
25	7.228E-058	.000
26	1.807E-060	.000
27	4.518E-063	.000
28	1.129E-065	.000
29	2.824E-068	.000
30	7.059E-071	.000
31	1.765E-073	.000
32	4.412E-076	.000
33	1.103E-078	.000
34	2.757E-081	.000
35	6.894E-084	.000
36	1.723E-086	.000
37	4.308E-089	.000
38	1.077E-091	.000
39	2.693E-094	.000
40	6.732E-097	.000
41	1.683E-099	.000
42	4.207E-102	.000
43	1.052E-104	.000
44	2.630E-107	.000
45	6.574E-110	.000
46	1.644E-112	.000
47	4.109E-115	.000
48	1.027E-117	.000
49	2.568E-120	.000
50	6.420E-123	.000
51	1.605E-125	.000
52	4.013E-128	.000

53	1.003E-130	.000
54	2.508E-133	.000
55	6.270E-136	.000
56	1.567E-138	.000
57	3.919E-141	.000
58	9.796E-144	.000
59	2.449E-146	.000
60	6.123E-149	.000
61	1.531E-151	.000
62	3.827E-154	.000
63	9.567E-157	.000
64	2.392E-159	.000
65	5.881E-162	.000
66	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de .000. La iteración actual es 130. La distancia mínima entre los centros iniciales es de 18.320.

Número de casos en cada conglomerado

Conglomerado	1	562.000
	2	748.000
Válidos		1310.000
Perdidos		.000

Anexo 23: “Resultado de la ejecución del “K medias” en el SPSS con el total de casos (k=3)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados		
	1	2	3
1	5.319	4.927	11.205
2	4.018	2.816	.061
3	3.127	1.943	.439
4	.414	.051	.164
5	.002	.001	.000
6	5.982E-006	3.541E-005	1.603E-007
7	2.275E-008	9.319E-007	1.584E-010
8	8.649E-011	2.452E-008	1.565E-013
9	3.310E-013	6.454E-010	2.062E-016
10	9.296E-016	1.698E-011	.000
11	3.469E-018	4.460E-013	.000
12	.000	1.094E-014	.000
13	.000	1.340E-015	.000
14	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de .000. La iteración actual es 14. La distancia mínima entre los centros iniciales es de 17.109.

Número de casos en cada conglomerado

Conglomerado	1	489.000
	2	49.000
	3	772.000
Válidos		1310.000
Perdidos		.000

Anexo 24: “Resultado de la ejecución del “K medias” en el SPSS con el total de casos (k=4)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados			
	1	2	3	4
1	5.319	7.266	6.694	9.744
2	1.021	4.878	1.187	1.940
3	.102	.056	.217	.095
4	.010	.006	.007	.008
5	.001	8.954E-006	.000	1.411E-005
6	.000	1.319E-008	8.894E-006	2.367E-008
7	1.021E-005	1.942E-011	3.067E-007	3.971E-011
8	1.021E-006	2.886E-014	1.058E-008	6.634E-014
9	1.021E-007	9.061E-017	3.647E-010	1.852E-016
10	1.021E-008	.000	1.258E-011	9.630E-026
11	1.021E-009	.000	4.342E-013	1.616E-028
12	1.021E-010	.000	1.444E-014	2.711E-031
13	1.021E-011	.000	4.303E-016	4.549E-034
14	1.020E-012	.000	1.628E-019	7.632E-037
15	1.030E-013	.000	5.615E-021	1.281E-039
16	1.137E-014	.000	1.936E-022	2.149E-042
17	7.396E-016	.000	6.677E-024	3.605E-045
18	7.758E-017	.000	2.302E-025	6.049E-048
19	.000	.000	7.939E-027	1.015E-050
20	.000	.000	2.738E-028	1.703E-053
21	.000	.000	9.440E-030	2.857E-056
22	.000	.000	3.255E-031	4.794E-059
23	.000	.000	1.122E-032	8.043E-062
24	.000	.000	3.870E-034	1.350E-064
25	.000	.000	1.335E-035	2.264E-067

26	.000	.000	4.602E-037	3.799E-070
27	.000	.000	1.587E-038	6.374E-073
28	.000	.000	5.472E-040	1.070E-075
29	.000	.000	1.887E-041	1.794E-078
30	.000	.000	6.507E-043	3.011E-081
31	.000	.000	2.244E-044	5.052E-084
32	.000	.000	7.737E-046	8.476E-087
33	.000	.000	2.668E-047	1.422E-089
34	.000	.000	9.200E-049	2.386E-092
35	.000	.000	3.172E-050	4.004E-095
36	.000	.000	1.094E-051	6.718E-098
37	.000	.000	3.772E-053	1.127E-100
38	.000	.000	1.301E-054	1.891E-103
39	.000	.000	4.485E-056	3.173E-106
40	.000	.000	1.547E-057	5.324E-109
41	.000	.000	5.333E-059	8.933E-112
42	.000	.000	1.839E-060	1.499E-114
43	.000	.000	6.342E-062	2.515E-117
44	.000	.000	2.187E-063	4.219E-120
45	.000	.000	7.541E-065	7.079E-123
46	.000	.000	2.600E-066	1.188E-125
47	.000	.000	8.966E-068	1.993E-128
48	.000	.000	3.092E-069	3.344E-131
49	.000	.000	1.066E-070	5.611E-134
50	.000	.000	3.676E-072	9.414E-137
51	.000	.000	1.268E-073	1.580E-139
52	.000	.000	4.371E-075	2.650E-142
53	.000	.000	1.507E-076	4.447E-145
54	.000	.000	5.198E-078	7.461E-148
55	.000	.000	1.792E-079	1.252E-150

56	.000	.000	6.181E-081	2.100E-153
57	.000	.000	2.131E-082	3.524E-156
58	.000	.000	7.349E-084	5.913E-159
59	.000	.000	2.534E-085	9.940E-162
60	.000	.000	8.739E-087	.000
61	.000	.000	3.013E-088	.000
62	.000	.000	1.039E-089	.000
63	.000	.000	3.583E-091	.000
64	.000	.000	1.236E-092	.000
65	.000	.000	4.260E-094	.000
66	.000	.000	1.469E-095	.000
67	.000	.000	5.066E-097	.000
68	.000	.000	1.747E-098	.000
69	.000	.000	6.024E-100	.000
70	.000	.000	2.077E-101	.000
71	.000	.000	7.162E-103	.000
72	.000	.000	2.470E-104	.000
73	.000	.000	8.517E-106	.000
74	.000	.000	2.937E-107	.000
75	.000	.000	1.013E-108	.000
76	.000	.000	3.492E-110	.000
77	.000	.000	1.204E-111	.000
78	.000	.000	4.152E-113	.000
79	.000	.000	1.432E-114	.000
80	.000	.000	4.937E-116	.000
81	.000	.000	1.702E-117	.000
82	.000	.000	5.871E-119	.000
83	.000	.000	2.024E-120	.000
84	.000	.000	6.980E-122	.000
85	.000	.000	2.407E-123	.000

86	.000	.000	8.300E-125	.000
87	.000	.000	2.862E-126	.000
88	.000	.000	9.869E-128	.000
89	.000	.000	3.403E-129	.000
90	.000	.000	1.174E-130	.000
91	.000	.000	4.047E-132	.000
92	.000	.000	1.395E-133	.000
93	.000	.000	4.812E-135	.000
94	.000	.000	1.659E-136	.000
95	.000	.000	5.721E-138	.000
96	.000	.000	1.973E-139	.000
97	.000	.000	6.803E-141	.000
98	.000	.000	2.346E-142	.000
99	.000	.000	8.089E-144	.000
100	.000	.000	2.789E-145	.000
101	.000	.000	9.619E-147	.000
102	.000	.000	3.317E-148	.000
103	.000	.000	1.144E-149	.000
104	.000	.000	3.944E-151	.000
105	.000	.000	1.360E-152	.000
106	.000	.000	4.690E-154	.000
107	.000	.000	1.617E-155	.000
108	.000	.000	5.576E-157	.000
109	.000	.000	1.923E-158	.000
110	.000	.000	6.630E-160	.000
111	.000	.000	2.288E-161	.000
112	.000	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de .000. La iteración actual es 224. La distancia mínima entre los centros iniciales es de 15.725.

Número de casos en cada conglomerado

Conglomerado	1	13.000
	2	637.000
	3	46.000
	4	614.000
Válidos		1310.000
Perdidos		.000

Anexo 25: “Centroides de los grupos formados por los algoritmos aplicados al total de datos”.

Herramianta	**SPSS**										
Algoritmo	**Conglomerado en dos fases**		**Kmedias k=2**		**Kmedias k=3**			**Kmedias k=4**			
Cluster #	1	2	1	2	1	2	3	1	2	3	4
Edad (años)	52	55	58	50	59	55	50	57	51	44	56
Talla (m)	1.71	1.56	1.55	1.69	1.55	1.68	1.68	1.58	1.7	1.66	1.56
Peso (kg)	82.88	71.85	70.29	82.54	68.78	80.84	82.45	75.23	83.08	86.96	70.59
IMC	28.71	29.69	29.3	29.17	28.88	28.67	29.48	30.02	28.92	31.6	29.34
Menarca (años)	0	12	12	1	12	1	2	12	0	3	12
Embarazos	0	3	3	0	4	0	0	9	0	0	3
Abortos	0	0	1	0	1	0	0	7	0	0	0
Menopausia	0	33	39	0	41	4	2	30	0	6	35
TGP (u)	32.16	25.64	24.58	32.05	23.44	28.23	32.3	22.01	28.24	101.39	24.18
Glicemia (mmol/L)	8.02	8.39	8.33	8.13	8.42	7.98	8.1	10.24	7.99	7.67	8.45
PPD (mmol/L)	8.19	8.66	8.7	8.22	8.88	8.12	8.16	10.14	8.13	7.56	8.76
Creatinina (mmol/L)	73.65	61.74	61.61	72.14	61.06	145.73	66.82	72.35	73.87	57.52	61.8
Microalbuminuria (g/L)	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.04	0.03	0.01	0.02
Eritro (mm/h)	19.73	31.78	33.6	20.1	32.94	28.82	21.24	31.6	19.69	23.19	32.42
Hemoglobina (g/L)	15.32	13.19	13.15	15.06	13.1	16.25	14.83	13.15	15.3	14.63	13.13
Triglicérido (mmol/L)	1.97	1.96	1.99	1.94	1.98	2.45	1.93	1.35	1.94	2.15	1.99
Ácido Úrico	303.26	273.15	276.06	297.21	271.88	372.04	293.11	284.4	301.09	322.24	272.22

(mmol/L)											
Colesterol (mmol/L)	5.42	5.34	5.35	5.18	5.26	5.89	5.21	4.03	5.19	5.84	5.3
Sexo	0	1	1	0	1	0	0	1	0	0	1
Fumador	0	0	0	0	0	0	0	0	0	0	0
Café	1	1	1	1	1	1	1	1	1	0	1
Bebidas alcohólicas	0	0	0	0	0	0	0	0	0	0	0
Otros hábitos tóxicos	0	0	0	0	0	0	0	0	0	0	0
Obesidad	1	1	1	1	1	1	1	1	1	1	1
Antecedentes obstétricos (malformaciones)	0	0	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (macrofetos)	0	0	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (muerte)	0	0	0	0	0	0	0	0	0	0	0
APF DM	1	1	1	1	1	1	1	1	1	1	1
APP HTA	1	1	1	1	1	1	1	1	1	1	1
APP hiperlipoproteinemia	0	0	0	0	0	0	0	0	0	0	0
APP cardiopatia isquémica	0	0	0	0	0	0	0	0	0	0	0
APP clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0	0

APP otros	0	0	0	0	0	0	0	0	0	0	0
APF HTA	0	0	0	0	0	0	0	1	0	0	0
APF hiperlipoproteinemia	0	0	0	0	0	0	0	0	0	0	0
APF cardiopatia	0	0	0	0	0	0	0	0	0	0	0
APF clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0	0
APF otros	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (RD)	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (ND)	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (NeuroD)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (CI)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (AVE)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (IAP)	0	0	0	0	0	0	0	0	0	0	0

Herramianta	**Weka**										
Algoritmo	**EM**		**Kmeans (k=2, Euclidiana)**		**Kmeans (k=3, Euclidiana)**			**Kmeans (k=4, Euclidiana)**			
Cluster #	1	2	1	2	1	2	3	1	2	3	4
Edad (años)	55	52	53	53	53	53	53	57	54	54	52
Talla (m)	1.56	1.7	1.63	1.59	1.69	1.6	1.61	1.56	1.58	1.56	1.7

Peso (kg)	72.65	82.46	78.19	72.12	81.82	72.57	76.72	73.39	71.93	73	82.46
IMC	29.88	28.64	29.5	28.71	28.47	28.7	29.84	30.29	28.88	29.96	28.62
Menarca (años)	11	0	5	10	1	10	7	11	11	11	0
Embarazos	3	0	1	2	0	2	1	3	3	3	0
Abortos	0	0	0	0	0	0	0	0	0	0	0
Menopausia	31	0	15	26	2	26	19	33	29	30	0
TGP (u)	25.04	32.41	28.63	26.89	29.92	27.18	28.07	24.21	25.37	25.96	31.95
Glicemia (mmol/L)	8.26	8.05	7.99	8.89	8.08	8.78	8	7.94	8.93	8.14	8.11
PPD (mmol/L)	8.38	8.17	8.13	8.89	8.17	8.79	8.17	8.19	8.72	8.29	8.22
Creatinina (mmol/L)	61.16	73.44	67.2	64	72.2	63.69	65.55	60.5	63.55	61.65	73.13
Microalbuminuria (g/L)	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.03
Eritro (mm/h)	31.37	19.53	25.97	26.98	21.2	26.26	27.85	32.76	28.05	31.28	19.5
Hemoglobina (g/L)	13.53	15.34	14.56	13.43	15.08	13.55	14.32	13.48	13.18	13.67	15.36
Triglicérido (mmol/L)	1.95	2.02	2	1.91	2.11	1.91	1.96	1.93	2.04	1.97	1.99
Ácido Úrico (mmol/L)	271.36	301.92	290.39	262.94	299.36	266.6	285.72	282.03	264.75	271.17	302.13
Colesterol (mmol/L)	5.25	5.78	5.44	5.64	5.47	5.73	5.4	4.82	5.99	5.49	5.52
Sexo	1	0	1	1	0	1	1	1	1	1	0
Fumador	0	0	0	0	1	0	0	0	0	0	0
Café	1	1	1	1	1	1	1	1	1	1	1
Bebidas alcohólicas	0	0	0	0	0	0	0	0	0	0	0
Otros hábitos tóxicos	0	0	0	0	0	0	0	0	0	0	0
Obesidad	1	1	1	1	1	1	1	1	1	1	1
Antecedentes obstétricos (malformaciones)	0	0	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (macrofetos)	0	0	0	0	0	0	0	0	0	0	0

Antecedentes obstétricos (muerte)	0	0	0	0	0	0	0	0	0	0	0
APF DM	1	1	1	1	0	1	1	0	1	1	1
APP HTA	1	1	1	1	1	1	1	1	1	1	1
APP hiperlipoproteinemia	0	0	0	0	0	0	0	0	1	0	0
APP cardiopatia isquémica	0	0	0	0	0	0	0	0	0	0	0
APP clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0	0
APP otros	0	0	0	0	0	0	0	0	0	0	0
APF HTA	0	0	0	1	0	1	0	0	1	0	0
APF hiperlipoproteinemia	0	0	0	0	0	0	0	0	0	0	0
APF cardiopatia	0	0	0	0	0	1	0	0	0	0	0
APF clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0	0
APF otros	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (RD)	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (ND)	0	0	0	0	0	0	0	0	0	0	0
Complicación micro (NeuroD)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (CI)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (AVE)	0	0	0	0	0	0	0	0	0	0	0
Complicación macro (IAP)	0	0	0	0	0	0	0	0	0	0	0

Herramianta	**Weka**								
Algoritmo	**Kmeans (k=2, Manhattan)**		**Kmeans (k=3, Manhattan)**			**Kmeans (k=4, Manhattan)**			
Cluster #	1	2	1	2	3	1	2	3	4
Edad (años)	53	53	57	54	51	57	53	54	51
Talla (m)	1.63	1.62	1.56	1.56	1.7	1.56	1.58	1.56	1.71
Peso (kg)	77.46	75.05	74.47	72.41	82.59	73.44	71.82	73.14	82.59
IMC	29.45	28.91	30.42	29.73	28.66	30.27	28.94	30.2	28.63
Menarca (años)	6	7	11	11	0	11	11	10	0
Embarazos	1	2	3	3	0	3	3	3	0
Abortos	0	0	0	0	0	0	0	0	0
Menopausia	17	19	32	30	0	33	29	30	0
TGP (u)	28.39	27.82	24.82	25.68	31.92	23.86	26.15	25.64	32.07
Glicemia (mmol/L)	8.05	8.61	7.83	8.31	8.09	7.94	9	7.93	8.08
PPD (mmol/L)	8.25	8.43	8.05	8.42	8.19	8.22	8.98	8.09	8.19
Creatinina (mmol/L)	66.71	65.95	58.15	61.62	73.87	60.22	62.79	61.1	73.52
Microalbuminuria (g/L)	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.03
Eritro (mm/h)	26.54	24.84	36.76	30.31	19.29	32.41	28.34	32.26	19.37
Hemoglobina (g/L)	14.42	13.98	13.57	13.55	15.36	13.47	13.21	13.74	15.39
Triglicérido (mmol/L)	1.98	1.99	2.11	1.94	2	1.93	1.91	2.02	2
Ácido Úrico (mmol/L)	288.17	272.16	297.51	267.51	301.53	280.67	260.02	274.79	302.76
Colesterol	5.48	5.48	5.49	5.43	5.53	4.91	5.72	5.48	5.52

(mmol/L)									
Sexo	1	1	1	1	0	1	1	1	0
Fumador	0	0	1	0	0	0	0	0	0
Café	1	1	1	1	1	1	1	1	1
Bebidas alcohólicas	0	0	0	0	0	0	0	0	0
Otros hábitos tóxicos	0	0	0	0	0	0	0	0	0
Obesidad	1	1	1	1	1	1	1	1	1
Antecedentes obstétricos (malformaciones)	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (macrofetos)	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (muerte)	0	0	0	0	0	0	0	0	0
APF DM	1	1	0	1	1	0	1	1	1
APP HTA	1	1	1	1	1	1	1	1	1
APP hiperlipoproteinemia	0	0	0	0	0	0	0	0	0
APP cardiopatia isquémica	0	0	0	0	0	0	0	0	0
APP clauditacion_intermitente	0	0	0	0	0	0	0	0	0
APP otros	0	0	0	0	0	0	0	0	0
APF HTA	0	1	0	0	0	0	1	0	0
APF	0	0	0	0	0	0	0	0	0

hiperlipoproteinemia									
APF cardiopatia	0	0	0	0	0	0	1	0	0
APF clauditacion_intermitente	0	0	0	0	0	0	0	0	0
APF otros	0	0	0	0	0	0	0	0	0
Complicación micro (RD)	0	0	0	0	0	0	0	0	0
Complicación micro (ND)	0	0	0	0	0	0	0	0	0
Complicación micro (NeuroD)	0	0	0	0	0	0	0	0	0
Complicación macro (CI)	0	0	0	0	0	0	0	0	0
Complicación macro (AVE)	0	0	0	0	0	0	0	0	0
Complicación macro (IAP)	0	0	0	0	0	0	0	0	0

Anexo 26: “Resultado de la aplicación de los índices de validación al total de datos”.

Herramienta	Medida de distancia	Algoritmos	Ball	CH	Dunn	RMSSTD	RS
Weka		EM	2.63E+08	5.3348	0.1261	83.733	0.0032
	Euclidiana	Kmeans2	2.03E+08	497.1127	1.4554	73.5968	0.2299
		Kmeans 3	1.35E+08	251.137	0.2249	73.5045	0.2319
		Kmeans4	1.23E+08	41.6764	0.1015	80.8827	0.0699
	Manhattan	Kmeans2	2.25E+08	287.4755	1.101	77.4471	0.1472
		Kmeans3	1.61E+08	78.3512	0.2879	80.1772	0.0861
		Kmeans4	1.07E+08	126.2242	0.1398	75.692	0.1855
SPSS		Conglomerado en dos fases	2.40E+08	17.7188	0.2253	89.9758	0.0133
		Kmedias 2	2.43E+08	0.1513	0.1763	90.5752	1.15E-04
		Kmedias 3	1.61E+08	4.2679	0.158	90.2881	0.0065
		Kmedias 4	1.20E+08	5.2913	0.1508	90.0389	0.012

Anexo 27: “: “Resultado de la ejecución del “EM” en el Weka con los datos de los hombres”.

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 1000 -N -1 -M 1.0E-6 -S 100
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    712
Attributes:   38
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
EM
==
Number of clusters selected by cross validation: 3
                                Cluster
Attribute                            0        1        2
                                (0.31)   (0.15)   (0.54)
============================================================
historia_clinica
  mean                           0.5077  -0.0961  -0.2612
  std. dev.                      0.8777   0.9556   0.9672
edad
  mean                          -0.0481  -0.2327   0.0942
  std. dev.                      1.0622   0.9944   0.9497
sexo_femenino
  0                            219.2363 111.0886 384.6751
  1                                   1        1        1
  [total]                      220.2363 112.0886 385.6751
talla
  mean                           -0.184  -0.0582   0.1213
  std. dev.                      0.9726   1.0069   0.9946
p_peso_inicial
  mean                           0.1668   -0.146   -0.053
  std. dev.                      1.1271   1.0873   0.8755
p_indice_masa_corporal
  mean                           0.2259   0.0136  -0.1324
  std. dev.                       1.168   1.1739    0.797
ht_fumador
  0                            162.3366   68.415 273.2484
  1                             57.8998  43.6735 112.4267
  [total]                      220.2363 112.0886 385.6751
ht_cafe
  0                             91.0522  40.5091 162.4387
  1                            129.1841  71.5795 223.2364
  [total]                      220.2363 112.0886 385.6751
ht_bebidas_alcoholicas
  0                            166.4973  76.5934 270.9093
  1                             53.7391  35.4951 114.7658
  [total]                      220.2363 112.0886 385.6751
ht_otros
  0                             219.229 111.0886 382.6824
  1                              1.0073        1   2.9927
  [total]                      220.2363 112.0886 385.6751
obesidad_debut
  0                             67.5229  33.3866 111.0904
  1                            152.7134  78.7019 274.5847
```

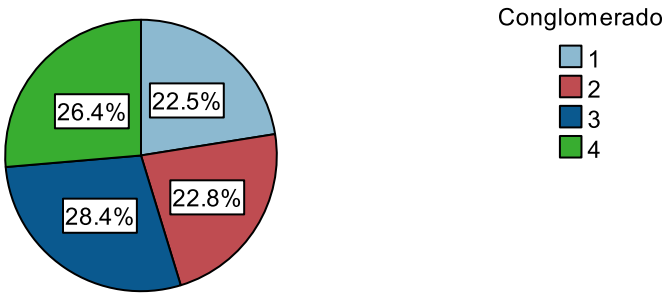
```
  [total]                          220.2363 112.0886 385.6751
apf_dm
  0                                 45.6002  27.6523  82.7475
  1                                174.6361  84.4362 302.9276
  [total]                          220.2363 112.0886 385.6751
app_HTA
  0                                  90.208  54.4806 173.3114
  1                                130.0283   57.608 212.3637
  [total]                          220.2363 112.0886 385.6751
app_hiperlipoproteinemia
  0                                208.0237  104.722 344.2543
  1                                 12.2126   7.3666  41.4208
  [total]                          220.2363 112.0886 385.6751
app_cardiopatia_isquemica
  0                                204.7527  97.0325 327.2147
  1                                 15.4836   15.056  58.4604
  [total]                          220.2363 112.0886 385.6751
app_clauditacion_intermitente
  0                                 219.236 110.0879 378.6761
  1                                  1.0003   2.0007    6.999
  [total]                          220.2363 112.0886 385.6751
app_otros
  0                                210.3297  88.7763  327.894
  1                                  9.9067  23.3122  57.7811
  [total]                          220.2363 112.0886 385.6751
apf_HTA
  0                                198.7639  88.2285 307.0076
  1                                 21.4724  23.8601  78.6675
  [total]                          220.2363 112.0886 385.6751
apf_hiperlipoproteinemia
  0                                216.2477 110.0143  373.738
  1                                  3.9886   2.0743  11.9371
  [total]                          220.2363 112.0886 385.6751
apf_cardiopatia_isquemica
  0                                204.4938  99.6452 333.8611
  1                                 15.7426  12.4434  51.8141
  [total]                          220.2363 112.0886 385.6751
apf_clauditacion_intermitente
  0                                 216.319 108.9735 373.7075
  1                                  3.9173    3.115  11.9676
  [total]                          220.2363 112.0886 385.6751
apf_otros
  0                                217.6322 105.6817 375.6862
  1                                  2.6042   6.4069    9.989
  [total]                          220.2363 112.0886 385.6751
tgp
  mean                              -0.0318   0.2657  -0.0582
  std. dev.                          0.7988   1.5644   0.8739
glicemia_inicio
  mean                               -0.306     0.42   0.0535
  std. dev.                          0.6529   1.3794   0.9787
ppd_inicio
  mean                              -0.2954   0.5331   0.0151
  std. dev.                          0.7062   1.4447   0.9161
creatinina
  mean                              -0.1523   0.2488   0.0152
  std. dev.                            0.757   1.1044   1.0705
microalbuminuria
```

```
  mean                              0.3384    0.956  -0.4668
  std. dev.                         0.7342    1.868        1
eritro
  mean                             -0.0697   0.3681   -0.066
  std. dev.                         0.8275   1.3481   0.9467
hemoglobina
  mean                              0.0342    0.261  -0.0944
  std. dev.                         1.0429   1.2333   0.8769
triglicerido
  mean                             -0.2059   0.7938  -0.1106
  std. dev.                         0.4399   1.9263   0.6791
acido_urico
  mean                               0.031   0.2466  -0.0884
  std. dev.                         0.8822   1.3837   0.9139
colesterol
  mean                              -0.128   0.4377  -0.0527
  std. dev.                         0.2477   2.3772   0.3601
c_micro_rd
  0                               209.1634 100.8252 341.0114
  1                                11.0729  11.2634  44.6637
  [total]                         220.2363 112.0886 385.6751
c_micro_nd
  0                               209.2103 105.9964 367.7932
  1                                 11.026   6.0921  17.8819
  [total]                         220.2363 112.0886 385.6751
c_micro_neurod
  0                               215.5859  108.656 374.7581
  1                                 4.6504   3.4326   10.917
  [total]                         220.2363 112.0886 385.6751
c_macro_ci
  0                               205.1873 101.0748 360.7379
  1                                 15.049  11.0137  24.9372
  [total]                         220.2363 112.0886 385.6751
c_macro_ave
  0                                216.218 111.0883 380.6938
  1                                 4.0184   1.0003   4.9813
  [total]                         220.2363 112.0886 385.6751
c_macro_iap
  0                               219.2231 111.0591 380.7179
  1                                 1.0133   1.0295   4.9573
  [total]                         220.2363 112.0886 385.6751
Time taken to build model (full training data): 235.28 seconds
=== Model and evaluation on training set ===

Clustered Instances
0      273 (38%)
1       92 (13%)
2      347 (49%)
Log likelihood: -26.32166
```

Anexo 28: ": "Resultado de la ejecución del "Conglomerado en dos fases" en el SPSS con los datos de los hombres".

Tamaños de conglomerados



Tamaño de conglomerado más pequeño	145 (22.5%)
Tamaño de conglomerado más grande	183 (28.4%)
Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño	1.26

Anexo 29: ": "Resultado de la ejecución del "K medias" en el Weka con los datos de los hombres (k=2)".

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 1000 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    712
Attributes:   38
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 7
Within cluster sum of squared errors: 2062.063615072958
Missing values globally replaced with mean/mode
Cluster centroids:
                                          Cluster#
Attribute                      Full Data          0          1
                                   (712)      (318)      (394)
=============================================================
historia_clinica                      0    -0.5313     0.4288
edad                                  0     0.1521    -0.1227
sexo_femenino                         0          0          0
talla                                 0     0.0405    -0.0327
p_peso_inicial                        0    -0.1304     0.1053
p_indice_masa_corporal                0    -0.1952     0.1575
ht_fumador                            0          1          0
ht_cafe                               1          1          0
ht_bebidas_alcoholicas                0          0          0
ht_otros                              0          0          0
obesidad_debut                        1          1          1
apf_dm                                1          1          1
app_HTA                               1          1          1
app_hiperlipoproteinemia              0          0          0
app_cardiopatia_isquemica             0          0          0
app_clauditacion_intermitente         0          0          0
app_otros                             0          0          0
apf_HTA                               0          0          0
apf_hiperlipoproteinemia              0          0          0
apf_cardiopatia_isquemica             0          0          0
apf_clauditacion_intermitente         0          0          0
apf_otros                             0          0          0
tgp                                   0    -0.0237     0.0191
glicemia_inicio                       0     0.0813    -0.0656
ppd_inicio                            0     0.0585    -0.0472
creatinina                            0     0.0431    -0.0348
microalbuminuria                      0    -0.0504     0.0407
eritro                                0    -0.0752     0.0607
hemoglobina                           0    -0.0483      0.039
triglicerido                          0     0.0414    -0.0334
acido_urico                           0    -0.0677     0.0546
```

```
colesterol                                  0      0.049    -0.0395
c_micro_rd                                  0          0          0
c_micro_nd                                  0          0          0
c_micro_neurod                              0          0          0
c_macro_ci                                  0          0          0
c_macro_ave                                 0          0          0
c_macro_iap                                 0          0          0
Time taken to build model (full training data): 0.37 seconds
=== Model and evaluation on training set ===
Clustered Instances
0      318 (45%)
1      394 (55%)
```

Anexo 30: ": "Resultado de la ejecución del "K medias" en el Weka con los datos de los hombres (k=3)".

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 1000 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    712
Attributes:   38
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 5
Within cluster sum of squared errors: 1765.1339284172136
Missing values globally replaced with mean/mode
Cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1           2
                              (712)       (211)       (274)       (227)
=============================================================================
historia_clinica                 0     -0.3802      0.3992     -0.1285
edad                             0       0.102     -0.1322      0.0648
sexo_femenino                    0           0           0           0
talla                            0      0.0014      -0.013      0.0144
p_peso_inicial                   0      0.1116      0.2838     -0.4463
p_indice_masa_corporal           0      0.0798      0.3354      -0.479
ht_fumador                       0           1           0           0
ht_cafe                          1           1           0           1
ht_bebidas_alcoholicas           0           0           0           0
ht_otros                         0           0           0           0
obesidad_debut                   1           1           1           0
apf_dm                           1           1           1           1
app_HTA                          1           1           1           0
app_hiperlipoproteinemia         0           0           0           0
app_cardiopatia_isquemica        0           0           0           0
app_clauditacion_intermitente    0           0           0           0
app_otros                        0           0           0           0
apf_HTA                          0           0           0           0
apf_hiperlipoproteinemia         0           0           0           0
apf_cardiopatia_isquemica        0           0           0           0
apf_clauditacion_intermitente    0           0           0           0
apf_otros                        0           0           0           0
tgp                              0     -0.0326      0.1469     -0.1471
glicemia_inicio                  0       0.024     -0.0502      0.0383
ppd_inicio                       0       0.012     -0.0352      0.0313
creatinina                       0      0.0743     -0.0337     -0.0284
microalbuminuria                 0     -0.0974     -0.0099      0.1025
eritro                           0     -0.0766      0.1489     -0.1085
hemoglobina                      0     -0.0544      0.0553     -0.0162
trigliceridо                     0      0.1399     -0.0074     -0.1211
acido_urico                      0       0.124      0.1851     -0.3386
colesterol                       0     -0.0091     -0.1038      0.1338
c_micro_rd                       0           0           0           0
c_micro_nd                       0           0           0           0
c_micro_neurod                   0           0           0           0
c_macro_ci                       0           0           0           0
c_macro_ave                      0           0           0           0
c_macro_iap                      0           0           0           0
```

```
Time taken to build model (full training data): 0.3 seconds

=== Model and evaluation on training set ===
Clustered Instances
0      211 (30%)
1      274 (38%)
2      227 (32%)
```

Anexo 31: “Resultado de la ejecución del “K medias” en el SPSS con los datos de los hombres (k=4)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados			
	1	2	3	4
1	9.243	8.829	5.448	9.448
2	.475	.640	.000	.102
3	.414	.609	.000	.087
4	.508	.418	.000	.101
5	.364	.324	.000	.095
6	.390	.210	.000	.141
7	.317	.282	.000	.193
8	.341	.284	.000	.263
9	.292	.254	.000	.283
10	.155	.133	.000	.162
11	.163	.158	.000	.170
12	.161	.210	.000	.187
13	.151	.218	.000	.138
14	.110	.100	.000	.120
15	.073	.066	.000	.078
16	.055	.054	.000	.092
17	.027	.025	.000	.040
18	.027	.033	.000	.029
19	.031	.018	.000	.046
20	.016	.083	.000	.061
21	.023	.157	.000	.090
22	.007	.000	.000	.012
23	.000	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. La iteración actual es 23. La distancia mínima entre los centros iniciales es de 15.062.

Número de casos en cada conglomerado

Conglomerado	1	309.000
	2	130.000
	3	5.000
	4	203.000
Válidos		647.000
Perdidos		65.000

Anexo 32: “Resultado de la aplicación de los índices de validación a los datos de los hombres”.

Herramienta	Algoritmos	Ball	CH	Dunn	RMSSTD	RS
Weka	**EM**	22.6311	90.9858	0.9003	0.0501	0.2042
	km2E	38.5563	75.5566	0.6776	0.0534	0.0962
	km3E	24.5765	55.723	0.5159	0.0522	0.1358
SPSS	**Conglomerado en dos fases**	0.0445	14.5551	0.0258	0.0028	0.0636
	Kmedias 4	0.0394	43.8315	0.0911	0.0026	0.1698

Anexo 33: “Centroides de los grupos formados por los mejores algoritmos aplicados a los datos de los hombres”.

Algoritmo	**EM**			**K medias k=4**			
Cluster #	1	2	3	1	2	3	4
Edad (años)	50	49	53	57	50	51	45
Talla (m)	1.69	1.7	1.71	1.69	1.71	1.7	1.73
Peso (kg)	86.27	80.02	80.11	74.46	79.85	69.4	97.8
IMC	30.17	28.79	27.32	26.27	27.72	24.32	33.12
TGP (u)	31.65	40.23	29.91	25.91	33.09	24.37	41.16
Glicemia (mmol/L)	6.69	10.11	8.68	7.02	12.75	10.16	6.55
PPD (mmol/L)	6.83	10.38	8.74	7.13	13.11	10.6	6.64
Creatinina (mmol/L)	68.38	77.25	76.15	75.91	75.03	47.7	69.65
Microalbuminuria (g/L)	0.03	0.08	0.01	0.03	0.02	0.09	0.03
Eritro (mm/h)	18.2	28.9	18	20.19	18.35	12.99	20.09
Hemoglobina (g/L)	15.57	16.03	15.01	15.14	14.71	14.53	16.01
Triglicérido (mmol/L)	1.66	3.42	1.88	1.64	2.54	3.16	2.1
Ácido Úrico (mmol/L)	307.97	336.35	287.98	285.44	255.95	199.66	362.2
Colesterol (mmol/L)	4.6	8.38	5.49	4.82	5.8	55.64	5.49
Sexo	0	0	0	0	0	0	0
Fumador	0	0	0	0	0	0	0
Café	1	1	1	1	1	0	1
Bebidas alcohólicas	0	0	0	0	0	0	0
Otros hábitos tóxicos	0	0	0	0	0	0	0
Obesidad	1	1	1	1	1	0	1
APF DM	1	1	1	1	1	1	1
APP HTA	1	1	1	1	1	0	1
APP hiperlipoproteinemia	0	0	0	0	0	0	0
APP cardiopatia	0	0	0	0	0	0	0

isquémica							
APP clauditacion_intermitente	0	0	0	0	0	0	0
APP otros	0	0	0	0	0	0	0
APF HTA	0	0	0	0	0	0	0
APF hiperlipoproteinemia	0	0	0	0	0	0	0
APF cardiopatia	0	0	0	0	0	0	0
APF clauditacion_intermitente	0	0	0	0	0	0	0
APF otros	0	0	0	0	0	0	0
Complicación micro (RD)	0	0	0	0	0	0	0
Complicación micro (ND)	0	0	0	0	0	0	0
Complicación micro (NeuroD)	0	0	0	0	0	0	0
Complicación macro (CI)	0	0	0	0	0	0	0
Complicación macro (AVE)	0	0	0	0	0	0	0
Complicación macro (IAP)	0	0	0	0	0	0	0

Anexo 34: “Resultado de la ejecución del “EM” en el Weka con los datos de las mujeres”.

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 1000 -N -1 -M 1.0E-6 -S 100
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    955
Attributes:   45
Ignored:              sexo_femenino
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
EM
==
Number of clusters selected by cross validation: 8
```

	Cluster							
Attribute	0	1	2	3	4	5	6	7
	(0.06)	(0.19)	(0.06)	(0.07)	(0.08)	(0.24)	(0.16)	(0.13)
historia_clinica								
mean	-1.2394	1.1519	-0.2708	0.7413	0.511	-0.2317	-1.042	0.0195
std. dev.	0.4566	0.3575	0.6412	0.8175	0.4976	0.8105	0.4119	0.7903
edad								
mean	-0.059	0.0897	-0.3834	0.9967	-0.0675	0.0513	0.4756	-1.1152
std. dev.	0.8886	0.8268	0.9094	0.6738	0.9155	0.9756	0.6715	0.7796
talla								
mean	0.1278	-0.297	0.2936	-0.5548	-0.0765	-0.0106	-0.0227	0.6317
std. dev.	0.9725	0.8509	0.7083	0.7751	0.9544	0.8527	0.8517	1.4156
p_peso_inicial								
mean	0.0865	0.4348	0.1875	-0.6011	-0.0034	-0.5805	-0.0755	0.7463
std. dev.	0.9584	0.9637	1.3679	0.5808	0.8401	0.6113	0.8461	1.0381
p_indice_masa_corporal								
mean	0.0257	0.6069	0.4315	-0.4682	-0.0236	-0.6394	-0.049	0.425
std. dev.	0.8564	1.0445	1.0384	0.6954	0.8679	0.6305	0.9228	0.9937
ht_fumador								
0	49.2433	150.2595	55.5508	58.2311	57.7715	184.8555	130.266	107.8221
1	9.7159	33.4002	6.119	15.0923	15.9226	49.8051	25.8458	21.099
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ht_cafe								
0	21.0332	95.4492	22.5397	27.9487	26.8088	74.5631	32.0012	58.6562
1	37.9261	88.2106	39.1301	45.3747	46.8854	160.0976	124.1107	70.2649
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ht_bebidas_alcoholicas								
0	56.6827	180.5636	57.9823	71.0773	72.0621	221.6257	149.1675	122.839
1	2.2766	3.0962	3.6876	2.2462	1.6321	13.035	6.9443	6.0822
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ht_otros								
0	57.9593	182.6597	60.6698	72.3234	72.6941	233.6606	155.1118	127.9211
1	1	1	1	1	1	1	1	1
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
obesidad_debut								
0	12.8054	19.0924	10.8157	37.9452	19.9959	107.3747	28.8701	23.1007
1	46.1539	164.5673	50.8541	35.3782	53.6982	127.286	127.2418	105.8205
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ao_menarca								
mean	0.4355	0.2278	-0.0718	0.6311	-0.5279	-0.0517	0.3385	-0.8605
std. dev.	0.723	0.7547	0.8431	0.7175	1.1276	0.8573	0.591	1.3878
ao_embarazos								
mean	1.1123	0.0356	-0.2129	0.5257	-0.2913	-0.2188	0.4224	-0.6927
std. dev.	1.6204	0.7676	0.7055	1.3138	0.6181	0.6332	1.0896	0.5768
ao_abortos								
mean	0.991	0.1223	0.0036	0.0897	-0.1482	-0.2044	0.0515	-0.2759
std. dev.	2.4027	0.9097	0.7162	0.9346	0.5589	0.5871	1.0203	0.5494
ao_malformaciones								
0	54.2393	182.5927	58.1042	72.316	71.6911	231.5905	146.0265	126.4398
1	4.72	1.0671	3.5656	1.0075	2.003	3.0702	10.0853	2.4813
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ao_macrofetos								
0	34.6115	175.3543	49.3445	64.7872	72.6398	201.9584	89.5604	119.744
1	24.3478	8.3055	12.3253	8.5362	1.0544	32.7022	66.5515	9.1772
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ao_muerte								
0	51.032	181.9318	56.9067	72.0489	71.6937	230.0829	142.7729	127.5311
1	7.9273	1.7279	4.7631	1.2745	2.0005	4.5777	13.339	1.39
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
ao_menopausia								
mean	-0.2141	0.1652	-0.5857	0.834	-0.1026	-0.0023	0.8327	-1.2826

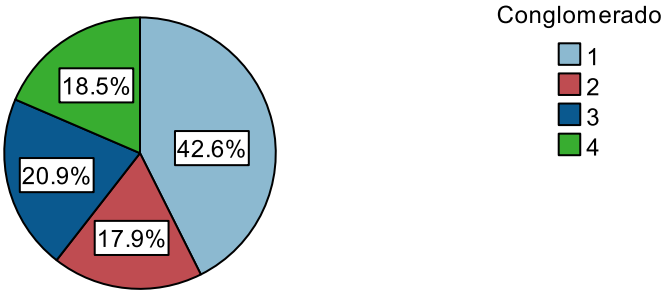
std. dev.	1.1203	0.9533	0.9065	0.2697	0.7313	0.9347	0.2662	0.3619
apf_dm								
0	9.2177	45.1941	12.5285	20.3921	11.5428	39.7333	31.3794	16.0122
1	49.7416	138.4657	49.1413	52.9313	62.1513	194.9274	124.7324	112.9089
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
app_HTA								
0	19.8437	39.6222	28.9804	17.2889	26.4171	110.3676	43.3602	66.1198
1	39.1156	144.0375	32.6894	56.0346	47.277	124.293	112.7516	62.8013
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
app_hiperlipoproteinemia								
0	38.3795	180.725	60.376	67.9814	64.2504	197.3206	130.1745	117.7926
1	20.5798	2.9347	1.2939	5.342	9.4437	37.34	25.9373	11.1286
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
app_cardiopatia_isquemica								
0	42.1555	174.0012	48.9635	67.1635	64.4533	198.6841	107.3945	115.1843
1	16.8038	9.6585	12.7063	6.1599	9.2408	35.9765	48.7174	13.7369
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
app_clauditacion_intermitente								
0	56.9004	182.6572	57.8259	69.2294	71.6697	227.2151	152.5592	126.9431
1	2.0589	1.0025	3.8439	4.094	2.0245	7.4455	3.5526	1.978
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
app_otros								
0	22.3606	182.6579	55.4348	71.0255	66.9575	211.2509	91.2001	120.1126
1	36.5986	1.0018	6.235	2.298	6.7367	23.4097	64.9117	8.8085
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
apf_HTA								
0	23.0484	180.0807	45.9482	69.9443	61.4033	162.013	90.8665	95.6956
1	35.9108	3.5791	15.7216	3.3792	12.2908	72.6476	65.2453	33.2256
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
apf_hiperlipoproteinemia								
0	41.8965	182.6463	60.5108	71.6309	69.5756	222.1332	148.5435	118.0632
1	17.0628	1.0135	1.159	1.6925	4.1185	12.5275	7.5683	10.8579
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
apf_cardiopatia_isquemica								
0	27.4468	180.6725	44.4705	69.8051	65.3622	175.505	104.4005	104.3373
1	31.5125	2.9872	17.1993	3.5183	8.3319	59.1556	51.7113	24.5838
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
apf_clauditacion_intermitente								
0	56.9412	182.6597	57.8543	70.2742	71.5582	226.5374	143.1994	123.9756
1	2.018	1	3.8155	3.0492	2.136	8.1232	12.9125	4.9456
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
apf_otros								
0	56.9079	182.6587	59.5865	72.2632	71.7965	228.8361	140.9311	121.0199
1	2.0514	1.001	2.0833	1.0602	1.8977	5.8245	15.1807	7.9012
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
tgp								
mean	-0.0413	0.1174	0.0863	-0.2979	0.0232	-0.2868	0.1139	0.3518
std. dev.	0.8507	0.9832	1.0173	0.6667	0.8041	0.626	1.037	1.5225
glicemia_inicio								
mean	1.0056	-0.3686	0.5666	0.3476	-0.2909	0.1666	-0.1444	-0.3512
std. dev.	1.6402	0.5894	1.0954	1.0913	0.6359	1.0969	0.6776	0.6873
ppd_inicio								
mean	0.7506	-0.2457	0.5546	0.5218	-0.3045	0.1446	-0.1387	-0.4638
std. dev.	1.4386	0.7936	0.9789	1.2494	0.6515	1.0967	0.6679	0.6375
creatinina								
mean	0.4201	-0.2045	-0.112	0.0793	0.675	-0.0163	-0.0152	-0.2206
std. dev.	1.618	0.5365	0.569	0.7321	2.3142	0.6806	0.7291	0.7079
microalbuminuria								
mean	0.5755	-0.0856	0.5794	0.2175	1.5616	-0.4934	-0.3371	-0.0987
std. dev.	1.4026	0.4152	1.5153	0.8855	1.995	1	0.2416	0.4824
eritro								
mean	-0.0067	0.2305	-0.0596	0.72	0.5493	-0.3104	0.0262	-0.4765
std. dev.	0.9407	1.0105	0.8336	1.3791	1.2622	0.7442	0.9511	0.5519
hemoglobina								
mean	-0.5088	0.0321	-0.3874	-0.4954	0.8442	-0.1236	-0.0923	0.5046
std. dev.	0.7532	0.7035	0.8645	0.6366	1.649	0.8123	0.8455	1.1572
triglicerido								
mean	-0.015	-0.128	0.2791	0.1843	0.9266	-0.1277	-0.0769	-0.2409
std. dev.	1.069	0.6784	1.4217	0.7752	2.1833	0.6631	0.665	0.6016
acido_urico								
mean	-0.2403	0.3143	-0.3165	0.0866	0.2617	-0.3351	0.2358	-0.0617
std. dev.	0.7227	0.8967	0.6627	1.0845	1.4073	0.8709	1.0114	0.9892
colesterol								
mean	-0.05	-0.1104	0.0725	0.1058	0.2136	-0.0028	0.0364	-0.0728
std. dev.	0.4614	0.217	3.5508	0.4388	0.5628	0.4432	0.6447	0.4626
c_micro_rd								
0	50.7447	178.9484	56.282	55.2075	64.877	215.185	147.8299	122.9255
1	8.2146	4.7113	5.3878	18.116	8.8171	19.4756	8.2819	5.9957
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
c_micro_nd								
0	53.4492	182.5904	59.8204	58.6757	67.5865	226.794	155.1114	124.9725
1	5.5101	1.0693	1.8495	14.6478	6.1077	7.8667	1.0004	3.9486
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211
c_micro_neurod								
0	55.9538	180.4421	55.5508	70.2926	69.9807	228.0416	154.155	125.5833
1	3.0054	3.2176	6.119	3.0309	3.7134	6.619	1.9568	3.3378
[total]	58.9593	183.6597	61.6698	73.3234	73.6941	234.6606	156.1118	128.9211

```
c_macro_ci
  0                          51.3917 162.5391  55.1892  57.3575  64.8421 210.5519 147.9431 125.1855
  1                           7.5676  21.1207   6.4806   15.966    8.852  24.1088   8.1688   3.7357
  [total]                    58.9593 183.6597  61.6698  73.3234  73.6941 234.6606 156.1118 128.9211
c_macro_ave
  0                          57.9394  182.636  60.6695  72.3195  72.6917 232.6326 154.1401 126.9712
  1                           1.0198   1.0238   1.0003    1.004   1.0025   2.0281   1.9717   1.9499
  [total]                    58.9593 183.6597  61.6698  73.3234  73.6941 234.6606 156.1118 128.9211
c_macro_iap
  0                          54.9138 180.6591  59.6738  69.3672  72.6551 232.6682 154.1183 126.9445
  1                           4.0455   3.0006   1.9961   3.9562    1.039   1.9924   1.9936   1.9766
  [total]                    58.9593 183.6597  61.6698  73.3234  73.6941 234.6606 156.1118 128.9211
```

```
Time taken to build model (full training data): 1162.18 seconds
=== Model and evaluation on training set ===
Clustered Instances
0       51 (5%)
1      196 (21%)
2       56 (6%)
3       70 (7%)
4       64 (7%)
5      203 (21%)
6      180 (19%)
7      135 (14%)
Log likelihood: -30.24555
```

Anexo 35: “Resultado de la ejecución del “Conglomerado en dos fases” en el SPSS con los datos de las mujeres”.

Tamaños de conglomerados



Tamaño de conglomerado más pequeño	120 (17.9%)
Tamaño de conglomerado más grande	285 (42.6%)
Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño	2.38

Anexo 36: “Resultado de la ejecución del “EM” en el Weka con los datos de las mujeres (k=3)”.

```
=== Run information ===
Scheme:       weka.clusterers.EM -I 500 -N 3 -M 1.0E-6 -S 100
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    955
Attributes:   45
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
EM
==
Number of clusters: 3
                                    Cluster
Attribute                                0        1        2
                                    (0.16)   (0.53)   (0.31)
============================================================
historia_clinica
  mean                             -1.0025  -0.1893    0.846
  std. dev.                         0.4876     0.93   0.5667
edad
  mean                              0.0408   0.0458  -0.0987
  std. dev.                         1.0156   0.9524   1.0596
sexo_femenino
  0                                      1        1        1
  1                                157.089 502.7091 298.2019
  [total]                          158.089 503.7091 299.2019
talla
  mean                              0.1039  -0.0292  -0.0052
  std. dev.                         0.8955   1.0264   1.0016
p_peso_inicial
  mean                             -0.1221  -0.1788    0.366
  std. dev.                         0.9513   0.8527   1.1452
p_indice_masa_corporal
  mean                             -0.1268  -0.1561   0.3301
  std. dev.                         0.9056   0.9036   1.1146
ht_fumador
  0                                134.1259 406.5542 248.3199
  1                                 23.9631  97.1549   50.882
  [total]                          158.089 503.7091 299.2019
ht_cafe
  0                                 46.4644 172.3693 135.1663
  1                                111.6246 331.3397 164.0356
  [total]                          158.089 503.7091 299.2019
ht_bebidas_alcoholicas
  0                                153.1121 482.5972 291.2907
  1                                   4.977  21.1119   7.9112
  [total]                          158.089 503.7091 299.2019
ht_otros
  0                                157.089 502.7091 298.2019
  1                                      1        1        1
  [total]                          158.089 503.7091 299.2019
```

```
obesidad_debut
  0                                38.8205 145.3992  70.7803
  1                                119.2686 358.3099 228.4216
  [total]                          158.089 503.7091 299.2019
ao_menarca
  mean                              0.3639   0.0174  -0.2205
  std. dev.                         0.6348   0.9291   1.1927
ao_embarazos
  mean                              0.5812  -0.0923  -0.1494
  std. dev.                         1.4436   0.7873   0.9246
ao_abortos
  mean                              0.3884  -0.1409   0.0339
  std. dev.                         1.7398   0.6904   0.8376
ao_malformaciones
  0                                146.1479 494.6665 297.1856
  1                                 11.9411   9.0426   2.0163
  [total]                          158.089 503.7091 299.2019
ao_macrofetos
  0                                101.8813 416.7014 284.4173
  1                                 56.2077  87.0076  14.7846
  [total]                          158.089 503.7091 299.2019
ao_muerte
  0                                145.4542 486.3416 297.2042
  1                                 12.6349  17.3675   1.9977
  [total]                          158.089 503.7091 299.2019
ao_menopausia
  mean                              0.1109   0.0396  -0.1251
  std. dev.                         1.0681   0.9702   0.9985
apf_dm
  0                                 28.8553  92.6805  59.4641
  1                                129.2337 411.0285 239.7378
  [total]                          158.089 503.7091 299.2019
app_HTA
  0                                 56.6691 194.3405  95.9905
  1                                  101.42 309.3686 203.2114
  [total]                          158.089 503.7091 299.2019
app_hiperlipoproteinemia
  0                                128.8326 442.2579 280.9095
  1                                 29.2564  61.4512  18.2924
  [total]                          158.089 503.7091 299.2019
app_cardiopatia_isquemica
  0                                114.5694 416.2776  282.153
  1                                 43.5196  87.4315  17.0489
  [total]                          158.089 503.7091 299.2019
app_clauditacion_intermitente
  0                                153.9177 487.8818 298.2005
  1                                  4.1714  15.8272   1.0014
  [total]                          158.089 503.7091 299.2019
app_otros
  0                                 95.9274 427.4905 292.5821
  1                                 62.1616  76.2185   6.6198
  [total]                          158.089 503.7091 299.2019
apf_HTA
```

```
  0                                77.8947 368.5281 277.5773
  1                                80.1944  135.181  21.6246
  [total]                           158.089 503.7091 299.2019
apf_hiperlipoproteinemia
  0                                138.2425 481.7633 289.9941
  1                                19.8465  21.9457   9.2078
  [total]                           158.089 503.7091 299.2019
apf_cardiopatia_isquemica
  0                                92.9404 390.4889 283.5707
  1                                65.1486 113.2202  15.6312
  [total]                           158.089 503.7091 299.2019
apf_clauditacion_intermitente
  0                                146.3022 484.7162 296.9816
  1                                11.7869  18.9928   2.2203
  [total]                           158.089 503.7091 299.2019
apf_otros
  0                                146.3562 486.1791 296.4647
  1                                11.7328    17.53   2.7372
  [total]                           158.089 503.7091 299.2019
tgp
  mean                              -0.016  -0.0889   0.1585
  std. dev.                         0.9811    0.831   1.2248
glicemia_inicio
  mean                              0.3463   0.0399  -0.2493
  std. dev.                         1.2401   1.0004   0.7649
ppd_inicio
  mean                              0.2819   0.0257  -0.1914
  std. dev.                          1.165   0.9844   0.8842
creatinina
  mean                              0.1397  -0.0289  -0.0247
  std. dev.                         1.1316   0.7517   1.2535
microalbuminuria
  mean                              0.3752  -0.4934   0.6359
  std. dev.                         1.0893        1   1.3029
eritro
  mean                             -0.1329  -0.0863   0.2154
  std. dev.                         0.8638    0.929   1.1371
hemoglobina
  mean                             -0.2487  -0.0987   0.2973
  std. dev.                           0.863   0.8572   1.2004
trigliceridо
  mean                             -0.0805   0.0513  -0.0443
  std. dev.                         0.9178    1.102   0.8414
acido_urico
  mean                             -0.0408  -0.0972   0.1855
  std. dev.                         0.9414   0.9581   1.0694
colesterol
  mean                             -0.0647   0.0423  -0.0374
  std. dev.                         1.3408   1.1307   0.3176
c_micro_rd
  0                                147.6434 463.2405 276.1161
  1                                10.4457  40.4686  23.0858
  [total]                           158.089 503.7091 299.2019
```

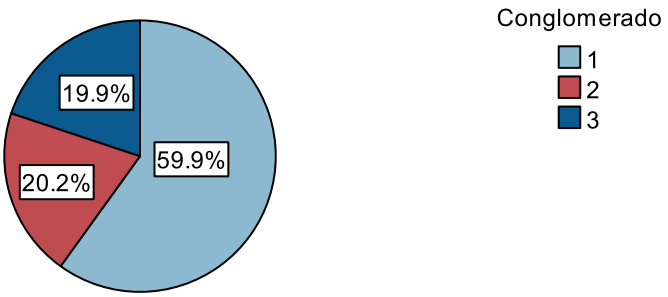
```
c_micro_nd
  0                                152.1913 487.8721 283.9366
  1                                  5.8978   15.837  15.2653
  [total]                           158.089 503.7091 299.2019
c_micro_neurod
  0                                152.7371 492.9012 289.3617
  1                                   5.352  10.8078   9.8402
  [total]                           158.089 503.7091 299.2019
c_macro_ci
  0                                 147.054 455.0583 267.8876
  1                                  11.035  48.6507  31.3143
  [total]                           158.089 503.7091 299.2019
c_macro_ave
  0                                157.0855 499.7212 298.1933
  1                                  1.0035   3.9879   1.0086
  [total]                           158.089 503.7091 299.2019
c_macro_iap
  0                                153.0962 498.4115 294.4922
  1                                  4.9928   5.2975   4.7097
  [total]                           158.089 503.7091 299.2019

Time taken to build model (full training data): 11.26 seconds
=== Model and evaluation on training set ===

Clustered Instances
0      155 (16%)
1      428 (45%)
2      372 (39%)
Log likelihood: -32.89967
```

Anexo 37: “Resultado de la ejecución del “Conglomerado en dos fases” en el SPSS con los datos de las mujeres (k=3)”.

Tamaños de conglomerados



Tamaño de conglomerado más pequeño	133 (19.9%)
Tamaño de conglomerado más grande	401 (59.9%)
Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño	3.02

Anexo 38: “Resultado de la ejecución del “K medias” en el Weka con los datos de las mujeres (k=2)”.

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 1000 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    955
Attributes:   45
Ignored:
              sexo_femenino
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 13
Within cluster sum of squared errors: 3033.648945031301
Missing values globally replaced with mean/mode
Cluster centroids:
                                            Cluster#
Attribute                     Full Data           0           1
                                (955)       (478)       (477)
=============================================================
historia_clinica                   0     -0.8619      0.8637
edad                               0       0.058     -0.0581
talla                              0      0.0529      -0.053
p_peso_inicial                     0     -0.2139      0.2143
p_indice_masa_corporal             0     -0.2152      0.2156
ht_fumador                         0           0           0
ht_cafe                            1           1           1
ht_bebidas_alcoholicas             0           0           0
ht_otros                           0           0           0
obesidad_debut                     1           1           1
ao_menarca                         0      0.2429     -0.2434
ao_embarazos                       0      0.1628     -0.1632
ao_abortos                         0      0.0091     -0.0091
ao_malformaciones                  0           0           0
ao_macrofetos                      0           0           0
ao_muerte                          0           0           0
ao_menopausia                      0      0.1056     -0.1058
apf_dm                             1           1           1
app_HTA                            1           1           1
app_hiperlipoproteinemia           0           0           0
app_cardiopatia_isquemica          0           0           0
app_clauditacion_intermitente      0           0           0
app_otros                          0           0           0
apf_HTA                            0           0           0
apf_hiperlipoproteinemia           0           0           0
apf_cardiopatia_isquemica          0           0           0
apf_clauditacion_intermitente      0           0           0
apf_otros                          0           0           0
tgp                                0     -0.0323      0.0324
```

```
glicemia_inicio                          0      0.2009    -0.2013
ppd_inicio                               0      0.1875    -0.1879
creatinina                               0      0.0722    -0.0723
microalbuminuria                         0     -0.1853     0.1857
eritro                                   0     -0.1299     0.1301
hemoglobina                              0     -0.2094     0.2098
trigliceridо                             0     -0.0703     0.0704
acido_urico                              0      -0.143     0.1433
colesterol                               0     -0.0292     0.0293
c_micro_rd                               0           0          0
c_micro_nd                               0           0          0
c_micro_neurod                           0           0          0
c_macro_ci                               0           0          0
c_macro_ave                              0           0          0
c_macro_iap                              0           0          0
Time taken to build model (full training data): 0.98 seconds

=== Model and evaluation on training set ===
Clustered Instances
0      478 (50%)
1      477 (50%)
```

Anexo 39: “Resultado de la ejecución del “K medias” en el Weka con los datos de las mujeres (k=3)”.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 1000 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    955
Attributes:   45
Ignored:      sexo_femenino
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 12
Within cluster sum of squared errors: 2789.534030758811
Missing values globally replaced with mean/mode
Cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1           2
                              (955)       (278)       (425)       (252)
=============================================================================
historia_clinica                 0      0.1342      -0.059     -0.0485
edad                             0     -0.8974      0.4767      0.1859
talla                            0      0.3208      -0.198     -0.0199
p_peso_inicial                   0       0.623     -0.0285     -0.6392
p_indice_masa_corporal           0       0.514      0.0875     -0.7147
ht_fumador                       0           0           0           0
ht_cafe                          1           1           1           1
ht_bebidas_alcoholicas           0           0           0           0
ht_otros                         0           0           0           0
obesidad_debut                   1           1           1           0
ao_menarca                       0      -0.515      0.2453      0.1544
ao_embarazos                     0     -0.3179      0.2376     -0.0501
ao_abortos                       0      0.0059      0.0019     -0.0098
ao_malformaciones                0           0           0           0
ao_macrofetos                    0           0           0           0
ao_muerte                        0           0           0           0
ao_menopausia                    0     -1.1394      0.6911      0.0913
apf_dm                           1           1           1           1
app_HTA                          1           1           1           1
app_hiperlipoproteinemia         0           0           0           0
app_cardiopatia_isquemica        0           0           0           0
app_clauditacion_intermitente    0           0           0           0
app_otros                        0           0           0           0
apf_HTA                          0           0           0           0
apf_hiperlipoproteinemia         0           0           0           0
apf_cardiopatia_isquemica        0           0           0           0
apf_clauditacion_intermitente    0           0           0           0
apf_otros                        0           0           0           0
tgp                              0      0.1772     -0.0141     -0.1717
glicemia_inicio                  0     -0.0113      0.0131     -0.0097
ppd_inicio                       0     -0.0576      0.0392     -0.0025
creatinina                       0     -0.1229       0.027      0.0901
microalbuminuria                 0      0.2221     -0.0984     -0.0791
eritro                           0     -0.0998       0.195     -0.2188
hemoglobina                      0      0.2073     -0.0489     -0.1462
triglicerido                     0     -0.0515      0.0417     -0.0136
acido_urico                      0     -0.0201      0.1226     -0.1845
```

colesterol	0	-0.0742	-0.0145	0.1064
c_micro_rd	0	0	0	0
c_micro_nd	0	0	0	0
c_micro_neurod	0	0	0	0
c_macro_ci	0	0	0	0
c_macro_ave	0	0	0	0
c_macro_iap	0	0	0	0

```
Time taken to build model (full training data): 1.06 seconds
=== Model and evaluation on training set ===

Clustered Instances
0       278 (29%)
1       425 (45%)
2       252 (26%)
```

Anexo 40: “Resultado de la ejecución del “K medias” en el Weka con los datos de las mujeres (k=4)”.

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 1000 -S 10
Relation:     diabetes_mellitus-
weka.filters.unsupervised.attribute.Standardize
Instances:    955
Attributes:   45
Ignored:              sexo_femenino
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 8
Within cluster sum of squared errors: 2465.6469622583354
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                 Full Data           0          1          2          3
                             (955)       (273)      (348)      (216)      (118)
=============================================================================
historia_clinica                 0      0.4078    -0.0401    -0.0297    -0.7709
edad                             0     -0.4579     0.0002     0.3061     0.4986
talla                            0      0.0863     0.0044    -0.1129    -0.0061
p_peso_inicial                   0      0.3786     0.1197    -0.7132     0.0765
p_indice_masa_corporal           0      0.3769     0.1469    -0.7605     0.0868
ht_fumador                       0           0          0          0          0
ht_cafe                          1           0          1          1          1
ht_bebidas_alcoholicas           0           0          0          0          0
ht_otros                         0           0          0          0          0
obesidad_debut                   1           1          1          0          1
ao_menarca                       0     -0.0202    -0.1288     0.1709     0.1138
ao_embarazos                     0     -0.1753     0.0302    -0.0514     0.4107
ao_abortos                       0     -0.0407    -0.0081    -0.0212      0.157
ao_malformaciones                0           0          0          0          0
ao_macrofetos                    0           0          0          0          0
ao_muerte                        0           0          0          0          0
ao_menopausia                    0     -0.3668     0.0621     0.1947     0.3091
apf_dm                           1           1          1          1          1
app_HTA                          1           1          1          1          1
app_hiperlipoproteinemia         0           0          0          0          0
app_cardiopatia_isquemica        0           0          0          0          1
app_clauditacion_intermitente    0           0          0          0          0
app_otros                        0           0          0          0          0
apf_HTA                          0           0          0          0          0
apf_hiperlipoproteinemia         0           0          0          0          0
apf_cardiopatia_isquemica        0           0          0          0          0
apf_clauditacion_intermitente    0           0          0          0          0
apf_otros                        0           0          0          0          0
tgp                              0      0.0972      0.062    -0.2209    -0.0032
glicemia_inicio                  0     -0.0235    -0.0325    -0.0357     0.2157
ppd_inicio                       0       0.049    -0.0676    -0.0164     0.1161
creatinina                       0     -0.0847     -0.064     0.1127     0.1785
microalbuminuria                 0      0.1389    -0.0237     -0.064    -0.1344
eritro                           0      0.0877    -0.0123    -0.2238     0.2431
hemoglobina                      0      0.0549     0.1066    -0.1693    -0.1315
trigliceridо                     0      0.0327    -0.0159    -0.0111    -0.0084
acido_urico                      0      0.0915     -0.013     -0.213     0.2166
colesterol                       0      -0.043    -0.0638     0.1146     0.0778
c_micro_rd                       0           0          0          0          0
c_micro_nd                       0           0          0          0          0
c_micro_neurod                   0           0          0          0          0
c_macro_ci                       0           0          0          0          0
```

c_macro_ave	0	0	0	0	0
c_macro_iap	0	0	0	0	0

```
Time taken to build model (full training data): 0.76 seconds
=== Model and evaluation on training set ===

Clustered Instances
0       273 (29%)
1       348 (36%)
2       216 (23%)
3       118 (12%)
```

Anexo 41: “Resultado de la ejecución del “K medias” en el SPSS con los datos de las mujeres (k=4)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados			
	1	2	3	4
1	5.946	6.749	6.010	7.150
2	1.705	.047	.581	.949
3	1.150	.095	.000	1.392
4	1.321	.186	1.439	1.091
5	.673	.172	.674	.576
6	.391	.129	.000	.353
7	.297	.204	.646	.418
8	.156	.249	.779	.361
9	.110	.275	.413	.355
10	.210	.261	.000	.306
11	.072	.177	.000	.166
12	.099	.160	.000	.129
13	.109	.098	.000	.058
14	.093	.076	.000	.064
15	.044	.077	.000	.053
16	.000	.068	.000	.048
17	.039	.038	.000	.028
18	.000	.025	.000	.017
19	.000	.038	.000	.026
20	.000	.023	.000	.015
21	.000	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de .000. La iteración actual es 21. La distancia mínima entre los centros iniciales es de 14.223.

Número de casos en cada conglomerado

Conglomerado	1	82.000
	2	230.000
	3	11.000
	4	346.000
Válidos		669.000
Perdidos		286.000

Anexo 42: “Resultado de la ejecución del “K medias” en el SPSS con los datos de las mujeres (k=8)”.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados							
	1	2	3	4	5	6	7	8
1	5.630	5.545	6.335	5.993	6.112	4.772	5.944	6.150
2	.000	.721	.412	1.132	.978	2.024	.387	.715
3	.618	1.357	.196	.000	.413	.000	.185	.372
4	.000	.541	.113	.000	.147	.000	.092	.195
5	.000	.805	.075	.000	.058	.000	.071	.072
6	.000	.527	.078	.000	.104	.000	.079	.107
7	.000	.609	.086	.000	.053	.000	.079	.059
8	.000	.592	.127	.000	.073	.000	.094	.076
9	.000	.449	.080	.000	.064	.000	.073	.050
10	.000	.262	.076	.000	.089	.000	.088	.058
11	.000	.310	.149	.000	.077	.000	.111	.000
12	.000	.134	.212	.000	.031	.000	.109	.000
13	.000	.110	.196	.000	.091	.000	.079	.039
14	.000	.068	.143	.000	.074	.000	.090	.124
15	.000	.147	.181	.000	.049	.000	.083	.109
16	.000	.000	.190	.000	.056	.000	.088	.079
17	.000	.075	.153	.000	.053	.000	.052	.031
18	.000	.103	.088	.000	.064	.000	.000	.057
19	.000	.000	.034	.000	.032	.000	.000	.000
20	.000	.000	.035	.000	.033	.000	.000	.000
21	.000	.000	.029	.000	.000	.000	.012	.000
22	.000	.000	.000	.000	.000	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. La iteración actual es 22. La distancia mínima entre los centros iniciales es de 10.077.

Número de casos en cada conglomerado

Conglomerado	1	8.000
	2	46.000
	3	109.000
	4	5.000
	5	118.000
	6	4.000
	7	258.000
	8	121.000
Válidos		669.000
Perdidos		286.000

Anexo 43: “Resultado de la aplicación de los índices de validación a los datos de las mujeres”.

Herramienta	Algoritmos	Ball	CH	Dunn	RMSSTD	RS
Weka	**EM**	1.42E+07	231.2896	0.1422	51.5014	0.6309
	EM(3)	6.11E+03	25.9626	0.3292	0.6603	0.0517
	km2	4.37E+07	2414	3.5122	45.099	0.717
	km3	1.02E+08	4.2421	0.0654	84.3961	0.0088
	km4	6.81E+07	42.4407	0.0549	79.6105	0.1181
SPSS	**Conglomerado en dos fases**	1.98E+06	10.6827	0.2642	16.5859	0.046
	C2F (3)	2.70E+06	8.5243	0.3031	16.7673	0.025
	Km4	1.99E+06	9.2573	0.1836	16.637	0.0401
	Km8	8.44E+05	21.5386	0.1913	15.324	0.1857

Anexo 44: “Centroides de los grupos formados por los mejores algoritmos aplicados a los datos de las mujeres”.

Herramienta	**Weka**		**SPSS**							
Algoritmo	K medias k=2		K medias k=8							
Cluster #	1	2	1	2	3	4	5	6	7	8
Edad (años)	55	54	58	42	44	62	52	64	62	57
Talla (m)	1.57	1.56	1.61	1.58	1.58	1.61	1.57	1.53	1.54	1.56
Peso (kg)	69.35	76.33	76.63	76.47	69.2	74	95.65	66.88	63.57	66.67
IMC	28.59	31.17	29.34	30.84	28.04	29.4	38.96	28.53	26.88	27.75
Menarca (años)	11	10	12	11	12	12	11	12	12	12
Embarazos	3	2	10	2	2	2	3	4	3	4
Abortos	0	0	8	0	1	0	0	1	0	1
Menopausia	32	28	31	10	5	42	31	40	48	38
TGP (u)	24.92	26.19	17.74	65.89	16.72	31.8	26.89	28.25	23.15	22.67
Glicemia (mmol/L)	8.9	7.54	8.34	8.54	6.89	7.88	7.47	7.8	6.85	13.93
PPD (mmol/L)	9.01	7.67	8.73	8.89	7.16	8.78	7.77	8.7	7.08	14.15
Creatinina (mmol/L)	63.33	59.67	71.75	52.23	60.05	58.4	60.28	111.5	61.6	66.46
Microalbuminuria (g/L)	0.02	0.04	0.05	0.02	0.02	0.01	0.03	0.04	0.02	0.02
Eritro (mm/h)	28.63	33.71	28	28.98	23.57	27.4	36.63	48	32.34	34.22
Hemoglobina (g/L)	13.18	13.94	13.43	13.55	13.31	13.77	13.22	13.73	13.33	12.56
Triglicérido (mmol/L)	1.9	2.05	1.25	1.9	1.58	5.3	1.8	10.33	1.93	2.18
Ácido Úrico (mmol/L)	257.5	286.5	286.72	270.78	234.84	293.03	337.77	418.08	282	220.13
Colesterol (mmol/L)	5.32	5.57	4.96	5.19	4.64	26.86	4.93	11.53	5.4	5.22
Sexo	1	1	1	1	1	1	1	1	1	1
Fumador	0	0	0	0	0	0	0	1	0	0
Café	1	1	1	0	1	1	1	1	1	1
Bebidas alcohólicas	0	0	0	0	0	0	0	0	0	0

Otros hábitos tóxicos	0	0	0	0	0	0	0	0	0	0
Obesidad	1	1	1	1	1	1	1	1	1	1
Antecedentes obstétricos (malformaciones)	0	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (macrofetos)	0	0	0	0	0	0	0	0	0	0
Antecedentes obstétricos (muerte)	0	0	0	0	0	0	0	0	0	0
APF DM	1	1	1	1	1	1	1	1	1	1
APP HTA	1	1	1	0	0	0	1	1	1	1
APP hiperlipoproteinemia	0	0	0	0	0	0	0	0	0	0
APP cardiopatia isquémica	0	0	0	0	0	0	0	0	0	0
APP clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0
APP otros	0	0	0	0	0	0	0	0	0	0
APF HTA	0	0	1	0	0	0	0	0	0	0
APF hiperlipoproteinemia	0	0	0	0	0	0	0	0	0	0
APF cardiopatia	0	0	1	0	0	1	0	0	0	0
APF clauditacion_intermitente	0	0	0	0	0	0	0	0	0	0
APF otros	0	0	0	0	0	0	0	0	0	0
Complicación micro (RD)	0	0	0	0	0	0	0	0	0	0
Complicación micro	0	0	0	0	0	0	0	0	0	0

(ND)										
Complicación micro (NeuroD)	0	0	0	0	0	0	0	0	0	0
Complicación macro (CI)	0	0	0	0	0	0	0	0	0	0
Complicación macro (AVE)	0	0	0	0	0	0	0	0	0	0
Complicación macro (IAP)	0	0	0	0	0	0	0	0	0	0

Anexo 45: “Clases propuestas para el conjunto de hombres”.

Historia Clínica	Clasificación
428	medio
430	medio
431	medio
432	medio
438	medio
442	medio
443	medio
445	medio
446	medio
448	medio
449	medio
452	medio
457	alto
458	alto
461	medio
462	medio
468	alto
471	alto
472	medio
475	medio
486	medio
487	medio
488	medio
489	medio
492	medio
495	medio
497	alto
499	medio
503	medio
506	medio
507	medio
511	medio
512	medio
520	medio

Historia Clínica	Clasificación
525	medio
530	medio
533	medio
536	medio
538	medio
539	medio
540	medio
544	medio
547	medio
550	medio
552	medio
553	medio
556	medio
562	medio
564	medio
565	alto
567	medio
571	medio
572	medio
573	alto
577	medio
579	alto
581	medio
582	medio
583	medio
590	medio
593	medio
596	alto
597	alto
598	medio
601	medio
602	alto
609	medio
610	medio

Historia Clínica	Clasificación
611	medio
613	medio
614	medio
618	bajo
624	medio
630	medio
634	medio
636	medio
646	medio
649	medio
657	medio
663	medio
665	medio
666	medio
667	medio
676	medio
678	medio
681	medio
682	alto
684	medio
686	medio
688	medio
691	medio
694	medio
699	alto
700	medio
704	medio
705	medio
710	medio
712	medio
715	medio
718	alto
724	medio
733	medio

Historia Clínica	Clasificación
734	medio
739	medio
741	bajo
748	medio
753	medio
754	medio
756	medio
762	medio
766	medio
767	bajo
770	medio
773	medio
777	alto
778	medio
784	medio
787	medio
789	bajo
792	medio
795	medio
805	medio
811	medio
813	medio
815	alto
817	medio
818	alto
825	alto
827	medio
828	medio
829	medio
831	alto
833	medio
836	medio
839	bajo
840	alto
844	medio
856	medio

Historia Clínica	Clasificación
860	alto
868	medio
872	alto
874	medio
877	medio
878	alto
885	medio
886	medio
887	medio
888	alto
892	medio
896	bajo
899	medio
908	medio
909	medio
910	medio
913	alto
916	medio
917	medio
920	medio
922	medio
924	bajo
927	bajo
931	bajo
932	alto
935	medio
936	medio
941	medio
942	medio
946	bajo
949	bajo
950	medio
961	alto
962	medio
963	medio
972	bajo

Historia Clínica	Clasificación
974	alto
976	alto
978	medio
979	medio
984	medio
985	medio
986	medio
988	medio
989	bajo
990	medio
993	alto
995	bajo
996	medio
997	bajo
998	medio
999	medio
1007	medio
1010	medio
1013	medio
1014	medio
1016	medio
1018	medio
1019	medio
1022	medio
1023	bajo
1026	medio
1027	medio
1028	medio
1029	medio
1033	medio
1037	medio
1041	medio
1043	bajo
1044	medio
1047	medio
1049	medio

Historia Clínica	Clasificación
1060	medio
1061	bajo
1063	medio
1064	medio
1065	medio
1068	medio
1069	medio
1073	medio
1074	medio
1078	medio
1079	medio
1081	medio
1082	medio
1092	medio
1094	medio
1096	bajo
1098	medio
1100	medio
1101	bajo
1102	medio
1104	medio
1106	bajo
1107	medio
1108	medio
1109	medio
1111	bajo
1115	medio
1119	medio
1121	medio
1123	medio
1126	bajo
1128	medio
1131	medio
1137	medio
1138	medio
1139	bajo

Historia Clínica	Clasificación
1141	medio
1145	bajo
1146	medio
1148	alto
1149	medio
1151	bajo
1152	medio
1153	bajo
1156	medio
1161	alto
1168	medio
1169	medio
1172	medio
1176	medio
1177	bajo
1179	medio
1180	medio
1181	medio
1182	medio
1183	medio
1184	medio
1187	bajo
1188	medio
1189	medio
1193	medio
1196	medio
1199	medio
1202	medio
1203	medio
1204	bajo
1212	medio
1221	bajo
1222	medio
1224	medio
1226	medio
1227	bajo

Historia Clínica	Clasificación
1238	alto
1240	alto
1247	medio
1248	medio
1249	medio
1250	medio
1258	medio
1259	alto
1261	alto
1262	medio
1263	medio
1267	bajo
1272	alto
1274	alto
1277	medio
1278	bajo
1280	medio
1283	alto
1288	alto
1289	bajo
1293	alto
1300	medio
1301	medio
1302	medio
1304	bajo
1306	bajo
1308	alto
1309	medio
1311	bajo
1312	bajo
1314	medio
1315	medio
1319	medio
1321	bajo
1323	bajo
1328	medio

Historia Clínica	Clasificación
1330	medio
1333	bajo
1334	medio
1335	bajo
1337	medio
1338	bajo
1340	bajo
1346	bajo
1347	medio
1349	medio
1351	alto
1353	bajo
1354	medio
1358	bajo
1359	alto
1360	alto
1361	bajo
1363	medio
1365	bajo
1369	medio
1377	alto
1380	bajo
1382	bajo
1383	medio
1385	bajo
1387	bajo
1388	medio
1389	medio
1391	medio
1393	alto
1394	alto
1395	medio
1397	bajo
1402	alto
1407	alto
1408	medio

Historia Clínica	Clasificación
1409	bajo
1410	medio
1412	medio
1413	medio
1418	medio
1422	medio
1424	medio
1430	bajo
1433	medio
1434	alto
1438	medio
1440	medio
1443	medio
1444	medio
1456	bajo
1458	medio
1461	medio
1462	bajo
1465	alto
1466	bajo
1469	medio
1471	bajo
1473	alto
1475	bajo
1476	bajo
1490	bajo
1491	bajo
1495	alto
1496	medio
1500	bajo
1507	bajo
1509	bajo
1510	medio
1512	alto
1514	medio
1515	bajo

Historia Clínica	Clasificación
1516	bajo
1517	alto
1519	medio
1523	bajo
1524	medio
1525	alto
1528	medio
1529	bajo
1533	alto
1535	alto
1538	alto
1539	medio
1540	medio
1541	bajo
1542	medio
1543	bajo
1547	medio
1548	bajo
1553	alto
1559	bajo
1560	medio
1563	medio
1564	bajo
1565	medio
1573	bajo
1574	bajo
1576	alto
1578	alto
1581	alto
1582	alto
1587	bajo
1593	medio
1597	medio
1600	medio
1604	alto
1606	alto

Historia Clínica	Clasificación
1610	bajo
1615	bajo
1618	bajo
1620	bajo
1621	bajo
1627	bajo
1628	bajo
1629	bajo
1631	medio
1632	bajo
1633	medio
1635	bajo
1641	alto
1642	bajo
1643	bajo
1645	bajo
1646	bajo
1647	alto
1648	alto
1649	bajo
1651	alto
1652	bajo
1654	medio
1655	bajo
1656	medio
1657	bajo
1660	bajo
1662	medio
1664	bajo
1665	bajo
1666	alto
1667	medio
1672	alto
1673	medio
1674	bajo
1675	bajo

Historia Clínica	Clasificación
1678	bajo
1679	bajo
1685	bajo
1687	bajo
1689	bajo
1691	bajo
1692	medio
1695	bajo
1697	alto
1699	alto
1705	bajo
1708	bajo
1711	medio
1716	bajo
1718	bajo
1721	bajo
1722	bajo
1725	bajo
1728	bajo
1729	medio
1730	bajo
1731	medio
1734	bajo
1735	bajo
1736	bajo
1737	bajo
1747	medio
1748	bajo
1755	bajo
1767	bajo
1769	bajo
1770	bajo
1773	medio
1774	bajo
1776	bajo
1778	medio

Historia Clínica	Clasificación
1783	medio
1784	bajo
1786	bajo
1787	bajo
1790	bajo
1791	medio
1795	medio
1798	alto
1799	medio
1802	bajo
1804	bajo
1808	medio
1809	bajo
1810	bajo
1811	medio
1813	alto
1819	medio
1822	bajo
1826	bajo
1828	bajo
1829	bajo
1830	bajo
1832	medio
1834	alto
1837	medio
1838	bajo
1839	medio
1843	bajo
1847	bajo
1853	bajo
1861	bajo
1863	bajo
1864	medio
1866	bajo
1872	bajo
1881	bajo

Historia Clínica	Clasificación
1883	bajo
1886	bajo
1890	bajo
1891	bajo
1893	bajo
1895	bajo
1896	bajo
1897	bajo
1898	bajo
1899	bajo
1907	bajo
1909	medio
1910	bajo
1919	bajo
1920	bajo
1921	bajo
1922	medio
1923	alto
1924	bajo
1928	bajo
1931	bajo
1932	medio
1937	bajo
1938	bajo
1939	bajo
1940	bajo
1941	bajo
1943	medio
1944	bajo
1946	bajo
1949	bajo
1955	alto
1963	bajo
1964	bajo
1965	bajo
1974	bajo

Historia Clínica	Clasificación
1977	bajo
1979	bajo
1981	medio
1982	medio
1983	bajo
1985	bajo
1987	medio
1988	medio
1997	bajo
1998	bajo
1999	bajo
2000	bajo
2002	medio
2005	bajo
2007	bajo
2011	alto
2012	bajo
2015	bajo
2018	medio
2020	medio
2026	bajo
2027	bajo
2028	bajo
2029	bajo
2032	bajo
2034	bajo
2035	bajo
2040	medio
2044	medio
2045	bajo
2046	medio
2047	bajo
2049	bajo
2050	bajo
2057	bajo
2061	alto

Historia Clínica	Clasificación
2065	medio
2068	medio
2074	medio
2075	bajo
2077	bajo
2080	bajo
2081	bajo
2082	bajo
2083	bajo
2084	bajo
2086	bajo
2087	medio
2090	alto
2092	medio
2096	bajo
2097	medio
2099	medio
2102	bajo
2109	medio
2113	medio
2114	bajo
2115	bajo
2123	medio
2127	bajo
2133	bajo
2135	bajo
2139	alto
2140	medio
2144	bajo
2145	bajo
2146	bajo
2147	bajo
2148	bajo
2151	alto
2153	bajo
2156	medio

Historia Clínica	Clasificación
2160	bajo
2161	medio
2166	alto
2168	alto
2170	bajo
2174	bajo
2175	bajo
2176	medio
2177	bajo
2178	bajo
2183	bajo
2190	bajo
2191	alto
2195	bajo
2196	medio
2197	bajo
2200	bajo
2202	medio
2208	bajo
2213	medio
2216	bajo
2217	alto
2219	bajo
2220	medio
2221	medio
2224	bajo
2225	alto
2226	bajo
2227	medio
2232	bajo
2233	bajo
2235	bajo
2240	medio
2243	bajo
2244	bajo
2245	medio

Historia Clínica	Clasificación
2246	alto
2247	bajo
2249	bajo
2250	bajo
2253	bajo
2255	alto
2256	medio
2263	medio
2265	bajo
2266	bajo
2269	bajo
2270	bajo
2272	bajo
2273	bajo
2274	bajo
2275	bajo
2277	medio
2279	medio
2280	bajo
2281	bajo
2282	bajo
2285	bajo
2286	bajo
2291	bajo
2294	bajo
2295	bajo
2296	bajo
2297	bajo
2298	bajo
2299	medio
2300	bajo
2306	medio
2308	bajo

Anexo 46: “Centroides de los grupos formados por el “Conglomerado en dos fases” en el conjunto de mujeres”.

Herramienta	**SPSS**			
Algoritmo	**Conglomerado en dos fases**			
Cluster #	**1**	**2**	**3**	**4**
Edad (años)	56	47	54	63
Talla (m)	1.57	1.57	1.56	1.52
Peso (kg)	74.52	82.4	64.09	64.24
IMC	30.52	33.48	26.52	27.69
Menarca (años)	12	11	12	12
Embarazos	3	2	3	3
Abortos	1	1	0	0
Menopausia	35	19	30	46
TGP (u)	26.56	31.84	23.81	19.61
Glicemia (mmol/L)	8.84	7.95	8.5	7.7
PPD (mmol/L)	8.8	8.71	8.69	8.24
Creatinina (mmol/L)	63.14	57.81	61.21	62.95
Microalbuminuria (g/L)	0.02	0.04	0.02	0.02
Eritro (mm/h)	31.06	33.08	25.45	39.33
Hemoglobina (g/L)	13.21	13.05	13.33	13.12
Triglicérido (mmol/L)	1.99	2.02	1.79	2.04
Ácido Úrico (mmol/L)	276.21	289.23	236.77	291.6
Colesterol (mmol/L)	5.58	4.9	5.18	5.4

Sexo	1	1	1	1
Fumador	0	0	0	0
Café	1	0	1	1
Bebidas alcohólicas	0	0	0	0
Otros hábitos tóxicos	0	0	0	0
Obesidad	1	1	0	1
Antecedentes obstétricos (malformaciones)	0	0	0	0
Antecedentes obstétricos (macrofetos)	0	0	0	0
Antecedentes obstétricos (muerte)	0	0	0	0
APF DM	1	1	1	1
APP HTA	1	1	0	1
APP hiperlipoproteinemia	0	0	0	0
APP cardiopatia isquémica	0	0	0	0
APP clauditacion_intermitente	0	0	0	0
APP otros	0	0	0	0
APF HTA	0	0	0	0
APF hiperlipoproteinemia	0	0	0	0
APF cardiopatia	0	0	0	0
APF clauditacion_intermitente	0	0	0	0
APF otros	0	0	0	0
Complicación micro (RD)	0	0	0	0

Complicación micro (ND)	0	0	0	0
Complicación micro (NeuroD)	0	0	0	0
Complicación macro (CI)	0	0	0	0
Complicación macro (AVE)	0	0	0	0
Complicación macro (IAP)	0	0	0	0

Anexo 47: “Clases propuestas para el conjunto de mujeres”.

Historia Clínica	Clasificación
359	R2
425	R4
427	R2
429	R2
433	R4
434	R2
435	R2
436	R2
437	R2
439	R4
440	R2
441	R2
444	R2
447	R2
450	R1
451	R2
453	R2
454	R2
455	R4
456	R2
459	R2
460	R2
463	R2
465	R2
466	R4
467	R2
469	R2
470	R2
473	R2
474	R2
476	R2
477	R2
478	R2
479	R2

Historia Clínica	Clasificación
481	R4
482	R2
483	R2
484	R3
485	R2
490	R2
491	R2
493	R2
494	R2
496	R2
498	R2
500	R2
501	R2
502	R2
504	R2
505	R2
508	R2
509	R2
510	R2
513	R4
514	R2
515	R2
516	R2
517	R2
518	R2
519	R2
521	R2
522	R2
523	R3
524	R2
526	R2
527	R2
528	R2
529	R2

Historia Clínica	Clasificación
531	R2
532	R2
534	R2
535	R2
541	R2
542	R2
545	R2
546	R1
548	R2
549	R2
551	R4
554	R2
555	R2
557	R2
558	R4
559	R2
560	R2
561	R1
563	R2
566	R2
568	R1
569	R2
570	R4
574	R2
575	R3
576	R1
578	R4
580	R2
584	R4
585	R2
586	R2
587	R2
589	R2
591	R2

Historia Clínica	Clasificación
592	R4
594	R4
595	R2
599	R2
600	R3
604	R4
605	R2
606	R2
607	R4
612	R4
615	R2
616	R2
619	R4
620	R3
621	R2
622	R2
623	R1
625	R4
626	R2
627	R2
629	R4
631	R2
632	R2
633	R2
635	R2
637	R3
642	R2
643	R4
647	R2
648	R3
651	R2
652	R4
656	R2
658	R4
659	R4
660	R3

Historia Clínica	Clasificación
661	R2
662	R2
664	R2
668	R2
669	R4
670	R2
671	R2
672	R2
673	R4
674	R2
679	R2
680	R3
683	R2
689	R2
690	R1
693	R2
695	R4
696	R3
697	R2
698	R2
701	R2
706	R2
708	R2
709	R2
713	R4
714	R4
716	R4
719	R2
721	R4
722	R4
727	R2
728	R2
729	R4
730	R2
731	R4
732	R2

Historia Clínica	Clasificación
735	R2
740	R2
742	R2
743	R4
744	R2
746	R4
751	R2
760	R4
761	R4
763	R2
765	R2
768	R2
769	R2
771	R2
772	R2
774	R4
776	R1
780	R3
782	R2
786	R1
790	R2
793	R2
794	R2
798	R4
799	R4
802	R2
803	R4
804	R4
806	R2
809	R4
810	R1
816	R2
819	R4
820	R2
821	R2
822	R2

Historia Clínica	Clasificación
823	R1
826	R2
830	R1
835	R1
838	R4
842	R2
843	R2
846	R2
847	R2
850	R2
853	R3
854	R2
857	R1
858	R2
859	R4
862	R2
865	R2
867	R2
871	R2
873	R4
875	R2
876	R2
879	R2
881	R4
883	R2
884	R2
889	R2
891	R2
894	R2
898	R4
900	R2
901	R2
903	R2
904	R2
906	R2
907	R4

Historia Clínica	Clasificación
911	R2
912	R4
914	R3
915	R2
918	R2
919	R2
921	R2
923	R2
925	R3
926	R3
928	R4
929	R2
930	R2
933	R4
934	R4
937	R4
939	R2
944	R1
947	R2
948	R3
951	R2
952	R2
953	R4
954	R4
956	R2
957	R2
958	R2
960	R2
964	R4
966	R2
968	R2
969	R2
970	R2
971	R2
973	R4
975	R2

Historia Clínica	Clasificación
977	R3
980	R4
981	R4
982	R2
983	R1
992	R2
994	R2
1000	R2
1003	R2
1004	R1
1005	R2
1006	R1
1009	R2
1011	R1
1012	R2
1015	R1
1017	R2
1020	R4
1025	R1
1030	R3
1032	R2
1036	R3
1038	R2
1039	R4
1040	R4
1042	R2
1045	R2
1048	R1
1050	R1
1051	R4
1053	R3
1056	R2
1057	R2
1058	R2
1059	R2
1062	R2

Historia Clínica	Clasificación
1067	R2
1070	R4
1071	R1
1072	R1
1076	R3
1080	R4
1083	R3
1085	R4
1087	R4
1088	R2
1090	R4
1091	R2
1093	R2
1095	R2
1097	R2
1099	R4
1102	R2
1105	R4
1110	R2
1112	R2
1113	R4
1114	R1
1116	R2
1118	R4
1120	R2
1122	R4
1124	R4
1127	R2
1132	R2
1133	R1
1134	R4
1135	R4
1136	R4
1140	R3
1142	R3
1144	R2

Historia Clínica	Clasificación
1150	R2
1154	R1
1155	R1
1158	R2
1159	R2
1160	R2
1162	R2
1163	R2
1164	R2
1165	R4
1166	R2
1167	R2
1171	R2
1173	R4
1174	R2
1175	R3
1178	R4
1185	R1
1186	R4
1190	R4
1192	R1
1195	R2
1197	R2
1198	R2
1200	R2
1205	R2
1206	R2
1208	R4
1209	R2
1210	R2
1211	R4
1213	R4
1214	R2
1215	R4
1216	R2
1217	R2

Historia Clínica	Clasificación
1218	R4
1219	R2
1220	R4
1225	R2
1228	R4
1229	R3
1231	R4
1233	R1
1235	R2
1239	R4
1241	R2
1243	R4
1244	R2
1254	R4
1256	R2
1257	R2
1265	R1
1268	R3
1269	R3
1270	R4
1273	R4
1276	R4
1279	R2
1281	R4
1284	R2
1291	R4
1294	R2
1295	R4
1296	R4
1297	R2
1298	R4
1299	R3
1305	R2
1307	R2
1310	R3
1313	R1

Historia Clínica	Clasificación
1325	R4
1348	R3
1823	R4
1827	R2
1831	R4
1833	R3
1835	R2
1836	R4
1841	R4
1842	R3
1845	R4
1846	R2
1848	R3
1849	R2
1850	R4
1855	R3
1856	R1
1857	R1
1858	R1
1859	R1
1860	R3
1865	R3
1867	R3
1868	R2
1869	R1
1870	R1
1873	R3
1874	R3
1875	R3
1876	R3
1877	R4
1878	R4
1879	R1
1880	R4
1884	R2
1885	R1

Historia Clínica	Clasificación
1887	R2
1888	R1
1889	R3
1892	R3
1894	R3
1900	R3
1901	R1
1902	R1
1903	R4
1904	R3
1905	R3
1908	R3
1911	R4
1912	R3
1913	R3
1914	R4
1915	R2
1916	R3
1917	R1
1918	R2
1926	R1
1927	R1
1929	R4
1930	R1
1933	R4
1945	R2
1947	R3
1948	R2
1950	R2
1951	R1
1952	R3
1953	R3
1954	R4
1956	R1
1957	R4
1958	R3

Historia Clínica	Clasificación
1959	R3
1960	R4
1961	R3
1966	R3
1970	R4
1971	R3
1972	R3
1973	R3
1975	R1
1976	R3
1980	R1
1984	R1
1990	R3
1991	R3
1992	R3
1993	R1
1994	R3
1995	R3
1996	R1
2001	R1
2003	R3
2004	R3
2006	R2
2008	R1
2009	R3
2010	R1
2013	R2
2014	R4
2016	R1
2017	R2
2021	R3
2022	R1
2023	R4
2024	R4
2030	R3
2031	R1

Historia Clínica	Clasificación
2033	R3
2036	R1
2038	R1
2039	R1
2041	R1
2042	R1
2043	R4
2048	R3
2048	R3
2051	R3
2052	R1
2053	R3
2054	R3
2056	R3
2058	R4
2069	R3
2071	R1
2072	R1
2073	R3
2076	R1
2078	R3
2079	R1
2085	R1
2088	R1
2089	R2
2091	R1
2093	R1
2094	R1
2095	R4
2100	R3
2101	R3
2103	R3
2104	R3
2105	R3
2106	R1
2108	R1

Historia Clínica	Clasificación
2110	R3
2111	R3
2112	R1
2118	R3
2119	R1
2120	R2
2121	R1
2122	R3
2124	R3
2125	R1
2126	R3
2128	R1
2129	R1
2130	R4
2131	R1
2132	R4
2134	R1
2138	R3
2141	R3
2142	R1
2143	R1
2149	R1
2150	R1
2152	R1
2154	R1
2155	R1
2157	R1
2158	R2
2159	R1
2162	R3
2163	R1
2164	R1
2165	R1
2167	R1
2169	R3
2171	R4

Historia Clínica	Clasificación
2172	R1
2173	R1
2179	R1
2182	R3
2184	R3
2185	R1
2186	R1
2187	R3
2188	R3
2189	R3
2192	R1
2193	R1
2194	R2
2199	R3
2201	R1
2203	R1
2204	R3
2205	R3
2206	R1
2207	R3
2209	R3
2210	R1
2211	R1
2214	R1
2215	R1
2218	R1
2222	R1
2223	R3
2228	R1
2231	R3
2234	R2
2236	R3
2237	R4
2238	R3
2242	R3
2248	R4

Historia Clínica	Clasificación
2251	R4
2252	R3
2254	R3
2258	R4
2259	R4
2260	R1
2261	R2
2262	R3
2264	R2
2267	R4
2268	R3
2271	R2
2276	R3
2278	R1
2283	R3
2284	R2
2287	R3
2288	R3
2289	R4
2292	R2
2293	R1
2301	R3
2302	R2
2303	R3
2305	R2
2307	R4
2309	R3