PROYECTO DE GRADO

Presentado ante la ilustre UNIVERSIDAD DE LOS ANDES como requisito final para obtener el Título de INGENIERO DE SISTEMAS

PREDICCIÓN DE RIESGOS DE FRACTURA DE CADERA EN MUJERES CON OSTEOPOROSIS USANDO MINERÍA DE DATOS

Por

Br. Gabriel Castañeda

Tutor: Prof. Gerardo Colmenares

Abril 2009

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

©2009 Universidad de Los Andes Mérida, Venezuela

Predicción de Riesgos de Fractura de Cadera en Mujeres con Osteoporosis Usando Minería de Datos

Br. Gabriel Castañeda

Proyecto de Grado — Investigación de Operaciones

Resumen: La minería de datos permite identificar reglas y patrones en los datos que se pueden definir en un modelo que permite ver las razones por las cuales suceden las cosas y predecir el futuro. El análisis de correspondencia es una técnica estadística que construye un diagrama cartesiano basado en la asociación entre las variables analizadas. En dicho gráfico se representan conjuntamente las distintas modalidades de las variables, de forma que la proximidad entre los puntos representados está relacionada con el nivel de asociación entre dichas modalidades de las variables. En este estudio se utilizó ésta técnica para clasificar mujeres osteoporóticas mayores de 50 años en un estudio retrospectivo de 10 años teniendo como variables la densidad mineral ósea, longitud y ángulo del cuello del fémur, edad, y como clase control la presencia de fractura de cadera durante el tiempo de estudio. El objetivo de este estudio es explorar a través de técnicas de Minería de Datos los riesgos que tiene una mujer mayor de 50 años de sufrir fractura de cadera al caer de su propia altura. Los resultados muestran un tendencia de tener fractura de cadera en mujeres mayores de 78 años, fumadoras, que no toman sol y no se suplementan con calcio.

Palabras clave: Minería de datos, análisis de correspondencia, osteoporosis, fractura de cadera.

A mis padres, por su gran apoyo e interés en mi crecimiento personal.

Gracias.

Índice

Índice de Tablas 6
Índice de Figuras……..…….……...7
Agradecimientos.....8
Capítulo 1 Introducción.....9
1.1 Antecedentes……...……10
1.2 Planteamiento del Problema……13
1.3 Justificación…….13
1.4 Objetivos……..14
1.4.1 Objetivo General…….14
1.4.2 Objetivos Específicos……..15
1.5 Metodología……15
1.5.1 Metodología de Trabajo…….15
1.5.2 Análisis de Correspondencias Múltiple……..17
Capítulo 2 Marco Teórico.....18
2.1 Conceptos Básicos.....19
2.2 Minería de Datos.....19
2.2.1 Minería de Datos y el Proceso de Extracción de Conocimiento en Bases de Datos (KDD)……19
2.3 Análisis de Correspondencias……..22
2.3.1 Análisis de Correspondencias Múltiple……...23

Capítulo 3 Minería de Datos para Determinar Variables Significativas….. ....**¡Error! Marcador no definido.**4

3.1 Naturaleza de los Datos..... ……**¡Error! Marcador no definido.**4

3.2 Limpieza de los datos……25

3.3 Generación de la Muestra……..25

3.4 Medición y Creación de Nuevas Variables……..…25

3.5 Categorización de las Variables……....26

Capítulo 4 Análisis de Resultados e Interpretación……..29

4.1 Aplicación de Estadísticas Fundamentales…….……29

4.2 Análisis de Correspondencias...……...…….35

Bibliografía……..44

Índice de Tablas

Tabla 3.1 Categorización de la variable Edad……....27

Tabla 3.2 Categorización de la variable Estatura……...27

Tabla 3.3 Categorización de la variable Peso…….27

Tabla 3.4 Categorización de la variable Longitud…….28

Tabla 3.5 Categorización de la variable Ángulo…….28

Tabla 3.5 Categorización de la variable Tscore……..28

Tabla 4.1 Frecuencia de Variable Edad……...36

Tabla 4.2 Frecuencia de Variable Estatura……..36

Tabla 4.3 Frecuencia de Variable Peso……36

Tabla 4.4 Frecuencia de Variable Longitud……36

Tabla 4.5 Frecuencia de Variable Ángulo……...36

Tabla 4.6 Frecuencia de Variable Tscore…….36

Tabla 4.8 Re codificación de Variable Edad……...36

Tabla 4.9 Re codificación de Variable Estatura……..36

Tabla 4.10 Re codificación de Variable Peso……..…….36

Tabla 4.11 Re codificación de Variable Longitud……..36

Tabla 4.12 Re codificación de Variable Ángulo…….……36

Tabla 4.13 Re codificación de Variable Ángulo……..…36

Tabla 4.14 Re codificación de las variables dicotómicas…….37

Índice de Figuras

Figura 2.1 Fases del KDD…….....21
Gráfico 4.1 Pacientes con Osteoporosis Definida con Longitud y Ángulo……...34
Gráfico 4.2 Pacientes con Osteoporosis Definida con Longitud y Ángulo……...35
Gráfico 4.3 Análisis de Correspondencias Múltiple (Factor1 y Factor2)……38
Gráfico 4.4 Análisis de Correspondencias Múltiple (Factor1 y Factor3)……39
Gráfico 4.5 Análisis de Correspondencias Múltiple (Factor2 y Factor3)……40
Gráfico 4.6 Análisis de Correspondencias Múltiples Variable Fractura de Cadera.40

Agradecimientos

Estoy consciente de que este proyecto no se habría podido realizar sin la ayuda y el soporte de muchas personalidades, expreso mis agradecimientos:

A Dios

A mis padres.

A la Escuela de Sistemas de la Universidad de los Andes

Al profesor Gerardo Colmenares por su disposición e interés en el desarrollo del proyecto.

A la profesora Janeth Mora por su infinita disposición y asesoría durante la realización de éste trabajo.

A mis compañeros de clase y amigos: José Urbaez, Carlos Medina, Carlos Rodríguez, Marco Ávila, Nelson Grimaldo, Mario Rincón, Dave Rojas, Ivan, Beatriz, Mariale, Jesús, Marco Gatti, Yenny, Carlos Raúl, Jean Pierre, Alfonso entre otros por su ayuda y amistad a lo largo de la carrera.

A los doctores José Rafael Salinas y Edgar Nieto.

Al resto de mis amigos, a todos gracias.

Capítulo 1

Introducción

Una fractura ósea ocurre cuando se aplica una presión superior a la que pueda soportar un hueso del cuerpo humano en donde éste se parte o se rompe. Cualquier ruptura de hueso se denomina fractura y si el hueso llegase a romper la piel se refiere a una fractura expuesta o fractura compuesta. (Berkow et al, 2005)

Para las mujeres entre 65 y 84 años, el 85% de las fracturas de cadera se le atribuyen a la osteoporosis que es un problema de salud pública, la mayoría de las fracturas por osteoporosis ocurren en las mujeres posmenopáusicas y más de la mitad de ellas tendrá una fractura. La mayoría de estas fracturas se precipitan por una caída. En España el número de éstas a lo largo de un año se sitúan por encima de 60 000 casos (Arellana et al, 2007), en Estados Unidos el 17,5% de las mujeres blancas tendrá una fractura de cadera durante su vida (Wasnich, 1999) en Venezuela el problema es similar con una cifra de 9.6 fracturas por día (16.3/100000 hab), con una mortalidad de 17% en los primeros 4 meses, según un estudio prospectivo realizado en UNILIME (Centro de Investigaciones de Osteoporosis y Metabolismo

de la Universidad de Carabobo) en 1995. La proyección al año 2030 será de 67 fracturas de cadera por día, además la morbilidad y mortalidad de estas lesiones suponen una importante carga socioeconómica ante la que muchos países todavía no están demostrando la necesaria receptividad.

La osteoporosis es una enfermedad sistémica esquelética caracterizada por una baja Masa Ósea (MO) y deterioro de la micro arquitectura del tejido óseo, originando fragilidad ósea y susceptibilidad a la fractura. Para medir la osteoporosis se usa la densitometría ósea que hace referencia a la medición de la MO con cualquier tecnología empleada. A la fecha, la Densidad Mineral Ósea (DMO) es el mejor factor de riesgo independiente para predecir fracturas. Las mujeres con DMO más baja tienen una Mayor predisposición a fracturas en la medida que su edad avanza. (Conferencia de Consenso desarrollada en 1993).

Este estudio propone crear una herramienta que prediga la probabilidad que tiene un nuevo paciente a sufrir nueva fractura de cadera con ayuda de técnicas y métodos que proporciona la minería de datos.

1.1 Antecedentes

Son muchos los casos de fractura ósea que han sido documentados, incluyendo aquellos de baja densidad ósea e historiales de fracturas. Sin embargo, se sabe muy poco acerca de fracturas a corto plazo y el tiempo que lleva fracturarse una mujer con diagnóstico de osteoporosis. Por lo tanto, se realizó un estudio prospectivo de 5 años para evaluar el riesgo y tiempo de incidencia de fracturas clínicas en 759 mujeres posmenopáusicas, con baja densidad mineral ósea. El estudio se realizó en diez poblaciones para seleccionar las mujeres posmenopáusicas que habían

padecido fracturas y los factores de riesgo para la incidencia de fracturas se determinaron mediante un análisis univariante, en donde los factores que resultaron significativos fueron incluidos en un análisis de regresión multivariante de Cox. Posteriormente, los factores de riesgo significativos permitieron construir algoritmos de aplicación práctica. Un 12.5% (con un intervalo de confianza de 10.1 – 14.9 y un grado de confianza de 95%) de las mujeres presentaron una nueva fractura clínica. Aquellas mujeres que hayan presentado fractura después de la menopausia y con una baja densidad mineral ósea BMD (T-score <-1.0) fue seleccionada como un factor significativo.

Mujeres con fracturas recientes (durante los 5 años del estudio) tienen un riesgo absoluto de 50.1%, mientras que aquellas que se fracturaron previo a dicho período tienen un riesgo absoluto de 21.2%. En las mujeres sin historia de fractura el riesgo absoluto era de 13.8% si la densidad mineral ósea era baja y 7% si la densidad mineral ósea era normal. El estudio concluyó que una de cada dos mujeres con fractura en el período de estudio presentará una nueva fractura clínica independiente de la densidad mineral ósea mientras que el riesgo absoluto para la primera fractura fue mucho menos y dependiente de la densidad mineral ósea. Geel et al, (2006)

Un estudio de casos y controles realizado en Mar de Plata, Argentina, para conocer la incidencia de fractura de fémur proximal por osteoporosis y los factores de riesgo asociados, se registraron todos los casos de fractura de fémur proximal por osteoporosis en personas mayores de 50 años de edad que acudieron a los centros de salud pública y privados. Se registraron un total de 246 casos y controles, 259 en mujeres y de 92 varones que fueron los controles; la mayor incidencia fue siempre más alta a mayor edad y sobre todo a partir de los 75 años. Los factores asociados con aumento del riesgo a fractura del fémur proximal con significancia estadística

fueron: antecedentes de enfermedades neurológicas, consumo de psicofármacos, consumo de alcohol, fracturas previas, enfermedades cardiovasculares y menor consumo de lácteos. No se observaron diferencias entre los casos y los controles con respecto a la edad de inicio de la menopausia, peso, talla, actividad previa, hábito de fumar o exposición al sol, como tampoco en el porcentaje de mujeres que habían tenido ooforectomías. Mosquera et al. (1998)

Se realizó una investigación para conocer si la etnia Mapuche es un factor de exposición para la variable de respuesta fractura de fémur proximal en el adulto mayor, considerando el efecto de otras variables. Se realizó una investigación de caso control. Los casos fueron individuos mayores de 50 años internados en el servicio de traumatología y urgencia del hospital regional de Temuco (Chile) por fractura de fémur proximal excluyendo aquellos cuyas fracturas fue producto de traumatismos mayores y atribuidos a compromiso neoplásico del hueso. Los controles sin fractura se obtuvieron de forma aleatoria de otros servicios del hospital excluyendo pediatría, neurología y psiquiatría pariados por edad a los casos. La variable de respuesta fue la fractura del fémur proximal y la variable de exposición fue la etnia (Mapuche o no Mapuche). Se realizó un análisis bivariado y estratificado. Utilizando el programa Stata 6.0 se modeló la asociación entre las variables de exposición y respuesta y se estableció que la condición etnica (Mapuche) es un factor protector contra la fractura de fémur proximal. También se le atribuye a la etnia otros factores: el menor riesgo de caída, mayor masa muscular y la geometría del cuello femoral . Sapunar et al. (2003)

No se dispone de suficientes datos sobre la arquitectura del hueso en poblaciones de diferentes razas y grupos étnicos. Un estudio examinó la geometría del fémur proximal en 1190 negros, hispanos, y blancos. Un análisis cross-seccional

indicó mayor fuerza ósea entre los hombres negros, que los hispanos con respecto a la edad. Mediante absorciometría dual y análisis estructural se obtuvieron las mediciones del cuello del fémur e ínter trocantes del fémur proximal. Los resultados mostraron mayor densidad mineral ósea, mayor área cross-sectional y mayor sección del módulo en la población de hombres negros comparada con hombres hispanos y la edad lo que puede indicar un elevado riesgo a fractura en los hombres hispanos de mayor edad en Estados Unidos. Travison et al (2007)

1.2 Planteamiento del Problema

La mayoría de las investigaciones relacionadas con fracturas de cadera en mujeres osteoporoticas y menopausicas mayores de 50 años han señalado como factores de riesgo: baja densidad mineral ósea, antecedentes de enfermedades neurológicas, consumo de psicofármacos, consumo de alcohol, fracturas previas, enfermedades cardiovasculares y menor consumo de lácteos. Hasta ahora no se han relacionado las medidas de longitud y área de la masa ósea del fémur, por lo tanto en este estudio se pretende determinar si estas medidas del fémur son significativas para predecir riesgo a fractura de cadera.

Los factores asociados con aumento del riesgo a fractura del fémur proximal con significancia estadística fueron: antecedentes de enfermedades neurológicas, consumo de psicofármacos, consumo de alcohol, fracturas previas, enfermedades cardiovasculares y menor consumo de lácteos. No se observaron diferencias entre los casos y los controles con respecto a la edad de inicio de la menopausia, peso, talla, actividad previa, hábito de fumar o exposición al sol, como tampoco en el porcentaje de mujeres que habían tenido ooforectomías (extirpación de los ovarios). (Mosquera et al. 1998)

1.3 Justificación

Los pacientes con osteoporosis tienen un alto grado de posibilidad a padecer de una fractura ósea. A pesar de que los factores de riesgo son diagnosticados muchos médicos no asesoran correctamente a sus pacientes de la enfermedad y de los riesgos que tienen de sufrir una nueva fractura en los próximos años. La medición de la densidad mineral ósea no es suficiente para la determinación del riesgo de fractura; algunos pacientes con baja densidad mineral ósea no padecen fracturas. Es por ello que se hace necesario determinar mediciones de longitud y áreas del hueso del fémur para facilitar la determinación del riesgo de fractura de cadera que pudieran hacerse de rutina en la densitometría ósea del fémur. La alteración de estas medidas de longitud indica la necesidad de instaurar un tratamiento preventivo.

1.4 Objetivos

1.4.1 Objetivo General

Realizar análisis exploratorio con la ayuda de técnicas basadas en minería de datos en mujeres mayores de 50 años con osteoporosis propensas a sufrir fractura de cadera al caer de su propia altura.

1.4.2 Objetivos Específicos

- Observar las características de un paciente que lo hacen propenso a sufrir fractura de cadera.

- Determinar si existe alguna relación significativa con la longitud del cuello femoral y su ángulo para recibir una fractura de cadera.
- Obtener base de datos con registro histórico de pacientes del Instituto de Osteoporosis y Densitometría Ósea de Mérida.
- Conseguir permisología de usuario para el software Norland X32 que permitirá manipular la base de datos antes descrita y tomar medidas correspondientes.
- Realizar limpieza de datos sobre la base de datos original a fin de tener exclusivamente pacientes de sexo femenino y con edades mayores a los 50 años.
- Tomar muestra de pacientes a través de muestreo estratificado.
- Emplear el software Norland X32 para tomar mediciones del cuello y ángulo del fémur de cada uno de los pacientes de la muestra.
- Conocer si los pacientes en cuestión han sufrido fractura de cadera en los últimos 5 años a través de una entrevista telefónica.

1.5 Metodología

1.5.1 Metodología de Trabajo

Tipo de Estudio:

El diseño de esta investigación es la siguiente: se trata de un estudio descriptivo, longitudinal, observacional, retrospectivo.

Es un estudio descriptivo porque se tomarán los datos y con finalidad puramente descriptiva. La secuencia de tiempo en este estudio es longitudinal ya que surgen nuevas variables en el transcurso del tiempo. Es un estudio

observacional ya que los factores de estudio no son controlados por los investigadores, sino que estos se limitan a observar, medir, y analizar las variables en los sujetos. En relación con la cronología de los hechos se considera retrospectivo porque el diseño del estudio es posterior a los hechos estudiados, los datos se obtuvieron de la base de datos del Centro de Metabolismo Óseo Mérida- Venezuela. (Argimon –Pallas et al 2005)

Población de Estudio:

El conjunto de mujeres mayores de 50 años de edad diagnosticadas con osteoporosis mediante la técnica de densitometría ósea, que acudieron a consulta del Centro de Metabolismo Óseo ubicado en la urbanización Alto Chama de la ciudad de Mérida durante el lapso 1996 – 2006.

Formación de Grupos de Estudio:

El proceso de formación de los grupos depende del tipo de estudio. En los diseños observacionales, se realiza en función de la existencia o no de la enfermedad de interés (estudios de casos y controles) o de la presencia o no de la exposición (estudios de cohortes). En los estudios experimentales la situación es muy distinta, ya que los sujetos se asignan a los diferentes grupos que se desea comparar por un procedimiento aleatorio.

Dado que en los estudios observacionales el investigador no decide a qué grupos serán asignados los sujetos, pueden aparecer limitaciones importantes en la comparabilidad inicial.

El grupo de casos y controles deben ser representativos de una misma población, formada conceptualmente por los miembros de una cohorte subyacente, definida por los criterios de inclusión y exclusión. Un tipo de estudio paradigmático en este sentido es el de casos y controles anidados en una cohorte, donde se

identifica claramente la cohorte subyacente, de la que proceden tanto los casos como los controles.

Grupo Control:

Conformado por las mujeres mayores de 50 años diagnosticadas con osteoporosis y que no presentaron fractura de cadera en el tiempo de estudio.

Grupo Casos:

Conformado por las mujeres mayores de 50 años diagnosticadas con osteoporosis y que presentaron fractura de cadera en el tiempo de estudio.

Medición de las variables:

- Mediciones del área del cuello del fémur.
- Mediciones del ángulo del cuello del fémur.
- Medición de la variable Tscore que indica el nivel de masa ósea.
- Presencia de fractura de cadera durante los 10 años del estudio prospectivo
- Presencia de fractura de cadera: (se excluirán las mujeres que se hayan fracturado por accidentes automovilísticos)

Las variables se midieron haciendo uso del programa Norland X 32. La información de presencia de fractura de cadera en un paciente se obtuvo por medio de llamada telefónica.

1.5.2 Análisis de Correspondencia Múltiple

El Análisis de Correspondencias es una técnica estadística que construye un diagrama cartesiano basado en la asociación entre las variables analizadas. En dicho gráfico se representan conjuntamente las distintas modalidades de las variables, de forma que la proximidad entre los puntos representados está relacionada con el nivel de asociación entre dichas modalidades de las variables. Figueras, M (2003)

La idea del análisis de correspondencia múltiple en este estudio fue descubrir las relaciones de atracción-repulsión entre las variables seleccionadas.

Capitulo 2

Marco Teórico

En esta sección se hablará de los conceptos fundamentales tanto a nivel clínico como a nivel de minería de datos necesarios para la comprensión y desarrollo de este proyecto.

2.1 Conceptos Básicos

Los conceptos que a continuación se describen fueron tomados de Yupanqui H (2003):

Osteoporosis: Es una enfermedad caracterizada por baja densidad mineral ósea y deterioro de la micro arquitectura del Tejido Óseo causando fragilidad ósea y propensidad a tener una fractura.

Densitometría Ósea: Se refiere a la medición de la masa ósea con cualquier tecnología empleada. La densidad mineral ósea es el mejor factor de riesgo para predecir fracturas. Mientras menor sea el valor de la DMO en mujeres mayor es la predisposición a tener fracturas a medida que la edad avanza.

Cadera y Fémur Proximal: Posee forma irregular y está formado por tres huesos: íleon, isquion y pubis.

El Fémur: Es el hueso más largo y pesado. Se extiende desde la articulación de la cadera hasta la articulación de la rodilla, sus Condillos se articulan con la tibia.
Fémur Proximal: Es región de interés en densitometría, une la cabeza femoral (epífisis) al cuerpo (diáfisis) femoral a través del cuello.
Norland X32: Tecnología utilizada en la medición de masa ósea.

2.2 Minería de Datos:

Orallo (2004) define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, que se busca encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo los patrones descubiertos deberían ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Por lo tanto dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, etc), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil de manera comprensible, por ello se hace uso de representaciones gráficas, convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos.

2.2.1 Minería de Datos y el Proceso de Extracción de Conocimiento en Bases de Datos (KDD)

El proceso de Extracción de Conocimiento en Bases de Datos o KDD (Knowledge Discovery in Data Bases) se utiliza conjuntamente con la minería de datos de un modo indiferente a pesar de que existen diferencias entre los dos. Según Fayyad (1996) KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. El conocimiento extraído debe tener las siguientes propiedades:

Válido: hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.

Novedoso: que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.

Potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.

Comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

El KDD es un proceso complejo que incluye no sólo la obtención de modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos.

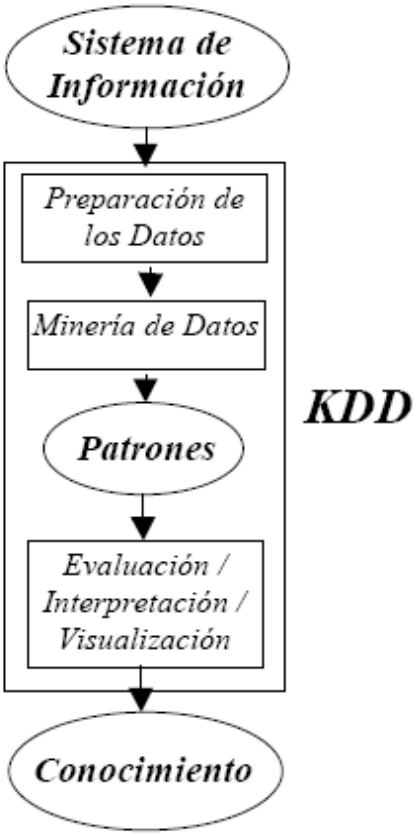

Figura 2.1 Fases del KDD

Los sistemas KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso. Esta definición del proceso clarifica claramente la relación entre minería de datos y KDD: el KDD es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos. Al ser fase de generación de modelos, comúnmente se asimila KDD con minería de datos.

2.3 Análisis de Correspondencia

El Análisis de Correspondencias según Gutierrez, G (2002) fue ideado por el estadístico francés Benzecri en 1973. Este análisis posibilita el profundizar en el análisis de la similaridad de modalidades, de la asociación entre atributos y de las relaciones de atracción y repulsión de sus modalidades en tablas de contingencia construidas comúnmente sobre variables de tipo cualitativo.

De forma simplificada, el Análisis de Correspondencias, trata de resolver el problema de explicar de forma simple como se atraen o repelen las modalidades de las variables cualitativas observadas. Además del análisis de atracción-repulsión entre modalidades de atributos (variables cualitativas) diferentes, la técnica del Análisis Factorial de Correspondencias también permite realizar estudios de proximidad (similaridad/disimilaridad) entre las modalidades de una misma variable; es decir, permite evaluar el parecido de las distribuciones de casos o individuos que presentan modalidades comparadas de una misma variable, proporcionándonos información necesaria para evaluar la homogeneidad de las mimas.

Así pues, los objetivos de esta técnica son:

- Descubrir las relaciones de atracción-repulsión existentes entre las distintas modalidades de diferentes variables cualitativas enfrentadas en una tabla de contingencia.

- Descubrir las relaciones de proximidad existentes entre las distintas modalidades de una misma variable cualitativa.

- Visualizar y caracterizar de forma simple las relaciones anteriores en un espacio de dimensión lo más reducida posible".

2.3.1 Análisis de Correspondencia Múltiple

El análisis de correspondencias múltiple cuantifica los datos nominales (categóricos) mediante la asignación de valores numéricos a los casos (objetos) y a las categorías, de manera que los objetos de la misma categoría estén cerca los unos de los otros y los objetos de categorías diferentes estén alejados los unos de los otros. Cada objeto se encuentra lo más cerca posible de los puntos de categoría para las categorías que se aplican a dicho objeto. De esta manera, las categorías dividen los objetos en subgrupos homogéneos. Las variables se consideran homogéneas cuando clasifican objetos de las mismas categorías en los mismos subgrupos. Manual SPSS 15 (2006)

El análisis de correspondencias múltiple se puede utilizar para representar gráficamente la relación entre la edad, la estatura, el peso, la alimentación de los pacientes y si estos han padecido de fractura de cadera y medir así el grado de repulsión o atracción de las variables.

Capitulo 3

Minería de Datos para Determinar Variables Significativas

En este capítulo se explica el procedimiento empleado para la obtención de las bases de datos de pacientes del Instituto de Densitometria Ósea de Mérida Venezuela, la cual cuenta con un banco de datos de diez años de antigüedad. Se habla sobre los criterios de limpieza aplicados sobre los registros así como la escogencia de las variables necesarias para este estudio. Se menciona el procedimiento empleado para la toma de la muestra. Se explica la manera como se aplicó estadística descriptiva sobre los datos y las técnicas de análisis de correspondencia para estudiar el grado de relación que tienen las variables en cuestión.

3.1 Naturaleza de los Datos

La base de datos originalmente consta de 27970 registros en donde se especifica el nombre del paciente y sus datos básicos como nombre, edad, peso, cédula de identidad, sexo, fecha de nacimiento, (a partir de la cual se re-calculó la edad actual) teléfono, dirección, entre otras variables que eran mediciones médicas realizadas para estudios como el tipo de tratamiento que se aplicó y otras más que no son de importancia para este estudio.

Con esta base de datos y la permisología necesaria fue extraida del Instituto de Osteoporosis y Densitometría Ósea y fue trabajada en Excel, en donde se procedió a explorar los registros en cuestión.

De las variables antes mencionadas se va a trabajar con Edad, Estatura y Peso, el resto de las variables contenidas en la base de datos original no son necesarias para este estudio.

3.2 Limpieza de Datos:

Los registros están conformados por mujeres mayores de 50 años por lo que de la base de datos original son eliminados aquellos registros que estaban formados por pacientes de sexo masculino y aquellas mujeres con edad inferior a los 50 años. Cabe aclarar que la edad de los registros ha sido actualizada al momento de comenzar este estudio, acción sencilla de lograr puesto que se contaba con la fecha de nacimiento de cada uno de los pacientes.

Una vez filtrada la base de datos se procedió a chequear cada una de las variables así como sus diferentes valores a lo largo de los registros con la finalidad de detectar y corregir valores faltantes y/o erróneos.

3.3 Generación de la Muestra.

Para la generación de la muestra se empleó muestreo estratificado que según Gullan J.A (1966) es aquel en el que se divide la población de N individuos, en k subpoblaciones o estratos del mismo tamaño. Luego, según la cantidad de elementos requeridos para la muestra final se toma la misma cantidad de cada uno de los estratos que sumada dará el valor de la muestra requerida.

Se empleó un programa codificado en Matlab 7.0 que se basa en muestreo estratificado sobre nuestra base de datos, dando como resultado una muestra de 770 registros.

3.4 Medición y creación de nuevas variables.

Nuevas variables son generadas en este estudio, entre ellas destacan la longitud del cuello femoral en centímetros Longitudcm, Angulo del cuello femoral (en grados) y el valor Tscore que es un indicador del estado de osteoporosis en donde se considera osteoporosis a partir de -2.5 y entre más negativo sea dicho valor más riesgo a fracturarse se tendrá. Estas tres variables fueron tomadas haciendo uso del programa Norland X32 en donde cada uno de los pacientes generados por la muestra era buscado para tomar las medidas correspondientes. Esta medición fue tomada registro a registro de forma manual.

Posteriormente se procedió a la entrevista telefónica de cada uno de las personas en la muestra en donde nuevas variables fueron creadas. A cada persona entrevistada se le preguntó:

Si se habían fracturado en los últimos cinco años

Si fumaban con regularidad

Si tomaban sol con regularidad

Si realizaban algún tipo de actividad física

Si consumían yogurt, leche, pescados y sardinas, calcio en tabletas.

Estas entrevistas dieron lugar a las creación de las variables SufrioFractura, Fuma, Toma Sol, Actividad Física, Pescado (esta variable incluye a las sardinas), Leche, Yogurt y Calcio. Cabe destacar que cada una de estas variables es de tipo binario con dos posibles valores (Si o No).

3.5 Categorización de las Variables

Para mejorar y facilitar el tratamiento y estudio de la información se han categorizado todas las variables en cuestión permitiendo estandarizar la información y que esta pertenezca a grupos exclusivos.

Las variables numéricas como Edad, Peso, Estatura, Longitud y Angulo se han categorizado en variables de tipo ordinal, por ejemplo, la variable Edad se ha categorizado en 9 grupos, asignándole el valor **1** a aquellas personas con edades menores e iguales a los 55 años de edad, el valor **2** perteneciente a aquellas personas con edades comprendidas entre los 56 y 60 años de edad, **3** para edades entre 61 y 65 años, y **4** para edades de 66 y 70, como se muestra en la tabla 3.1.

A continuación se resume las variables numéricas transformadas a categorías ordinales:

Variable: Edad	
Categoría	Código
<= 55	1
56 - 60	2
61 - 65	3
66 - 70	4
71 - 75	5
76 - 80	6
81 - 85	7
86 - 90	8
91+	9

Tabla 3.1 Categorización de la variable Edad

Variable: Estatura	
Categoría	Código
<= 135.00	1
135.01 - 140.00	2
140.01 - 145.00	3
145.01 - 150.00	4
150.01 - 155.00	5
155.01 - 160.00	6
160.01 - 165.00	7
165.01 - 170.00	8
170.01+	9

Tabla 3.1 Categorización de la variable Estatura

Variable: Peso	
Categoría	Código
<= 45.00	1
45.01 - 55.00	2
55.01 - 65.00	3
65.01 - 75.00	4
75.01 - 85.00	5
85.01+	6

Tabla 3.2 Categorización de la variable Peso

Variable: Longitudcm	
Categoría	Código
<= 1.40	1
1.41+	2

Tabla 3.3 Categorización de la variable Longitud

Variable: Ángulos	
Categoría	Código
<= 20.00	1
20.01 - 30.00	2
30.01 - 40.00	3
40.01+	4

Tabla 3.4 Categorización de la variable Ángulo

Variable: Tscore	
Categoría	Código
<= -2.50	1
-2.49 - 0	2
.01+	3

Tabla 3.5 Categorización de la variable Tscore

Por otra parte tenemos las variables binarias Sufrio Fractura, Toma Sol, Pescados, Leche, Yogurt, Actividad Física y Fuma que hacen referencia a los hábitos del paciente en cuestión cuyos valores posibles son: Sí o No. Para mejorar el procesamiento de estos datos sus valores de cadena se han modificado a los valores binario 0,1 (No= 0 , Si =1).

Capitulo 4

Análisis de Resultados e Interpretación

A continuación se expone los análisis estadísticos expuestos en el capitulo anterior dentro de un contexto aplicado a los datos trabajados. Se estudia la naturaleza de los datos ya categorizados, se les aplica estadística descriptiva simple a cada uno de los registros con respecto a las variables y posteriormente se aplicará análisis de correspondencias sobre los datos.

4.1 Aplicación de Estadísticas Fundamentales

Al explorar los datos ya categorizados, se obtienen una serie de gráficas en donde se representa la frecuencia de cada una de las variables a lo largo de todos los registros.

De la tabla 4.1 se observa que la gran mayoría de los pacientes corresponden a una edad entre los 55 y los 75 años de edad mientras que personas mayores a los 76 años se comienzan a mostrar con menor frecuencia. Al final de la tabla se observa que tan solo 13 pacientes poseen edades mayores a los 90 años.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
<= 55	107	14,2	14,2
56-60	142	18,8	32,9
61-65	141	18,7	51,6
66-70	115	15,2	66,8
71-75	110	14,6	81,3
76-80	68	9,0	90,3
81-85	43	5,7	96,0
86-90	17	2,2	98,3
>90	13	1,7	100,0
Total	756	100,0	

Tabla 4.1 Frecuencia de Variable Edad

A continuación en la tabla 4.2 observamos la estatura en donde la mayoría de las mujeres examinadas en la muestra poseen una estatura entre 150 a 155 centímetros de longitud con una frecuencia de 233 que representa el 30.80% de la muestra. Es importante notar que existen mujeres de baja estatura, inferiores a 145 cm con un porcentaje de 9.8%, tan solo 5 personas poseen una estatura superior a los 170 cm.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
<= 135.00	2	,3	,3
135.01 - 140.00	13	1,7	2,0
140.01 - 145.00	59	7,8	9,8
145.01 - 150.00	166	22,0	31,7
150.01 - 155.00	233	30,8	62,6
155.01 - 160.00	181	23,9	86,5
160.01 - 165.00	80	10,6	97,1
165.01 - 170.00	17	2,2	99,3
>170.00	5	,7	100,0
Total	756	100,0	

Tabla 4.2 Frecuencia de Variable Estatura

Con respecto a la variable peso notamos que el 39.9% de la muestra tiene un peso entre los 55 a 60 kilogramos, seguido por un 26.5% que poseen un peso entre los 65 y 75 kilogramos. En los extremos encontramos 16 pacientes con un peso menor a los 45 Kg. que representa el 2.1% de la muestra y 30 personas con un peso superior a los 85 Kg. que representan el 4% de la muestra.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
<= 45.00	16	2,1	2,1
45.01 - 55.00	145	19,2	21,3
55.01 - 65.00	302	39,9	61,2
65.01 - 75.00	200	26,5	87,7
75.01 - 85.00	63	8,3	96,0
>85.00	30	4,0	100,0
Total	756	100,0	

Tabla 4.3 Frecuencia de Variable Peso

En la tabla 4.4 se detalla la frecuencia de la variable longitud (del cuello femoral). Claramente se observa que el 68.5% de los datos tienen una longitud de cuello femoral de 1.5 cm, una longitud de 1.4 cm para el 7% de la muestra y 10.6% para una longitud de 1.3 cm. Tan solo 5 personas poseen una longitud de 0.9 cm.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
,9	5	,7	,7
1,0	30	4,0	4,6
1,1	23	3,0	7,7
1,2	47	6,2	13,9
1,3	80	10,6	,7
1,4	53	7,0	4,6
1,5	518	68,5	7,7
Total	756	100,0	13,9

Tabla 4.4 Frecuencia de Variable Longitud

El ángulo del cuello femoral muestra una tendencia a estar comprendido entre los 30 y 40 grados que representan un 63.2% de los registros. Un 29.2% de los datos poseen ángulo entre los 20 y 30 grados, un 0.5% con ángulos menor a los 20° y un 7% con ángulos mayores a los 40 grados.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
<= 20.00	4	,5	,5
20.01 - 30.00	221	29,2	29,8
30.01 - 40.00	478	63,2	93,0
>40.00	53	7,0	100,0
Total	756	100,0	

Tabla 4.5 Frecuencia de Variable Ángulo

Otra de las variables que fue medida en este trabajo fue el Tscore como indicador del nivel de DMO (densidad mineral ósea). Se observa en la tabla 4.6 que del total de 756 pacientes 275 tienen osteoporosis definida, 449 tienen baja densidad mineral ósea lo que los hace propensos a tener osteoporosis y tan solo 32 personas poseen una buena densidad mineral ósea.

Categoría	Frecuencia	Porcentaje	Porcentaje Acumulado
Osteoporosis Definida	275	36,4	36,4
Baja Densidad Mineral Osea	449	59,4	95,8
Buena Densidad Mineral Osea	32	4,2	100,0
Total	756	100,0	

Tabla 4.6 Frecuencia de Variable Tscore

Finalmente se resumirán las variables binarias, es decir, aquellos cuyos únicos valores posibles son Si o No. Estas variables fueron tomadas en la entrevista telefónica hecha a cada uno de los pacientes de la muestra. La tabla 4.7 resume la información.

Variable	No	Si	Total
Sufrió Fractura de Cadera	713 (94.3%)	43 (5.7%)	756
Toma Sol	611 (80.8%)	145 (19.2%)	756
Hábito de Fumar	717 (94.8%)	39 (5.2%)	756
Actividad Física	133 (17.6%)	623 (82.4%)	756
Consume Pescados	107 (14.2%)	649 (85.8%)	756
Consume Leche	65 (8.6%)	691 (91.4%)	756
Consume Queso	14 (1.9%)	742 (98.1%)	756
Consume Yogurt	118 (15.6%)	638 (84.4%)	756
Consume Calcio Capsulas	46 (6.1%)	710 (93.9%)	756

Tabla 4.7 Frecuencia de Variables Dicotómicas

Al ver la tabla 4.7 podemos decir que 5.7% de la muestra si sufrió fractura de cadera al caerse de su propia altura, que constituye 43 pacientes del total de 756. La gran mayoría de las personas realiza actividad física, esta variable debe ser entendida como no sedentarismo, es decir, que las personas que afirmaron realizar actividades físicas se referían a que caminaban constantemente durante sus actividades diarias: mientras se dirigían al mercado, cuando salían a tomar el autobús, caminaban dentro de la casa haciendo sus labores diarias, regando el jardín, atendiendo nietos e hijos, etc. Otra variable importante de notar es el hábito de fumar en donde tan sólo el 5.2% de la muestra afirmaron ser fumadoras. El resto de las variables se denotan claramente en la tabla 4.7.

4.1.1 Gráfico de Barras

A continuación presentamos un gráfico de barras obtenido de SPSS 15.0 en donde se observan los pacientes con osteoporosis definida, los ángulos de cuello femoral así como su longitud.

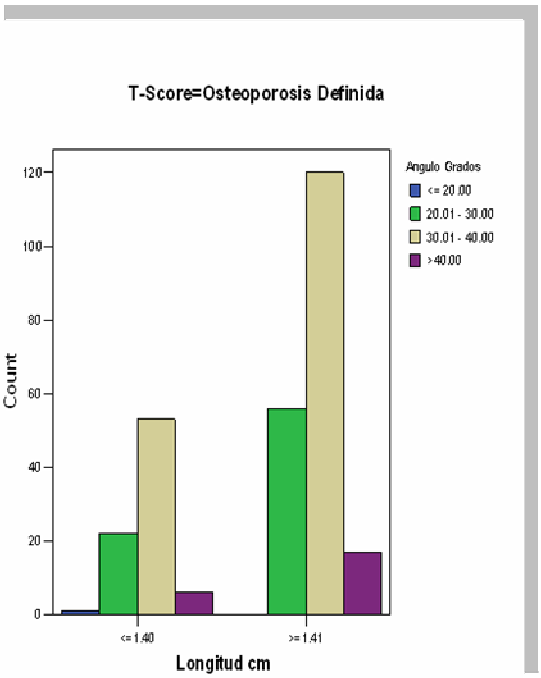

Gráfico 4.1 Gráfico de Barras. Pacientes con Osteoporosis Definida con Longitud y Ángulo

Claramente se observa que la gran mayoría de los pacientes en la muestra poseen un ángulo mayor a 1.41 cm y ángulos entre los 30 y 40 grados. Seguido de aquellos pacientes con ángulos entre los 20-30 grados y una longitud mayor a 1.41cm. Debido a que este grupo posee osteoporosis definida es el más propenso a sufrir fracturas.

El gráfico 4.2 muestra la misma relación que estamos discutiendo solo que rotando los ejes. Igualmente se llega a las mismas conclusiones.

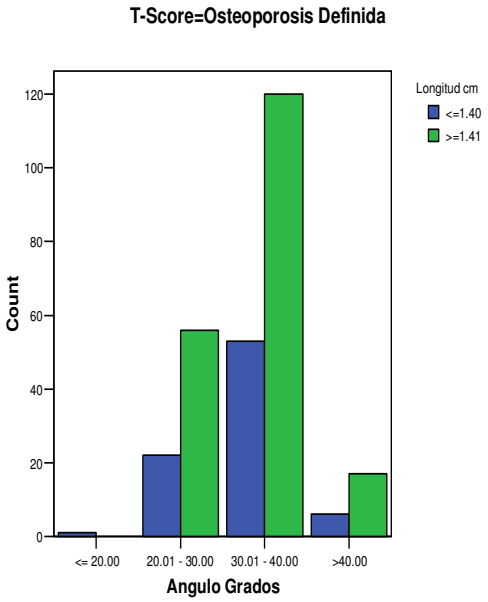

Gráfico 4.2 Gráfico de Barras. Pacientes con Osteoporosis Definida con Longitud y Ángulo

4.3 Análisis de Clúster

En este apartado realizamos un análisis de clúster y un análisis factorial a los datos en cuestión para empezar a observar las asociaciones de las variables sobre un mapa perceptual, por lo que fue necesario realizar las siguientes clasificaciones sobre los datos:

Variable	Categoría	Código
Edad	<= 55	edad1
	56-60	edad2
	61-65	edad3
	66-70	edad4
	71-75	edad5
	76-80	edad6
	81-85	edad7
	86-90	edad8
	>90	edad9

Tabla 4.8 Re codificación de Variable Edad

Variable	Categoría	Código
Estatura	<= 135.00	ta1
	135.01 - 140.00	ta2
	140.01 - 145.00	ta3
	145.01 - 150.00	ta4
	150.01 - 155.00	ta5
	155.01 - 160.00	ta6
	160.01 - 165.00	ta7
	165.01 - 170.00	ta8
	>170.00	ta9

Tabla 4.9 Re codificación de Variable Estatura

Variable	Categoría	Código
Peso	<= 45.00	pes1
	45.01 - 55.00	pes2
	55.01 - 65.00	pes3
	65.01 - 75.00	pes4
	75.01 - 85.00	pes5
	>85.00	pes6

Tabla 4.10 Re codificación de Variable Peso

Variable	Categoría	Código
Longitud	≤1.40	lon1
	≥1,41	lon2

Tabla 4.11 Re codificación de Variable Longitud

Variable	Categoría	Código
Ángulo	<= 20.00	ang1
	20.01 - 30.00	ang2
	30.01 - 40.00	ang3
	>40.00	ang4

Tabla 4.12 Re codificación de Variable Ángulo

Variable	Categoría	Código
Tscore	Osteoporosis Definida	oste
	Baja Densidad Mineral Osea	bdmo
	Buena Densidad Mineral Osea	buen

Tabla 4.13 Re codificación de Variable Tscore

A continuación se muestran la re codificación de las variables binarias:

Variable	No	Si
Sufrió Fractura de Cadera	nsfc	ssfc
Toma Sol	nsol	ssol
Hábito de Fumar	nhdf	shdf
Actividad Física	nafi	safi
Consume Pescados	ncpe	scpe
Consume Leche	ncle	scle
Consume Queso	ncqu	scqu
Consume Yogurt	ncyo	scyo
Consume Calcio Capsulas	nccc	sccc

Tabla 4.14 Re codificación de las variables dicotómicas

Por medio de un análisis de clúster y análisis factorial podemos llegar a las siguientes asociaciones de cada una de las siguientes variables. Las figuras se pueden observar con en los anexos.

Con una asociación fuerte tenemos:

1. glo con sol
2. cpe – cle
3. sfc-nhdf

Con una asociación menos acentuada tenemos:

1. gan-gts
2. ged-ges-gpe

Como una asociación espurea tenemos a

1. afi-cyo

4.2 Análisis de Correspondencias Múltiple

Con la ayuda de la estadística descriptiva a estas alturas tenemos un conocimiento básico y claro de las variables trabajadas para cada uno de los pacientes. En este apartado comenzaremos a realizar el análisis de correspondencias múltiples que nos permitirá tener una visión más clara y detallada de los datos de donde queremos detectar grupos o conglomerados y detectar variables significativas que nos permitan concluir acerca de los riesgos para sufrir fracturas de cadera.

Para realizar el análisis de correspondencias se utilizó el SPAT 4.50 y se re codificaron las variables de la siguiente manera como muestran en las tablas 4.8 a 4.13 que ya descritas en el aparatado anterior.

Al aplicar análisis de correspondencias a la muestra en cuestión obtenemos el gráfico que se muestra a continuación en donde los rombos rosados representan a cada variable y cada punto azul es una paciente.

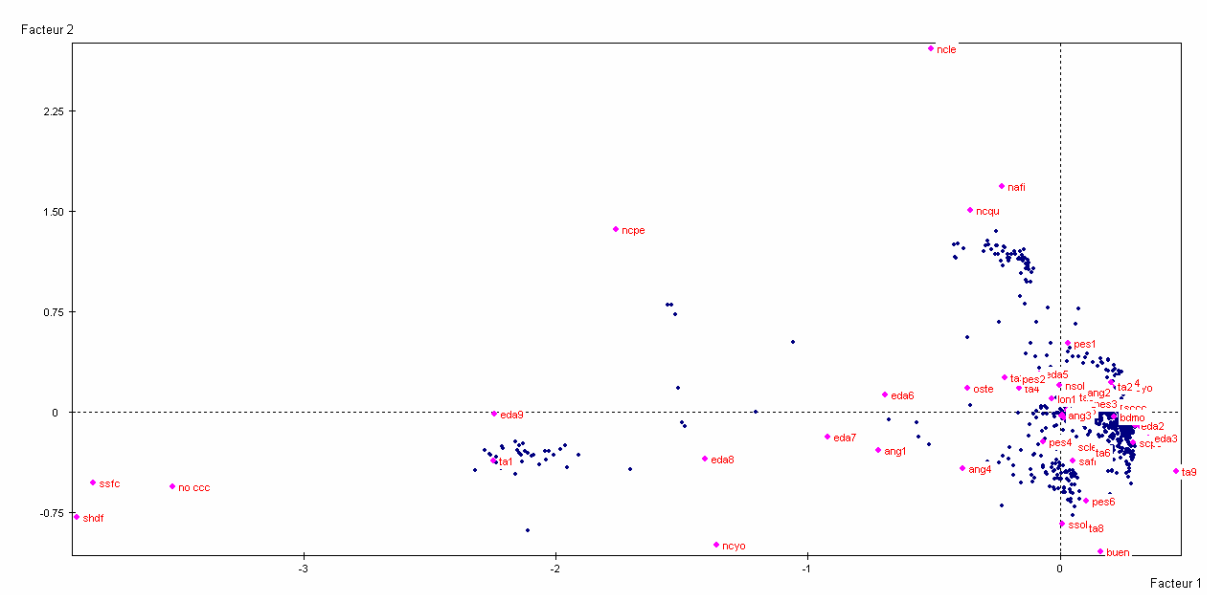

Figura 4.1 Análisis de Correspondencias Múltiple (Factor1 y Factor2)

De manera muy bien definida se observa en el extremo inferior izquierdo que las variables noccc (no consume calcio en cápsulas) shdf (si posee hábito de fumar) y ssfc (si se fracturó la cadera) están muy bien relacionadas entre sí, además de estar muy próximas al factor 2.

También notamos que próximo a estas tres variables que acabamos de mencionar se encuentra las variables ta1 (talla menor o igual a los 135 cm) y edad9 (mayores a los 90 años) y alrededor de estas dos variables un grupo de puntos azules que representan un grupo de individuos. Esto se interpreta como individuos con una altura menor a los 135 cm y mayores de 90 años, lo que tiene mucho sentido ya que a medida que la edad avanza el ser humano tiende a encogerse.

Junto al factor 1 notamos a las variables buen (buena densidad mineral ósea), ta8 (tallas entre 165 y 170 cm) y ssol (si recibe sol) y pes6 (pesos mayores a los 85 Kg.). El indicador de tener una buena densidad mineral ósea está relacionado con valores positivos del tscore y por lo general el sistema (de densitometria ósea) les asigna dichos valores a personas con sobrepeso debido al tejido que poseen fijo en los huesos.

En el extremo superior derecho completamente opuesto a la variable buen (buena densidad mineral ósea) encontramos la variable ncle (no consume leche) y un poco mas por debajo de la misma nafi (no realiza actividad física) y ncqu (no consume queso).
En el gráfico 4.4 se muestra el mismo análisis de correspondencias pero esta vez con los factores 1 y 3. En el gráfico 4.5 se muestra con los factores 2 y 3. En ambos casos se llegan a las mismas conclusiones.

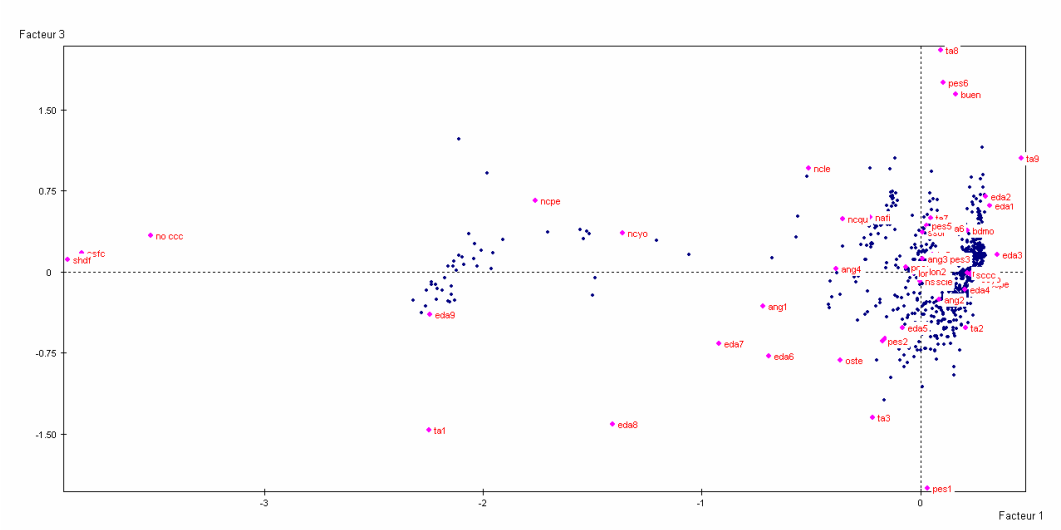

Figura 4.2 Análisis de Correspondencias Múltiple (Factor1 y Factor3)

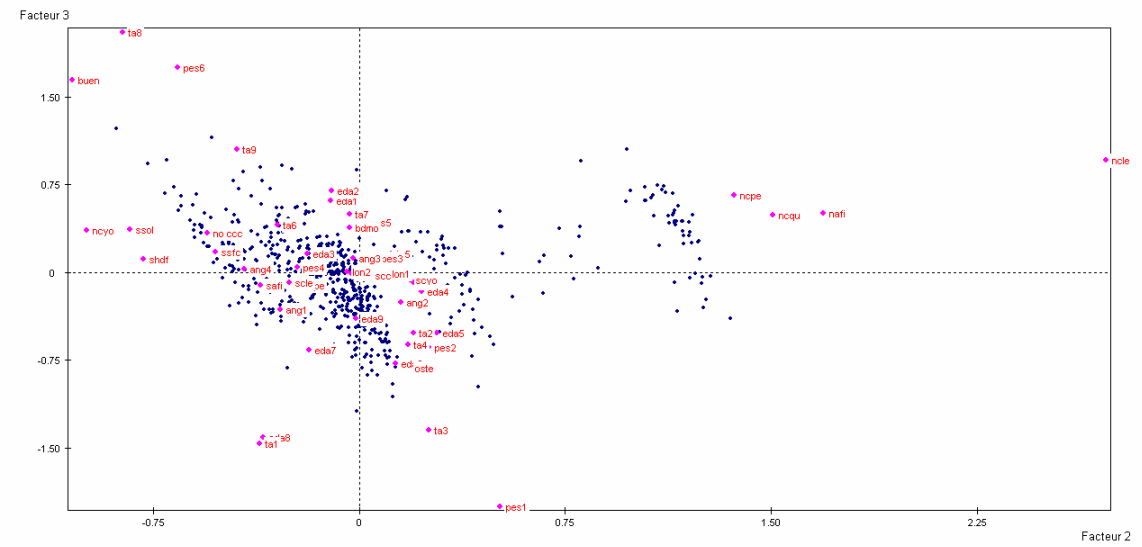

Figura 4.3 Análisis de Correspondencias Múltiple (Factor2 y Factor3)

Nuestra variable de interés es si el paciente se fracturó o no la cadera, por lo que a continuación se muestra el análisis de correspondencias en donde los 1s representan aquellas personas que si recibieron una fractura de cadera y los 2s aquellas personas que no recibieron fractura de cadera.

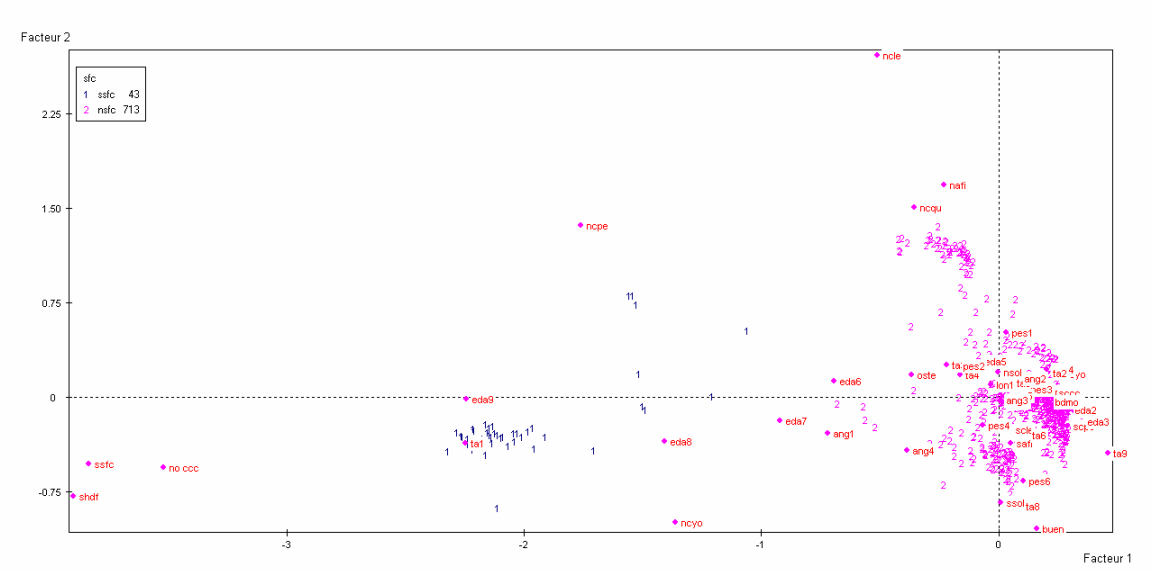

Figura 4.4 Análisis de Correspondencias Múltiples con Variable Fractura de Cadera

Nótese que los 1s de color azul son aquellas personas que recibieron fractura de cadera y se encuentran próximas a las variables noccc (no consume calcio en cápsulas) shdf (si posee hábito de fumar) ssfc (si se fracturó la cadera) ta1 (talla menor o igual a los 135 cm) edad9 (mayores a los 90 años) ncpe (no consume pescado) edad8 (edades entre 86 y 90 años) y ncyo (no consume yogurt).

Una vez observado el mapa perceptual de cada una de las categorías de cada variable sobre los factores procedemos a conglomerarlos por grupos como se muestra en la figura 4.7.

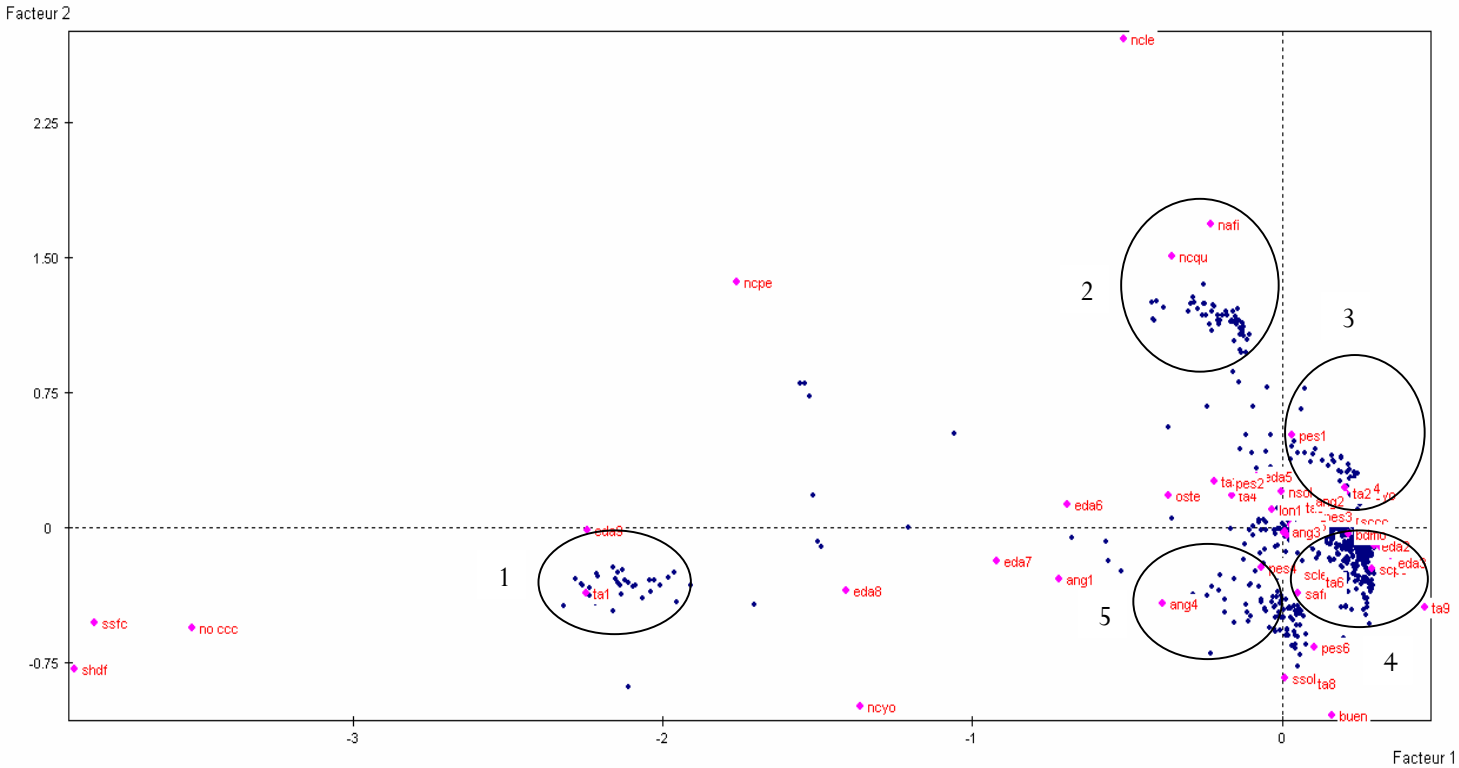

Figura 4.7 Análisis de Correspondencias Múltiple con conglomerados

En el gráfico 4.7 se observa que cada uno de los conglomerados se encuentra organizado con respecto a cada una de las categorías de las variables como sigue (tabla 4.15) en donde cada categoría de variable (cuya leyenda se encuentra en las tablas tablas 4.8 a 4.13 del presente capitulo) está contenida dentro de los límites del conglomerado.

Conglomerado N°	Categorías de variable asociadas
1	edad 9 talla 1
2	nafi, ncqu
3	pes1, ta2, scyo,nhf,sccc,pes3,ta5,pes5
4	bdmo,edad2,edad3,ta6,safi,scle
5	ang4, pes4

Tabla 4.15 Categorías de variable asociadas a cada conglomerado

Con el criterio de la tabla 4.15 se logró clasificar 191 registros que constituyen el 25.26% de la data. El resto de los registros se terminó de clasificar asignando a cada

conglomerado las categorías de variables que más próximas y mayor relación tuvieran con cada uno de los conglomerados como se muestra en la tabla 4.16.

Conglomerado N°	Categorías de variable asociadas	Nuevas Categorías de variable asociadas
1	edad 9 talla 1	edad8, ncyo, ssfc, shdf, noccc
2	nafi, ncqu	ncpe, cle, oste
3	pes1, ta2, scyo,nhf,sccc,pes3,ta5,pes5	ta3,pes2,ta4,edad6
4	bdmo,edad2,edad3,ta6,safi,scle	ta9,edad3,edad2,edad1,bdmo, scpe,pes6,ta8,buen,ssol,ta6,ang3,ta7
5	ang4, pes4	edad7, ang1

Tabla 4.16 Nuevas Categorías de variable asociadas a cada conglomerado

Con esta segunda clasificación se terminó de clasificar toda la data restante de modo que todos los puntos pertenecen a uno de los 5 conglomerados.

Basándonos en esta última clasificación se procedió a asignarles una etiqueta o nombre que describe a cada conglomerado como sigue:

Conglomerado N°	Etiqueta
1	Pacientes con osteoporosis fracturados y/o con mucho riesgo a fractura
2	Pacientes propensos a sufrir fractura
3	Pacientes propensos a contraer osteoporosis
4	Pacientes con menor propensidad a fracturarse
5	Pacientes media propensidad a fracturarse

Tabla 4.17 Etiquetas asignadas a cada conglomerado

4.3 Aplicación de Máquinas de Soporte Vectorial. (MSV)

En este apartado se construyó un modelo matemático basado en máquinas de soporte vectorial con la librería LIBSVM de linux con la intención de tener una herramienta que permita clasificar nuevos registros (pacientes) dado los valores en cada una de las variables y así poder predecir a que conglomerado o grupo pertenecerán.

Para construir el modelo de máquina de soporte vectorial se utilizó como función de kernel tipo radial en donde se probó con diferentes valores de parámetros hasta conseguir el mejor modelo, es decir, el modelo que prediga mejor. Para escoger los datos de entrenamiento y validación se empleó muestreo estratificado, (igual con la primera muestra original de la base de datos), en donde los datos de entrenamiento representan el 90% de la data y los datos de validación el 10% restante.

A continuación resumimos los diferentes valores de los parámetros con sus respectivos aciertos y porcentaje de precisión.

C	g	Aciertos
1	0.01	75% (57/76)
1	0.1	77.63% (59/76)
1	1.0	82.89% (63/76)
1	10	61.84% (47/76)
1	100	61.84% (47/76)
10	0.01	76.31% (58/76)
10	0.1	86.84% (66/76)
10	1.0	89.47% (68/76)

10	10	61.84% (47/76)
10	100	61.84% (47/76)
100	0.01	81.57% (62/76)
100	**0.1**	**94.73% (72/76)**
100	1.0	89.47% (68/76)
100	10	61.84% (47/76)
100	100	61.84% (47/76)

Tabla 4.18 Parámetros de entrenamiento para MSV

El mejor modelo para máquina de soporte vectorial se consiguió con los parámetros C= 100 y gamma= 0.1 con una precisión de 94.73%. La información detallada de esta validación se exhibe mejor a través de una matriz de confusión en donde las filas representas los valores reales de cada instancia y cada columna los valores asignados por el modelo matemático obtenido.

	1	2	3	4	5
1	15	0	0	0	0
2	0	30	**1**	0	0
3	0	**2**	10	0	0
4	0	0	**1**	16	0
5	0	0	0	0	1

Tabla 4.19 Matriz de confusión

De la matriz de confusión se aprecia que el modelo ha entrenado bien ya que la validación es buena. Una instancia perteneciente a la clase 2 la ha clasificado como de clase 3. Dos instancias de clase 3 han sido clasificadas como de tipo 2 y una instancia de tipo 4 ha sido clasificada como de tipo3. De 76 instancias utilizadas para probar y validar el modelo 74 fueron clasificadas correctamente.

Capitulo 5

Conclusiones y Recomendaciones

5.1 Conclusiones

La aplicación de análisis KDD a los datos iniciales (banco de datos) con más de 30 mil registros permitió arreglarlos de un modo mucho más cómodo para la vista del ser humano.

La selección de la muestra con muestreo estratificado resultó ser significativa para llegar a los objetivos planteados al inicio de este estudio y la estadística descriptiva aplicada que incluyó histogramas, tablas de frecuencia, gráfico de barras permitió observar la tendencia de cada una de las variables y sus rangos de valores para proceder luego a la fase de limpieza que es la parte más incómoda de la minería de datos ya que es esta fase donde se detectan datos espurios que pudieran ser eliminados, corregidos o reemplazados.

La categorización de las variables permitió observar diferencias entre cada una de ellas de manera observable a cortos pasos.

El paso de de variables cuantitativas a cualitativas para realizar el análisis de correspondencias múltiples nos dio un mapa perceptual bastante cómodo para interpretar y diferenciar cada una de las categorías de variables. Este es el paso más importante porque nos permite concluir acerca de la relación de atracción y repulsión de las variables de las variables y como resultaron bien agrupadas por conglomerados bien definidos podemos llegar a las conclusiones siguientes:

Personas mayores a los 86 años que no consumen yogurt, que poseen el hábito de fumar, que no consumen calcio en capsulas, que no toman sol y que presentan una etapa avanzada de osteoporosis de hecho sufrieron una fractura de cadera al caerse de su propia altura y aquellas que no se fracturaron poseen un alto riesgo a fracturarse.

Pacientes que no realizan actividades físicas, no consumen queso, pescados, leche y que tienen definida la osteoporosis se encuentran propensos a sufrir fractura de cadera al caerse de su propia altura.

Los pacientes con menor riesgo a sufrir fractura de cadera son aquellos que poseen los siguientes hábitos y características: buena densidad mineral ósea, toman sol, si realizan actividades físicas, consumen leche y pescados, calcio en cápsulas. Poseen ángulo de cuello femoral entre los 30° y 40° y edades entre 55 y 65 años.

Se encontró una fuerte asociación entre la longitud del cuello femoral y el hábito de tomar sol al menos 1 vez al día. También con fuerte asociación tenemos los hábitos de consumir pescados y sardinas y consumir leche. El hecho de sufrir una fractura de cadera tiene una relación fuerte con el hábito de fumar.
Con una asociación menos fuerte tenemos el ángulo del cuello femoral con la densidad mineral ósea. La edad, estatura y peso resultaron también asociados así como consumir calcio en cápsulas y queso. Por otra parte no se determinó asociación con realizar actividades físicas ni con el hábito de consumir yogurt.

El modelo matemático basado en máquina de soporte vectorial entrenó y validó muy bien la data de prueba por lo que este modelo puede ser usado en el futuro para clasificar instancias de nuevos pacientes.

5.2 Recomendaciones

- Diseñar un sistema que permita tomar de manera automática medidas del cuello femoral y de su ángulo como de otras áreas de interés médico en la tecnología norland x32 para sistematizar la recolección de datos para futuros estudios.

- Realizar el mismo análisis de este estudio pero con muestras de otras poblaciones de otros estados y ciudades del país y comparar resultados.

- Realizar una reunión en donde los resultados de este trabajo sean discutidos en medios relacionados con la medicina.

- Realizar otros modelos matemáticos con otros métodos y tecnologías para luego comparar con los obtenidos en este apartado.

Bibliografía

Berkow, R., Beers, M., Fletcher, A. (2005)., Manual de Merck. Océano. Grupo Editorial S.A. España.

Arellana, J., Ferrández L., (2007). Guía de Buena práctica Clínica en Geriatría. Anciano Afecto de Fractura de Cadera. Rovi. España.

Wasnich, R. (1999). No Espere tener una fractura. Identifique la Osteoporosis. *QBGYN.net Latina*. Disponible en http://latina.obgyn.net/sp/articles/enero%2000/osteo.htm

Consensus Development Conference. Diagnosis, prophylaxis, and Treatment of osteoporosis Am J. Med (1993) 94:646 – 50.

Argimon – Pallas, J. Jiménez Villa, J. (2005). Métodos de Investigación Clínica. Tercera Edición. Elsevier. Madrid España.

Yupanqui, H. Castillo, J. Calvo, E. Soto, C (2003). Densitometría Ósea En Colombia. Biogen. Colombia.

Orallo, J. Ramírez, M. Ferri, C. (2004). Introducción a la Minería de Datos. Pearson Prentice Hall. Madrid. España.

Salvador Figueras, M (2003): "Análisis de Correspondencias", [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/correspondencias

Fayyad, U. M; Piatetsky-Shapiro, G; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, pp 1-34, AAAI/MIT Press 1996.

Gutierrez Montoya G (2002): "Análisis de Correspondencias", [en línea] www.aulafacil.com/curso-spss-gratis-2/curso/Temario.htm,

SPSS para Windows. Versión 15.0.1 (22 Nov 2006) Vista Hotfix Applied. SPSS Inc.

J.A Gullan (1966), Manual 3. Manual de Métodos de Muestreo y Estadísticos para la Biología Pesquera - Parte 1. Métodos de Muestreo. Laboratorio de pesca Lowestoft (Inglaterra)

Geel, A., Geusens, P., Nagtzaam, I. Schreurs, C., van der Voort, D., Rinkens, P., Kester, A., Dinant G. (2006). Timing and risk factors for clinical fractures among postmenopausal women: a 5-year prospective study. *BMC Medicine* (4:24)

Mosquera, M., Laurel, D., Pavon, S. Incidencia y factores de riesgo de la fractura de fémur proximal por osteoporosis. *Rev Panam Salud Pública*. 3(4): 211-219.q

Sapunar, J.,Bravo, P., Schneider, C., Jiménez, M. 2003 ¿Es la etnia mapuche un factor de riesgo para padecer fracturas de fémur proximal en adultos mayores?. Rev. Méd. Chile 131 (10): 1135-1141.

Travison, T.G, Beck T.J. Esche G.R. McKinlay, J.B. 2007. Age trends in proximal femur geometry in men: variation by race and ethnicity. Osteoporosis Internacional.19 (3:277-287)