SECCIÓN ARTÍCULOS ORIGINALES
REVISTA UNIVERSIDAD Y SALUD
Año 2012 Vol. 14(2) Págs. 117 - 129

La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino

Applied data mining patterns to discovery survival in women with invasive cervical cancer

Ricardo Timarán Pereira[1], María Clara Yépez Chamorro[2]

1. Doctor en Ingeniería énfasis Ciencias de la Computación. Director Grupo de Investigación GRIAS. Profesor Asociado Departamento de Sistemas. Facultad de Ingeniería. Universidad de Nariño. San Juan de Pasto, Colombia. e-mail: ritimar@udenar.edu.co
2. Magister en Ciencias Biomédicas. Directora Centro de Estudios en Salud Universidad de Nariño. Profesora Asociada Facultad de Ciencias de la Salud. Universidad de Nariño. San Juan de Pasto, Colombia. e-mail: mcych@udenar.edu.co

Fecha de recepción: Marzo 30 - 2012 Fecha de aceptación: Diciembre 18 - 2012

Timarán Pereira R, Yépez Chamorro MC. La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. Rev Univ. salud. 2012; 14(2):117 - 129

RESUMEN

En este artículo se presentan los resultados del proyecto de investigación cuyo objetivo fue extraer patrones de supervivencia en mujeres con diagnóstico de cáncer invasivo de cuello uterino utilizando técnicas de minería de datos a partir de la información almacenada en el Registro Poblacional de Cáncer del Municipio de Pasto (Colombia), durante el periodo de 1998 a 2007. De acuerdo a los resultados obtenidos aplicando la técnica de clasificación basada en árboles de decisión, el tiempo de supervivencia de estas mujeres es mayor que 37 meses, contados a partir de la fecha de diagnóstico hasta la fecha última de observación de este estudio. Aplicando la tarea de Asociación se conocieron los principales factores socioeconómicos y clínicos asociados a la supervivencia de este grupo poblacional. El conocimiento generado permitirá soportar la toma de decisiones eficaces de los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de políticas públicas y programas de protección a las mujeres con esta enfermedad.

Palabras clave: Cáncer de cuello uterino, minería de datos. (Fuente: DeCs, BIREME)

ABSTRACT

In this paper the results of the research project whose goal was to extract survival patterns in diagnosed women with invasive cervical cancer with data mining techniques from data reported in Population-based Cancer Registry of the Municipality of Pasto (Colombia), during a period between 1998 and 2007 are presented. According to the

results,which were obtained by applying the classification technique based on decision trees, the survival time of these women is greater than 37 months, counted from the date of diagnosis to the last date of observation of the present study. By applying the Association task, the main socioeconomic and clinical factors associated with survival of this population group were met. The generated knowledge will support effective decision making of government agencies and private health sector in relation to the approach of public policies and protection programs for women with this disease.

Key words: Data mining, cervical cancer. (Source: MeSH, NLM)

INTRODUCCIÓN

Tradicionalmente, el análisis estadístico ha sido una herramienta que a través de la confirmación de hipótesis ha permitido dentro del campo de la salud generar conocimientos y comprender ciertos fenómenos en beneficio del desarrollo de tecnología y de protocolos para el cuidado de las enfermedades y de la recuperación de la salud.[1] Mediante el análisis estadístico, se consideran fundamentalmente variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento de los datos más complejo, que solo es posible con la minería de los mismos.[2]

La minería de datos surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos.[3] De forma general, los datos son la materia prima bruta y en el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces se refiere a conocimiento.[4,5] En este contexto, la minería de datos emerge como el siguiente paso evolutivo en el proceso de su análisis.

Utilizando la información que los especialistas médicos consideran importante, con la minería de datos se pretende encontrar patrones y relaciones entre las variables, que permitan predecir de antemano, en el caso específico del cáncer uterino, los factores que inciden en la supervivencia de las mujeres que padecen esta enfermedad. En este contexto, el objetivo del trabajo es caracterizar y clasificar la población de pacientes con cáncer mediante técnicas de minería de datos, esperando encontrar relaciones subyacentes que no pueden identificarse mediante un tratamiento estadístico clásico.

En las aplicaciones médicas, donde no se puede obviar la importancia del componente temporal, las técnicas de minería de datos han adquirido gran relevancia.[6] Las aplicaciones de estas técnicas van desde la visualización inteligente de grandes cantidades de datos médicos, hasta el control de calidad en centros hospitalarios.

En detección de enfermedades como el cáncer, el conjunto de técnicas y métodos de minería de datos resultan ser un instrumento muy útil en ciertas partes del proceso de detección y diagnóstico de cáncer donde se requiere clasificar información y encontrar conocimiento en grandes volúmenes de datos.[7]

Reparaz y cols reportan el uso de la minería de datos para determinar la eficacia de la braquiterapia en tratamiento de cáncer de próstata.[8] En este contexto, el objetivo del trabajo fue caracterizar y clasificar la población de 206 pacientes tratados con braquiterapia prostática utilizando la técnica de minería de datos *clustering* con siete variables. Los resultados muestran que el 83% de los pacientes con cáncer de próstata están directamente asociados a los casos de tratamiento exitoso.

Hernández y Lorente describen la utilización de la herramienta de minería de datos Weka en la detección de cáncer de mama.[9] El objetivo de este trabajo fue detectar posibles casos de cáncer de mama mediante minería a partir del análisis en la base de datos *Wisconsin Breast Cancer Database*, la cual consta de 699 casos, cada uno con 9 atributos correspondientes a observaciones subjetivas de los tumores, más dos atributos correspondientes a la identificación del caso y a la clasificación del tumor (benigno o maligno). Los resultados obtenidos determinan que si bien todos los atributos pueden afectar en mayor o menor medida a la clasificación de un tumor como benigno o maligno, se cumple que a menor valor del atributo mayor probabilidad de que se trate de clase benigna.

De igual manera, Hurtado y cols presentan los resultados de aplicar técnicas de minería de datos en el estudio epidemiológico del cáncer mamario.[10] El trabajo consistió en levantar y analizar los procesos asociados al tratamiento de los pacientes con cáncer mamario e identificar *clusters* epidemiológicos con los registros almacenados en el periodo comprendido entre los años 2004 y 2006. El análisis de los datos entregó que el 75% de los pacientes son evaluados en niveles precoces, donde la edad se centra entre 51-60 años, predominan los pacientes de la zona central de Viña del Mar (Chile). También las pacientes operadas por tumorectomía con disección axilar, con tamaños tumorales pequeños y ganglios regionales en etapas tempranas, obtienen resultados positivos a diferencia de las no operadas y en etapas avanzadas.

Bellaachia y Guven realizaron un análisis de la predicción de la tasa de supervivencia de los pacientes con cáncer de mama usando técnicas de minería de datos.[11] Los datos usados fueron tomados de la base de datos del SEER Public-Use Data y correspondieron a 151.886 registros y 16 atributos almacenados en el periodo de 1973 a 2002. Se aplicaron tres técnicas de minería de datos: Naive Bayes, árboles de decisión con el algoritmo C4.5 y redes neuronales. Los experimentos realizados mostraron que los árboles de decisión con C4.5 presentan mejor rendimiento y exactitud para predecir la tasa de supervivencia de estos pacientes que las otras dos técnicas. De acuerdo a ello, si el tiempo de supervivencia es mayor o igual a 60 meses, el paciente vive. El 76.8 % de todos los pacientes se clasifica de esta manera. En el caso contrario el paciente muere si la causa de la muerte es cáncer de mama. El 23.2% de todos los pacientes se clasifica así.

López diseñó un sistema para descubrir el comportamiento de los factores de riesgo de cáncer de cuello uterino utilizando minería de datos en la Secretaría Distrital de Salud (SDS) de Bogotá (Colombia), específicamente el detectar patrones y relaciones entre factores de la muestra de citología, el resultado de la muestra, los métodos de planificación y establecer tendencias sobre el comportamiento del cáncer de cuello uterino.[12] El sistema permitió determinar que factores como el fumar, el iniciar a tener relaciones tempranamente, número de compañeros sexuales, información sobre la planificación, número de partos son factores de riesgo de cáncer de cuello. Se encontró también que es fácil predecir la presencia de cáncer de cuello uterino si se da la ocurrencia de los factores: número de compañeros sexuales, edad de la paciente, edad de inicio de las relaciones sexuales, método de planificación, número de partos, si fuma o no y ser gestante.

En este artículo se presentan los resultados del proyecto de investigación cuyo objetivo fue extraer patrones de supervivencia en mujeres con diagnóstico de cáncer invasivo de cuello uterino utilizando técnicas de minería de datos a partir de la información almacenada en el Registro Poblacional de Cáncer del Municipio de Pasto (Colombia), durante el periodo de 1998 a 2007 y monitoreada en diferentes bases de datos hasta el 2010. Los hallazgos encontrados

son importantes para soportar la toma de decisiones en los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de políticas públicas y programas que permitan mejorar la atención a las mujeres con cáncer de cuello uterino, con el fin de disminuir las tasas de mortalidad y como consecuencia aumentar el tiempo de supervivencia.

MATERIALES Y MÉTODOS

El Descubrir Conocimiento en Bases de Datos (DCBD) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos para que el usuario los analice.[13,14] Este proceso implica generalmente seleccionar, limpiar y transformar los datos, hacer minería de datos (*data mining*), evaluar y presentar resultados.[15,16] Los resultados obtenidos con las técnicas de minería de datos no pueden utilizarse para confirmar o rechazar hipótesis, porque puede conducir a errores fatales. Su función es otra, explorar datos, darles sentido, convertir un volumen de datos, que poco o nada aportan a la descripción, en conocimiento para interpretar un fenómeno y para adoptar decisiones de acuerdo con las necesidades.

Etapa de selección de datos

El objetivo de esta etapa es obtener las fuentes internas y externas de datos que sirven de base para el proceso de minería de ellos. Como fuente interna, se seleccionó la base de datos REGCANDB del aplicativo REGCAMP del Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP), donde se recopila periódicamente y de manera confiable datos sobre la incidencia de cáncer en el municipio de Pasto. En la ventana de observación de este estudio (1998-2007) se encuentra almacenada la información de 17.350 casos de diferentes tipos de cáncer. Como fuentes externas principales se seleccionaron la base de datos del Registro Individual de Prestación de Servicios de Salud (RIPS) y la base de datos del Sistema de Identificación de Beneficiarios Potenciales de Programas Sociales (SISBEN) del municipio de Pasto.

En RIPS se almacena el conjunto de datos mínimos y básicos que el Sistema General de Seguridad Social en Salud requiere para los procesos de dirección, regulación y control. Los datos de este registro se refieren a la identificación del prestador del servicio de salud, del usuario que lo recibe, de la prestación del servicio propiamente dicho y del motivo que originó su prestación: diagnóstico y causa externa.

En SISBEN se almacenan todos los datos socioeconómicos contemplados en la encuesta realizada a los beneficiarios con el fin de determinar un puntaje de clasificación socioeconómica que ubica al beneficiario en un determinado nivel (nivel 1 a nivel 6).

Otras fuentes externas que se seleccionaron para complementar, validar u obtener información fueron: clínicas privadas, hospitales públicos y privados, empresas sociales del estado, empresas prestadoras de servicios de salud, laboratorios de patología e instituciones de servicios especializados de salud y la Registraduría Nacional del Estado Civil Colombiano.

De los 17.350 casos de cáncer almacenados en la base de datos REGCANDB se seleccionaron inicialmente 3.151 registros correspondientes a mujeres con cáncer de cuello uterino, de los cuales únicamente 507 registros corresponden a las mujeres con cáncer invasivo de cuello uterino residentes en la ciudad de Pasto, objeto de esta investigación. Estos casos fueron almacenados en la base de datos CANCERDB, construida con el sistema gestor de base de datos PostgreSQL,[17,18] específicamente en la tabla CANCER la cual consta de 39 atributos y 507 registros. Esta tabla

servirá de base para las subsiguientes etapas del proceso de descubrimiento de patrones de supervivencia.

Etapa de limpieza de datos

El objetivo de esta etapa es obtener datos limpios, es decir, datos sin valores nulos o anómalos que permitan obtener patrones de calidad. Por medio de consultas SQL ad-hoc o a través de histogramas, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de la tabla CANCER.

Teniendo en cuenta la relevancia de ciertos atributos para la investigación, los valores nulos de estos atributos fueron actualizados con los valores encontrados en fuentes externas. En la tabla 1 se muestra este proceso y la técnica utilizada para reemplazar sus valores nulos. Por otra parte, los atributos con un alto porcentaje de valores nulos tales como *quimio* (88,76% nulos), que determina si el paciente recibió quimioterapia, *braqui* (100% nulos), que determina si el paciente recibió braquiterapia y *paliativo* (100% nulos) que determina si el paciente recibió un tratamiento paliativo, fueron eliminados por la imposibilidad de obtener estos valores con las fuentes externas o utilizando técnicas estadísticas tales como la media, mediana y la moda o derivando sus valores a través de otros.

Tabla 1. Limpieza de atributos de la tabla CANCER

Atributo	Técnica
LUGARN	Actualización con datos de la Registraduría
FECHADEFUNCION	Actualización con datos de la Registraduría
REGIMEN	Actualización con datos del SISBEN
ESCOLARIDAD	Actualización con datos del SISBEN
NUMEROHIJOS	Actualización con datos del SISBEN
PARENTESCOJEFE	Actualización con datos del SISBEN
OCUPACION	Actualización con datos del SISBEN
FECHA_CONSULTA_RIPS	Actualización con datos de RIPS
DIAG_PRINCIPAL_RIPS	Actualización con datos de RIPS

Etapa de transformación de datos

El objetivo de esta fase es transformar la fuente de datos en un conjunto listo para aplicar las diferentes técnicas de minería de datos. Para facilitar la extracción de patrones, se crearon en la tabla CANCER nuevos atributos a partir de otros atributos. En la tabla 2 se muestran estos nuevos atributos. Con el fin de generar conocimiento acerca de los factores sociales, económicos y clínicos que inciden en la supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino, se seleccionaron, de la tabla CANCER, los atributos más representativos para cada factor y se crearon en la base de datos CANCERDB, nuevas tablas con estos atributos. El resto de atributos de la tabla CANCER se eliminó. Las tablas de CANCERDB se muestran en la tabla 3.

La descripción de los 22 atributos más relevantes para la realización de esta investigación conforman la tabla t507a22all y se muestran en la tabla 4. Los 15 atributos a considerar para los factores socioeconómicos forman la tabla t507a15econ y son los 13 primeros atributos de la tabla t507a22all y los dos últimos (atributos *nmeses* y *vivo_muerto*). Los 9 atributos a considerar para los factores clínicos forman la tabla t507a09clinico y son los 9 últimos atributos de la tabla t507a22all. Los 22 atributos de la tabla t334a22vivos son los que se describen en la tabla 4.

Tabla 2. Descripción de nuevos atributos de la tabla CANCER

Atributo	Descripción
BARRIO	Barrio de residencia del paciente en el municipio de Pasto en el momento del diagnóstico del cáncer
COMUNA	Comuna del municipio de Pasto a la cual pertenece el barrio
ESTRATO	Estrato socioeconómico al cual pertenece el paciente en el momento del diagnóstico
REGIÓN	Región de nacimiento del paciente
EDADDX	Edad del paciente en el momento del diagnóstico de cáncer
CABEZAFLIA	Determina si el paciente es cabeza de familia o no
TIPOVIVIENDA	Tipo de vivienda que habita el paciente: casa, apartamento, cuarto, otro tipo de vivienda
FUENTEDEAGUA	Sistema de obtención de agua del paciente en su vivienda: acueducto, pozo, rio o similares
PUNTAJESISBEN	Puntaje obtenido por el paciente para su clasificación en el SISBEN
NIVELSISBEN	Nivel de clasificación del SISBEN
VIVOMUERTO	Determina si el paciente está vivo o muerto
NMESES	Número de meses de vida del paciente desde el momento del diagnóstico de cáncer hasta la ventana de observación del estudio (2007)

Tabla 3. Tablas base de datos CANCERDB

Tabla	Descripción
t507a22all	Tabla general con los 507 casos de cáncer de cuello uterino (334 vivos y 173 muertos) y los 22 atributos a considerar en el estudio
t507a15econ	Tabla de todos los 507 casos de cáncer de cuello uterino y los 15 atributos a considerar para los factores sociales y económicos
t507a09clinico	Tabla de todos los 507 casos de cáncer de cuello uterino y los 9 atributos a considerar para los factores clínicos
t334a22vivos	Tabla de los 334 casos de mujeres vivas con 22 atributos a considerar en el estudio

Tabla 4. Descripción de los atributos más relevantes para el estudio

Nº	Atributo	Descripción
1	*REGION*	Región de nacimiento del paciente
2	*COMUNA*	Comuna del municipio de Pasto a la cual pertenece el barrio
3	*ESTRATO*	Estrato socioeconómico al cual pertenece el paciente en el momento del diagnóstico
4	*EDADDX*	Edad del paciente al momento del diagnóstico de cáncer
5	*ESTADOCIVIL*	Estado civil del paciente en el momento del diagnóstico
6	*OCUPACION*	Ocupación del paciente en el momento del diagnóstico
7	*ESCOLARIDAD*	Escolaridad del paciente en el momento del diagnóstico
8	*REGIMEN*	Régimen al cual pertenece el paciente en el momento del diagnóstico
9	*NIVELSISBEN*	Es el nivel de clasificación del SISBEN
10	*CABEZAFLIA*	Determina si el paciente es cabeza de familia o no
11	*TIPOVIVIENDA*	Tipo de vivienda que habita el paciente: casa, apartamento, cuarto, otro tipo de vivienda
12	*FUENTEDEAGUA*	Sistema de obtención de agua del paciente en su vivienda: acueducto, pozo, rio o similares
13	*DISCAPACIDAD*	Si el paciente tiene una discapacidad o no
14	*FUENTE*	Organización donde se diagnosticó el cáncer de cuello uterino
15	*METODODX*	Método utilizado para el diagnóstico del cáncer
16	*MORFOLOGIA*	Morfología del tumor
17	*LOCESP*	Localización específica del tumor
18	*RADIO*	Existencia o no de tratamiento de radioterapia al paciente
19	*CIRUGIA*	Determina si el paciente ha tenido como tratamiento cirugía
20	*BIOPSIA*	Determina si al paciente se le realizo o no una biopsia
21	*NMESES*	Número de meses de vida del paciente desde el momento del diagnóstico de cáncer hasta la ventana de observación del estudio (2007)
22	*VIVOMUERTO*	Determina si el paciente está vivo o muerto hasta la ventana del estudio

Etapa de minería de datos

El objetivo de la etapa de minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés aplicando tareas de descubrimiento tales como *clasificación*,[19-21] *clustering*,[22,23] *patrones secuenciales*,[24] *asociaciones*[25] entre otras.

Las tareas de minería de datos escogidas para el proceso de descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino fueron clasificación y asociación, teniendo en cuenta que con los valores del atributo *vivomuerto* se puede construir un modelo de clasificación que determine las características de las pacientes que viven y de las que mueren y además se pueden identificar relaciones no explícitas entre los atributos del conjunto de datos pertenecientes a las mujeres que sobreviven por medio de asociaciones.

La tarea de clasificación de datos permite obtener resultados a partir de un proceso de aprendizaje supervisado. La clasificación es el proceso por medio del cual se encuentra propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases, de acuerdo al modelo de clasificación.[26] La técnica de clasificación utilizada fue árboles de decisión. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender.[27-29] El conocimiento se representa en forma de árboles, en los que cada nodo representa una condición sobre una variable. Después de una serie de evaluaciones en los nodos intermedios se alcanza una conclusión definitiva al llegar a las hojas o nodos finales, cada una de las cuales se asocia a una de las clases consideradas por el sistema.[30]

La tarea de asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos en un conjunto de datos determinado. En este problema, se da un conjunto de atributos y una colección de registros de una base de datos. La tarea es encontrar relaciones entre los atributos de esa base de datos para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de soporte y confianza.[31]

Las reglas de clasificación se obtuvieron con la herramienta Weka (*Waikato Environment for Knowledge Analysis*). Weka fue desarrollada en la Universidad de Waikato (Nueva Zelanda) bajo licencia GPL. Esta herramienta permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.[32] Weka es una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años.

El algoritmo en Weka utilizado para obtener las reglas de clasificación con árboles de decisión fue J48, el cual implementa al algoritmo C.45.[33] El algoritmo J48 se basa en la utilización del criterio ratio de ganancia (*gain ratio*). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido.[34] Este algoritmo se aplicó utilizando los repositorios de datos descritos en la tabla 3.

Para obtener las reglas de clasificación generales que inciden en la supervivencia de las mujeres con cáncer invasivo de cuello uterino, se utilizó el repositorio t507a22all y se escogió como clase, el atributo *vivomuerto* (334 vivos y 173 muertos), con una confianza C=0.7 y con un número de registros por nodo M=20. En la figura 1 se muestran los resultados obtenidos en Weka. De igual manera, se utilizaron los repositorios t507a15econ y t507a09clinico para determinar, respectivamente, los factores socioeconómicos

y clínicos que inciden en la supervivencia de las mujeres con cáncer invasivo de cuello uterino. Se escogió como clase, el atributo *vivomuerto* con una confianza C=0.7 y con un número de registros por nodo M=10. Las reglas de clasificación más relevantes se muestran en la sección de resultados.

Para obtener las reglas de asociación de las mujeres que sobrevivieron al cáncer invasivo de cuello uterino, se utilizó el repositorio t334a22vivos, con un soporte de 0.2 y una confianza de 0.8. En la figura 2 se muestran los parámetros utilizados en Weka. Las reglas más representativas de asociación se muestran en la sección de resultados.

Etapa de interpretación y evaluación de resultados

En esta etapa se interpretan los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto. Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Los resultados de esta etapa se analizan en la siguiente sección.

Figura 1. Resultados de la Clasificación en Weka

```
weka.classifiers.trees.J48 -C 0.7 -M 20
=== Classifier model (full training set) ===
J48 pruned tree
------------------
nmeses <= 37
|   cabezaflia <= 0
|   |   nivelsisben = 1: VIVO (50.0/20.0)
|   |   nivelsisben = 7: MUERTO (165.0/38.0)
|   |   nivelsisben = 2: VIVO (33.0/6.0)
|   |   nivelsisben = 3: VIVO (4.0/1.0)
|   |   nivelsisben = 4: VIVO (1.0)
|   cabezafamilia > 0: VIVO (42.0/4.0)
nmeses > 37: VIVO (212.0/15.0)

Number of Leaves  :        7
Size of the tree  :        10
```

RESULTADOS Y DISCUSIÓN

De acuerdo a la definición del diccionario de cáncer del *National Cáncer Institute at the National Institutes of Health*, "la supervivencia se enfoca en la salud y la vida de una persona después del tratamiento de cáncer hasta el final de la vida".[35]

Por esta razón, para descubrir los patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino con clasificación basada en arboles de decisión, el conjunto de datos t507a22all se clasificó tomando como clase el atributo *vivomuerto*. Este atributo determina si la paciente está viva o muerta hasta la ventana del estudio.

Figura 2. Parámetros utilizados en Asociación en Weka

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 0 -C
0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c 19
Relation:     QueryResult-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.NumericToNominal-R6-
weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-
R5-weka.filters.unsupervised.attribute.Remove-R4
Instances:    334
Attributes:   14
              region
              comuna
              edaddx
              nmeses_2007
              cabezafamilia
              regimen
              nivelsisben
              fuenteagua
              tipovivienda
              estrato
              escolaridad
              estadocivil
              ocupacion
              discapacidad
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.2 (67 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 16
```

Los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino descubiertos son:

Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta la ventana de observación del estudio (2007) es mayor que

37 entonces la mujer se consideró sobreviviente. El 42% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 63% de las mujeres vivas cumplen con este patrón.

Las mujeres que no cumplieron el periodo de sobrevivencia tuvieron el siguiente patrón: "no ser madre cabeza de familia, no estar clasificada en el SISBEN ni pertenecer a un régimen de salud". El 32.5% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasificó de esta manera y el 95.4% de las mujeres muertas cumple con este patrón.

Entre las reglas de asociación más representativas, que permiten identificar relaciones no explícitas entre los atributos del conjunto de datos t334a22vivos que almacena los casos de mujeres que sobreviven al cáncer invasivo de cuello uterino, con una confianza mínima del 80% y un soporte mínimo del 20%, están:

El 100% de las mujeres que poseen servicio público de acueducto y no tienen ninguna discapacidad, pertenece al régimen de salud subsidiado. El 43% de todas las mujeres que sobrevive tiene estas características.

El 100% de las mujeres que viven en casa o apartamento y su ocupación es hogar, pertenece al régimen de salud subsidiado. El 30.2% de todas las mujeres que sobrevive tiene estas características.

El 100% de las mujeres cuya escolaridad es primaria y no tienen ninguna discapacidad, pertenece al régimen de salud subsidiado. El 30% de todas las mujeres que sobrevive tiene estas características.

El 93% de las mujeres que no son cabeza de familia cuya escolaridad es primaria y su ocupación es hogar, pertenece al régimen de salud subsidiado. El 20.4% de todas las mujeres que sobrevive tiene estas características.

Teniendo en cuenta los resultados obtenidos en la etapa de minería de datos, en la cohorte 1998-2007 tomada para este estudio, el patrón general es que la supervivencia de las mujeres después de haber sido diagnosticadas con cáncer invasivo de cuello uterino sobrepasa los 37 meses. Comparando este resultado con los resultados obtenidos en un estudio anterior de supervivencia, aplicando la técnica Kaplan-Meier en la cohorte 1998-2002 con 203 casos, la mediana de supervivencia de las mujeres con cáncer invasivo de cuello uterino fue de 36.8 meses,[36] valor que se asemeja al patrón descubierto con técnicas de minería de datos. Teniendo en cuenta el número total de casos de cáncer de cuello uterino analizados en este estudio, el 65.9% sobrevive y de estos el 63.5% sobrepasa el umbral de 37 meses de diagnosticado el cáncer.

Entre los factores socioeconómicos asociados a la supervivencia de las mujeres con cáncer invasivo de cuello uterino están: tener servicio público de acueducto, vivir en casa propia, ser cabeza de familia, no tener ninguna discapacidad, pertenecer a un sistema de salud subsidiado, una escolaridad de primaria y ocupación hogar. En el estudio mencionado anteriormente,[37] se estableció que las mujeres diagnosticadas con cáncer de cuello uterino tenían las siguientes características: El 82% de ellas fueron procedentes de zona urbana, el 70% conviven con una pareja, el 65% tenían baja escolaridad, el 78% tenían aseguramiento en salud. A este conocimiento previo se le adiciona los factores socioeconómicos descubiertos con minería de datos.

Entre los factores clínicos que se consideran patrón de supervivencia de más de 37 meses de las mujeres con cáncer invasivo de cuello uterino están: el método de diagnóstico a través de Histología y con diagnóstico de Tumor Primario;el tratamiento a través de cirugía y la fuente de diagnóstico, un hospital o clínica.

Los resultados del estudio muestran que existe un patrón asociado a condiciones socioeconómicas de las mujeres con cáncer invasivo de cuello uterino del Municipio de Pasto Colombia, reafirmando los hallazgos realizados en otros estudios en los cuales este tipo de cáncer se asocia con la clase social.

En el Municipio de Pasto, como se muestra en los resultados, el mayor porcentaje de mujeres con cáncer invasivo de cuello uterino tienen como régimen de aseguramiento el denominado Régimen Subsidiado que es el mecanismo mediante el cual la población más pobre del país sin capacidad de pago tiene acceso a los servicios de salud a través de un subsidio que ofrece el Estado. Este régimen es un indicador de la situación social de estas mujeres quienes al pertenecer a estratos sociales bajos tienen mayor riesgo de enfermar y morir por cáncer de cuello uterino.[38]

CONCLUSIONES Y RECOMENDACIONES

Los resultados obtenidos a través de las técnicas clasificación por árboles de decisión y asociación indican que estas son capaces de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en las bases de datos, para este caso, en el Registro Poblacional de Cáncer del Municipio de Pasto.

Una de las grandes dificultades que se presenta en esta clase de estudios es la mala calidad de los datos que muchas veces, después del proceso de limpieza, hace que se descarten ciertas variables por la imposibilidad de obtener sus valores y que de alguna manera influye en los resultados de la minería de datos.

Se ha obtenido un patrón general de supervivencia basado en el número de meses que transcurren desde el momento del diagnóstico del cáncer hasta la fecha final del periodo de observación de este estudio. Se han determinado factores socioeconómicos y clínicos asociados a la supervivencia de las mujeres con cáncer invasivo de cuello uterino. La evaluación, análisis y utilidad de estos patrones permitirá soportar la toma de decisiones eficaces de los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de políticas públicas y programas de protección a las mujeres con esta enfermedad.

Como trabajos futuros están el aplicar otras técnicas de minería de datos como *clustering* con el fin de determinar similitudes entre los factores socioeconómicos y clínicos de las mujeres que sobreviven y de las que fallecieron, utilizando los repositorios de datos descritos en la tabla 3.

Agradecimientos

Al Sistema de Investigaciones de la Universidad de Nariño por financiar esta investigación.

REFERENCIAS

1. Hernández E, Lorente R. Minería de datos aplicada a la detección de cáncer de mama. Universidad Carlos III, Madrid, España. [Consultado en octubre de 2011]. Disponible en: URL:http://www.it.uc3m.es/jvillena/irc/practicas/08-9/14.pdf.
2. Mora R. El papel de la minería de datos en la detección y diagnóstico de cáncer. Universidad de Salamanca. Salamanca, España. [Consultado en octubre de 2011]. Disponible en: URL: http://sistemaminergescon.blogspot.com/.
3. Febles JP, González A. Aplicación de la minería de datos en la bioinformática. En: SciELO ACIMED Vol 10(2). ISSN 1024-9435. Ciudad de La Habana, Cuba; 2002.
4. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. Discovering data mining from concept to implementation, Prentice Hall; 1997.

5. Timarán R. Una mirada al descubrimiento de conocimiento en bases de datos. En: revista Ventana Informática No 20, Centro de Investigaciones, Desarrollo e Innovación, Facultad de Ingeniería, Universidad de Manizales, ISSN 0123-9678:39-58. Armenia, Colombia; 2009.
6. Hernández J, Ramírez MJ, Ferri C. Introducción a la minería de datos. Editorial Pearson Prentice Hall, ISBN 84-205-4091-9. Madrid, España; 2005.
7. Reparaz D, Merlino H, Rancan C, Rodríguez D, Britos P, García-Martínez R. Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basada en minería de datos. En: Memorias WIIC; 2008.
8. Hurtado V, Silva V, Salas R, Lobos A. Técnicas de minería de datos en el estudio epidemiológico del cáncer mamario. En: Memorias XXVII Jornadas Chilenas de Salud Pública, Universidad de Chile. Santiago de Chile; 2008.
9. Hernández E, Lorente R. Minería de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III, Madrid, España. [Consultado en octubre de 2011]. Disponible en: URL:http://www.it.uc3m.es/jvillena/irc/practicas/08-9/14.pdf.
10. Hurtado V, Silva V, Salas R, Lobos A. Técnicas de minería de datos en el estudio epidemiológico del cáncer mamario. En: Memorias XXVII Jornadas Chilenas de Salud Pública, Universidad de Chile. Santiago de Chile; 2008.
11. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. The George Washington University. Washington D.C; 2005.
12. López RA. Descubrir el comportamiento de los factores de riesgo de cáncer de cuello uterino utilizando minería de datos. En: Tesis de Maestría, Biblioteca Digital Repositorio Institucional, Universidad Nacional de Colombia. Bogotá D.C., Colombia; 2001.
13. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: VLDB Conference. Santiago de Chile, Chile; 1994.
14. Chen M, Han J, Yu P. Data mining: An overview from database perspective. In: Journal IEEE Transactions on Knowledge Data Engineering. ISSN: 1041-4347. New Jersey, USA; Vol.8(6); 1996: 866-883.
15. Piatetsky-Shapiro G, Brachman R, Khabaza T. An Overview of issues in developing industrial data mining and knowledge discovery applications. In: The Second International Conference on Knowledge Discovery and Data Mining KDD, AAAI Press. Portland, Oregon, USA; 1996:89-95.
16. Han J, Kamber M. Data mining concepts and techniques. Morgan Kaufmann Publishers. San Francisco, USA; 2001.
17. Stonebraker M, Rowe LA. The design of postgres, In: Proceedings of the ACM-SIGMOD Conference. Washington D.C., USA; 1986.
18. Momjian B. PostgreSQL: Introduction and concepts. Addison-Wesley, ISBN: 0-2001-70331-9. New York, USA; 200:455.
19. Wang M, Iyer B, Scott VJ. Scalable mining for classification rules in relational databases. In: International Database Engineering and Application Symposium - IDEAS. Cardiff, Wales, U.K.; 1998: 58-67
20. Witten I, Frank E. Data mining: Practical machine learning tools and techniques with Java Implementations. Morgan Kaufmann Publishers. ISBN: 1-55860-552-5. San Francisco, CA, USA; 2000:365.
21. Ng R, Han J. Efficient and effective clustering method for spatial data mining. In: The 20th International Conference on Very Large Data Bases VLDB 94. Santiago de Chile, Chile; 1994:144-155.
22. Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: ACM SIGMOD International Conference on Management of Data. Montreal, Canada, 1996:103-114.
23. Agrawal R, Srikant R. Mining sequential patterns. In: The 11th International Conference on Data Engineering ICDE. Taipei, Taiwan; 1995: 3-14.
24. Agrawal R, Ghosh S, Imielinski T, Iyer B, Swami A. An interval classifier for database mining applications, In: Proceedings VLDB Conference. Vancouver, Canada; 1992.

25. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: VLDB Conference. Santiago de Chile, Chile; 1994.
26. Hernández J, Ramírez MJ, Ferri C. Introducción a la minería de datos. Editorial Pearson Prentice Hall, ISBN 84-205-4091-9. Madrid, España; 2005.
27. Wang M, Iyer B, Scott VJ. Scalable mining for classification rules in relational databases. In: International Database Engineering and Application Symposium – IDEAS. Cardiff, Wales, U.K.; 1998:58-67
28. Sattler K, Dunemann O. SQL Database primitives for decision tree classifiers. In: CIKM. Atlanta, Georgia, USA; 2001.
29. Timarán R, Millán M. New algebraic operators and SQL primitives for mining classification rules. In: proceedings of The Five IASTED International Conference on Computational Intelligence (CI 2006), International Association of Science and Technology for Development, ISBN No 0-88986-603-1. San Francisco, USA; 2006.
30. Reparaz D, Merlino H, Rancan C, Rodríguez D, Britos P, Garcia-Martinez R. Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basada en minería de datos. En: Memorias WIIC; 2008.
31. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: VLDB Conference. Santiago de Chile, Chile; 1994.
32. Garcia D. Manual de Weka. [Consultado en abril de 2011]. Disponible en: URL: http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf.
33. Quinlan JR. C4.5: Programs for machine learning. Morgan Kaufmann Publishers. ISBN 1-55860-238-0. San Francisco, CA, USA; 1993:229.
34. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. The George Washington University. Washington D.C; 2005.
35. Instituto Nacional del Cáncer de los Institutos Nacionales de la Salud de EE.UU. [Consultado en abril de 2011]. Disponible en: URL: http://www.cancer.gov/diccionario?cdrid=445089.
36. Yépez MC, Cerón E, Hidalgo A, Cerón C. Supervivencia de mujeres con cáncer de cuello uterino, Municipio de Pasto. En: Revista Universidad y Salud, Año 11 Vol. 2 No. 14 Pasto, Colombia; 2011: 7 - 18.
37. Ibid
38. Arias S. Inequidad y cáncer: una revisión conceptual. En: revista Facultad Nacional de Salud Pública Vol. 27(3). Universidad de Antioquia. Medellín, Colombia; 2009.