**Cómo citar este artículo:** Timarán-Pereira, R & Yépez-Chamorro, M.C. (2016). Caracterización de la supervivencia de mujeres con cáncer invasivo de cuello uterino usando minería de datos. Rev.investig.desarro.innov, 7(1), 127-139.
doi: 10.19053/20278306.v7.n1.2016.4315

# Caracterización de la supervivencia de mujeres con cáncer invasivo de cuello uterino usando minería de datos

# Characterizing the survival of women with invasive cervical cancer by using data mining

**Ricardo Timarán-Pereira**[1]
**María Clara Yépez-Chamorro**[2]

Recibido: marzo 11 de 2016
Aceptado: junio 28 de 2016

**Resumen**

En este artículo se presenta uno de los resultados del proyecto de investigación denominado: Detección de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino con técnicas de minería de datos, utilizando como fuente principal la información almacenada en la base de datos del Registro Poblacional de Cáncer del Municipio de Pasto (Colombia). Aplicando la metodología para proyectos de minería de datos CRISP-DM, se construyó, limpió y transformó un repositorio de datos con la información de las mujeres que fueron diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2002, con una ventana de observación hasta el 2007. Se detectaron los principales factores socioeconómicos y clínicos asociados con la supervivencia de este grupo poblacional, utilizando las tareas de minería de datos: clasificación, asociación y agrupación. El patrón principal descubierto es aquel que caracteriza a una mujer con cáncer invasivo de cuello uterino como sobreviviente, si sobrepasa los 52 meses después del momento del diagnóstico del cáncer.

**Palabras clave:** cáncer de cuello uterino, CRISP-DM, patrones de supervivencia, minería de datos.

**Abstract**

In this paper, one of the results of the research project entitled: Detection of survival patterns in diagnosed women with invasive cervical cancer with data mining techniques, using as the main source the information stored in the database of Cancer Registry of the Municipality of Pasto (Colombia) is presented here. Applying the CRISP-DM methodology, a data repository with information from diagnosed women with invasive cervical cancer during the period between 1998 and 2002 with an observation window until 2007, was built, cleaned, and transformed. The main socioeconomic and clinical factors related to survival of this population group, using classification, association, and clustering tasks were detected. The principal pattern discovered was that if a woman exceeds 52 months after the time of diagnosis of invasive cervical cancer, she will be characterized as a cancer survivor.

**Keywords:** cervical cancer, CRISP-DM, survival patterns, data mining.

[1] **Ingeniero de Sistemas, Doctor en Ingeniería énfasis Ciencias de la Computación, Universidad de Nariño, Colombia. E-mail: ritimar@udenar.edu.co**
[2] **Licenciada en Enfermería, Magister en Ciencias Biomédicas, Universidad de Nariño, Colombia. E-mail: mcych@udenar.edu.co**

## 1. Introducción

Según los reportes de la Agencia Internacional de Investigación en Cáncer (IARC), a nivel mundial, el cáncer de cuello uterino es el tercer tipo de cáncer más común en las mujeres, y el séptimo entre todos los tipos de cáncer; la tasa de incidencia ajustada por edad para el quinquenio 2004-2008 se calculó en 15,2 por 100.000 mujeres y la de mortalidad de 7,8. Para el año 2008, se estimaron 530.000 nuevos casos y 275.000 muertes, de las cuales el 88% ocurrieron en África, Asia, América Latina y el Caribe. Más del 85% de la carga de la enfermedad a nivel mundial se produce en los países en desarrollo, donde representa el 13% de todos los cánceres femeninos (Ferlay et al., 2010).

En América Latina el cáncer de cuello uterino disminuye la expectativa de vida de las mujeres más que el SIDA, la tuberculosis o las enfermedades asociadas al embarazo y al parto. Las muertes por el cáncer de cuello uterino son mucho más frecuentes en aquellas mujeres que por desconocimiento o por falta de acceso a los servicios de salud no se someten a los estudios periódicos para el diagnóstico precoz de las lesiones que conducen a este cáncer, lo que explica por qué el 80% de las muertes que provoca ocurren en los países pobres donde los programas de detección no están debidamente implementados o no son efectivos (Ciencia Hoy, 2006).

En el análisis realizado por el Ministerio de Salud y el Instituto Nacional de Cancerología, como base para el diseño del plan de control del cáncer 2012-2021, en Colombia se ubica geográficamente el riesgo de mortalidad por cáncer de cuello uterino en las habitantes de departamentos alejados, zonas de frontera y riberas de ríos, mientras que socialmente el riesgo es para aquellas mujeres pobres pertenecientes a régimen subsidiado en aseguramiento en salud. En el país, la incidencia estimada por edad para el periodo 2002-2006 de cáncer de cuello uterino fue de 28,2 por 100.000 habitantes y la tasa de mortalidad observada por edad para el mismo periodo fue de 10,0 por 100.000 habitantes (Pardo & Cendales, 2010). La tasa estandarizada por edad de mortalidad en cáncer de cuello uterino durante 2010 fue de 7,9 por 100.000 mujeres; la meta propuesta de reducción en el Plan Nacional de Salud Pública para ese mismo año fue una tasa de mortalidad de 7,6 por 100.000 y de 4,5 por 100.000 en 2019.

Para el Departamento de Nariño, el Instituto Nacional de Cancerología reporta en el período 2002-2006 una tasa de incidencia anual (TAE) de cáncer de cuello uterino de 26,4 y de mortalidad de 9,8 / 100.000 habitantes (Pardo & Cendales, 2010).

Los anteriores estudios se basan en información procesada mediante un análisis estadístico básico, donde se consideran fundamentalmente variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que se pueden descubrir utilizando un tratamiento de los datos más complejo, que es posible con la minería de datos.

Mientras que la Estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles, la minería de datos descubre patrones a partir de los datos que mediante su interpretación propone, en el caso del cáncer invasivo de cuello uterino, patrones de supervivencia no previstos desde la Estadística.

En este artículo se presenta el proceso de descubrimiento de patrones de supervivencia en casos de mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2003 y observados hasta el año 2007, a partir de los datos almacenados en el Registro Poblacional de Cáncer del municipio de Pasto, aplicando la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) y las tareas de minería de datos clasificación, asociación y agrupación.

El resto del artículo se organiza en secciones, así: en la siguiente sección se describe y desarrolla la metodología para proyectos de minería de datos denominada CRISP-DM, aplicada a la detección de

patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. En la sección tres se presentan los resultados obtenidos y su discusión y finalmente, en la última sección se presentan las conclusiones y trabajo futuro.

## 2. Materiales y métodos

La investigación se desarrolló bajo el enfoque cuantitativo, de tipo descriptivo, aplicando un diseño no experimental. Se utilizó la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*), por ser uno de los modelos principalmente utilizados en los ambientes académico e industrial y la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos (Hernández, Ramírez, & Ferri, 2005). CRISP-DM contempla seis fases: análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación.

### 2.1 Análisis del problema

En esta fase se requiere comprender con exactitud el problema al cual se le va a dar solución utilizando la minería de datos. Esto permitirá recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Gallardo, 2009).

Algunos estudios (Asport & Rivero, 2004; Castro, Vera, & Posso, 2006; Ferlay, Bray, Pisani, & Parkin 2004), muestran que cuando el cáncer de cuello uterino es detectado y atendido en etapa temprana, por lo general se puede curar. El índice de supervivencia de cinco años para el cáncer cervical pre invasivo es del 100 por ciento y para el cáncer invasivo en etapa temprana es del 91%. El índice de supervivencia de cinco años de los cánceres cervicales en todas las etapas combinadas baja al 70% (Ferlay et al., 2004).

Un estudio realizado por el Registro Poblacional de Cáncer de Cali (Colombia), muestra una probabilidad de supervivencia para 5 años del 45%, dato similar encontrado en un estudio para el Municipio de Pasto en la cohorte 1998-2002 (Yépez, Cerón, Hidalgo-Troya, & Cerón, 2011). Otro estudio reporta que el pronóstico del cáncer de cuello uterino es dependiente de las características socioeconómicas y demográficas de la paciente, del estadío clínico al momento del diagnóstico, del esquema tratamiento y del tiempo transcurrido entre el diagnóstico y el tratamiento y de su continuidad, variables que inciden en la supervivencia (Merle, 2004).

El problema de la supervivencia de las mujeres que han sido diagnosticadas con cáncer invasivo de cuello uterino en el municipio de Pasto, se convirtió en un problema a resolver con minería de datos.

### 2.2 Análisis de los datos

En esta fase se realiza la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis (Gallardo, 2009).

Se definieron las fuentes internas y externas de datos con el fin de construir posteriormente un conjunto de datos unificado que sirva de base para aplicar las técnicas de minería de datos y obtener los patrones de supervivencia de mujeres con cáncer invasivo de cuello uterino. Como fuente interna, se seleccionó la base de datos del Registro Poblacional de Cáncer del municipio de Pasto, donde se encuentran almacenados los datos de 17.350 casos de diferentes tipos de cáncer desde el año 1998 hasta el 2007, correspondiente al periodo de observación de este estudio. Como fuentes externas principales se seleccionaron las bases de datos del Registro Individual de Prestación de Servicios de Salud (RIPS) y la del Sistema de Identificación de Beneficiarios Potenciales de Programas Sociales SISBEN del municipio de Pasto.

Otras fuentes externas que se seleccionaron para complementar u obtener datos fueron las bases

de datos de 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.

De los 17.350 casos de cáncer se seleccionaron inicialmente 3.151 registros correspondientes a mujeres con cáncer de cuello uterino. De este conjunto, se seleccionaron 507 registros correspondientes a las mujeres con cáncer invasivo de cuello uterino, y finalmente, de los 507 casos, se escogieron únicamente 235 que pertenecen a mujeres diagnosticadas en el periodo comprendido entre 1998 y 2002 y cuyo seguimiento u observación se les hizo hasta 2007, con el fin de determinar, en un periodo de cinco años, su supervivencia.

Utilizando el método de ranqueo de atributos basado en la ganancia de información, se seleccionaron de 48 atributos, inicialmente los 36 atributos más representativos. Como resultado de esta etapa se obtuvo el repositorio de datos T235A36 con 507 registros y 36 atributos, que sirvió de base para las subsecuentes fases.

2.3 Preparación de los datos

Una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se aplicarán. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Gallardo, 2009).

Por medio de consultas SQL ad-hoc e histogramas, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos del conjunto de datos T235A36. Como resultado de este proceso, los valores nulos de nueve atributos fueron actualizados con los valores encontrados en fuentes externas. Adicionalmente, para facilitar la extracción de patrones, se discretizaron los valores numéricos a valores nominales, se crearon doce nuevos atributos en reemplazo de otros y se eliminaron la llave primaria, los atributos que se utilizaron para crear nuevos atributos, los atributos no relevantes y los atributos con un alto porcentaje de nulos por la imposibilidad de obtener sus valores. Como resultado de esta fase se obtuvo el repositorio limpio y transformado con 22 atributos y 235 registros denominado T235A22, listo para aplicarle las técnicas de minería de datos. La descripción de los 22 atributos que forman este repositorio se muestra en la tabla 1.

| Atributos | Descripción |
|---|---|
| Región | Región de nacimiento del paciente |
| 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. |
| edaddx | Edad del paciente en el momento del diagnóstico. |
| estadocivil | Estado civil del paciente en el momento del diagnóstico |
| ocupacion | Ocupación del paciente en el momento del diagnóstico |
| escolaridad | Escolaridad del paciente en el momento del diagnóstico |
| regimen | Régimen al cual pertenece el paciente en el momento del diagnóstico. |
| nivelsisben | El nivel de clasificación en el SISBEN de acuerdo al puntaje obtenido |
| cabezafamilia | Determina si el paciente es cabeza de familia o no. |
| tipovivienda | Tipo de vivienda que habita el paciente |
| fuenteagua | Si la residencia cuenta con servicio de agua |
| discapacidad | Si el paciente tiene una discapacidad o no |

| | |
|---|---|
| fuente | Organización donde se diagnosticó el cáncer de cuello uterino |
| metododx | Método utilizado para el diagnóstico del cáncer |
| morfologia | Morfología del tumor. |
| locesp | Localización específica del tumor |
| radio | Existencia o no de tratamiento de radioterapia al paciente. |
| cirugia | Determina si el paciente ha tenido como tratamiento cirugía |
| biopsia | Determina si al paciente se le realizó o no una biopsia |
| nmeses | Número de meses de vida del paciente desde el momento del diagnóstico |
| vivomuerto | Determina si el paciente está vivo o muerto |

**Tabla 1.** Descripción de los atributos del repositorio T507A22.

## 2.4 Modelado

En esta fase se seleccionan las tareas más apropiadas para el proyecto de minería de datos. Se seleccionaron las tareas de minería de datos clasificación, asociación y agrupamiento para descubrir conocimiento sobre la supervivencia de mujeres con cáncer invasivo de cuello uterino a partir de los datos del repositorio T235A22.

i) *Tarea de clasificación*

La clasificación es el proceso por medio del cual se encuentran 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 (Hernández, et al., 2005). Tomando como clase los valores del atributo *vivomuerto* del conjunto de datos T235A22, se construyó un modelo de clasificación que determinó las características de las pacientes que sobrevivieron al cáncer y las que no. La técnica de clasificación utilizada fue árboles de decisión. Esta técnica, es probablemente la más utilizada y popular por su simplicidad y facilidad para entender (Han & Kamber, 2001; Sattler & Dunemann, 2001; Timarán & Millán, 2006). La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y solo una hoja, asignando una única clase a la predicción (Hernández & Lorente, 2009).

Las reglas de clasificación se obtuvieron utilizando el algoritmo J48 que implementa el conocido algoritmo de árboles de decisión C4.5 (Quinlan, 1993).

Para la poda del árbol se tuvo en cuenta el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido (García & Álvarez, 2010) y el mínimo número de instancias o registros por nodo del árbol M (Witten & Frank, 2000).

Para evaluar la calidad del modelo y su validez, dividiendo el repositorio de datos en dos conjuntos: entrenamiento y prueba, se escogió el método validación cruzada con n pliegues (*n-fold cross validation*) (Hernández, et al., 2005). En este estudio se utilizó n=10 particiones, que es el valor que comúnmente se emplea y que se ha probado que da buenos resultados (Hernández, et al., 2005).

Por otra parte, se estimó el coste del clasificador para el conjunto de datos con la matriz de confusión. La matriz de confusión (*Confusion Matrix*) representa de forma detallada el número de instancias que son predichas por clase (Fernández, 2009).

Teniendo en cuenta los parámetros de evaluación anteriores se procedió a construir diferentes árboles de decisión con el algoritmo J48 con el fin de obtener el mejor. Para tal efecto, se varió el factor confianza C de 0,1 hasta 0,5 incrementando en 0,1 y el número de registros por nodo M de 2 hasta 20 con un incremento de 2. De acuerdo con los resultados obtenidos, el árbol construido con los parámetros M=2 y C=0,2 fue el mejor con un porcentaje de 93,2 % de instancias correctamente clasificadas correspondiente a 219 instancias de 235. En la figura 1 se presentan los mejores resultados de esta prueba con los parámetros M=2 C=0,2.

```
J48 pruned tree
------------------

nmeses <= 52
| locesp = Cuello uterino SAI                  : MUERTO (110.63/14.0)
| locesp = EndocÃ©rvix
| | nmeses <= 16: VIVO (3.32/0.32)
| | nmeses > 16: MUERTO (2.05)
| locesp = ExocÃ©rvix                         : MUERTO (0.0)
nmeses > 52
| metododx = Certifi. Defuncion solo    : MUERTO (4.0)
| metododx = Citologia o Hematologia    : VIVO (1.0)
| metododx = Histologia Tumor Primario   : VIVO (113.0/2.0)
| metododx = Imagenologia         : VIVO (0.0)
| metododx = Solamente Clinico       : VIVO (1.0)

Number of Leaves :              9

Size of the tree :      13


=== Summary ===

Correctly Classified Instances     219       93.1915 %
Incorrectly Classified Instances    16       6.8085 %
Kappa statistic              0.8638
Mean absolute error          0.1233
Root mean squared error        0.2467
Relative absolute error        24.9358 %
Root relative squared error       49.6151 %
Coverage of cases (0.95 level)     99.1489 %
Mean rel. region size (0.95 level)   74.2553 %
Total Number of Instances       235

=== Confusion Matrix ===
```

**Figura 1.** Mejores resultados generados por el algoritmo J48 con el conjunto de datos T235A22.

ii) *Tarea de asociación*

La tarea de asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos en un conjunto de datos determinado (Agrawal & Srikant, 1994). Para evaluar una regla de asociación se utiliza el soporte y la confianza, dos métricas que permiten conocer la calidad de la regla. El soporte o cobertura de una regla se define como el número de instancias en las que la regla se puede aplicar. La confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar (Hernández et al., 2005).

Tomando el conjunto de datos de las mujeres que sobrevivieron al cáncer de cuello uterino, se extrajeron reglas que determinaron ciertas características que aparecen juntas en este tipo de mujeres. Para obtener las reglas de asociación se utilizó el algoritmo Apriori (Agrawal & Srikant, 1994). Se fijó como mínima confianza el 80% (0,8), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0,1, un incremento de 0,5 y un número de reglas a generar de 25. También se filtraron las reglas para obtener solo aquellas, donde el atributo *vivomuerto* se encuentre como consecuente de la regla. Las mejores reglas resultantes fueron aquellas con un soporte mínimo del 10% (0,1). Las mejores 25 reglas generadas con una confianza del 100% se muestran en la figura 2.

iii) *Tarea de agrupación*

En esta tarea se trata de encontrar grupos similares entre un conjunto de datos basado en el concepto de distancia (Han & Kamber, 2001; Hernández, et al., 2005). Los *clusters* tienen una alta homogeneidad interna (dentro del *cluster*) y una alta heterogeneidad externa (entre *cluster*) (Chen, Han, & Yu, 1996). Tomando el conjunto de datos de todos los registros de mujeres que padecen cáncer invasivo de cuello uterino, se obtuvo grupos homogéneos de mujeres con esta enfermedad.

```
## Best rules found:
##
## 1. cabezafamilia=0 nmeses=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 2. edaddx=3 metododx=Histologia Tumor Primario regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 3. edaddx=3 biopsia=si regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 4. estadocivil=CASADO discapacidad=NINGUNA fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 5. estadocivil=CASADO metododx=Histologia Tumor Primario fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 6. estadocivil=CASADO biopsia=si fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 7. discapacidad=NINGUNA cabezafamilia=0 nmeses_2007=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 8. metododx=Histologia Tumor Primario regimen=SUBSIDIADO region=PASTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 9. metododx=Histologia Tumor Primario cabezafamilia=0 nmeses=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 10. fuente=HOSPITALES locesp=Cuello uterino SAI nmeses=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 11. morfologia=Squamous cell carcinoma, large cell, nonkeratinizing, NOS cirugia=no
regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 12. radio=si cabezafamilia=0 nmeses=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 13. biopsia=si cabezafamilia=0 nmeses=(84,96] 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 14. tipovivienda=CASA O APARTAMENTO fuenteagua=ACUEDUCTO cabezafamilia=0 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 15. edaddx=3 metododx=Histologia Tumor Primario radio=si regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 16. edaddx=3 radio=si biopsia=si regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 17. escolaridad=PRIMARIA metododx=Histologia Tumor Primario locesp=Cuello uterino SAI
nivelsisben=1 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 18. escolaridad=PRIMARIA metododx=Histologia Tumor Primario biopsia=si nivelsisben=1 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 19. escolaridad=PRIMARIA metododx=Histologia Tumor Primario nivelsisben=1 regimen=SUBSIDIADO
24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 20. escolaridad=PRIMARIA locesp=Cuello uterino SAI tipovivienda=CASA O APARTAMENTO
regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 21. escolaridad=PRIMARIA biopsia=si nivelsisben=1 regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 22. estadocivil=CASADO discapacidad=NINGUNA radio=si fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 23. estadocivil=CASADO discapacidad=NINGUNA fuenteagua=ACUEDUCTO regimen=SUBSIDIADO
24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 24. estadocivil=CASADO metododx=Histologia Tumor Primario radio=si fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
## 25. estadocivil=CASADO metododx=Histologia Tumor Primario fuenteagua=ACUEDUCTO
regimen=SUBSIDIADO 24
## ==> vivo_muerto=VIVO 24 conf:(1)
```

**Figura 2.** Mejores reglas generadas con el algoritmo Apriori con el conjunto de datos T235A22.

Para la tarea de agrupación se utilizó la técnica particional con el algoritmo K-means (Han & Kamber, 2001), en el cual se configura el número de grupos (*NumClusters*) a formar y la semilla (*seed*), que se utiliza en la generación de un número aleatorio, el cual es usado para hacer la asignación inicial de instancias a los grupos. Para evaluar los resultados del agrupamiento, se utilizó el propio conjunto de entrenamiento (*Use training set*), que indica que porcentaje de instancias se van a cada grupo. Se configuró K-means para encontrar K=2, K= 4 y K=6 *clusters* con una semilla por defecto de 10. Para evaluar los resultados del agrupamiento se utilizó el propio conjunto de entrenamiento (*Use training set*) igual al 66%. Con el parámetro K=2 se encontró los dos *clusters* más homogéneos: en el *cluster* 0 se agrupó a 112 casos de mujeres que sobrevivieron y en el *cluster* 1 se agrupó a 123 mujeres que no sobrevivieron. De esta manera, se pudo encontrar cuales son las similitudes particulares de cada uno de estos grupos, en un proceso no supervisado, donde no se especificó la clase *vivomuerto,* como se hizo en la tarea de clasificación. Los resultados se muestran en la tabla 2.

| Atributo | Full data (235) | Cluster 0 (112) | Cluster 1 (123) |
|---|---|---|---|
| edaddx | 3 | 4 | 3 |
| escolaridad | PRIMARIA | PRIMARIA | PRIMARIA |
| estadocivil | CASADO | CASADO | CASADO |
| ocupacion | HOGAR | HOGAR | HOGAR |
| discapacidad | NINGUNA | NINGUNA | NINGUNA |
| metododx | Histologia_Tumor_Primario | Histologia_Tumor_Primario | Histologia_Tumor_Primario |
| fuente | HOSPITALES | HOSPITALES | HOSPITALES |
| morfologia | Squamous_cell_carcinoma, large_cell, nonkeratinizing,NOS | Squamous_cell_carcinoma,large cell,nonkeratinizing,NOS | Squamous_cell_carcinoma,NOS |
| locesp | Cuello_uterino_SAI | Cuello_uterino_SAI | Cuello_uterino_SAI |
| cirugia | no | no | no |
| radio | si | si | si |
| biopsia | si | si | si |
| estrato | 2 | 2 | 2 |
| tipoovivienda | CASA O APARTAMENTO | CASA_O_ APARTAMENTO | CASA_ O_ APARTAMENTO |
| fuenteagua | ACUEDUCTO | ACUEDUCTO | ACUEDUCTO |
| nivelsisben | 7 | 1 | 7 |
| regimen | SUBSIDIADO | SUBSIDIADO | VINCULADO |
| cabezafamilia | 0 | 0 | 0 |
| comuna | 5 | 5 | 5 |
| region | PASTO | OCCIDENTAL ANDINA | PASTO |
| nmeses | [0,12] | [60,72] | [0,12] |
| vivo_muerto | VIVO | VIVO | MUERTO |

**Tabla 2.** *Clusters* resultantes con K-means con el repositorio T235A22ALL.

2.5 Evaluación

En esta fase 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. La interpretación de los patrones descubiertos y su discusión se hará en la sección 3.

2.6 Explotación o implementación

En esta fase, el conocimiento obtenido se transforma en acciones dentro del proceso de negocio. Se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización y difundir informes sobre el conocimiento extraído. El conocimiento descubierto se incorporará al existente y se integrará a los procesos de toma de decisiones 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.

## 3. Resultados y discusión

En esta sección se realiza una evaluación e interpretación de los resultados obtenidos con los datos de las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo comprendido entre los años 1998 y 2002, observados hasta el 2007 y almacenados en el conjunto de datos T235A22.

3.1 Análisis de patrones de clasificación

Analizando los resultados obtenidos de la prueba de clasificación realizada con el conjunto de datos T235A22, donde se almacenan los datos de 235 mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998 al 2002 y cuyo objetivo fue descubrir los factores socioeconómicos y clínicos que inciden en la supervivencia de estas pacientes, se puede observar que el árbol de decisión resultante (ver figura 1), clasifica 219 instancias correctamente, que corresponde a un porcentaje de precisión del 93,1%, y 16 instancias incorrectamente, correspondiente a un porcentaje de error del 6,8%.

Teniendo en cuenta la distribución de los valores del atributo clase *vivomuerto* del repositorio T235A22 que es de 130 vivos y 105 muertos y evaluando el modelo con la matriz de confusión, este clasifica correctamente a 103 casos de las pacientes que murieron por cáncer y 116 casos de las sobrevivientes. Además, clasifica incorrectamente a 2 casos de las que no sobreviven y 14 casos de las que sobreviven. Esto significa que el modelo clasifica correctamente al 98,1% de las pacientes muertas y el 89,2% de las pacientes que sobreviven.

Por otra parte, el estadístico Kappa, que mide la coincidencia de la predicción con la clase real de este modelo, es de 0,8638, que se considera excelente, pues 1,0 significa que ha habido coincidencia absoluta.

Los porcentajes de instancias correctamente clasificadas, presentados tanto en el árbol como en la matriz de confusión, indican que el modelo tiene una precisión alta y es por lo tanto confiable y eficiente, para clasificar nuevos casos.

De acuerdo con este modelo, ver figura 1, los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino descubiertos son:

*Si el número de meses de vida de las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002 y observadas hasta 2007, contados a partir de la fecha de diagnóstico, es mayor que 52 y el método de diagnóstico fue una Histología de Tumor Primario, entonces la mujer se consideró sobreviviente. El 47,2% de los 235 casos de mujeres consideradas en este estudio, se clasifican de esta manera y el 85,4% de las 130 mujeres sobrevivientes al cáncer cumplen con este patrón.*

*Si el número de meses de vida de las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002 y observadas hasta 2007, contados a partir de la fecha de diagnóstico, es menor o igual a 37 y la localización específica del cáncer fue "Cuello uterino SAI" entonces la mujer no se consideró sobreviviente. El 40,9 % de los 235 casos de mujeres consideradas en este estudio, se clasifican de esta manera y el 91,4% de las 105 mujeres que no sobreviven cumplen con este patrón.*

3.2 Análisis de patrones de Asociación

Se utilizó el conjunto de datos T235A22ALL para obtener reglas de asociación que relacionen los factores socioeconómicos y clínicos de 235 mujeres con la supervivencia de estas. Las pacientes fueron diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998 al 2002 y observadas hasta finales del año 2007.

Las 25 reglas de asociación generadas y que se muestran en la figura 2, tienen una confianza del 100% y un soporte mínimo del 10%, lo que las convierte en reglas fuertes (*strong rules*) y por lo tanto interesantes, significativas y con una alta precisión. Entre las reglas de asociación de tamaño más representativas están:

*Regla 2*. El 100% de las mujeres que sobreviven tienen una edad entre 15 y 28 años, fueron diagnosticadas mediante una Histología de Tumor Primario y pertenecen al régimen de salud subsidiado. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

*Regla 4*. El 100% de las mujeres que sobreviven son casadas, no tienen ninguna discapacidad y poseen servicio público de acueducto. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

*Regla 7*. El 100% de las mujeres que sobreviven no tienen ninguna discapacidad, son cabeza de familia y el número de meses de vida contados a partir de la fecha de diagnóstico está entre 84 y 96 meses. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

*Regla 10*. El 100% de las mujeres que sobreviven fueron diagnosticadas en un hospital, con un cáncer localizado en el Cuello uterino SAI y el número de meses de vida contados a partir de la fecha de diagnóstico está entre 84 y 96 meses. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

*Regla 11*. El 100% de las mujeres que sobreviven tienen un cáncer con una morfología de tipo *Squamous cell carcinoma, large cell, nonkeratinizing* NOS, no se le practicó cirugía como tratamiento y pertenecen al régimen de salud subsidiado. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

*Regla 18*. El 100% de las mujeres que sobreviven tienen un nivel de escolaridad primaria, fueron diagnosticadas mediante una Histología de Tumor Primario, se les practicó una biopsia de la lesión y pertenecen al nivel 1 del SISBEN. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

3.3 Análisis de patrones de agrupación

Se utilizó el conjunto de datos T235A22ALL, para aplicarle la técnica de *clustering*, con el fin de encontrar similitudes entre todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2002, observadas hasta el año 2007, formando grupos similares, que relacionen los factores socioeconómicos y clínicos de estas mujeres.

Como muestran los resultados de la tabla 2, se formaron dos *clusters* (parámetro K=2). En el *cluster* 0 se clasificaron 112 mujeres que supuestamente sobrevivieron al cáncer y en el *cluster* 1 se agrupó

a 123 mujeres que posiblemente no sobrevivieron. En el repositorio T235A22ALL hay realmente 130 mujeres que sobrevivieron al cáncer y 105 que murieron. Por esa razón, los resultados indican que 18 pacientes, que realmente sobrevivieron, están en el grupo de las que murieron, que por sus características tienen mayor similitud con ese grupo. Midiendo la precisión del modelo, se puede decir que este agrupa correctamente al 86,2% de las mujeres sobrevivientes al cáncer, con una tasa de error general del 7,7% que indica que el modelo es bueno.

De acuerdo a las características o atributos que diferencian al *cluster* 0 del *cluster* 1, se pueden obtener los siguientes patrones:

*Cluster 0*. El 48% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002 y observadas hasta el 2007, sobreviven y son aquellas cuya edad está entre 55 y 67 años, su cáncer tiene una morfología de tipo *Squamous cell carcinoma*, *large cell, nonkeratinizing* NOS, son de nivel 1 del SISBEN, pertenecen a un régimen de salud subsidiado, provienen de la región Occidental Andina de Nariño y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 60 y 72 meses.

*Cluster 1*. El 52% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002 y observadas hasta el 2007 no sobreviven al cáncer invasivo de cuello uterino y son aquellas cuya edad está entre 42 y 54 años, su cáncer tiene una morfología de tipo *Squamous cell carcinoma*, NOS, no están en el sistema SISBEN, pertenecen al sistema general de seguridad social como vinculados, son de Pasto y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 0 y 12 meses.

3.4 Discusión de resultados

De acuerdo con los resultados obtenidos en las diferentes pruebas realizadas en la etapa de minería de datos con las tareas de clasificación, asociación y agrupación, en la cohorte 1998-2002, donde se analizan 235 casos, el patrón de supervivencia de las mujeres, después de haber sido diagnosticadas del cáncer, es mayor que 52 meses. Comparando estos resultados, con los 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 (Yépez et al., 2011), valor que se diferencia del patrón encontrado en la misma cohorte con técnicas de minería de datos.

En la cohorte 1998-2002 se analizaron 235 casos, de los cuales 130 corresponden a casos de mujeres que sobrevivieron a este tipo de cáncer y 105 que no. Teniendo en cuenta estas cifras, el 53,3% de todas las mujeres diagnosticadas con este tipo de cáncer sobrevive y de estos, el 86,9% sobrepasan el umbral de 52 meses de vida, después del diagnóstico.

Entre los factores socioeconómicos asociados a la supervivencia de las mujeres con cáncer invasivo de cuello uterino en esta cohorte están: poseer servicio público de acueducto, ser de nivel 1 del sistema SISBEN, vivir en casa o apartamento, ser cabeza de familia, no tener ninguna discapacidad, pertenecer a un sistema de salud subsidiado, tener un nivel de escolaridad de primaria y ocupación hogar.

En el estudio realizado por Yépez et al. (2011), 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 adicionan los factores socioeconómicos descubiertos con minería de datos.

Entre los factores clínicos asociados a la supervivencia de las mujeres con cáncer invasivo de cuello uterino están: el método de diagnóstico a través de Histología de Tumor Primario, el tratamiento a

través de cirugía y como 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. Los factores clínicos no inciden tanto como los socioeconómicos.

En el Municipio de Pasto, como se muestra en los resultados, el mayor porcentaje de mujeres con cáncer invasivo de cuello uterino tiene 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 (Arias, 2009).

## 4. Conclusiones

Con las tareas de clasificación, asociación y agrupación, se han obtenido patrones socioeconómicos y clínicos asociados a la supervivencia de las mujeres con cáncer invasivo de cuello uterino, a partir de los datos almacenados en el Registro Poblacional de Cáncer del municipio de Pasto. El patrón general de supervivencia descubierto es el número de meses mayor a 52 que transcurren desde el momento del diagnóstico del cáncer, entre los años 1998 y 2002, hasta la fecha final del periodo de observación de este estudio: 2007.

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 extraer patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino en el periodo 2003 -2007 con una ventana de observación hasta el 2012 y comparar los resultados con los obtenidos en este estudio.

## Agradecimientos

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

## Referencias

Agrawal, R., & Srikant, R. (Septiembre de 1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*. Conferencia llevada a cabo en Santiago de Chile, Chile.

Arias, S.A. (2009). Inequidad y cáncer: una revisión conceptual. *Revista Facultad Nacional de Salud Pública, 27* (3), 341-348. Recuperado de: https://aprendeenlinea.udea.edu.co/revistas/index.php/fnsp/article/view/2060

Asport, S., & Rivero, T. (2004). *Plan nacional de control de cáncer de cuello uterino 2004-2008*. Ministerio de Salud y Deportes de Bolivia. Recuperado de: http://saludpublica.bvsp.org.bo/textocompleto/ncc23332.pdf

Castro, M., Vera, L., & Posso, H. (2006). Epidemiología del cáncer de cuello uterino: estado del arte. *Revista Colombiana de Obstetricia y Ginecología*, *57* (3) 182-189. Recuperado de: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74342006000300006

Chen, M., Han, J., & Yu, P. (1996). Data mining: an overview from database perspective. *IEEE Transactions on Knowledge Data Engineering*, *8* (6), 866-883. doi: 10.1109/69.553155

Ciencia Hoy (2006*). Nuevas vacunas que salvarán millones de vidas: cáncer del cuello uterino*. Revista Ciencia Hoy en Línea, Vol. 16, No. 95. Recuperado de: http://www.cienciahoy.org.ar/ch/ln/hoy95/cancer.htm

Ferlay, J., Bray, F., Pisani, P., & Parkin, D.M. (2004). *GLOBOCAN 2002: Cancer incidence, mortality and prevalence worldwide*. Lyon, Francia: IARC Press.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., & Parkin, D.M. (2010). *GLOBOCAN 2008: Cancer incidence and mortality worldwide*. Lyon, Francia: IARC Press.

Fernández, G. (2009). *Extracción de Información de la web usando técnicas de minería de datos.* Recuperado de: http://www.tdg-seville.info/Download.ashx?id=48.

Gallardo, J. (2009). *Metodología para el desarrollo de proyectos en minería de datos CRISP-DM*. Recuperado de: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf.

García, M., & Álvarez, A. (2010). *Análisis de datos en WEKA –Pruebas de selectividad*. Recuperado de: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf.

Han, J., & Kamber, M. (2001). *Data mining concepts and techniques*. San Francisco, Estados Unidos: Morgan Kaufmann Publishers.

Hernández, E., & Lorente, R. (2009). *Minería de datos aplicada a la detección de cáncer de mama*. Recuperado de: http://ww.it.uc3m.es/jvillena/irc/practicas/08-09/14.pdf.

Hernández, J., Ramírez, M. J., & Ferri, C. (2005). *Introducción a la minería de datos*. Madrid, España: Editorial Pearson Prentice Hall.

Merle, J. L. (2004). *Análisis de la situación del cáncer cérvico-uterino en América Latina y el Caribe*. Washington, Estados Unidos: OPS.

Pardo, C., & Cendales, R. (2010). *Incidencia estimada y mortalidad por cáncer en Colombia: 2002-2006*. Bogotá, Colombia: Instituto Nacional de Cancerología E.S.E. Ministerio de Salud y Protección Social.

Quinlan, J.R. (1993). *C4.5: Programs for machine leraning*. San Francisco, Estados Unidos: Morgan Kaufmann Publishers.

Sattler, K., & Dunemann, O. (2001). SQL Database primitives for decision tree classifiers. *Proceedings of the Tenth International Conference on Information and Knowledge Management*. Conferencia llevada a cabo en Atlanta, Estados Unidos.

Timarán, R., & Millán, M. (2006). *New algebraic operators and sql primitives for mining classification rules*. *Proceedings of the Five IASTED International Conference on Computational Intelligence*. Conferencia llevada a cabo en San Francisco, Estados Unidos.

Witten, I., & Frank, E. (2000). *Data mining: practical machine learning tools and techniques with java implementations*. San Francisco, Estados Unidos: Morgan Kaufmann Publishers.

Yépez, M.C., Cerón, E., Hidalgo-Troya, A., & Cerón, C. (2011). Supervivencia de mujeres con cáncer de cuello uterino, Municipio de Pasto. *Revista Universidad y Salud, 2* (14), 7-18.