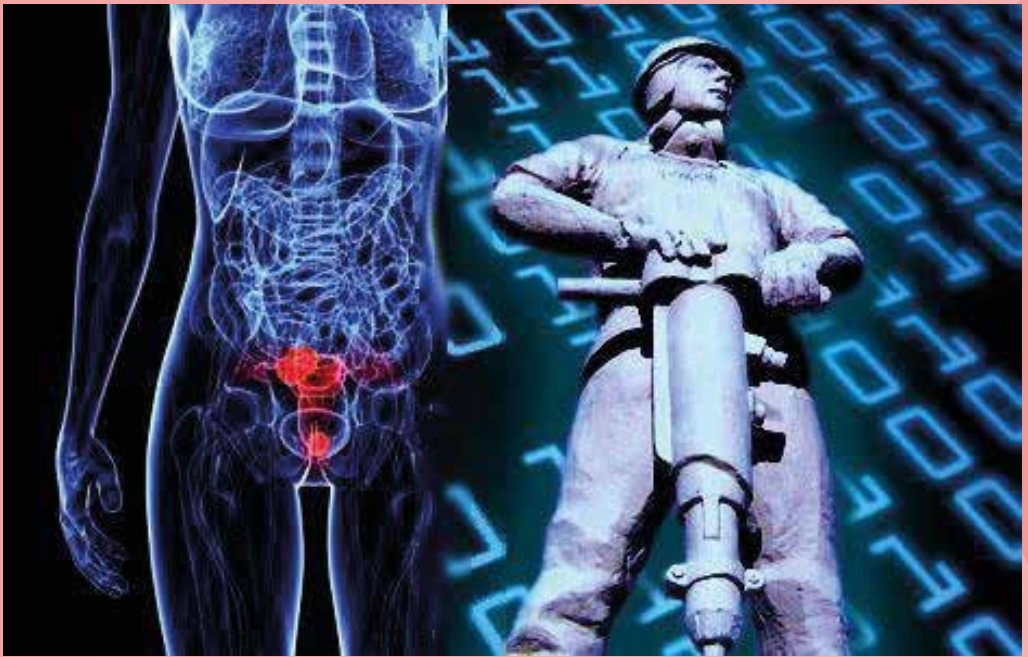


La Minería de Datos Aplicada a las bases de Datos de Cáncer Invasivo de Cuello Uterino



Ricardo Timarán Pereira
María Clara Yépez



Editorial
Universidad de **Nariño**



Editorial
Universidad de **Nariño**

LA MINERÍA DE DATOS APLICADA
A LAS BASES DE DATOS DE CÁNCER
INVASIVO DE CUELLO UTERINO

Ricardo Timarán Pereira
María Clara Yépez

LA MINERÍA DE DATOS APLICADA
A LAS BASES DE DATOS DE CÁNCER
INVASIVO DE CUELLO UTERINO



Editorial
Universidad de **Nariño**

Timarán Pereira, Ricardo

La minería de datos aplicada a las bases de datos de cáncer invasivo de cuello uterino / Ricardo Timarán Pereira, María Clara Yépez. -- 1ª. ed. -- San Juan de Pasto : Editorial Universidad de Nariño, 2021

159 p. : il. col., byn., tablas

Incluye referencias bibliográficas p. 150-158

ISBN: 978-958-5123-40-3

1. Neoplasma de cuello uterino--Bases de datos--Pasto--Nariño 2. Cáncer en cuello uterino--Procesamiento de datos--Investigaciones--Pasto 3. Minería de datos--Investigaciones--Pasto 4. Neoplasma de cuello uterino--Estadísticas--Pasto 5. Neoplasma de cuello uterino--prevención y control I. Yépez, María Clara

WP480 T582 - National Library of Medicine Classification Biblioteca Alberto Quijano Guerrero

La minería de datos aplicada a las bases de datos de cáncer invasivo de cuello uterino

© Ricardo Timarán Pereira
siritimar@gmail.com

María Clara Yépez

© Editorial Universidad de Nariño

ISBN: 978-958-5123-40-3

Primera edición

Impresión:

Graficolor Pasto SAS

Calle 18 No. 29-67 Tel. 7310652

graficolorpasto@hotmail.com

Marzo de 2021

San Juan de Pasto, Nariño, Colombia

Prohibida la reproducción total o parcial, por cualquier medio o con cualquier propósito, sin autorización escrita de los autores o de la Editorial Universidad de Nariño.

Agradecimientos

Al Sistema de Investigaciones de la Universidad de Nariño por financiar esta investigación y la publicación de este libro.

CONTENIDO

CAPÍTULO I	13
INTRODUCCIÓN	13
1.1 EL PROBLEMA DEL CÁNCER INVASIVO DE CUELLO UTERINO	14
1.2 METODOLOGIA	19
1.3 ORGANIZACIÓN DEL LIBRO	20
CAPÍTULO II	21
CÁNCER DE CUELLO UTERINO	21
2.1 EPIDEMIOLOGIA DEL CÁNCER DE CUELLO UTERINO	21
2.2 FACTORES DE RIESGO Y EVOLUCIÓN DE LA ENFERMEDAD	23
2.3 SUPERVIVENCIA	28
2.4 CASOS REGISTRABLES	28
CAPÍTULO III	29
MINERIA DE DATOS Y SUS APLICACIONES MÉDICAS	29
CAPÍTULO IV	38
DESCUBRIMIENTO DE PATRONES DE SUPERVIVENCIA EN MUJERES CON CÁNCER INVASIVO DE CUELLO UTERINO	38
4.1 ETAPA DE SELECCIÓN	38
4.2 ETAPA DE PREPROCESAMIENTO DE DATOS	43
4.3 ETAPA DE TRANSFORMACIÓN	43
4.4 ETAPA DE MINERÍA DE DATOS	61
4.4.1 Tarea de Clasificación	62
4.4.1.1 Parámetros de Evaluación del Modelo de Clasificación	62
4.4.1.2 Descubrimiento de Patrones con Árboles de Decisión	65
4.4.2 Tarea de Asociación	78
4.4.2.1 Descubrimiento de Reglas de Asociación	79
4.4.3 Tarea de Clustering o Segmentación	84
4.4.3.1 Descubrimiento de Patrones con Clustering	86
4.5 Descubrimiento de Patrones para los Casos de Cáncer de la Cohorte 1998-2002	89
4.5.1 Tarea de Clasificación	89
4.5.2 Tarea de Asociación	98
4.5.3 Tarea de Clustering	104
4.6 ETAPA DE EVALUACION E INTERPRETACIÓN DE RESULTADOS	108

4.6.1 Evaluación e Interpretación de resultados para la cohorte 1998-2007.....	108
4.6.1.1 Modelos de Clasificación	108
4.6.1.2 Modelos de Asociación.....	117
4.6.1.2 Modelos de Clustering	120
4.6.2 Evaluación e Interpretación de resultados en la cohorte 1998-2002.....	124
4.6.2.1 Modelos de Clasificación	124
4.6.2.2 Modelos de Asociación.....	130
5.6.2.3 Modelos de Clustering	134
CAPÍTULO V	138
DISCUSIÓN DE RESULTADOS	138
CAPÍTULO VI	147
CONCLUSIONES.....	147
REFERENCIAS BIBLIOGRÁFICAS	150

LISTA DE TABLAS

Tabla 1. Fuentes de Datos Externas	39
Tabla 2. Atributos iniciales de la Tabla CÁNCER	40
Tabla 3. Atributos Tabla CÁNCER con un alto porcentaje de nulos	44
Tabla 4. Limpieza de Atributos de la Tabla CÁNCER	45
Tabla 5. Nuevos Atributos de la Tabla CÁNCER	46
Tabla 6. Valores del Atributo COMUNA de acuerdo a los barrios que agrupa	47
Tabla 7. Valores del Atributo ESTRATO	53
Tabla 8. Valores del Atributo REGION de acuerdo a los municipios que agrupa ..	53
Tabla 9. Valores del Atributo EDADDX	54
Tabla 10. Valores del Atributo NIVELISISBEN de acuerdo a los puntajes obtenidos ..	54
Tabla 11. Valores del Atributo FUENTE de acuerdo al tipo de entidad	55
Tabla 12. Morfologías de cáncer de cuello uterino	55
Tabla 13. Atributos Tabla CÁNCER eliminados	56
Tabla 14. Descripción de tablas de la base de datos CÁNCERDB	57
Tabla 15. Atributos seleccionados para la investigación	58
Tabla 16. Atributos socio-económicos para la investigación	59

Tabla 17. Atributos clínicos para la investigación	60
Tabla 18. Clusters resultantes con K-means con el repositorio T507A22ALL	87
Tabla 19. Clusters resultantes con K-means con el repositorio T507A15ECON	88
Tabla 20. Clusters formados con K-means con el repositorio T507A09CLINICO	90
Tabla 21. Clusters resultantes con K-means con el repositorio T235A22ALL	105
Tabla 22. Clusters resultantes con K-means con el repositorio T235A15ECON	106
Tabla 23. Clusters formados con K-means con el repositorio T235A09CLINICO	107
Tabla 24. Valoración del coeficiente Kappa	109

LISTA DE FIGURAS

Figura 1. Historia natural de cáncer de cuello uterino.	27
Figura 2. Etapas del proceso de minería de datos.	30
Figura 3. Calidad de datos del atributo ESCOLARIDAD	44
Figura 4. Calidad de datos del atributo ESTADOCIVIL	45
Figura 5. Mejores resultados de la prueba 1 con el repositorio T507A22ALL y con $M=10$ y $C=0.3$	67
Figura 6. Mejores resultados de la prueba 2 con el repositorio T507A22ALL y con $M=5$ $C=0.2$	68
Figura 7. Mejores resultados de la prueba 3 con el repositorio T507A22ALL y con $M=2$ $C=0.2$	69
Figura 8. Mejores resultados de la prueba 4 con el repositorio T507A22ALL y con $M=2$ $C=0.4$	70
Figura 9. Mejor árbol de las pruebas con el repositorio T507A22ALL y con $M=2$ $C=0.4$	71
Figura 10. Resultados factores socioeconómicos con el repositorio T507A15ECONOM y con $M=2$ y $C=0.4$	72
Figura 11. Árbol factores socioeconómicos del repositorio T507A15ECONOM con $M=2$ y $C=0.4$	73
Figura 12. Resultados para los factores clínicos del repositorio T507A09CLINICO con $M=2$ y $C=0.4$	74
Figura 13. Árbol para los factores clínicos del repositorio T507A09CLINICO con $M=2$ y $C=0.4$	75
Figura 14. Resultados factores socioeconómicos con el repositorio T507A15ECONOM sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$	76
Figura 15. Árbol para los factores socioeconómicos con el repositorio T507A15ECONOM sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$	76

Figura 16. Resultados factores clínicos con el repositorio T507A09CLINICO sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.	77
Figura 17. Parámetros de ejecución del algoritmo Apriori con el repositorio T507A22ALL.	80
Figura 18. Mejores reglas generadas con Apriori con el conjunto de datos T507A22ALL.	82
Figura 19. Parámetros de ejecución del algoritmo Apriori con los factores socioeconómicos del repositorio T507A15ECON.	83
Figura 20. Mejores reglas generadas con Apriori con el conjunto de datos T507A15ECON.	84
Figura 21. Parámetros y reglas generadas con Apriori con el conjunto de datos T507A09CLINICO.	85
Figura 22. Mejores resultados para los casos de cáncer del repositorio T235A22ALL con $M=2$ y $C=0.2$.	92
Figura 23. Mejor árbol para los casos de cáncer del repositorio T235A22ALL con $M=2$ y $C=0.2$.	92
Figura 24. Mejores resultados para los factores socioeconómicos del repositorio T235A15ECON.	93
Figura 25. Mejor árbol para los factores socioeconómicos del repositorio T235A15ECON.	93
Figura 26. Resultados para los factores clínicos del repositorio T235A09CLINICO con $M=2$ y $C=0.2$.	94
Figura 27. Árbol para los factores clínicos del repositorio T235A09CLINICO con $M=2$ y $C=0.2$.	95
Figura 28. Resultados factores socioeconómicos con el repositorio T235A15ECON sin el atributo nmeses y con los parámetros $M=2$ y $C=0.4$.	96
Figura 29. Resultados factores clínicos con el repositorio T235A09CLINICO sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.	97
Figura 30. Parámetros de ejecución del algoritmo Apriori para el repositorio T235A22ALL.	98
Figura 31. Mejores reglas generadas con Apriori con el conjunto de datos T235A22ALL.	100
Figura 32. Parámetros de ejecución del algoritmo Apriori para los factores socioeconómicos del repositorio T235A15ECON.	101
Figura 33. Parámetros de ejecución del algoritmo Apriori para los factores socioeconómicos del repositorio T235A15ECON.	103
Figura 34. Parámetros y reglas generadas con Apriori para los factores clínicos con el conjunto de datos T235A09CLINICO.	104

CAPÍTULO I

INTRODUCCIÓN

El Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP) en Colombia tiene como misión “recopilar periódicamente y de manera confiable, datos sobre la incidencia de cáncer en el municipio de Pasto que ayude a los entes gubernamentales en la toma de decisiones, en las políticas públicas tendientes a la implementación de programas de prevención y rehabilitación y como base para la investigación”. Desde su creación en el año 1998 como proyecto de investigación permanente del Centro de Estudios en Salud de la Universidad de Nariño (CESUN) en convenio con la Secretaría Municipal de Salud del municipio de Pasto, el RPCMP ha funcionado de manera continua con el objetivo de proporcionar datos sobre incidencia de cáncer para la toma de decisiones en las entidades de salud y la implementación de programas de promoción y prevención. El RPCMP, forma parte de la Asociación Internacional de Registros de Cáncer (IACR) y de la Red de Registros poblacionales de Colombia junto con los Registros de Cali, Bucaramanga, Manizales y Barranquilla (Yépez et al., 2012a).

Para cumplir con este propósito, el RPCMP requería de la construcción de un sistema de información transaccional que permita almacenar los datos de una manera confiable, íntegra y eficiente, que proporcione información de calidad y que conjuntamente con herramientas de inteligencia de negocios, como la minería de datos, pueda detectar patrones ocultos que ayuden a soportar lo toma de decisiones en todo lo relacionado con la incidencia, supervivencia, mortalidad y prevalencia de cáncer en el municipio de Pasto.

Por dar solución a este requerimiento, entre los años 2010-2013 se ejecuta el proyecto de investigación denominado “Detección de Patrones de Supervivencia en Mujeres con Cáncer Invasivo de Cuello Uterino con Técnicas de Minería de datos”, financiado por el sistema de investigaciones de la Universidad de Nariño, en cuya primera fase se construyó el sistema transaccional y en la segunda fase, a partir de los datos almacenados en la base de datos de este sistema, se descubrieron patrones de supervivencia de mujeres diagnosticadas con cáncer invasivo de cuello utilizando técnicas de minería de datos. El proyecto se delimitó a la información disponible y depurada por el RPCMP en el momento de realizar este estudio y que corresponde al periodo comprendido entre los años 1998 y 2007 y a una supervivencia a cinco años. El RPCMP depura la información de acuerdo a la metodología propuesta por la Asociación Internacional de Registros de Cáncer (IARC) para estimar la incidencia, mortalidad y prevalencia de cáncer por quinquenios, con el fin de ser comparables con los datos publicados por otros registros en el mundo.

1.1 EL PROBLEMA DEL CÁNCER INVASIVO DE CUELLO UTERINO

Según la Organización Mundial de la Salud (OMS), el cáncer es una de las principales causas de morbilidad y mortalidad en todo el mundo. Para el año 2012, se estimaron un total de 14,09 millones de nuevos casos y causó 8,01 millones de defunciones (Globocan, 2012). Se espera para el año 2030 un crecimiento significativo en la magnitud de este importante problema de salud pública, como producto del crecimiento demográfico y el envejecimiento de la

población. Se estima que los casos nuevos ascenderán a 21 millones y que la mortalidad por cáncer aumentará 45% ocasionando aproximadamente 13 millones muertes (Yepez et al., 2012b).

En Colombia se presentan altas tasas de incidencia y mortalidad por cáncer, relacionados directamente con el incremento en la esperanza de vida, el envejecimiento de la población, el desplazamiento poblacional hacia centros urbanos y las limitaciones en el acceso a los servicios de salud (Piñeros, Murillo & Porras, 2011). En información estadística disponible en el Instituto Nacional de Cancerología (INC) en Colombia refiere que, para el periodo 2002-2006 se estimaron anualmente 70.887 casos de cáncer: 32.316 en hombres y 38.571 en mujeres (INC, 2012). En hombres, la tasa de incidencia anual ajustada por edad (TAE) para cáncer en todas las localizaciones (excepto piel) fue de 186.6 por 100.000 habitantes. En mujeres, la TAE para cáncer en todas las localizaciones (excepto piel) fue de 196.9 por 100.000 habitantes. Las primeras cuatro localizaciones de cáncer en hombres fueron, en orden decreciente, próstata, estómago, pulmón y colon, recto y linfoma no hodking. En mujeres, las principales localizaciones fueron mama, cuello uterino, tiroides, estómago, colon, recto y ano. En el periodo de 2007-2011 el INC reporta un total de 62,812 casos nuevos anuales de cáncer, 29,734 casos en hombres y 33,084 casos en mujeres. En hombres, la mayor incidencia de tumores se presentó en: próstata, estómago y colon y recto. En mujeres: mama, cuello del útero, y colon y recto. En el mismo periodo se registraron 32,653 muertes por cáncer, 16,081 en hombres y 16,572 en mujeres. Las principales causas de mortalidad por cáncer en hombres fueron los tumores de: estómago, próstata y pulmón; en mujeres los tumores de: mama, cuello del útero y estómago. Según este informe, para el Departamento de Nariño, donde se ubica Pasto, el cáncer de estómago tanto en hombres como en mujeres produce la tasa más alta de mortalidad (Yépez et al., 2018).

En el caso del cáncer invasivo de cuello uterino, aunque es una enfermedad potencialmente prevenible sigue siendo un problema de salud pública en el mundo, siendo en el 2012, según la OMS, el cuarto tipo de cáncer más frecuente en mujeres, con un estimado de 527.624 nuevos casos, correspondiente al 7.9% del total de in-

cidencia de cáncer en mujeres y una mortalidad del 7.5% del total de casos de muertes por cáncer (Globocan, 2012). En los países suramericanos, en este mismo año, el cáncer de cuello uterino fue la segunda causa de muerte en mujeres por cáncer con una incidencia de 45.008 nuevos casos (Globocan, 2012).

La tasa estimada de mortalidad estandarizada por edad por cáncer de cuello uterino para el año 2012 en Colombia fue de 8 por 100.000 mujeres (Globocan, 2012) y según la descripción epidemiológica en el Plan Decenal para el control del cáncer 2012-2021, el comportamiento de éste tipo de cáncer está relacionado socialmente con poblaciones que viven en los lugares más alejados del país y pertenecientes a régimen subsidiado (MinSalud, 2012).

Para el Departamento de Nariño (Colombia), las últimas estadísticas del INC reportan en el período 2002-2006 una tasa ajustada por edad (TAE) de incidencia estimada anual de cáncer en todas las localizaciones, excepto piel, en mujeres de 145.6 x 100.000 habitantes y una mortalidad de 57.2 por 100.000 habitantes (INC, 2012). En este mismo período la tasa de incidencia anual (TAE) de cáncer de cuello uterino fue de 26.4 y de mortalidad de 9.8 x 100.000 habitantes (Pardo & Cendales, 2010).

Para el Municipio de Pasto, el RPCMP en el período 2003-2007, reportó las tasas de incidencia de cáncer en todas las localizaciones ajustadas por edad (TAE) excepto piel, de 141.6 x 100.000 habitantes en hombres y en mujeres de 152.4/ 100.000 habitantes, siendo el cáncer de estómago en hombres el de más incidencia con una TAE de 31.7 hombres/año y en mujeres el de cuello uterino con una TAE de 27.3 x 100.000 mujeres/año. La tasa de mortalidad para el período mencionado fue de 6.6 por 100.000 mujeres año. (Yépez et al., 2012a). Sin embargo, se desconoce la situación de las mujeres del Municipio de Pasto, una vez se les ha diagnosticado el cáncer invasivo de cuello uterino en el cual se aumenta la probabilidad de muerte.

Algunos estudios (Castro et al., 2006), (Asport & Rivero, 2004), (Ferlay et al., 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. En este último estudio, se encontró además que la mediana de supervivencia de las mujeres con cáncer invasivo de cuello uterino fue de 36.8 meses (Yépez et al., 2011).

Otro estudio reporta que el pronóstico del cáncer de cuello uterino es dependiente de las características socioeconómicas y demográficas de la paciente, del estadio clínico al momento del diagnóstico, del esquema de tratamiento y del tiempo transcurrido entre el diagnóstico y el tratamiento y de su continuidad, variables que inciden en la supervivencia (Merle, 2004).

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 únicamente se pueden descubrir utilizando un tratamiento de los datos más complejo, que solo es posible con la minería de datos.

En la última década, la minería de datos y el análisis estadístico han sido ampliamente utilizados en la industria del cuidado de la salud. Cuando estos métodos son utilizados junto a información extraída de grandes cantidades de datos, pueden ayudar a los médicos para tomar decisiones y mejorar el servicio (Hernández & Lorente, 2009).

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 interpre-

tación propone, en el caso del cáncer invasivo de cuello uterino, patrones de supervivencia no previstos desde la estadística.

La minería de datos surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, los datos son la materia prima bruta. 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. En este contexto, la minería de datos emerge como el siguiente paso evolutivo en el proceso de análisis de datos.

En general, los especialistas médicos no cuentan con herramientas objetivas, que los ayuden a tomar una decisión respecto al tratamiento óptimo para un paciente. Utilizando la información que los especialistas consideran importante, con la minería de datos se pretende encontrar patrones y relaciones entre las variables, de forma de poder 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 en los datos que no pueden identificarse mediante un tratamiento estadístico clásico.

En la Universidad de Nariño, en el municipio de Pasto e incluso en el departamento de Nariño, no se han planteado proyectos de investigación que permitan construir modelos a partir de los datos registrados y utilizando técnicas de minería de datos, detectar de manera temprana algún tipo de cáncer.

En este libro se presentan los resultados del proyecto de investigación que tuvo como objetivo detectar patrones de supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino utilizando técnicas de descubrimiento de conocimiento a partir de los datos reportados en el Registro Poblacional de Cáncer del municipio de Pasto en el periodo 1998-2007, que generen cono-

cimiento acerca de los diferentes factores (sociales, económicos y clínicos) asociados a la supervivencia de este grupo poblacional y que ayude a soportar la toma de decisiones eficaces de los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de programas de prevención y control de las mujeres con este tipo de cáncer.

1.2 METODOLOGÍA

La investigación se desarrolló bajo el enfoque cuantitativo, de tipo descriptivo, y un diseño no experimental. Contempló dos fases, cada una con duración de un año.

En la primera fase, se construyó un nuevo sistema de información transaccional orientado a la web para el RPCMP. Se diseñó y construyó una nueva base de datos utilizando el Sistema Gestor de Bases de Datos PostgreSQL y se pobló con los datos del aplicativo anterior denominado CANREG y con otras fuentes de datos que se necesitaron. Se seleccionó el conjunto de datos objeto de estudio y se definió el periodo de análisis, dependiendo de la disponibilidad de los datos.

En la segunda fase, basados en el proceso de Descubrimiento de Conocimiento en bases de Datos (DCBD) se analizaron y seleccionaron entre los diferentes atributos de la base de datos del RPCMP, los que mayor incidencia tenían en la supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino. A este conjunto de datos se le aplicaron las dos subsiguientes etapas del DCBD como lo son preprocesamiento y transformación de datos, y se obtuvo un conjunto de datos limpios y listos para aplicarle las técnicas y algoritmos de minería de datos. Se analizaron y aplicaron las técnicas y algoritmos de minería de datos más adecuados para resolver el problema, utilizando la herramienta de minería de datos libre Weka que permitió obtener los patrones de supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino y que fueron el objeto de esta investigación.

1.3 ORGANIZACIÓN DEL LIBRO

Este libro está organizado de la siguiente manera:

- En el segundo capítulo se describe brevemente la epidemiología del cáncer de cuello uterino.
- En el tercer capítulo se presenta una visión general de la minería de datos y sus aplicaciones en el ámbito médico.
- En el cuarto capítulo se describe el proceso de descubrimiento de patrones de supervivencia en mujeres con cáncer de cuello uterino.
- En el quinto capítulo se hace la discusión de resultados.
- En el último capítulo se dan las conclusiones y trabajos futuros.

CAPÍTULO II

CÁNCER DE CUELLO UTERINO

2.1 EPIDEMIOLOGÍA DEL CÁNCER DE CUELLO UTERINO

Aunque el cáncer invasivo de cuello uterino es una enfermedad potencialmente prevenible sigue siendo un problema de salud pública (Cannistra & Niloff, 1996), (Ferlay et al., 2004).

Según los reportes de la Agencia Internacional de Investigación en Cáncer (IARC), a nivel mundial el cáncer de cuello uterino es el cuarto tipo de cáncer más común en las mujeres, y el séptimo entre todos los tipos de cáncer. Para el año 2012, se estimaron 528.000 nuevos casos y 266.000 muertes que corresponden al 7.5% de todas las muertes de mujeres por cáncer. Más del 85% de la carga de la enfermedad a nivel mundial se producen en los países en desarrollo donde representa el 12% de todos los cánceres femeninos (Globocan, 2012).

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).

Según la Organización Panamericana de la Salud OPS, en el año 2012, en las Américas, fueron diagnosticadas con cáncer de cuello uterino más de 83.000 mujeres y casi 36.000 fallecieron por ésta enfermedad. Si se mantiene ésta tendencia el número de muertes se aumentará en un 45% en el 2030. En América Latina y el Caribe, las tasas de mortalidad son 3 veces más altas que en Norte América, haciendo evidente las desigualdades en salud. Este problema se debe enfrentar de forma integral a través de programas de salud sexual y reproductiva, salud de los adolescentes, inmunización y control de este cáncer, por cuanto la historia natural de la enfermedad ofrece la oportunidad de intervenir a lo largo de la vida de la mujer mediante estrategias efectivas de prevención primaria y secundaria, así como de diagnóstico y tratamiento, incluyendo el acceso a cuidados paliativos (OPS, s.f.).

Los programas de tamizaje que introdujeron la citología del cuello uterino como herramienta fundamental en las actividades de prevención, seguido del tratamiento de las lesiones precancerosas identificadas, ya han demostrado su efectividad en los países desarrollados (Sankaranarayanan et al., 2001). La probabilidad de cambio con el tiempo de una lesión observada en una citología es dependiente de su grado de compromiso inicial. El pronóstico de las mujeres con cáncer de cuello uterino varía según las características del tumor, del agente y de la paciente.

La vacunación contra el virus del papiloma humano (VPH) de chicas adolescentes puede prevenir aproximadamente un 70% de los casos de cáncer de cuello uterino (OPS, s.f.).

En Colombia 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, ubican geográficamente el

riesgo de mortalidad por cáncer de cuello uterino a las habitantes de departamentos alejados, zonas de frontera y riberas de ríos y socialmente en 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 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 ajustada por edad (TAE) de incidencia estimada anual de cáncer en todas las localizaciones excepto piel en mujeres de 145.6 x 100.000 habitantes y una mortalidad de 57.2 por 100.000 habitantes. En este mismo período la tasa de incidencia anual (TAE) de cáncer de cuello uterino fue de 26.4 y de mortalidad de 9.8 x 100.000 habitantes (Pardo & Cendales, 2010).

Para el Municipio de Pasto, el Registro Poblacional de Cáncer (RPC) en el período 2003-2007, reportó las tasas de incidencia de cáncer en todas las localizaciones ajustadas por edad (TAE) excepto piel, de 141.6 X 100.000 habitantes en hombres y en mujeres de 152.4/ 100.000 habitantes, siendo el cáncer de estómago en hombres el de más incidencia con una TAE de 31.7 hombres/año y en mujeres el de cuello uterino con una TAE de 27.3 x 100.000 mujeres/año. La tasa de mortalidad para el período mencionado fue de 6.6 por 100.000 mujeres año (Yépez et al., 2012a).

2.2 FACTORES DE RIESGO Y EVOLUCIÓN DE LA ENFERMEDAD

El factor principal de riesgo del cáncer del cuello uterino es la infección con el virus del papiloma humano VPH (Heard, Palefsky & Kazatchkine, 2004). Estudios realizados con técnicas de biología molecular de alta sensibilidad y con muestras biológicas

adecuadas, detectan VPH de alto riesgo en prácticamente el 100% de los casos de carcinoma escamoso de cérvix (Tejada et al., 2007).

Esta enfermedad se puede transmitir de una persona a otra durante las relaciones sexuales. El sostener relaciones sexuales sin protección, especialmente a una edad temprana, hace que sea más probable infectarse con el virus de VPH (Walboomers, Jacobs & Manos, 1999). La infección genital por VPH es una de las enfermedades de transmisión sexual más comunes; la infección cervical con VPH con genotipos de alto riesgo, es un evento temprano necesario para la carcinogénesis.

Esta neoplasia es una de las más prevenible, dado su bien conocida causa viral y su lenta progresión después de la persistencia de la infección por el virus de papiloma humano, (VPH) de alto riesgo. La prevalencia del VPH en muestras cervicales de mujeres mayores de 30 años en países con una alta incidencia de cáncer cervical, es de aproximadamente 11 al 40%, y en países con baja incidencia es de 5 al 10% (Sankaranarayanan, Budukh & Rajkumar, 2001).

Aunque la evidencia epidemiológica ha sugerido que el virus de papiloma humano (VPH) es causa necesaria para el desarrollo del cáncer de cuello uterino, sin embargo, no se puede catalogar como una causa suficiente ya que existen otros factores asociados como son la persistencia de la infección, estados de inmunodepresión, uso de anticonceptivos orales y el hábito de fumar, que si bien no tienen definido un papel específico si se ha comprobado su relación con la patología (Castro, 2006), (Villa, 2003), (Arends et al., 1998).

Estos cofactores, que ayudan a definir grupos de riesgo y que se involucran en el proceso de carcinogénesis formando parte de la red de causalidad y, dependiendo de su acción, pueden clasificarse en directos (aquellos que por mecanismos propios ayudan a la transformación celular) e indirectos (aquellos que aumentan el riesgo de infección por VPH).

Dentro del grupo de factores directos se encuentran el tabaquismo, la influencia hormonal, la infección por *Chlamydia trachomatis*, el tipo de HLA, la inmunosupresión y la susceptibilidad genética. Los carcinógenos del tabaco pueden encontrarse en el moco cervical de las mujeres fumadoras y, por tanto, el consumo de tabaco es una variable obligatoria en los estudios de causalidad (Prokopczyk et al., 1997), (Franco et al., 2001). La asociación entre el uso de contraceptivos orales y el carcinoma cervical no es consistente (Madeleine et al., 2001) y pierde significación cuando se ajusta por la presencia de infección por VPH (Chichareon et al., 1998).

Los factores indirectos más frecuentemente estudiados son el inicio temprano de la actividad sexual, el número de compañeros sexuales, el tener un compañero sexual de alto riesgo, el antecedente de enfermedades sexualmente transmitidas, el estado civil y el antecedente de citología anormal (Stuver & Adami, 2002), (Herrero et al., 1990). En un estudio de poblaciones de alto riesgo se concluyó que el riesgo del cáncer cervical aumentaba con el inicio precoz de la actividad sexual, el número de compañeros sexuales y el antecedente de infección gonocócica (Herrero et al., 1990). La escolaridad muestra una asociación fuerte con el cáncer de cuello uterino hasta el punto de considerarse una forma para definir grupos de riesgo (Corral et al., 1996).

Otro factor de riesgo es la edad, el riesgo de contraer este cáncer es muy bajo entre las niñas menores de quince años, pero aumenta entre los últimos años de la adolescencia y la mitad de la tercera década de vida. A diferencia de muchos otros tipos de cáncer, que rara vez afectan a las mujeres adultas jóvenes, el cáncer del cuello uterino puede afectar a las jóvenes en su tercera década de vida e incluso a las adolescentes. Aunque el riesgo de cáncer del cuello uterino no aumenta mucho después de los 40, tampoco disminuye. Por último, hay estudios que consideran que una baja condición socioeconómica es un factor importante de riesgo en problemas de la salud, y en particular en cáncer de cuello uterino. Se ha encontrado que, en mujeres de baja condición económica,

con bajos ingresos ayudan a que el problema de cáncer de cuello se presente con cierta frecuencia, dada su baja concientización sobre los problemas de salud (Muñoz, Bosch & De Sanjosé, 2003).

Los datos compartidos de ocho estudios de casos y controles sobre el cáncer invasivo del cuello uterino y dos estudios sobre el carcinoma in situ efectuados en cuatro continentes, sugieren que las mujeres con tres o cuatro embarazos a término tenían un riesgo 2 a 6 veces más alto de aparición de cáncer cérvico uterino que aquellas que nunca habían dado a luz; las mujeres con siete partos o más, presentaron un riesgo 3 a 8 veces mayor (Casado, Vidart & Coronado, 2004).

La historia natural de la enfermedad muestra que existe un período prepatogénico en el cual se describen los factores relacionados con el cáncer de cuello uterino que probablemente son: la infección con el virus del papiloma humano (VPH) como el agente causal, las características del huésped relacionadas con el grupo étnico, tabaquismo, hábitos sexuales, percepción del riesgo a enfermar, edad, estado inmune, herencia, uso de servicios de salud y escolaridad, entre otros, los factores ambientales como el nivel de ingresos, medidas de salud pública y acceso a los servicios de salud. La persistencia del agente causal, propicia el período de patogénesis inicial que da origen a lesiones de bajo grado de patogenicidad o de alto grado, cuya recuperación está relacionada con el tratamiento oportuno, las creencias y actitudes frente a la enfermedad y la respuesta individual a los tratamientos. El período de patogénesis tardía desencadena el carcinoma invasivo de cuello uterino cuya recuperación está relacionada con los mismos factores de la patogénesis inicial, del tipo de VPH y de las características clínicas e histológicas del tumor. El pronóstico para las mujeres que no se recuperan es la cronicidad, la discapacidad y la muerte (Ashford & Collymore, 2005).

Según Castro et al. (Castro, Vera & Posso, 2006), fundamentados en los hallazgos de otros estudios, la descripción de la historia natural de la enfermedad la resumen en la figura 1.

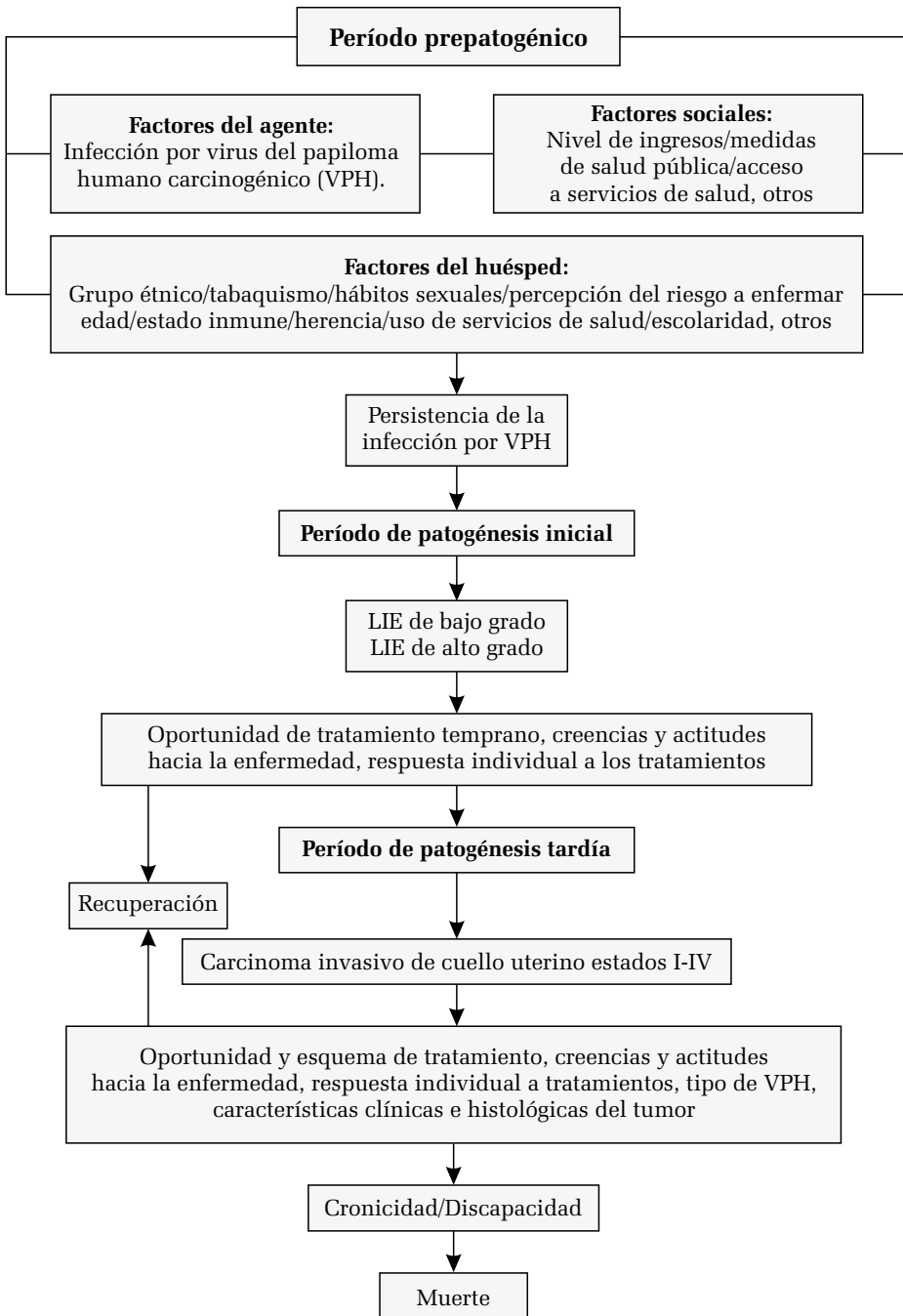


Figura 1. Historia natural de cáncer de cuello uterino.

Fuente: Castro et al, 2006. Estado del arte de la epidemiología del cáncer de cuello uterino

2.3 SUPERVIVENCIA

De acuerdo a la definición de supervivencia en el diccionario de cáncer del National Cancer Institute at the National Institutes of Health, la supervivencia se enfoca en la salud y la vida de una persona después del diagnóstico o el tratamiento del cáncer hasta el final de la vida (NCI, s.f).

En las enfermedades crónicas, tales como el cáncer, la supervivencia se mide como una probabilidad de permanecer vivo durante una determinada cantidad de tiempo. La supervivencia al año o a los 5 años es a menudo expresada como indicador de la severidad de una enfermedad y como pronóstico. Típicamente, el pronóstico del cáncer se valora determinando el porcentaje de pacientes que sobrevive a lo menos cinco años después del diagnóstico. A fin de obtener tasas de supervivencia a 5 años, los médicos tienen que analizar a personas que fueron tratadas al menos 5 años atrás. Los avances en el tratamiento desde entonces pueden producir un pronóstico más favorable para personas que hoy día son diagnosticadas con cáncer de cuello uterino.

El tiempo de supervivencia se define como el tiempo transcurrido desde el acontecimiento o estado inicial que puede ser el diagnóstico de la enfermedad hasta el estado final que puede ser la muerte.

El estado inicial debe ser definido de manera que la fecha en que se produjo el evento pueda ser conocida exactamente. El acontecimiento o suceso estudiado también debe estar perfectamente definido para poder determinar exactamente la fecha del mismo. Este evento está casi siempre asociado a la muerte del paciente.

2.4 CASOS REGISTRABLES

El RPCMP define como caso registrable a todo tumor maligno diagnosticado en pacientes residentes en el Municipio de Pasto a partir del año 1998, mediante cualquier método válido de diagnóstico (histología, citología, imágenes, endoscopia, clínico o por certificado de defunción). Se considera como fecha de incidencia de los casos de cáncer, la relacionada con el primer evento cronológico de confirmación diagnóstica de la enfermedad o en caso de no encontrar más información se utiliza la fecha de defunción (Yépez et al., 2012a).

CAPÍTULO III

MINERÍA DE DATOS Y SUS APLICACIONES MÉDICAS

Los recientes avances tecnológicos hacen que las capacidades para generar y almacenar datos se incrementen día a día. La disponibilidad de grandes volúmenes de datos crea la necesidad de nuevas técnicas y herramientas capaces de asistir a los usuarios de forma inteligente y automática en el análisis de datos que generen conocimiento para soportar la toma de decisiones. Todo este explosivo crecimiento generó la aparición de la minería de datos (Pérez & Santin, 2006).

La minería de datos es el proceso de descubrir nuevas y útiles correlaciones, patrones y tendencias dentro de grandes cantidades de datos almacenadas en repositorios, utilizando tecnología para el reconocimiento de patrones, así como técnicas matemáticas y estadísticas, con el fin de que sean analizados por los usuarios y ayuden a la toma de decisiones de las organizaciones (Fayyad, Piatetsky-Shapiro & Smyth, 1996). Este proceso implica preprocesar los datos, hacer minería de datos y presentar resultados (Agrawal & Srikant, 1994), (Chen et al., 1996), (Piatetsky-Shapiro et al, 1996), (Timarán, 2009), (Han & Kamber, 2012).

La minería de datos en realidad es el núcleo de todo un proceso llamado Descubrimiento de Conocimiento en Base de Datos - KDD (del inglés Knowledge Discovery in Databases), el cual es un proceso metodológico, interactivo e iterativo, para encontrar patrones válidos, novedosos, útiles y entendibles al usuario a partir de los datos (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Un patrón es válido si este sigue siendo preciso para datos nuevos. Es novedoso porque aporta algo desconocido. Es útil porque debe conducir a acciones que reporten algún beneficio para el usuario y es entendible porque se puede interpretar, revisar, validar y usar en la toma de decisiones (Han & Kamber, 2012).

En la figura 2 se muestra como la minería de datos es un proceso compuesto por un conjunto de etapas o fases, en las cuales se establece el problema, se seleccionan los datos adecuados para dar solución al problema, se exploran y limpian los datos, se transforman para aplicarles luego las técnicas de modelado que dan como resultado patrones que deben interpretarse y evaluarse para obtener conocimiento.

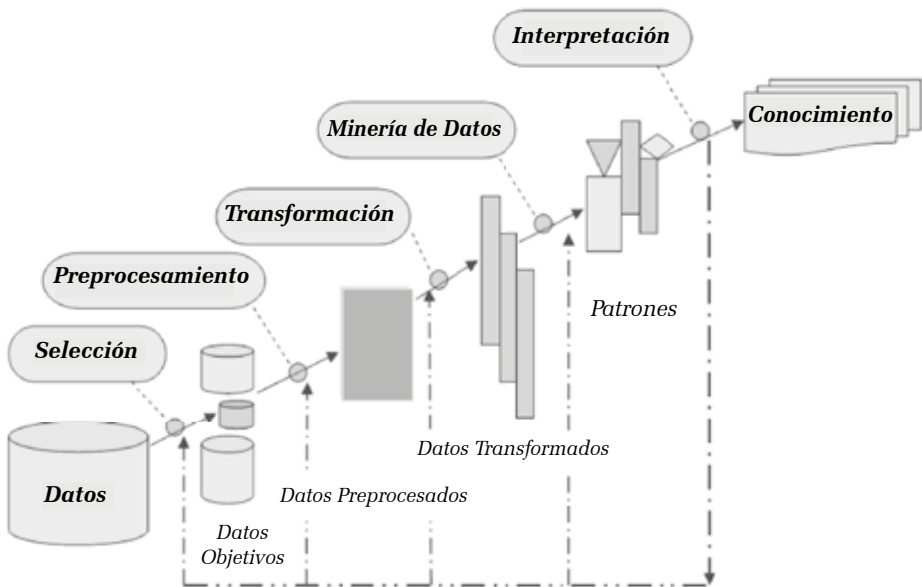


Figura 2. Etapas del proceso de minería de datos.

Fuente: Timarán et al, 2016, Descubrimiento de patrones de desempeño académico.

Como metodología para desarrollar proyectos de minería de datos se utilizan las diferentes etapas del proceso KDD. En Timarín et al. (2016) se describe cada una de estas etapas:

En la etapa de Selección, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso de minería de datos, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de éste, sobre el cual se va a realizar el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio. En la etapa de Preprocesamiento/Limpieza (Data cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con el usuario o analista. En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos. El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés aplicando tareas de descubrimiento tales como clasificación (Quinlan, 1986), (Wang et al., 1998), clustering (Ng & Han, 1994), (Zhang, Ramakrishnan & Livny, 1996), patrones secuenciales (Agrawal & Srikant, 1995) y asociaciones (Agrawal & Srikant, 1994), (Srikant & Agrawal, 1996) entre otras y finalmente en la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. 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. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones, o simplemente para documentarlo y reportarlo a las partes interesadas, así como también para verificar y

resolver conflictos potenciales con el conocimiento previamente descubierto.

Hablar de la tecnología informática en la medicina no es nada nuevo, sin embargo, los recientes avances en Medicina Molecular, han abierto muchas áreas de oportunidad para la aplicación de técnicas de análisis de datos que permiten explorar grandes conjuntos de los mismos, generados a partir de recientes descubrimientos en genómica y proteómica, producidos por las investigaciones del genoma y proteoma humanos respectivamente. Las áreas de la Medicina Molecular donde la Informática ha encontrado un lugar para su aplicación son muy variadas, se puede encontrar desde la producción de “fármacos perfectos personalizados” hasta la detección de enfermedades como el cáncer. Es aquí donde la minería de datos, resulta 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 (Mora, 2001).

La minería de datos es un conjunto de técnicas de análisis de datos que permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos, así como para predecir comportamientos futuros (Hernández, Ramírez & Ferri, 2005). Desde el punto de vista médico, sirve para establecer índices pronósticos y diagnósticos que permitan obtener información útil para predecir incidencias y ajustar los protocolos de actuación clínica (Zamarrón et al., 2006). En este proceso de minería de datos es importante el uso de modelos de representación y razonamiento temporal, tanto por la naturaleza intrínsecamente evolutiva del seguimiento de pacientes, como por la unidireccionalidad temporal de las relaciones causa/efecto.

Se han realizado numerosas aplicaciones de minería de datos al estudio de enfermedades, diagnósticos y tratamientos. Algunas de ellas son: cáncer de próstata, enfermedades cardiovasculares, hipertensión arterial, cáncer de mama, parkinson y enfermedades tumorales, cáncer de cuello uterino, diabetes, dengue, entre otras (Oviedo et al., 2015).

En las aplicaciones médicas, donde no se puede obviar la importancia de la componente temporal, las técnicas de minería de datos han adquirido gran relevancia (Hernández, Ramírez & Ferri, 2005). Las aplicaciones de estas técnicas van desde la visualización inteligente de grandes cantidades de datos médicos, hasta el control de calidad en centros hospitalarios.

CancerGrid es una iniciativa que involucra científicos de la Universidades de Oxford y Cambridge para reducir el costo de la investigación clínica a través de una efectiva compartición de datos. Ha desarrollado estándares y herramientas para la producción automática de los sistemas necesarios para soportar estudios clínicos y de investigaciones trasnacionales, es decir, el traslado del conocimiento de la investigación médica a la práctica clínica (Davies et al., 2012).

En el ámbito médico la aplicación de la minería de datos tiene interés en varios campos (Zamarrón et al., 2006):

- En el ámbito clínico resulta de ayuda para la identificación y diagnóstico de patologías. Así mismo tiene importancia para el descubrimiento de posibles interrelaciones entre diversas enfermedades.
- Al nivel de medicina preventiva, resulta de interés para la detección de pacientes con factores de riesgo para sufrir una patología.
- Al nivel de gestión hospitalaria, se puede usar para obtener predicciones temporales que permitiesen optimizar los recursos disponibles y priorizar el uso de los diversos tratamientos para una misma patología.

Una de las técnicas de minería de datos más utilizadas para el desarrollo de sistemas de diagnóstico médico son los árboles de decisión. Se trata de clasificadores que se basan en algoritmos inductivos para extraer conocimiento a partir de un conjunto de casos y definir las relaciones que aparecen en el proceso que se intenta simular. 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 (Daedalus, 2009).

En casos específicos de medicina, la aplicación de técnicas de minería de datos puede plantearse para el estudio de las alteraciones respiratorias durante el sueño, el cual se puede desarrollar en tres etapas bien diferenciadas: en una etapa inicial se pueden aplicar técnicas convencionales de minería de datos, posteriormente técnicas de minería de datos supervisada, de modo que el médico guía la búsqueda de hallazgos de interés mediante su asociación con eventos fisiopatológicos conocidos; por último, se pueden utilizar técnicas de minería de datos no supervisada, de modo que se busca una agregación en la identificación de manifestaciones a partir de la definición de índices de similitud (Zamarrón et al., 2006).

Kumari y Sunila (2011) presentan un estudio predictivo que compara las técnicas de clasificación de la minería de datos aplicada a las enfermedades cardiovasculares. Para realizar esta comparación se utilizó un conjunto de datos de pacientes con enfermedades cardiovasculares que cuenta con 14 atributos y 303 registros. Como resultado del estudio se obtuvieron mejores resultados al utilizar las máquinas de soporte vectorial como método clasificador. Solarte y Castro (2012), también presentan un estudio predictivo para el diagnóstico de enfermedades cardiovasculares. En este estudio se utilizan las redes bayesianas y los árboles de decisión para realizar tareas de predicción relacionadas con esta enfermedad. Como resultado del estudio se presenta una alternativa para determinar si se debe realizar un procedimiento clínico a un paciente basándose en variables como la presión arterial.

Hernández (2014) presenta una aplicación de la minería de datos a pacientes pre-diabéticos. El objetivo del estudio es detectar patrones de comportamiento para apoyar la toma de decisiones en futuros casos de personas con esta enfermedad. Para ello, el

estudio, extrajo información de los expedientes clínicos para descubrir y conocer las características como edad, sexo, niveles de colesterol, triglicéridos, glucosa e insulina. Para el análisis de datos se utilizó un clustering con el método K-means y una predicción con árboles de decisión.

Shaikh (2014) presenta un prototipo para la predicción de Parkinson y enfermedades tumorales primarias, utilizando técnicas de minería de datos. En este estudio se plantea una comparación de los resultados utilizando técnicas como redes neuronales, árboles de decisión y métodos bayesianos, obteniendo un mejor resultado al utilizar las redes neuronales.

Pérez (2014) presenta un estudio cuyo objetivo es realizar la predicción de la hipertensión arterial, este estudio toma importancia al conocer que el 30% de las muertes a nivel mundial son producidas por enfermedades cardiovasculares. Para realizar este estudio se tomó una muestra de 138 personas entre los 20 y 34 años de edad, la técnica utilizada en este caso fueron las redes neuronales.

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 (Mora, 2009).

En Saiz & Lau (2009) se muestra un estudio sobre análisis de datos médicos de cáncer pediátrico por medio de minería de datos. Se estudian modelos de minería de datos aplicados a datos de pacientes infantiles aquejados de cáncer. El trabajo forma parte de un sistema integral de control y seguimiento a este tipo de pacientes, para analizar de manera compleja su reacción a situaciones y estímulos. Se pretende crear las bases para una vez acumulados los datos, puedan utilizarse técnicas de minería de datos para encontrar patrones ocultos en los mismos.

En Reparaz et al. (2008) se reporta el uso de la minería de datos para determinar la eficacia de la braquiterapia en tratamiento

de cáncer de próstata. En este contexto, el objetivo del trabajo es caracterizar y clasificar la población de pacientes con cáncer de próstata mediante técnicas de minería de datos, esperando encontrar relaciones subyacentes en los datos que no pueden identificarse mediante un tratamiento estadístico clásico.

En Hernández & Lorente (2009) se describe la utilización de la herramienta de minería de datos Weka en la detección de cáncer de mama. El objetivo de este trabajo es realizar el análisis de los datos correspondientes a un conjunto de mujeres para poder obtener ciertos patrones a la hora de poder aplicarlos para detectar posibles casos de cáncer de mama mediante minería de datos.

De igual manera en Hurtado et al (2008) se presentan los resultados de aplicar técnicas de minería de datos en el estudio epidemiológico del cáncer mamario. Los resultados obtenidos indican que las diferentes técnicas empleadas, son capaces de extraer modelos consistentes con la realidad observada y el respaldo teórico, pese a que éstos no tienen conocimiento previo del área clínica, es decir no manejan información del cáncer ni su comportamiento, aun así, discriminan de manera satisfactoria grupos de pacientes con la patología y generan modelos que se adecuan a dichos datos. Igualmente, Molero, Céspedes y Campaña (2013) presentan un estudio predictivo con una caracterización y análisis de base de datos de cáncer de mama del programa de vigilancia, epidemiología y resultados finales (SEER) del Instituto Nacional de Cáncer de los Estados Unidos por medio de series de tiempo temporales. Uno de los objetivos del estudio consiste en establecer los factores que más influyen en la enfermedad. Como resultado del estudio se resalta que el cáncer de mama es una de las principales causas de muerte en mujeres menores de 30 años.

En Bellaachia & Guven (2005) se realiza un análisis de la predicción de índice de supervivencia de los pacientes con cáncer de mama usando técnicas de minería de datos. Los datos usados han sido tomados de SEER Public-Use Database y corresponde a 151.886 registros y 16 atributos. Se aplicaron tres técnicas de minería de datos: NaiveBayes, árboles de decisión y redes neuronales.

En López (2011) se diseñó un sistema para descubrir el comportamiento de los factores de riesgo de cáncer de cuello uterino utilizando minería de datos en la Secretaria 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. 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, si es gestante.

Timarán y Yépez (2012) realizaron un estudio descriptivo para descubrir patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. En este estudio se utiliza la asociación para determinar los factores socioeconómicos y clínicos asociados a la supervivencia de mujeres con esta enfermedad. El objetivo de este estudio es apoyar la toma de decisiones relacionadas con políticas públicas y programas de protección para las mujeres que padecen esta enfermedad.

CAPÍTULO IV

DESCUBRIMIENTO DE PATRONES DE SUPERVIVENCIA EN MUJERES CON CÁNCER INVASIVO DE CUELLO UTERINO

Para el descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino, se siguieron cada una de las etapas que conforman el proceso de descubrimiento de conocimiento en bases de datos, descrito en el capítulo III. Como se especificó en el capítulo I, este proyecto se delimitó a la información disponible y depurada por el RPCMP en el momento de realizar este estudio y que corresponde al periodo comprendido entre los años 1998 y 2007. En este estudio, la supervivencia de estas mujeres se tomó a cinco años, teniendo en cuenta que el RPCMP depura la información de acuerdo a la metodología internacionalmente establecida para estimar la incidencia, mortalidad y prevalencia de cáncer por quinquenios.

4.1 ETAPA DE SELECCIÓN

El objetivo de esta etapa es obtener las fuentes de datos internas y externas que sirven de base para el proceso de Minería de datos. Como fuente interna, se seleccionó la base de datos REGCANDB del aplicativo REGCAMP 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.

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 (1-6).

Otras fuentes externas que se seleccionaron para complementar u obtener información faltante se muestran en la tabla 1.

De los 17.350 casos registrados de cáncer en la base de datos REGCANDB, se seleccionaron inicialmente 3.151 registros, correspondientes a mujeres con cáncer de cuello uterino (invasivo y no invasivo), y de estos únicamente 550 casos corresponden a mujeres con cáncer invasivo de cuello uterino, objeto de esta investigación. Estos casos fueron almacenados en la base de datos CÁNCERDB, construida con el sistema gestor de base de datos PostgreSQL (Stonebraker & Rowe, 1986), (Momjian, 2001) específicamente en la tabla CÁNCER la cual consta de 36 atributos y 550 registros. Esta tabla servirá de base para las subsiguientes etapas del proceso de descubrimiento de patrones de supervivencia.

La descripción de los atributos que conforman esta tabla se puede ver en la tabla 2.

Tabla 1. Fuentes de Datos Externas

No.	NOMBRE DE LA FUENTE
1	CLÍNICA MARIDÍAZ ISS
2	SALUDCOOP
3	I.C.N.
4	PATÓLOGOS ASOCIADOS
5	HOSPITAL SAN PEDRO
6	CLÍNICA PALERMO
7	HOSPITAL DEPARTAMENTAL
8	UNIDAD RADIOLOGÍA H. DEPARTAMENTAL

No.	NOMBRE DE LA FUENTE
9	COOMEVA
10	DIRECCIÓN MUNICIPAL DE SALUD
11	EMSSANAR
12	CRUZ ROJA
13	CLÍNICA SAN JUAN
14	F. U. DE NARIÑO
15	IPS PROSALUD
16	CÓNDOR
17	CLÍNICA FÁTIMA
18	CLÍNICA LOS ANDES SALUDCOOP
19	LIGA DE LUCHA CONTRA EL CÁNCER
20	HOSPITAL CIVIL DE IPIALES
21	REGISTRADURÍA NACIONAL DEL ESTADO CIVIL
22	FOSYGA
23	PLANEACIÓN MUNICIPAL

Tabla 2. Atributos iniciales de la Tabla CÁNCER

CAMPO	TABLA CÁNCER
	DESCRIPCIÓN Y VALORES
identificacion	Número de identificación del paciente.
tipoid	Contiene el tipo de identificación del paciente: CC cédula, TI tarjeta identidad, OO otro.
papellido	Campo que contiene el primer apellido del paciente.
sapellido	Campo que contiene el segundo apellido del paciente.
apellidocasada	Campo que contiene el apellido de casada
pnombre	Campo que contiene el primer nombre del paciente.
snombre	Campo que contiene el segundo nombre del paciente.
sexo	Campo que contiene el sexo del paciente.
fechan	Fecha de nacimiento del paciente
lugarn	Lugar de nacimiento del paciente
direcciondx	Dirección del paciente en el momento del diagnóstico de cáncer
escolaridad	Escolaridad del paciente: primaria, secundaria, profesional, ninguna, sin dato

CAMPO	TABLA CÁNCER
	DESCRIPCIÓN Y VALORES
estadocivil	Estado civil del paciente en el momento del diagnóstico: casado, viudo, unión libre, soltero, divorciado/separado, desconocido
ocupacion	Ocupación del paciente en el momento del diagnóstico: hogar, servicios domésticos, comerciante, modista, empleada, docente, desconocido
discapacidad	Discapacidad que tenga el paciente: sordera, ceguera, dificultad para movilizarse por sí mismo, ninguna
metododx	Método diagnóstico del cáncer: Solamente clínico, histología de tumor primario, imagenealogía, citología, hematología, certificado de defunción. Este último método de diagnóstico se debe a que la paciente nunca fue diagnosticada en un hospital o clínica y únicamente este caso fue reportado por medicina legal como causa de su muerte.
diag_principal_rips	Diagnóstico principal de la última consulta del paciente
régimen	Régimen al cual pertenece el paciente en el momento del diagnóstico. Los posibles valores son: Contributivo, subsidiado, particular, vinculado, sin dato
numerohijos	Número de hijos del paciente
parentescojefe	El tipo de parentesco del paciente con respecto al núcleo familiar
fechadefuncion	Fecha de fallecimiento del paciente
fechadx	Almacena la fecha de diagnóstico de la neoplasia maligna.
fecha_ultimoc	Fecha último contacto con el paciente
fecha_consulta_rips	Fecha de la última consulta realizada por el paciente al sistema de salud
fuelle	Organización fuente de diagnóstico: hospital, clínica, centro de salud, laboratorio, IPS, EPS, certificado defunción, desconocido

CAMPO	TABLA CÁNCER
	DESCRIPCIÓN Y VALORES
Morfología	Morfología del tumor de cáncer de cuello uterino Adenocarcinoma NOS, Carcinoma NOS, Clear cell adenocarcinoma NOS, Squamous cell carcinoma, keratinizing NOS, Squamous cell carcinoma microinvasive, Squamous cell carcinoma NOS, Squamous cell carcinoma small cell nonkeratinizing, Squamous cell carcinoma large cell, nonkeratinizing NOS, Endometrioid adenocarcinoma NOS, Adenosquamous carcinoma, Malignant tumor small cell type, Malignant lymphoma small B lymphocytic NOS
locesp	Localización específica del cáncer: cuello uterino SAI, Exocérvis, Endocérvis
cirugía	Determina si al paciente se le ha realizado alguna cirugía: si, no, desconocido
radio	Determina si al paciente ha recibido tratamiento de radioterapia: si, no, desconocido
quimio	Determina si al paciente ha recibido tratamiento de quimioterapia: si, no, desconocido
hormono	Determina si al paciente ha recibido tratamiento de hormonoterapia: si, no, desconocido
inmuno	Determina si al paciente ha recibido tratamiento de inmunoterapia: si, no, desconocido
braqui	Determina si al paciente ha recibido tratamiento de braquiterapia: si, no, desconocido
paliativo	Determina si al paciente ha recibido tratamiento de paliativo: si, no, desconocido
biopsia	Determina si al paciente se le realizó una biopsia: si, no
labbio	Laboratorio donde se realizó la biopsia

4.2 ETAPA DE PREPROCESAMIENTO DE DATOS

El objetivo de esta etapa es obtener datos limpios, i.e. datos sin valores nulos o anómalos que permitan obtener patrones de calidad. Por medio de consultas 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 CÁNCER. En las figuras 3 y 4 se presenta una muestra del resultado del análisis de la calidad de los datos, específicamente del atributo escolaridad y estado civil respectivamente. En la tabla 3 se muestra los atributos de la tabla CÁNCER con un alto porcentaje de valores nulos.

Con el fin de obtener datos limpios para aquellos atributos que se consideraron relevantes para la investigación, se reemplazaron sus valores nulos con los valores que se obtuvieron al realizar un cruce con las fuentes externas de datos. En la tabla 4 se muestra este proceso y la técnica utilizada para reemplazar los valores nulos.

Como resultado de la limpieza de datos, se detectaron que 43 registros no pertenecen a casos de cáncer de cuello uterino de mujeres del municipio de Pasto, por lo tanto, se tuvieron que eliminar, quedando así la tabla CÁNCER con 507 registros válidos.

4.3 ETAPA DE TRANSFORMACIÓN

La transformación de datos engloba cualquier proceso que modifique la forma de los datos. El objetivo de esta etapa es transformar la fuente de datos en un conjunto de datos listo para aplicar las diferentes técnicas de minería de datos. Entre las operaciones que se realizan para transformar los datos están: eliminación de los atributos menos relevantes, creación de nuevos atributos derivándolos de otros (manteniendo o sustituyendo esos atributos), cambiando el tipo de los atributos (discretizándolos o numerizándolos).

Con el fin de relacionar los factores sociales y económicos que pueden incidir en la supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino, fue necesario crear en la tabla CÁNCER nuevos atributos a partir de otros atributos ya existentes. En la tabla 5 se describen estos nuevos atributos y las acciones que se realizaron para su creación.

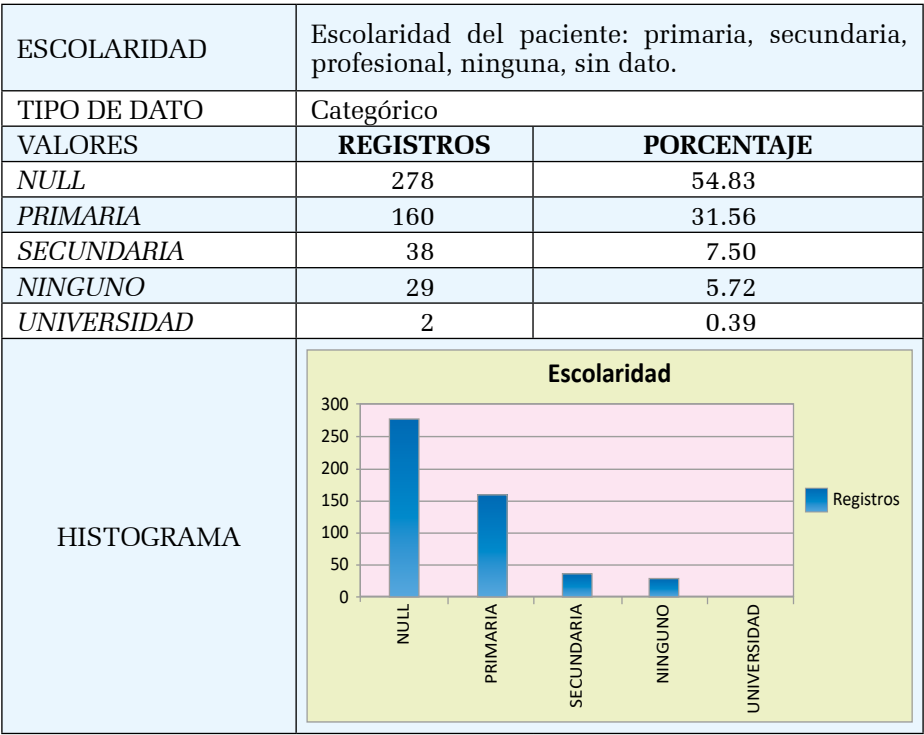


Figura 3. Calidad de datos del atributo ESCOLARIDAD

Tabla 3. Atributos Tabla CÁNCER con un alto porcentaje de nulos

ATRIBUTO	VALORES NULOS	PORCENTAJE %
BRAQUI	550	100.00
PALIATIVO	550	100.00
HORMONO	546	99.27
INMUNO	546	99.27
QUIMIO	487	88.55
CIRUGIA	410	74.55
APELLIDOCASADA	367	66.73
RADIO	366	66.55
FECHADEFUNCION	300	54.55
DIRECCIONDX	286	52.00
DISCAPACIDAD	286	52.00
PARENTESCOJEFE	286	52.00

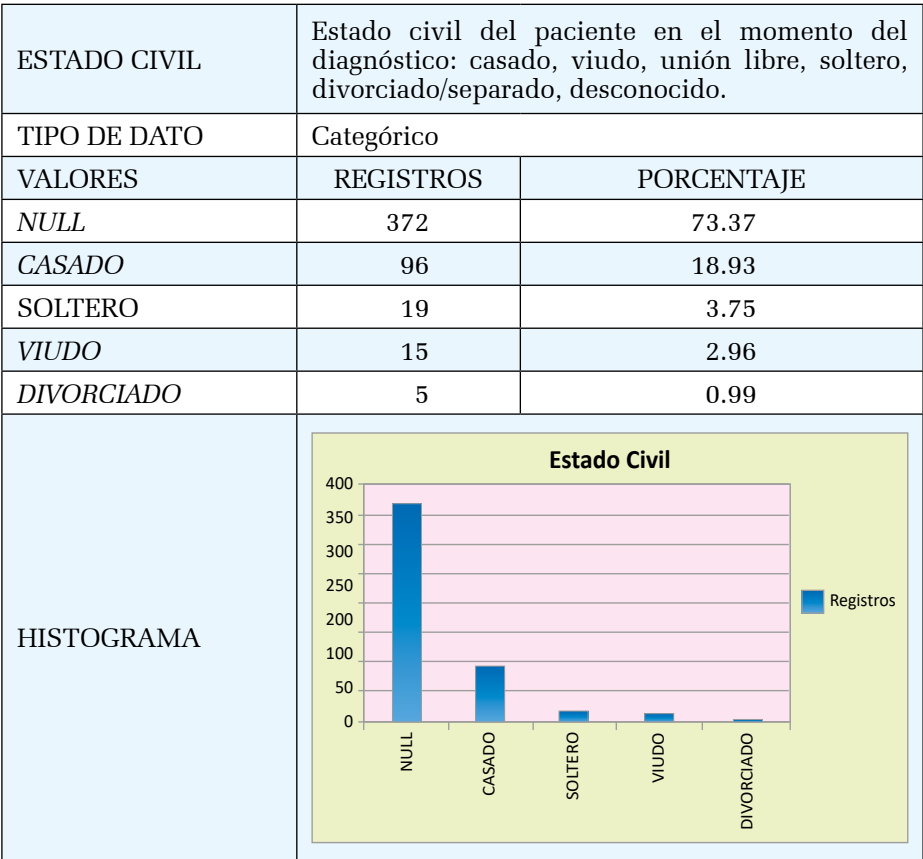


Figura 4. Calidad de datos del atributo ESTADO CIVIL

Tabla 4. Limpieza de Atributos de la Tabla CÁNCER

ATRIBUTO	TÉCNICA DE LIMPIEZA
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

Tabla 5. Nuevos Atributos de la Tabla CÁNCER

ATRIBUTO	DESCRIPCIÓN	ACCIÓN
BARRIO	Barrio de residencia del paciente en el municipio de Pasto en el momento del diagnóstico del cáncer.	Se extrae del atributo DIRECCIONDX
COMUNA	Comuna del municipio de Pasto a la cual pertenece el barrio. Los posibles valores son comuna 1 hasta la comuna 12 y sector rural.	Se obtiene a partir del BARRIO y se ubica en la base de datos de Planeación Municipal
ESTRATO	Estrato socioeconómico al cual pertenece el paciente en el momento del diagnóstico. Valores :1-bajo bajo, 2-bajo, 3-medio bajo, 4- medio medio, 5- medio alto, 6- alto y 9-desconocido	Se obtiene a partir del atributo BARRIO, se ubica en la base de datos de Planeación Municipal
REGION	Región de nacimiento del paciente	Se obtiene a partir del atributo LUGARN agrupando los municipios en regiones
EDADDX	Edad del paciente en el momento del diagnóstico de cáncer	Se obtiene a partir de los atributos FECHAN y FECHADX
CABEZAFAMILIA	Determina si el paciente es cabeza de familia o no.	Se obtiene a partir del atributo PARENTESCOJEFE
TIPOVIVIENDA	Tipo de vivienda que habita el paciente: casa, apartamento, cuarto, otro tipo de vivienda	Se obtiene a partir del cruce de información con la base de datos del SISBEN
FUENTEDEAGUA	Sistema de obtención de agua del paciente en su vivienda: acueducto, pozo, río o similares.	Se obtiene a partir del cruce de información con la base de datos del SISBEN
PUNTAJESISBEN	Es el puntaje obtenido por el paciente para su clasificación en el SISBEN	Se obtiene a partir del cruce de información con la base de datos del SISBEN
NIVELISISBEN	Es el nivel de clasificación del SISBEN	Se obtiene a partir del atributo PUNTAJESISBEN y el cruce de información con la base de datos del SISBEN
VIVOMUERTO	Determina si el paciente está vivo o muerto así: 1 – muerto, 2-vivo	Se obtiene a partir de los atributos FECHA_DEFUNCIÓN para los muertos y FECHA_CONSULTA_RIPS o FECHA_ULTIMOC para los vivos

ATRIBUTO	DESCRIPCIÓN	ACCIÓN
NMESES	Número de meses de vida del paciente desde el momento del diagnóstico de cáncer hasta la fecha de defunción o hasta el año de la ventana de observación. Este atributo toma el valor de 0 en el caso de que su muerte fue reportada por medicina legal y nunca la paciente fue diagnosticada con cáncer	Se obtiene a partir de los atributos FECHADX y FECHADEFUNCION para muertos y FECHA_CONSULTA_RIPS o FECHAULTIMOC para vivos

Para asignar los valores de los nuevos atributos de la tabla CÁNCER, se procedió de la siguiente manera:

Para los valores del atributo COMUNA, se tomaron de Planeación Municipal, los barrios y corregimientos del municipio de Pasto que pertenecen a cada una de ellas. Esta organización se presenta en la tabla 6.

Tabla 6. Valores del Atributo COMUNA de acuerdo a los barrios que agrupa

COMUNA	BARRIOS	COMUNA	BARRIOS
COMUNA 1	Avenida Santander Bombona Caracha Centro Condominio Santiago El Churo El Cilindro El Parque El Portalito Hullaguanga La Pendería Las Américas Los Dos Puentes Marcos De La Rosa San Agustín Centro San Andrés San Andresito San Andresito San José San José Obrero Santiago	COMUNA 2	Aire Libre Aalhambra Atahualpa Avenida Boyacá Avenida Champagnat Avenida Colombia Bella Vista Casa Bella Coliseo Cubierto El Olivo El Prado El Recuerdo Fátima Javeriano Julián Bucheli La Gran Colombia Las Lunas I Las Lunas II Las Violetas I Las Violetas II Las Violetas III

COMUNA	BARRIOS	COMUNA	BARRIOS
			Las Violetas IV Los Abedules Los Álamos Los Balcones Los Olivos Medardo Bucheli Navarrete Normandía Parque Bolívar Salomón San Miguel Sector San Juan Bosco Villa Lucía
COMUNA 3	Alejandría Arnulfo Guerrero Calcedonia Camilo Torres Casa Loma El Ejido Guamuez José Antonio Galán La Esmeralda La Estrella Las Brisas Las Lajas Las Mercedes Los Pinos Mercedario Pie de Cuesta Pinar del Río Popular Pucalpa I Pucalpa II Pucalpa III Rosal del Oriente Santa Bárbara Santa Catalina Santa Mónica Villa Oriente Villaflores I Villaflores II	COMUNA 4	Albergue del Sol Altos de Lorenzo Altos del Campo Avenida Idema Belén Bernal Betania Chile Doce de Octubre I Doce de Octubre II El Porvenir El Rosario El Tejar El Triunfo La Paz Laureano Gómez Lorenzo de Aldana Los Eliseos Miraflores I Miraflores II Praga Puerta del Sol Rincón Colonial San Germán San Juan de los Pastos Santa Fe I Santa Fe II Santacruz Sendoya Siete de Agosto Villa Olímpica Villa Victoria Villa Docente

COMUNA	BARRIOS	COMUNA	BARRIOS
COMUNA 5	Altos de Chapalito I Altos de Chapalito II Altos de Chapalito III Antonio Nariño Cantarana Chambú I Chambú II Chapal Chapal II El Pilar El Progreso El Remanso Emilio Botero I Emilio Botero II Emilio Botero III Emilio Botero IV La Minga La Rosa La Vega Las Ferias Los Cristales Los Robles Madrigal María Isabel I María Isabel II María Isabel III Potrerillo Prados del Sur Salida al Sur-Urbano San Martín Santa Clara Venecia Villa del Río Vivienda Cristiana	COMUNA 6	Agualongo Altamira Bachué Caicedo Ciudad Jardín Cooperativa Popular Nariñense El Estadio El Libertador Gilberto Pabon Granada I Granada II Granada III Granada IV Inem Jerusalén La Cruz La Palma Los Fundadores Luis Carlos Galán Majitayo II Mijitayo Mijitayo Bajo Mirador De Niza Niza I Niza II Niza III Nueva Colombia Prados del Niza Quillacinga Quito López I Quito López II Quito López III San Carlos San Miguel de Jongovito San Sebastián Santa Isabel Santanita Sumatambo Tamasagra I Tamasagra II Tequendama Valle Lili Villa de los Ríos
COMUNA 7	Achalay Avenida Panamericana Capusigra Castillos del Norte Conjunto El Parque El Bosque El Edén El Rincón de la Aurora	COMUNA 8	Altamira - San Vicente Altavista Altos de la Colina Anganoy Arcos Iris Avenida Panamericana Balcones de Mariluz Bello Horizonte Colón

COMUNA	BARRIOS	COMUNA	BARRIOS
	El Rincón de la Panamericana Francisco de la Villota La Aurora La Primavera Las Acacias Las Camelias Los Andes Los Hexágonos Rosales II Rosales I San Felipe San Ignacio Santa María Villa Aurora Villa Campanela Villa Sofía Villa Vergel		Colpatria Conjunto San Diego El Remanso del Norte Gualcaloma Jorge Giraldo La Castellana La Cuesta La Pradera Las Margaritas Los Frailejones Los Héroes Los Laureles Los Urapanes Mariluz I Mariluz II Mariluz III Mirador de San Juan Miravalle Montemor Panamericano Panamericano II Panamericano I Portal de la Colina Prados del Oeste Quintas de San Pedro Remansos del Norte Salazar Mejía San Diego San Juan de Anganoy San Juan de Dios I San Juan de Dios II San Pedro San Vicente Sindamanoy Torres de Pubenza Veracruz Villa Jardín Villas de San Rafael
COMUNA 9	Alto Juanoy Av. los Estudiantes Briceño Calatrava Camino Real Castilla Chapultepec Colonial Colonial	COMUNA 10	Aranda Avenida Aranda Avenida Oriental Avenida Oriental Río Pasto Buenos Aires Cementerio Condominio Bellavista

COMUNA	BARRIOS	COMUNA	BARRIOS
	Colonial Condominio Morasurco Conjunto Torobajo El Aljibe El Cerámico El Dorado El Mirador El Polvorín El Refugio Figueroa José Ignacio Zarama Juan XXIII Juanoy La Colina La Riviera Las Cuadras La Victoria Los Nogales Los Sauces Luis Brand Manacá Maridíaz Marsella Morasurco Nuevo Amanecer Palermo Pandiaco Paraná Parque Infantil Pinos del Norte Portal de la Colina Riviera San Antonio de Juanoy Santa Ana Santa Rita Sañudo Tequendama Terranova Terrazas de Briceño Titán Torobajo Universitario Urbanización Zarama Valle de Atriz Versalles Villa María Villa Campestre		Corazón de Jesús II Cra. 27A del Barrio Centenario Destechados El Futuro El Portal de Aranda El Portal del Norte El Rincón de Pasto El Rincón del Rosario Juan Pablo II La Esperanza La Floresta Libertad Loma del Carmen Marquetalia Niño Jesús de Praga Nueva Aranda Nuevo Horizonte Nuevo Sol Ocho de Marzo Prados del Norte Quebarada Gallinacera 2da Parte Quillotocito Rincón de Aranda Río Blanco Río Blanco que Termina Ojo de Agua San Albano Santa Matilde Sector Pedagógico Avenida Oriental Sol de Oriente Termina Sector Peatonal 28 Tescual Villa del Rosario Villa Guerrero Villas del Norte

COMUNA	BARRIOS	COMUNA	BARRIOS
COMUNA 11	Alameda Alameda II Aquine I Aquine II Aquine III Aquine Alto Aquine IV Aquine Reservado Belalcázar Centenario Chico Ciudad Real Corazón de Jesús El Calvario El Común El Corralito Hospital Civil La Floresta La Lomita Los Alcázares Rincón del Paraíso Salsipuedes Villa Elena Villa Jazmín	COMUNA 12	Altos de la Carolina Balcones del Este Campaña de Oriente Carlos Pizarro Cujacal Bajo El Manantial El Paraíso Fray Ezequiel Moreno Díaz Gualcalá La Carolina La Florida La Josefina Las Orquídeas María Paz María Paz II Montserrat Parque de Baviera San Diego Norte San Mateo Sena Simón Bolívar Sindagua Viila Colombia Villa Adriana María Villa Ángela Villa del Prado Villa Recreo Villa Rocío
RURAL	Buesaquillo Cabrera Catambuco El Encano El Socorro Genoy Gualmatán Jamondino Jongovito La Caldera La Laguna Mapachico Mocondino Morasurco Obonuco San Fernando Santa Bárbara		

Para los valores del atributo ESTRATO, se tomaron de Planeación Municipal los valores aplicables al municipio de Pasto. Estos valores y su descripción se muestran en la tabla 7.

Tabla 7. Valores del Atributo ESTRATO

ESTRATO	DESCRIPCIÓN
1	Bajo bajo
2	Bajo
3	Medio bajo
4	Medio medio
5	Medio alto
6	Alto
9	Desconocido

Los valores para el atributo REGIÓN y la forma como se asignan de acuerdo a los diferentes municipios de Nariño, Putumayo y otros departamentos de Colombia se muestran en la tabla 8.

Tabla 8. Valores del Atributo REGIÓN de acuerdo a los municipios que agrupa

REGIÓN	MUNICIPIOS
NORTE	Arboleda, Belén, Colón, El Rosario, El Tablón de Gómez, La Cruz, La Unión, Leiva, Policarpa, San Bernardo, San José de Albán, San Lorenzo, San Pablo, San Pedro de Cartago y Taminango.
SUR	Aldana, Contadero, Córdoba, Cumbal, Cuaspud, Funes, Guachucal, Gualmatán, Iles, Ipiales, Potosí, Puerres y Pupiales
PUTUMAYO	Colon, Orito, Puerto Asís, Puerto Caicedo, Puerto Guzmán, Puerto Leguizamo, San Francisco, San Miguel, Santiago, Sibundoy, Villagarzon,
PACIFICA	Tumaco, Barbacoas, Olaya Herrera, El Charco, La Tola, Magui Payán, Santa Bárbara, Pizarro, Mosquera, Roberto Payán
OCCIDENTAL ANDINA	Ancuya, Cumbitara, Guaitarilla, Imúes, La Llanada, Los Andes Sotomayor, Linares, Mallama, Ospina, Providencia, Ricaurte, Samaniego, Santacruz, Sapuyes, Túquerres.
CENTRO	Buesaco, Chachagüí, Consacá, El Peñol, El Tambo, Nariño, La Florida, Sandoná, Tangua y Yacuanquer.
PASTO	San Juan de Pasto
OTROS	Municipios de otros departamentos de Colombia

En algunos casos, por conveniencia, es necesario convertir los valores de un atributo numérico a valores nominales, proceso que se le conoce con el nombre de discretización (Hernández, Ramírez, & Ferri, 2005). Por esta razón, teniendo en cuenta los diferentes valores numéricos del atributo EDADDX, estos se discretizaron

a valores nominales, teniendo en cuenta intervalos del mismo tamaño y utilizando el máximo y mínimo como referencia.

En la discretización del atributo EDADDX se utilizaron seis intervalos del mismo tamaño (Ti) los cuales se calcularon teniendo en cuenta la diferencia entre el valor máximo (Vmax) de 93 años y el valor mínimo (Vmin) de 15 años y este resultado dividiéndolo entre el número de intervalos (Ni) igual a 6 así:

$$\begin{aligned} Ti &= (Vmax - Vmin) / Ni \\ Ti &= (93 - 15) / 6 \\ Ti &= 13 \end{aligned}$$

Los resultados de este proceso se muestran en la tabla 9.

Tabla 9. Valores del Atributo EDADDX

EDADDX	RANGO
1	15-28
2	29-41
3	42-54
4	55-67
5	68-80
6	81-93

Los valores del atributo NIVELSISBEN dependen del puntaje que haya obtenido la persona en la respectiva encuesta y que están registrados en el atributo PUNTAJESISBEN. Los niveles y la forma como se asignan se muestran en la tabla 10.

Tabla 10. Valores del Atributo NIVELSISBEN de acuerdo a los puntajes obtenidos

Nivel Sisben	Puntaje Zona Urbana	Puntaje Zona Rural
1	0-36	0-18
2	37-47	19-30
3	48-58	31-45
4	59-69	46-61
5	70-86	62-81
6	87-100	82-100
7	SIN SISBEN	SIN SISBEN

El atributo FUENTE especifica la organización donde se diagnosticó al paciente con cáncer de cuello uterino. En la tabla 11 se especifica cómo se agruparon estas entidades.

Tabla 11. Valores del Atributo FUENTE de acuerdo al tipo de entidad

Fuente	Entidades
Organizaciones	Dirección Municipal de Salud Liga de Lucha Contra el Cáncer Cruz Roja I.C.N.
Hospitales	Hospital Departamental Hospital San Pedro Clínica Fátima Clínica Palermo Clínica los Andes Clínica San Juan Clínica Maridíaz ISS
Eps	Fondo de Salud de la Universidad de Nariño Emssanar Salud Cóndor Saludcoop Ips Prosalud Cooameva
Particular	Consultorio Particular Laboratorio Particular Laboratorio de Referencia Patólogos Asociados
Medicina Legal	Certificado de Defunción

Dado el sinnúmero de morfologías se seleccionaron únicamente las de cuello uterino. En la tabla 12 se muestran estas morfologías.

Tabla 12. Morfologías de cáncer de cuello uterino

Tipo	Morfología de Cuello Uterino
Morfología	Adenocarcinoma, NOS Carcinoma, NOS Clear cell adenocarcinoma, NOS Squamouscell carcinoma, keratinizing, NOS Squamous cell carcinoma, microinvasive Squamous cell carcinoma, NOS Squamous cell carcinoma, small cell, nonkeratinizing Squamous cell carcinoma, large cell, nonkeratinizing, NOS Endometrioid adenocarcinoma, NOS Adenosquamous carcinoma Malignant tumor, small cell type Malignant lymphoma, small B lymphocytic, NOS

Finalmente, teniendo en cuenta los métodos de reemplazo de valores faltantes propuestos por Hernández, Ramírez, & Ferri (2005), los valores nulos de los atributos DISCAPACIDAD, RADIO y CIRUGÍA por ser valores nominales (S/N) fueron reemplazados por la moda.

Como resultado de este proceso de transformación se eliminaron de la tabla CÁNCER, la llave primaria que hace único cada registro, 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. Los atributos eliminados se muestran en la tabla 13.

Tabla 13. Atributos Tabla CÁNCER eliminados

ATRIBUTO	CAUSA
IDENTIFICACIÓN	Llave primaria
TIPOID	No relevante
PAPELLIDO	No relevante
SAPELLIDO	No relevante
APELLIDOCASADA	No relevante
PNOMBRE	No relevante
SNOMBRE	No relevante
SEXO	Son todas mujeres
LUGARN	Se utilizó para crear el atributo REGIÓN
DIRECCIONDX	Se utilizó para crear el atributo BARRIO
BARRIO	Se utilizó para crear los atributos ESTRATO y COMUNA
FECHAN	Se utilizó para crear el atributo EDADDX
FECHADX	Se utilizó para crear los atributos EDADDX y NMESES
PARENTESCOJEFE	Se utilizó para crear el atributo CABEZAFAMILIA
PUNTAJESISBEN	Se utilizó para crear el atributo NIVELSI-BEN
FECHADEFUNCION	Se utilizó para crear los atributos VIVO MUERTO y NMESES

ATRIBUTO	CAUSA
FECHA_CONSULTA_RIPS	Se utilizó para crear los atributos VIVO MUERTO y NMESES
FECHA_ULTIMOC	Se utilizó para crear los atributos VIVO MUERTO y NMESES
DIAG_PRINCIPAL_RIPS	No relevante
HORMONO	99.27 % valores nulos
INMUNO	99.27 % valores nulos
BRAQUI	100 % valores nulos
PALIATIVO	100% valores nulos

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 almacenados en el Registro Poblacional de Cáncer del municipio de Pasto en el periodo 1998-2007, se seleccionaron, de la tabla CÁNCER, los 507 casos correspondientes a este periodo y los atributos más representativos para cada factor y se crearon en la base de datos CÁNCERDB, nuevas tablas con estos atributos y registros. Por otra parte, de los 507 casos, se escogieron 235 que pertenecen a mujeres diagnosticadas con cáncer de cuello uterino en el periodo comprendido entre 1998 y 2002 y cuyo seguimiento u observación se les hizo hasta 2007 y se creó con estos una nueva tabla. La descripción de las tablas que conforman la base de datos CÁNCERDB se presenta en la tabla 14.

Tabla 14. Descripción de tablas de la base de datos CÁNCERDB

No.	TABLAS	DESCRIPCIÓN
1	CÁNCER	Tabla inicial con 540 casos de cáncer de cuello uterino y 36 atributos iniciales, de las mujeres diagnosticadas en el periodo 1998-2007
2	T507A22ALL	Tabla con los 507 casos de cáncer de cuello uterino (334 vivos y 173 muertos) registrados en el periodo 1998-2007 y los 22 atributos a considerar en el estudio.
3	T507A15ECON	Tabla de todos los 507 casos de cáncer de cuello uterino y los 15 atributos a considerar para los factores sociales y económicos.

No.	TABLAS	DESCRIPCIÓN
4	T507A09CLINICO	Tabla de todos los 507 casos de cáncer de cuello uterino y los 9 atributos a considerar para los factores clínicos
5	T235A22ALL	Tabla de los 235 casos de cáncer de cuello uterino con los 22 atributos a considerar de las mujeres diagnosticadas en el periodo 1998-2002 observadas hasta el 2007
6	T235A15ECON	Tabla de todos los 235 casos de cáncer de cuello uterino y los 15 atributos a considerar para los factores sociales y económicos de las mujeres diagnosticadas en el periodo 1998-2002 observadas hasta el 2007
7	T235A09CLINICO	Tabla de todos los 235 casos de cáncer de cuello uterino y los 9 atributos a considerar para los factores clínicos de las mujeres diagnosticadas en el periodo 1998-2002 observadas hasta el 2007

La descripción de los 22 atributos que se consideraron en esta investigación y que forman parte de las tablas T507A22ALL y T235A22ALL se muestran en la tabla 15. Los 15 atributos seleccionados para el estudio de los factores sociales y económicos y que forman parte de las tablas T507A15ECON y T235A15ECON se enumeran en la tabla 16. Los 9 atributos seleccionados para el estudio de los factores clínicos y que forman parte de las tablas T507A09CLINICO y T235A09CLINICO se presentan en la tabla 17.

Tabla 15. Atributos seleccionados para la investigación

TABLA T507A22ALL	
CAMPO	DESCRIPCIÓN Y VALORES
region	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

TABLA T507A22ALL	
CAMPO	DESCRIPCIÓN Y VALORES
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
fuelleagua	Si la residencia cuenta con servicio de agua
discapacidad	Si el paciente tiene una discapacidad o no
fuelle	Organización donde se diagnosticó el cáncer de cuello uterino
metododx	Método utilizado para el diagnóstico del cáncer
morfologia	Morfología del tumor.
loresp	Localización específica del tumor
radio	Existencia o no de tratamiento de radioterapia al paciente.
cirugia	Determina si el paciente ha tenido como tratamiento cirugía o no
biopsia	Determina si al paciente se le realizó o no una biopsia
nmeses	Número de meses de vida de la paciente desde el momento del diagnóstico hasta su muerte o hasta la ventana de observación. El valor mínimo es 0 si fue diagnosticado por certificado de defunción expedido por Medicina Legal y nunca fue diagnosticada con cáncer.
vivomuerto	Determina si la paciente sobrevivió o no hasta la ventana de observación

Tabla 16. Atributos socio-económicos para la investigación

TABLA T507A15ECON	
CAMPO	DESCRIPCIÓN Y VALORES
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.

CAMPO	DESCRIPCIÓN Y VALORES
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
cabezaflia	Determina si el paciente es cabeza de familia o no.
tipovivienda	Tipo de vivienda que habita el paciente
fuelleagua	Si la residencia cuenta con servicio de agua
discapacidad	Si el paciente tiene una discapacidad o no
nmeses	Número de meses de vida del paciente desde el momento del diagnóstico
vivomuerto	Determina si el paciente está vivo o muerto

Tabla 17. Atributos clínicos para la investigación

TABLA T507A09CLINICO	
CAMPO	DESCRIPCIÓN Y VALORES
fuelle	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

4.4 ETAPA DE MINERÍA DE DATOS

El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés aplicando tareas de descubrimiento tales como clasificación (Quinlan, 1986), (Wang, Iyer and Scott, 1998), (Witten and Frank, 2000), clustering (Ng and Han, 1994), (Zhang, Ramakrishnan and Livny, 1996), patrones secuenciales (Agrawal and Srikant, 1995), asociaciones (Agrawal and Srikant, 1994), (Srikant and Agrawal, 1996) entre otras.

Teniendo en cuenta que el objetivo de esta investigación es descubrir patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino a partir de los datos registrados en el RPCMP en el periodo 1998-2007, se exploraron y evaluaron tanto tareas de minería de datos predictivas como descriptivas. Las predictivas se aplican a problemas en los que hay que predecir nuevos datos para uno o más ejemplos que van acompañados de una salida denominada clase (Hernández, Ramírez & Ferri, 2005). Para este tipo de tarea se evaluó la tarea de clasificación utilizando la técnica de árboles de decisión. Las descriptivas se aplican a problemas en los que hay que describir los valores de los datos existentes (Hernández, Ramírez & Ferri, 2005). Se evaluaron las tareas de reglas de asociación utilizando el algoritmo A priori (Agrawal, Imielinski & Swami, 1993) y de agrupamiento o clustering con el algoritmo k-means. 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, asociación y clustering. Con clasificación se pretende obtener un modelo que permita predecir para los nuevos casos de cáncer invasivo de cuello uterino su supervivencia. Con asociación se pretende identificar cuales factores socioeconómicos y clínicos están sucediendo frecuentemente juntos y que están relacionados con la supervivencia de este tipo de pacientes y con agrupación se pretende obtener grupos homogéneos de mujeres que padecen este tipo de enfermedad, con características similares en lo relacionado a la supervivencia. Se debe esperar que los resultados sean diferentes por el fin que se pretende obtener en cada tarea.

Se escogió la herramienta Weka (Waikato Environment for Knowledge Analysis) para realizar estas tareas (Garcia, s.f). 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. Weka es una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años.

A continuación, se describe la fase de minería de datos, en la cual se detectaron patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino a partir de los datos registrados en el RPCMP en el periodo 1998-2007, objeto de esta investigación, aplicando las tareas de clasificación, asociación y clustering. Como trabajo adicional, se realizó el mismo proceso, pero con los datos registrados en el RPCMP en la cohorte 1998-2002, con el fin de comparar los resultados obtenidos en este quinquenio, con los obtenidos en los dos quinquenios en conjunto. Este proceso se encuentra en la sección 5.5.

4.4.1 Tarea de Clasificación

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 (Hernández, Ramírez & Ferri, 2005). Tomando como clase los valores del atributo vivomuerto del conjunto de datos de todos los casos de mujeres con cáncer de cuello uterino, se construyó un modelo de clasificación que determinó las características de las pacientes que sobrevivieron al cáncer y las que no.

4.4.1.1 Parámetros de Evaluación del Modelo de Clasificación

La técnica de clasificación utilizada para el descubrimiento de patrones de supervivencia en mujeres con cáncer de cuello uterino 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 (Hang & 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).

El algoritmo de la herramienta Weka utilizado para la obtener el modelo de clasificación con árboles de decisión fue J48, el cual implementa al algoritmo C.45, (Quinlan, 1993). 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 (Hernández & Lorente, 2009). El parámetro más importante que se debe tener en cuenta para la poda es el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Para cada operación de poda, define la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños (García & Álvarez, 2010). Otra forma de variar el tamaño del árbol es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol, si bien es menos importante puesto que depende del número absoluto de instancias en el conjunto de datos de partida (Witten, et al., 2011).

Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba. Existen diferentes medidas de evaluación del clasificador en Weka (Hall, Frank & Witten, 2011):

- Usar el conjunto de datos de entrenamiento (Use training set): se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (esta técnica puede ser muy buena para ese conjunto de datos, pero puede ser poco precisa para nuevos datos).
- Proveer un conjunto de datos de prueba (Supplied test set): se emplea un conjunto de datos para entrenar y otro conjunto independiente al universo de los datos con los que se está trabajando para prueba (corriendo el riesgo que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplearon para entrenar el modelo).
- Porcentaje de Partición (Percentage Split): se emplea un % aleatorio de datos para entrenar y otro % para probar, este método difiere del anterior en que ambos conjuntos pertenecen al universo de datos con el que se está trabajando por lo que se elimina el riesgo que corre el anterior.
- Validación cruzada (Cross validation): este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición (Hernández, Ramirez y Ferri, 2005). Para este caso particular se utiliza el método de evaluación validación cruzada con n pliegues (n -fold cross validation). Esta es la opción por defecto y la más comúnmente utilizada. Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (folds) de forma aleatoria. El número de subconjuntos se puede introducir en el campo Folds. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes $n-1$ (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones.

Otra ventaja de la validación cruzada es que la varianza de los n errores de muestra parciales, permite estimar la variabilidad

del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (10-fold cross validation) (Hernández, Ramirez & Ferri, 2005).

Por otra parte, es bastante sencillo evaluar o estimar el coste de un clasificador para un determinado conjunto de ejemplos si se dispone de la matriz de confusión. La matriz de confusión (Confusion Matrix) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i , $i = 1 \dots n$ constituyen el número de instancias que realmente pertenecen a la clase i . Similarmente la sumatoria de los ejemplos o registros en cada columna j , $j = 1 \dots n$ son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y el resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra) (Fernández, 2009).

Teniendo en cuenta los parámetros de evaluación anteriores y los repositorios de datos descritos en la tabla 14 se procedió a construir los diferentes árboles de decisión con el algoritmo J48.

4.4.1.2 Descubrimiento de Patrones con Árboles de Decisión

Con el fin de detectar los patrones más confiables de supervivencia de las mujeres con cáncer invasivo de cuello uterino utilizando árboles de decisión, para todos los casos registrados entre 1998 y 2007, se fijó como parámetro de confiabilidad del modelo una precisión por encima del 60%. Con este fin, se realizaron diferentes pruebas para obtener los mejores parámetros de poda y por ende obtener el árbol con el mayor número y porcentaje de instancias correctamente clasificadas. Para realizar estas pruebas se utilizó la librería RWeka del paquete estadístico R que contiene el algoritmo J48, por la facilidad de generar y presentar reportes de tipo pdf.

Se utilizó el repositorio T507A22ALL, descrito en la tabla 15, para descubrir los factores socioeconómicos y clínicos que inciden en la supervivencia de estas pacientes. Se escogió como clase, el atributo vivomuerto, cuyos valores se distribuyen en 334 para vivos y 173 para muertos. Para evaluar la calidad del modelo y su validez se escogió el método de validación cruzada y específicamente la validación cruzada con 10 pliegues por los

mejores resultados que se obtienen. Las pruebas realizadas fueron las siguientes:

1. Prueba 1. Se mantuvo constante el número de registros por nodo M y variable el factor confianza C . El factor M se mantuvo en 10 y se varió el factor C en el rango entre $C=0.1$ a $C=0.5$. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros $M=10$ y $C=0.3$ fue el mejor con un porcentaje de 84.4181 % de instancias correctamente clasificadas correspondiente a 428 instancias de 507. En la figura 5 se muestran los mejores resultados de esta prueba con los parámetros $M=10$ $C=0.3$.
2. Prueba 2. Se continuó con el mismo criterio que la prueba anterior, es decir, se mantuvo constante el número de registros por nodo M y variable el factor confianza C . El factor M se mantuvo ahora en 5 y se varió el factor C en el rango entre $C=0.1$ a $C=0.5$. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros $M=5$ y $C=0.2$ fue el mejor con un porcentaje de 85.4043 % de instancias correctamente clasificadas correspondiente a 433 instancias de 507. En la figura 6 se presentan los mejores resultados de esta prueba con los parámetros $M=5$ $C=0.2$.
3. Prueba 3. Se estableció como parámetro constante el factor confianza C en 0.2 y se varió el factor M de 2 en 2 desde $M=2$ hasta $M=20$. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros $M=2$ y $C=0.2$ fue el mejor con un porcentaje de 86.1933 % de instancias correctamente clasificadas correspondiente a 437 instancias de 507. En la figura 7 se muestran los mejores resultados de esta prueba con los parámetros $M=2$ $C=0.2$.
4. Prueba 4. Se variaron tanto el factor confianza C como el número de registros por nodo M . Se estableció el parámetro C de 0.1 hasta 0.5 incrementando en 0.1 y M de 2 hasta 20 con un incremento de 2. De acuerdo a los resultados obtenidos, el árbol construido con los parámetros $M=2$ y $C=0.4$ fue el mejor con un porcentaje de 86.785 % de instancias correctamente clasificadas correspondiente a 440 instancias de 507. En la figura 8 se presentan los mejores resultados de esta prueba con los parámetros $M=2$ $C=0.4$.

```
J48 pruned tree
-----
nmeses <= 37
| cabezafamilia = 0
| | nmeses <= 0: MUERTO (29.0)
| | nmeses > 0
| | | nivelsisben = 1
| | | | cirugia = no
| | | | | tipovivienda = CASA O APARTEMENTO: VIVO (18.86/6.29)
| | | | | tipovivienda = CUARTO: MUERTO (17.14/7.43)
| | | | | tipovivienda = OTRO TIPO DE UNIDAD DE VIVIENDA (CARPA,
| | | | | BARCO, REFUGIO NATURAL, PUENTE): VIVO (0.0)
| | | | cirugia = si: VIVO (11.0/1.0)
| | | nivelsisben = 2: VIVO (33.0/6.0)
| | | nivelsisben = 3: VIVO (4.0/1.0)
| | | nivelsisben = 4: VIVO (1.0)
| | | nivelsisben = 7
| | | | fuente = EPS: MUERTO (4.12/1.0)
| | | | fuente = HOSPITALES: MUERTO (117.38/31.0)
| | | | fuente = MEDICINA LEGAL: MUERTO (0.0)
| | | | fuente = ORGANIZACIONES: VIVO (6.18/1.18)
| | | | fuente = PARTICULAR: MUERTO (11.33/1.0)
| | cabezafamilia = 1: VIVO (42.0/4.0)
nmeses > 37: VIVO (212.0/15.0)

Number of Leaves:      15
Size of the tree:      22

=== Summary ===

Correctly Classified Instances      428      84.4181 %
Incorrectly Classified Instances     79      15.5819 %
Kappa statistic                   0.6529
Mean absolute                     0.2286
Root mean squared                  0.3373
Relative absolute error             50.8345 %
Root relative squared              71.1499 %
Coverage of cases (0.95 level)     100 %
Mean rel. region size (0.95 level) 97.0414 %
Total Number of Instances          507

=== Confusion Matrix ===

a b <-- classified as
133 40 | a = MUERTO
39 295 | b = VIVO
```

Figura 5. Mejores resultados de la prueba 1 con el repositorio T507A22ALL y con $M=10$ y $C=0.3$.

J48 pruned tree

nmeses <= 37

| nmeses <= 0: MUERTO (30.0)

| nmeses > 0

| | nivelsisben = 1

| | | regime = CONTRIBUTIVO: MUERTO (5.0/1.0)

| | | regimen = NINGUNO: MUERTO (3.0/1.0)

| | | regimen = PARTICULAR: VIVO (0.0)

| | | regimen = SUBSIDIADO: VIVO (62.0/14.0)

| | | regimen = VINCULADO: VIVO (0.0)

| | nivelsisben = 2: VIVO (47.0/6.0)

| | nivelsisben = 3: VIVO (6.0/1.0)

| | nivelsisbenv = v4: VIVO (1.0)

| | nivelsisben = 7

| | | locesp = Cuello uterino SAI: MUERTO (132.77/35.0)

| | | locesp = EndocÃ©rvix: VIVO (6.18/1.18)

| | | locesp = ExocÃ©rvix: MUERTO (2.06)

nmeses > 37: VIVO (212.0/15.0)

Number of Leaves: 13

Size of the tree: 18

=== Summary ===

Correctly Classified Instances	433	85.4043 %
Incorrectly Classified Instances	74	14.5957 %
Kappa statistic		0.6753
Mean absolute error		0.2329
Root mean squared error		0.3411
Relative absolute error		51.7777 %
Root relative squared error		71.9392 %
Coverage of cases (0.95 level)		100 %
Mean rel. region size (0.95 level)		96.7456 %
Total Number of Instances		507

=== Confusion Matrix ===

a b <-- classified as

136 37 | a = MUERTO

37 297 | b = VIVO

Figura 6. Mejores resultados de la prueba 2 con el repositorio T507A22ALL y con $M=5$ $C=0.2$.

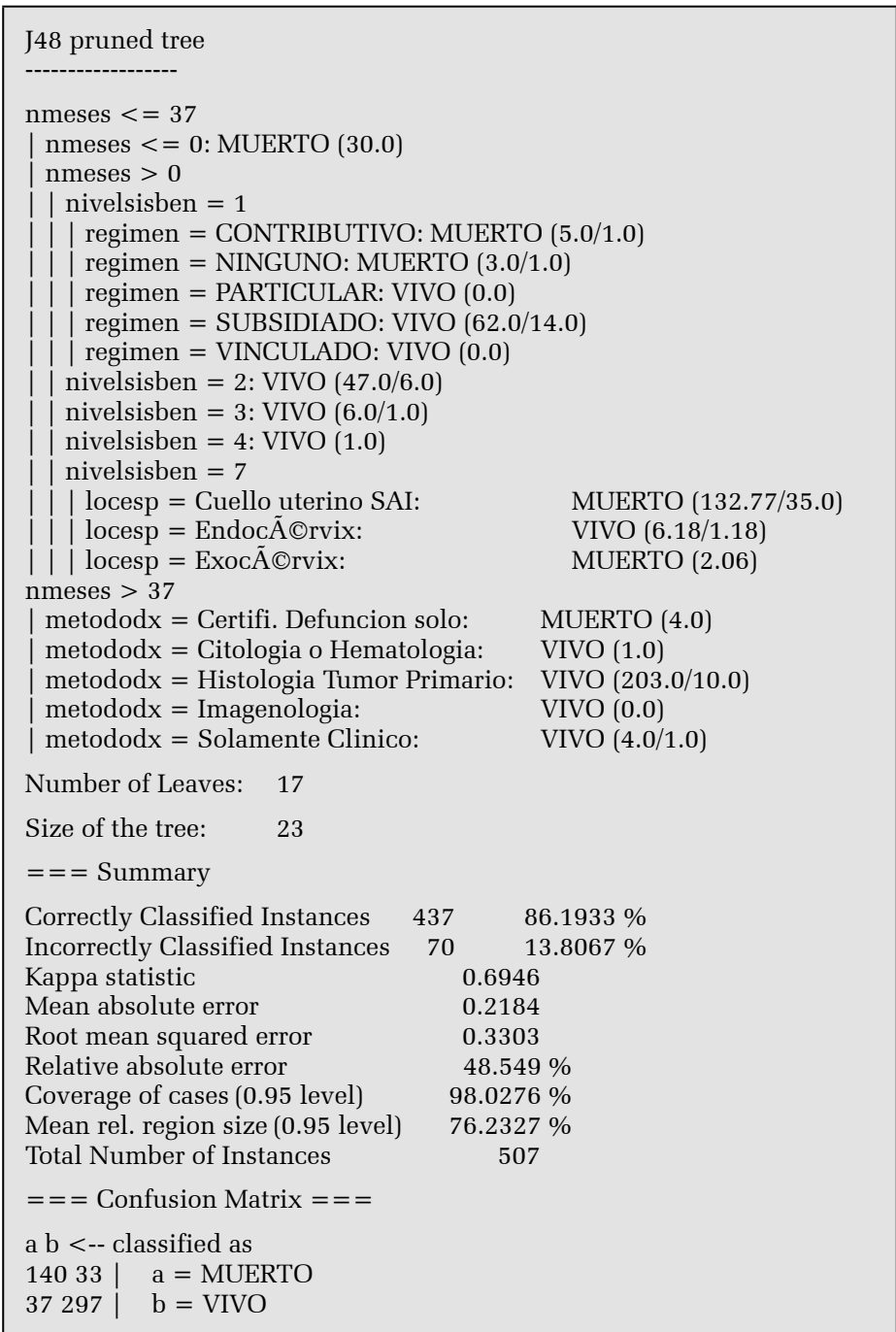


Figura 7. Mejores resultados de la prueba 3 con el repositorio T507A22ALL y con $M=2$ $C=0.2$.

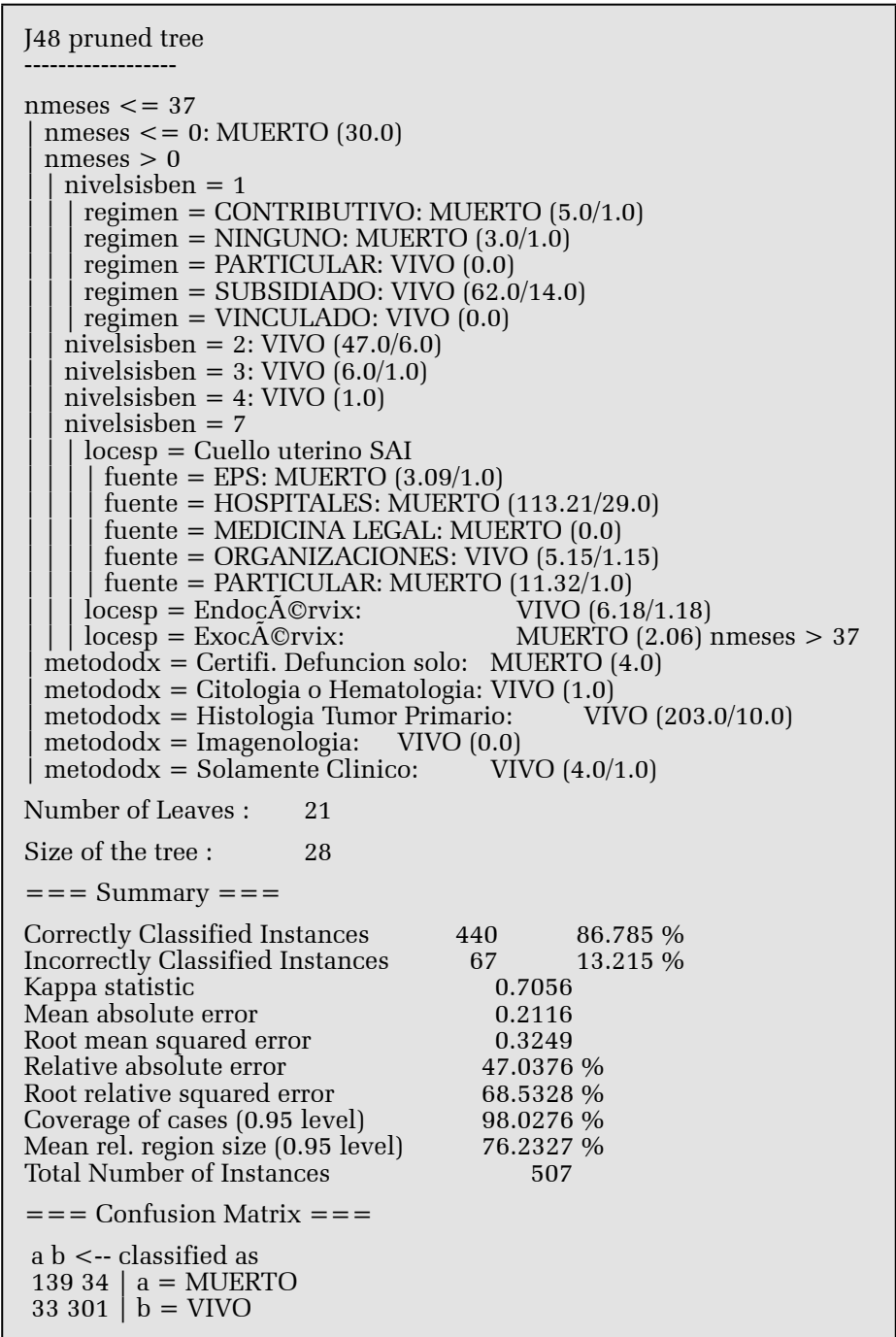


Figura 8. Mejores resultados de la prueba 4 con el repositorio T507A22ALL y con $M=2$ $C=0.4$.

Analizando los resultados obtenidos de las pruebas anteriores, el árbol de la prueba 4 con un nivel de confianza $C=0.4$ y un número de registros por nodo de $M=2$, es el que mejor clasifica los casos de mujeres con cáncer de cuello uterino, con un porcentaje de clasificación de 86.785% correspondiente a 440 casos correctamente clasificados de 507 existentes. En la figura 9 se muestra este árbol.

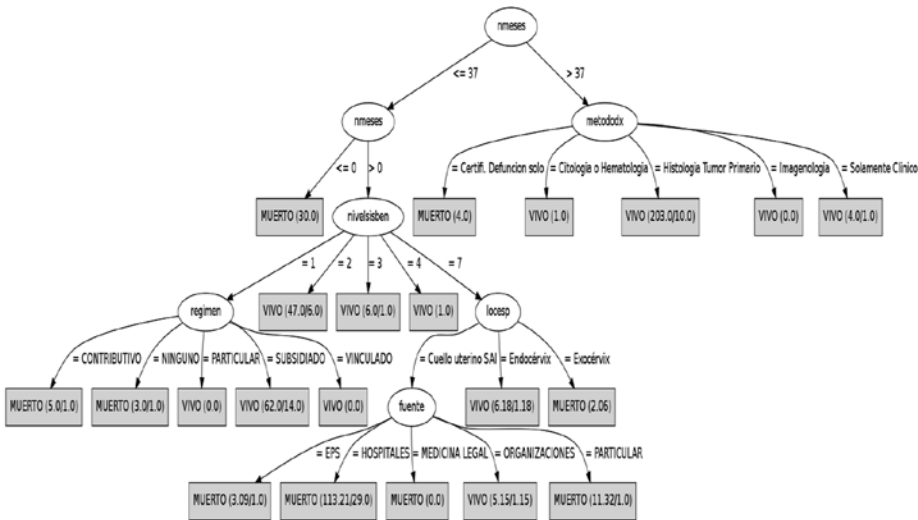


Figura 9. Mejor árbol de las pruebas con el repositorio T507A22ALL y con $M=2$ $C=0.4$.

Una vez se determinaron los mejores parámetros de poda del árbol para obtener el árbol que mejor clasifica los casos de supervivencia de mujeres con cáncer de cuello uterino, se procedió, utilizando estos parámetros y los repositorios T507A15ECON y T507A09CLINICO, a determinar, respectivamente, únicamente los factores socioeconómicos o clínicos relacionados con la supervivencia de las mujeres con cáncer invasivo de cuello uterino. Se escogió como en las pruebas anteriores el atributo vivomuerto como atributo clase con un nivel de confianza $C=0.4$ y con un número de registros por nodo $M=2$.

Los resultados para los factores socioeconómicos muestran un porcentaje de 84.8126 % de instancias correctamente clasificadas correspondiente a 430 instancias de 507. Estos resultados se

presentan en la figura 10 y el respectivo árbol de decisión, en la figura 11.

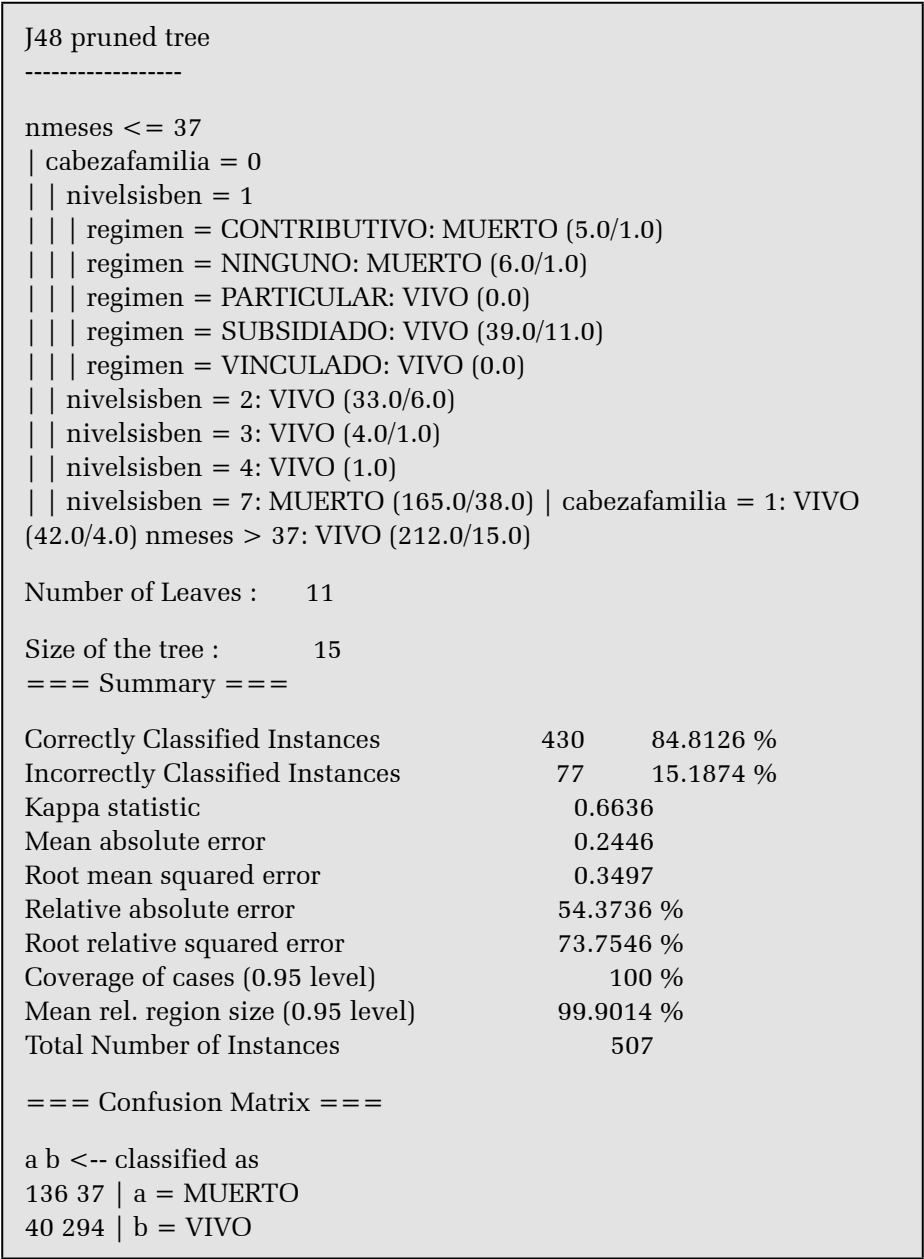


Figura 10. Resultados factores socioeconómicos con el repositorio T507A15ECONOM y con M=2 y C=0.4.

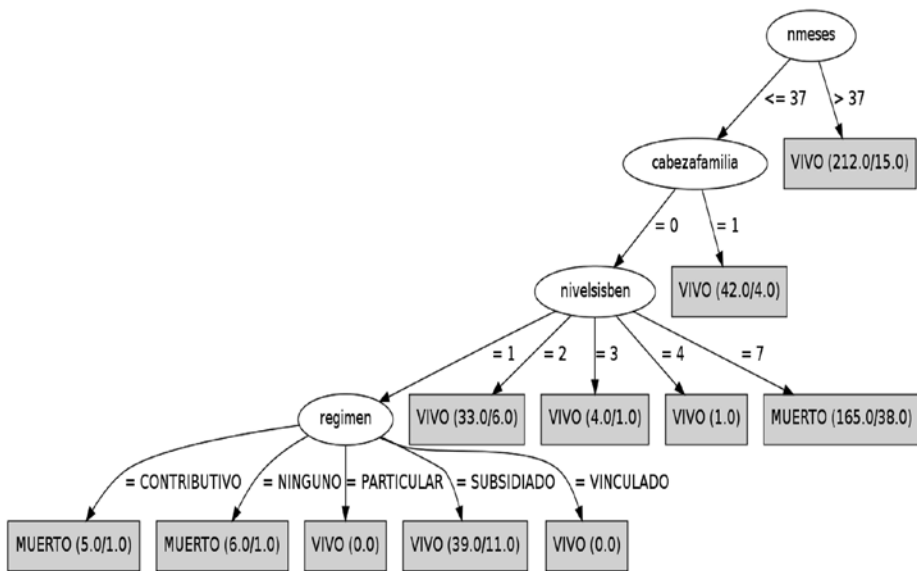


Figura 11. Árbol factores socioeconómicos del repositorio T507A15ECONOM con $M=2$ y $C=0.4$.

Los resultados para los factores clínicos muestran un porcentaje de 79.0927 % de instancias correctamente clasificadas correspondiente a 401 instancias de 507. Estos resultados se presentan en la figura 12 y el respectivo árbol de decisión, en la figura 13.

Teniendo en cuenta los resultados obtenidos para los factores socioeconómicos y clínicos con los conjuntos de datos T507A15ECON y T507A09CLINICO respectivamente y dada la relación directa del atributo nmeses en la generación de los árboles, se planteó realizar otras pruebas quitando este atributo. Aplicando los mismos criterios que la prueba 4, el árbol construido con los parámetros $M=2$ y $C=0.5$ para los factores socioeconómicos, fue el que mejor resultados obtuvo con un porcentaje de 79.0927 % de instancias correctamente clasificadas correspondiente a 401 instancias de 507. Estos resultados se presentan en la figura 14 y el respectivo árbol de decisión, en la figura 15.

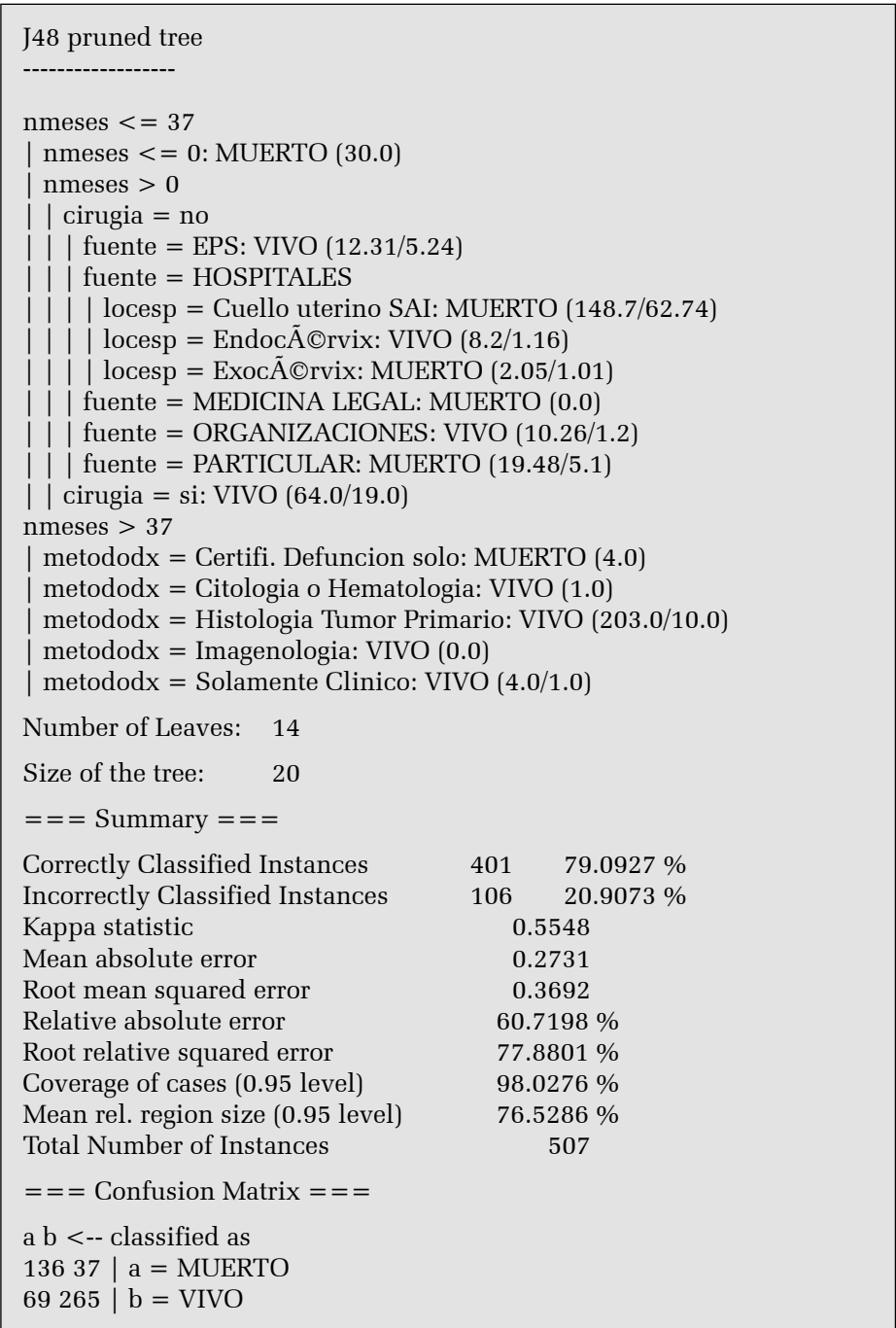


Figura 12. Resultados para los factores clínicos del repositorio T507A-09CLINICO con $M=2$ y $C=0.4$.

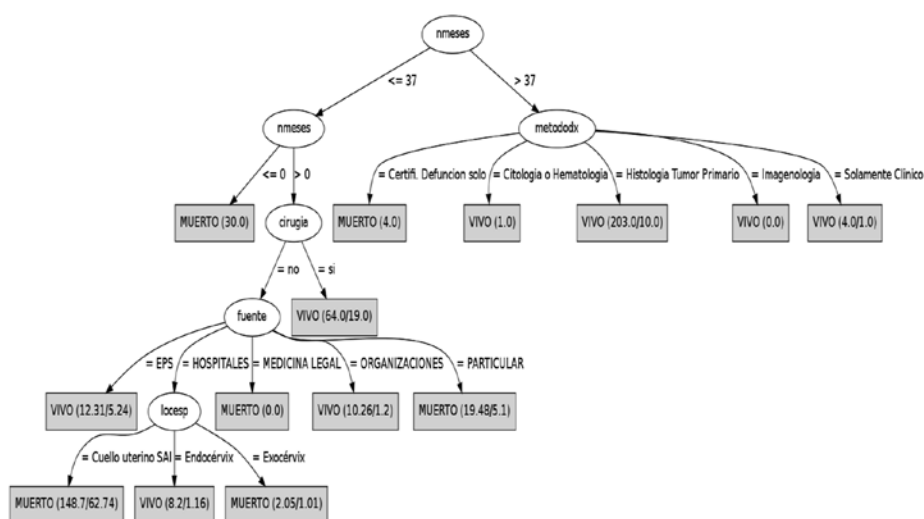


Figura 13. Árbol para los factores clínicos del repositorio T507A09CLINICO con $M=2$ y $C=0.4$.

J48 pruned tree

```

regimen = CONTRIBUTIVO: VIVO (103.0/38.0)
regimen = NINGUNO: MUERTO (88.0/25.0)
regimen = PARTICULAR: MUERTO (1.0)
regimen = SUBSIDIADO: VIVO (232.0/19.0)
regimen = VINCULADO
| edaddx = 1: VIVO (2.0)
| edaddx = 2: MUERTO (22.0/9.0)
| edaddx = 3: MUERTO (20.0/4.0)
| edaddx = 4
| | region = CENTRO: MUERTO (2.0)
| | region = NORTE: MUERTO (3.0)
| | region = OCCIDENTAL ANDINA: VIVO (4.0/1.0)
| | region = OTRO: VIVO (1.0)
| | region = PACIFICA: MUERTO (0.0)
| | region = PASTO: MUERTO (11.0/4.0)
| | region = PUTUMAYO: VIVO (1.0)
| | region = SUR: VIVO (3.0/1.0)
edaddx = 5: MUERTO (9.0/4.0)
edaddx = 6: MUERTO (5.0/1.0)

```

Number of Leaves : 17

Size of the tree : 20

=== Summary ===

Correctly Classified Instances	401	79.0927 %
Incorrectly Classified Instances	106	20.9073 %
Kappa statistic	0.5271	
Mean absolute error	0.2952	
Root mean squared error	0.3842	
Relative absolute error	65.625 %	
Root relative squared error	81.0271 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	99.0138 %	
Total Number of Instances	507	
=== Confusion Matrix ===		
a b <-- classified as		
114 59	a = MUERTO	
47 287	b = VIVO	

Figura 14. Resultados factores socioeconómicos con el repositorio T507A15ECONOM sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.

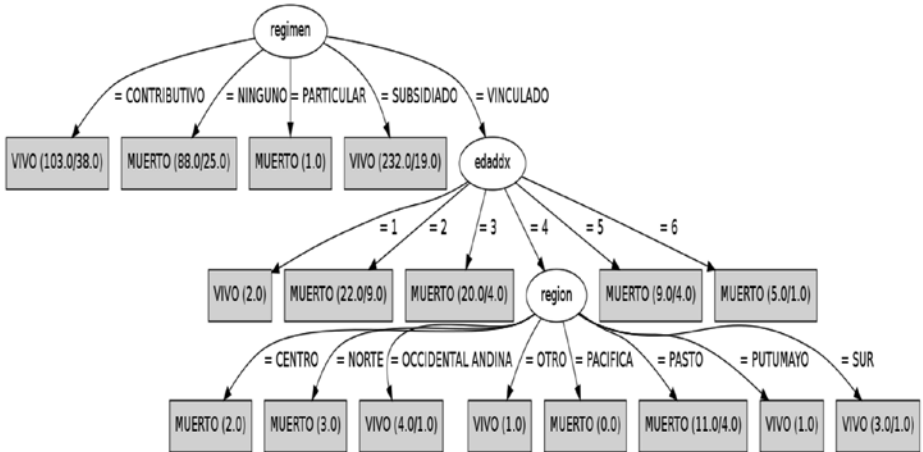


Figura 15. Árbol para los factores socioeconómicos con el repositorio T507A15ECONOM sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.

Para los factores clínicos, el árbol construido con los parámetros $M=2$ y $C=0.5$, fue el que mejor resultados obtuvo con un porcentaje de 72.1893 % de instancias correctamente clasificadas correspondiente a 366 instancias de 507. Estos resultados se presentan en la figura 16.



Figura 16. Resultados factores clínicos con el repositorio T507A09CLINICO sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.

4.4.2 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. 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 (Agrawal & Srikant, 1994). 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 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, Ramírez & Ferri, 2005).

En esta investigación, para obtener las reglas de asociación, se utilizó la herramienta de minería de datos Weka que provee el paquete WEKA.associations.Apriori que contiene la implementación del algoritmo de aprendizaje de reglas de asociación Apriori (Agrawal & Srikant, 1994). Este algoritmo es muy simple y popular. Este algoritmo permite encontrar el conjunto de ítems frecuentes (aquellos cuyo soporte es mayor o igual al soporte mínimo establecido). Para ello, en primer lugar, se construyen simplemente los conjuntos formados por un solo ítem que superen el soporte mínimo. Este conjunto de ítems frecuentes de tamaño 1 se utiliza para construir el conjunto de conjuntos de dos ítems y encontrar luego el conjunto de ítems frecuentes de tamaño 2, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de ítems con el soporte requerido (Hernández, Ramírez & Ferri, 2005).

Una vez seleccionados los conjuntos de ítems frecuentes, el siguiente paso es extraer las reglas de asociación que cumplan una confianza mínima.

En Weka se puede configurar este algoritmo con varias opciones: con la opción `UpperBoundMinSupport` se indica el límite superior de soporte requerido para aceptar un conjunto de ítems. Si no se encuentran conjuntos de ítems suficientes para generar las reglas requeridas se va disminuyendo el límite hasta llegar al límite inferior (opción `LowerBoundMinSupport`). Con la opción `minMetric` se indica la confianza mínima (u otras métricas dependiendo del criterio de ordenación) para mostrar una regla de asociación; y con la opción `numRules` se indica el número de reglas que se desea generar. La ordenación de estas reglas se puede configurar mediante la opción `MetricType`, algunas opciones que se pueden utilizar son: confianza de la regla, lift (confianza dividido por el número de ejemplos cubiertos por la parte derecha de la regla), y otras más elaboradas (Hernández & Ferri, 2005).

4.4.2.1 Descubrimiento de Reglas de Asociación

Para obtener reglas de asociación interesantes y significativas con una alta precisión, se utilizó el repositorio T507A22ALL descrito en la tabla 15, con el fin de detectar patrones que relacionen los factores socioeconómicos y clínicos de todas las mujeres que sobrevivieron al cáncer invasivo de cuello uterino entre los años 1998 y 2007. Se fijó como mínima confianza (`minMetric`) el 80% (0.8), un soporte mínimo superior (`UpperBoundMinSupport`) de 1.0, un soporte mínimo inferior (`LowerBoundMinSupport`) de 0.1, un Δ de 0.5 y un número de reglas (`numRules`) a generar de 25. Además, 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). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 17 y las mejores 25 reglas generadas, con una confianza del 100%, en la figura 18.

```
##
## Apriori
## =====##
## Minimum support: 0.1 (51 instances)
## Minimum metric <confidence>: 0.8
## Number of cycles performed: 18##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 46##
## Size of set of large itemsets L(2): 309
##
## Size of set of large itemsets L(3): 1022##
## Size of set of large itemsets L(4): 1984
##
## Size of set of large itemsets L(5): 2441##
## Size of set of large itemsets L(6): 1948
##
## Size of set of large itemsets L(7): 989##
## Size of set of large itemsets L(8): 298
##
## Size of set of large itemsets L(9): 45##
## Size of set of large itemsets L(10): 2
```

Figura 17. Parámetros de ejecución del algoritmo Apriori con el repositorio T507A22ALL.

```
## Best rules found:
##
## 1. edaddx=3 locesp=Cuello uterino SAI fuenteagua=ACUEDUCTO
51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 2. escolaridad=PRIMARIA discapacidad=NINGUNA
cabezafamilia=0 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 3. morfologia=Squamous cell carcinoma, large cell, nonkeratinizing,
NOS biopsia=si tipovivienda=CASA O APARTAMENTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 4. tipovivienda=CASA O APARTEMENTO
```

```

fuenteagua=ACUEDUCTO cabezafamilia=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 5. edaddx=3 discapacidad=NINGUNA biopsia=si fuenteagua=
ACUEDUCTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 6. edaddx=3 locesp=Cuello uterino SAI radio=si fuenteagua =
ACUEDUCTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 7. edaddx=3 locesp=Cuello uterino SAI biopsia=si fuenteagua =
ACUEDUCTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 8. edaddx=3 locesp=Cuello uterino SAI fuenteagua=ACUEDUCTO
regimen=SUBSIDIADO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 9. escolaridad = PRIMARIA ocupacion = HOGAR discapacidad =
NINGUNA fuente=HOSPITALES 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 10. escolaridad=PRIMARIA discapacidad=NINGUNA cirugia=no
tipovivienda=CASA O APARTAMENTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 11. escolaridad=PRIMARIA discapacidad = NINGUNA regimen =
SUBSIDIADO cabezafamilia=0 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 12. ocupacion=HOGAR discapacidad=NINGUNA fuenteagua =
ACUEDUCTO nivelsisben=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 13. discapacidad=NINGUNA fuente=HOSPITALES tipovivienda
= CASA O APARTEMENTO cabezafamilia=0 51 ## ==> vivo_
muerto_2007 = VIVO 51 conf:(1)
## 14. discapacidad=NINGUNA morfologia=Squamous cell carcinoma,
NOS locesp=Cuello uterino SAI
fuenteagua=ACUEDUCTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 15. discapacidad=NINGUNA cirugia=no fuenteagua=ACUEDUCTO
cabezafamilia=0 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 16. metododx=Histologia Tumor Primario morfologia=Squamous
cell carcinoma, large cell, nonkeratinizing, NOS biopsia=si tipovivienda
= CASA O APARTEMENTO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 17. fuente=HOSPITALES tipovivienda=CASA O APARTAMENTO
fuenteagua=ACUEDUCTO cabezafamilia=0 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 18. morfologia=Squamous cell carcinoma, large cell, nonkeratinizing,
NOS biopsia=si tipovivienda = CASA O

```

```

APARTEMENTO regimen=SUBSIDIADO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 19. radio=si tipovivienda=CASA O APARTAMENTO fuenteagua =
ACUEDUCTO cabezafamilia=1 51
## ==> vivo_muerto_2007 = VIVO 51 conf:(1)
## 20. biopsia=si tipovivienda=CASA O APARTAMENTO nivelsisben=2
regimen = SUBSIDIADO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 21. tipovivienda = CASA O APARTAMENTO fuenteagua =
ACUEDUCTO regimen=SUBSIDIADO cabezafamilia=1 51
## ==> vivo_muerto_2007 = VIVO 51 conf:(1)
## 22. edaddx=3 discapacidad = NINGUNA radio = si biopsia=si ##
== > vivo_muerto_2007 = VIVO 51 conf:(1)
## 23. edaddx=3 discapacidad = NINGUNA biopsia = si fuenteagua =
ACUEDUCTO ## ==> vivo_muerto_2007 = VIVO 51 conf:(1)
## 24. edaddx=3 locesp = Cuello uterino SAI radio= si biopsia=si ##
== > vivo_muerto_2007 = VIVO 51 conf:(1)
## 25. edaddx=3 locesp = Cuello uterino SAI radio= si fuenteagua=A##
== > vivo_muerto_2007 = VIVO 51 conf:(1)

```

Figura 18. Mejores reglas generadas con Apriori con el conjunto de datos T507A22ALL.

Por otra parte, se obtuvieron reglas de asociación que relacionen solo los factores socioeconómicos con la supervivencia de las mujeres con cáncer de cuello uterino. Estas se obtuvieron utilizando el conjunto de datos T507A15ECON y con los mismos parámetros de la prueba anterior. Se encontraron las mejores reglas con una confianza del 80% y un soporte mínimo del 10 % (0.1). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 19 y las mejores 11 reglas generadas en la figura 20.

```

##
## Apriori
## =====##
## Minimum support: 0.1 (51 instances)
## Minimum metric <confidence>: 0.8
## Number of cycles performed: 18
##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 29##

```

```
## Size of set of large itemsets L(2): 84
##
## Size of set of large itemsets L(3): 95 ##
## Size of set of large itemsets L(4): 50
##
## Size of set of large itemsets L(5): 13##
## Size of set of large itemsets L(6): 2
```

Figura 19. Parámetros de ejecución del algoritmo Apriori con los factores socioeconómicos del repositorio T507A15ECON.

Se obtuvieron reglas de asociación que relacionen solo los factores clínicos con la supervivencia de las mujeres con cáncer de cuello uterino. Estas se obtuvieron utilizando el conjunto de datos T507A09CLINICO y con los mismos parámetros de las pruebas anteriores, pero bajando la confianza mínima al 50% (0.5). De esta manera se lograron encontrar reglas con una confianza mínima del 50% y un soporte mínimo del 10 % (0.1). Los parámetros de ejecución del algoritmo Apriori y las 2 reglas generadas, con una confianza del 53%, se presentan en la figura 21.

```
## Best rules found:
##
## 1. ocupacion=HOGAR fuenteagua=ACUEDUCTO nivelsisben=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 2. escolaridad=PRIMARIA discapacidad=NINGUNA cabezafamilia
= 0 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 3. tipovivienda=CASA O APARTEMENTO fuenteagua=ACUEDUCTO
cabezafamilia=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 4. ocupacion=HOGAR fuenteagua=ACUEDUCTO nivelsisben=1
regimen=SUBSIDIADO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 5. escolaridad=PRIMARIA discapacidad=NINGUNA fuenteagua =
ACUEDUCTO nivelsisben=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 6. escolaridad = PRIMARIA discapacidad = NINGUNA regimen =
SUBSIDIADO cabezafamilia=0 51 ## ==> vivo_muerto_2007=VIVO
51 conf:(1)
```

```

## 7. tipovivienda=CASA O APARTEMENTO fuenteagua=ACUEDUCTO
regimen=SUBSIDIADO cabezafamilia=1 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 8. escolaridad=PRIMARIA discapacidad=NINGUNA
fuenteagua=ACUEDUCTO nivelsisben=1
regimen=SUBSIDIADO 51
## ==> vivo_muerto_2007=VIVO 51 conf:(1)
## 9. region=OCCIDENTAL ANDINA regimen=SUBSIDIADO 59
## ==> vivo_muerto_2007=VIVO 51 conf:(0.86)
## 10. escolaridad=PRIMARIA tipovivienda=CUARTO 59
## ==> vivo_muerto_2007=VIVO 51 conf:(0.86)
## 11. region=PASTO ocupacion=HOGAR regimen=SUBSIDIADO
cabezafamilia=0 62 ## ==> vivo_muerto_2007=VIVO 51 conf:(0.82)

```

Figura 20. Mejores reglas generadas con Apriori con el conjunto de datos T507A15ECON.

4.4.3 Tarea de Clustering o Segmentación

El clustering o segmentación es una de las tareas más frecuentes en la minería de datos. A diferencia de la tarea de clasificación, en el clustering no se conoce la clase a la cual pertenecen los datos, por esa razón se dice que es un proceso de aprendizaje no supervisado. En esta tarea se trata de encontrar grupos similares entre un conjunto de datos basado en el concepto de distancia (Han & Kamber, 2001), (Hernández, Ramirez & Ferri, 2005).

La tarea de clustering permite agrupar objetos físicos o abstractos en clases de objetos similares (Chen, Han & Yu, 1996). Los clusters tienen una alta homogeneidad interna (dentro del cluster) y una alta heterogeneidad externa (entre cluster) basadas en el concepto de distancia, en la cual individuos similares (ceranos) van al mismo grupo. Tomando el conjunto de datos de todos los registros de mujeres que padecen cáncer de cuello uterino, se obtuvo grupos homogéneos de mujeres con esta enfermedad.

El algoritmo más popular de los métodos de agrupamiento particionales es el K-means (Han & Kamber, 2001). Este nombre

es debido a que los centros de los clusters, denominados centroides, son las medias de las observaciones asignadas a cada grupo (Pérez & Santín, 2007). La idea de este algoritmo es tomar como entrada el parámetro k y particionar el conjunto de datos en k grupos o clusters, donde los casos o instancias pertenecientes al mismo grupo tengan características similares. Todo caso nuevo es comparado con los grupos formados y asociado a aquél que sea el más próximo, en los términos de una distancia que normalmente es la Euclidiana (Sharma, Bajpai & Litoriya, 2012).

```
## Apriori
## =====
##
## Minimum support: 0.1 (51 instances)
## Minimum metric <confidence>: 0.5
## Number of cycles performed: 18
##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 18
##
## Size of set of large itemsets L(2): 61
##
## Size of set of large itemsets L(3): 110
##
## Size of set of large itemsets L(4): 114
##
## Size of set of large itemsets L(5): 65
##
## Size of set of large itemsets L(6): 17
##
## Size of set of large itemsets L(7): 1
##
## Best rules found:
#### 1. metododx=Histologia Tumor Primario fuente=HOSPITALES
loesp=Cuello uterino SAI cirugia=no radio=si nmeses_2007 = [0,12] 97
## ==> vivo_muerto_2007=VIVO 51 conf:(0.53)
## 2. metododx=Histologia Tumor Primario fuente=HOSPITALES
loesp=Cuello uterino SAI cirugia=no nmeses_2007=[0,12] 98
## ==> vivo_muerto_2007=VIVO 51 conf:(0.52)
```

Figura 21. Parámetros y reglas generadas con Apriori con el conjunto de datos T507A09CLINICO.

En la herramienta Weka, para ejecutar K-means se escoge la opción del algoritmo SimpleKmeans y se configura número de clusters (NumClusters) o grupos 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 clusters. Para evaluar los resultados del agrupamiento, está la opción Cluster Mode. Lo más simple es utilizar el propio conjunto de entrenamiento, Use training set, que indica que porcentaje de instancias se van a cada grupo. El resto de opciones realizan un entrenamiento con un conjunto, sobre el que construyen los clusters y a continuación aplican estos clusters para clasificar un conjunto independiente que puede proporcionarse aparte (Supplied test), o ser un porcentaje del conjunto de entrada (split). Existe también la opción de comparar los clusters con un atributo de clasificación (Classes to clusters evaluation) que no se considera en la construcción de los clusters.

4.4.3.1 Descubrimiento de Patrones con Clustering

Para esta tarea, se utilizó el conjunto de datos T507A22ALL descrito en la tabla 15, con el fin de encontrar grupos similares, que relacionen los factores socioeconómicos y clínicos, entre todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2007. La idea de la prueba es encontrar clusters donde como mínimo haya un grupo de mujeres sobrevivientes y otro de no sobrevivientes. De esta manera, se pueden encontrar cuales son las similitudes particulares de cada uno de estos grupos, en un proceso no supervisado, donde no se especifica la clase vivomuerto como se hizo en tarea de clasificación. Por otra parte, se puede comparar el número de casos que la tarea de clustering agrupa de mujeres que sobrevivieron al cáncer, con el número de sobrevivientes reales que se encuentran etiquetados como vivos en el conjunto de datos T507A22ALL. De igual manera para el grupo de las mujeres que murieron. Así se podría conocer la precisión del modelo, aunque este no es el objetivo de la prueba.

Se configuró el número de clusters en 2, 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%. Como se puede observar en la tabla 18, el algoritmo

K-means agrupó a los 507 casos en dos grupos: en el cluster 0 a 304 vivos y en el cluster 1 a 203 muertos. Los atributos que los determinan corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

Tabla 18. Clusters resultantes con K-means con el repositorio T507A22ALL

ATRIBUTO	FULL DATA (507)	CLUSTER 0 (304)	CLUSTER 1 (203)
edaddx	3	3	4
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
fuelle	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
tipooovienda	CASA O APARTAMENTO	CASA_O_APARTAMENTO	CASA_O_APARTAMENTO
fuelleagua	ACUEDUCTO	ACUEDUCTO	ACUEDUCTO
nivelsisben	7	1	7
regimen	SUBSIDIADO	SUBSIDIADO	NINGUNO
cabezafamilia	0	0	0
comuna	13	13	13
region	PASTO	PASTO	PASTO
nmeses_2007	[0,12]	[24,36]	[0,12]
vivo_muerto_2007	VIVO	VIVO	MUERTO

Para encontrar grupos similares, que relacionen únicamente los factores socioeconómicos entre todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2007, se utilizó el conjunto de datos T507A15ECON descrito en la tabla 15. Se configuró el número de clusters en 2, con una semilla por defecto de 10. Para evaluar los resultados del agrupamiento se utilizó el propio conjunto de entrenamiento Use tranining set igual al 66%. En la tabla 19 se pueden observar los resultados con el algoritmo K-means clasificando en el cluster 0 a 298 vivos y en el cluster 1 a 130 muertos. De igual manera que el caso anterior, los atributos que los determinan corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

Tabla 19. Clusters resultantes con K-means con el repositorio T507A15ECON

ATRIBUTO	FULL DATA (507)	CLUSTER 0 (298)	CLUSTER 1 (209)
region	PASTO	PASTO	PASTO
comuna	13	13	13
edaddx	3	3	4
estadocivil	CASADO	CASADO	CASADO
ocupacion	HOGAR	HOGAR	HOGAR
escolaridad	PRIMARIA	PRIMARIA	PRIMARIA
discapacidad	NINGUNA	NINGUNA	NINGUNA
tipooivienda	CASA_O_ APARTAMENTO	CASA_O_ APARTAMENTO	CASA_O_ APARTAMENTO
estrato	2	2	2
fuelleagua	ACUEDUCTO	ACUEDUCTO	ACUEDUCTO
nivelsisben	7	1	7
regimen	SUBSIDIADO	SUBSIDIADO	CONTRIBUTIVO
cabezafamilia	0	0	0
nmeses_2007	[0,12]	[0,12]	[0,12]
vivo_muerto_2007	VIVO	VIVO	MUERTO

De igual manera, con el fin de encontrar grupos similares, que relacionen únicamente los factores clínicos entre todas las mujeres con cáncer invasivo de cuello uterino diagnosticadas entre los años 1998 y 2007, se utilizó el conjunto de datos T507A09CLINICO descrito en la tabla 15. Como el objetivo de las pruebas era encontrar por lo menos un cluster que agrupe a las pacientes que sobrevivieron y otro que agrupe a las que no sobrevivieron, en este caso solo se encontró grupos de vivos y uno de muertos cuando se configuró el número de clusters en 7, con una semilla por defecto de 10. Para evaluar los resultados del agrupamiento se utilizó el propio conjunto de entrenamiento Use tranining set igual al 66%. Como en los casos anteriores, los atributos de los grupos que los determinan corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005). En la tabla 20 se pueden observar los grupos formados con el algoritmo K-means con sus respectivos casos.

4.5 Descubrimiento de Patrones para los Casos de Cáncer de la Cohorte 1998-2002

4.5.1 Tarea de Clasificación

Con el fin de detectar los patrones más confiables de supervivencia de las mujeres con cáncer invasivo de cuello uterino utilizando árboles de decisión, para los casos registrados por el RPCMP entre 1998 y 2002, se utilizó el repositorio T235A22ALL, descrito en la tabla 15, para descubrir los factores socioeconómicos y clínicos que inciden en la supervivencia de estas pacientes. Se escogió como clase, el atributo vivomuerto, cuyos valores se distribuyen en 130 para vivos y 105 para muertos.

Para evaluar la calidad del modelo y su validez se escogió el método de validación cruzada y específicamente la validación cruzada con 10 pliegues por los mejores resultados que se obtienen.

Teniendo en cuenta las pruebas anteriores realizadas para la cohorte 1998-2007, para este caso, se realizó una prueba igual a la Prueba 4, es decir, se variaron tanto el factor confianza C como el número de registros por nodo M. Se estableció el parámetro C de 0.1 hasta 0.5 incrementando en 0.1 y M de 2 hasta 20 con un incremento de 2. De acuerdo a los resultados obtenidos, el árbol

construido con los parámetros $M=2$ y $C=0.2$ fue el mejor con un porcentaje de 93.1915 % de instancias correctamente clasificadas correspondiente a 219 instancias de 235. En la figura 22 se presentan los mejores resultados de esta prueba con los parámetros $M=2$ $C=0.2$ y en la figura 23 se muestra el árbol que mejor clasifica estos casos.

De igual manera como en las pruebas anteriores, una vez se determinaron los mejores parámetros de poda del árbol para obtener el árbol que mejor clasifica los casos de supervivencia de mujeres con cáncer de cuello uterino registrados entre los años 1998 y 2002, se procedió, utilizando estos parámetros y los repositorios T235A15ECON y T235A09CLINICO, a determinar, respectivamente, únicamente los factores socioeconómicos o clínicos que inciden en la supervivencia de estas pacientes. Se escogió como en las pruebas anteriores el atributo vivomuerto como atributo clase con un nivel de confianza $C=0.2$ y con un número de registros por nodo $M=2$.

Tabla 20. Clusters formados con K-means con el repositorio T507A09CLINICO

ATRIBUTO	FULL DATA (507)	CLUSTER 0 (121)	CLUSTER 1 (11)
metododx	Histologia_Tumor_Primario	Histologia_Tumor_Primario	Histologia_Tumor_Primario
fuelle	HOSPITALES	HOSPITALES	HOSPITALES
morfologia	Squamous_cell_carcinoma, large_cell, nonkeratinizing, NOS	Squamous_cell_carcinoma, large_cell, nonkeratinizing, NOS	Squamous_cell_carcinoma, small cell, nonkeratinizing, NOS
loresp	Cuello_uterino_SAI	Cuello_uterino_SAI	Cuello_uterino_SAI
cirugia	no	si	si
radio	si	si	si
biopsia	si	si	si
nmeses_2007	[0,12]	[24,36]	[108,120]
vivo_muerto_2007	VIVO	VIVO	VIVO
metododx	Histologia_Tumor_Primario	Histologia_Tumor_Primario	Histologia_Tumor_Primario
fuelle	HOSPITALES	HOSPITALES	HOSPITALES

ATRIBUTO	FULL DATA (507)	CLUSTER 0 (121)	CLUSTER 1 (11)
morfologia	Squamous_cell_carcinoma, keratinizing,NOS	Squamous_cell_carcinoma,NOS	Squamous_cell_carcinoma, large_cell, nonkeratinizing,NOS
locesp	Cuello_uterino_SAI	Cuello_uterino_SAI	Cuello_uterino_SAI
cirugia	no	no	no
radio	si	si	si
biopsia	si	si	si
nmeses_2007	[96,108]	[12,24]	[0,12]
vivo_muerto_2007	VIVO	VIVO	VIVO
ATRIBUTO	CLUSTER 5 (11)	CLUSTER 6 (33)	
metododx	Histologia_Tumor_Primario	Certifi.Defunción	
fuelle	ORGANIZACIONES	MEDICINA_LEGAL	
morfologia	Squamous_cell_carcinoma,NOS	Neoplasm,malignant	
locesp	Cuello_uterino_SAI	Cuello_uterino_SAI	
cirugia	no	no	
radio	si	si	
biopsia	si	no	
nmeses_2007	[12,24]	[0,12]	
vivo_muerto_2007	VIVO	MUERTO	

J48 pruned tree ----- nmeses <= 52 loesp = Cuello uterino SAI : MUERTO (110.63/14.0) loesp = EndocÁ©rvix nmeses <= 16: VIVO (3.32/0.32) nmeses > 16: MUERTO (2.05) loesp = ExocÁ©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

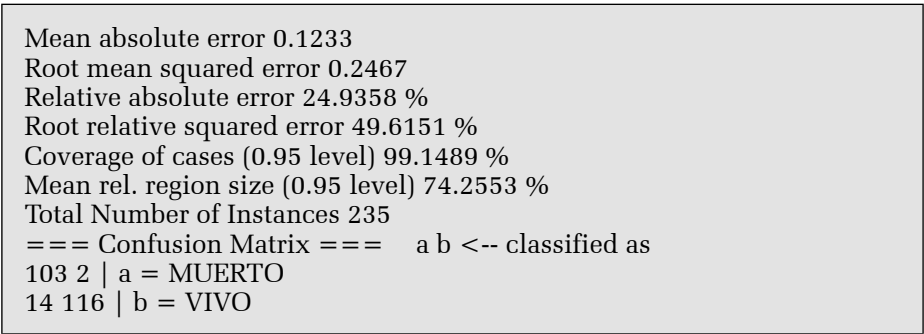


Figura 22. Mejores resultados para los casos de cáncer del repositorio T235A22ALL con $M=2$ y $C=0.2$.

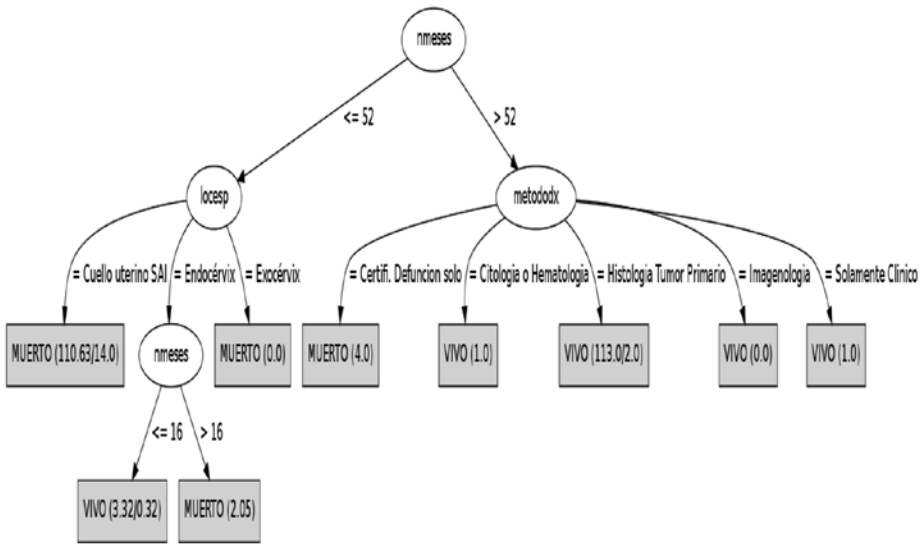


Figura 23. Mejor árbol para los casos de cáncer del repositorio T235A22ALL con $M=2$ y $C=0.2$.

Los resultados para los factores socioeconómicos muestran un porcentaje de 90.2128 % de instancias correctamente clasificadas correspondiente a 212 instancias de 235. Dado la simplicidad del árbol, se probó, como en las pruebas anteriores, construir árboles variando los parámetros $M=2$ hasta $M=20$ incrementando en 2 y $C=0.1$ hasta $C=0.5$ con un incremento de 0.1 pero los resultados fueron los mismos. Estos resultados se presentan en la figura 24 y el respectivo árbol de decisión, en la figura 25.

En cambio, la prueba realizada para los factores clínicos con los parámetros $M=2$ y $C=0.2$ si generaron resultados interesantes que muestran un porcentaje de 93.1915 % de instancias correctamente clasificadas correspondiente a 219 instancias de 235. Estos resultados se muestran en la figura 26 y el respectivo árbol de decisión, en la figura 27.

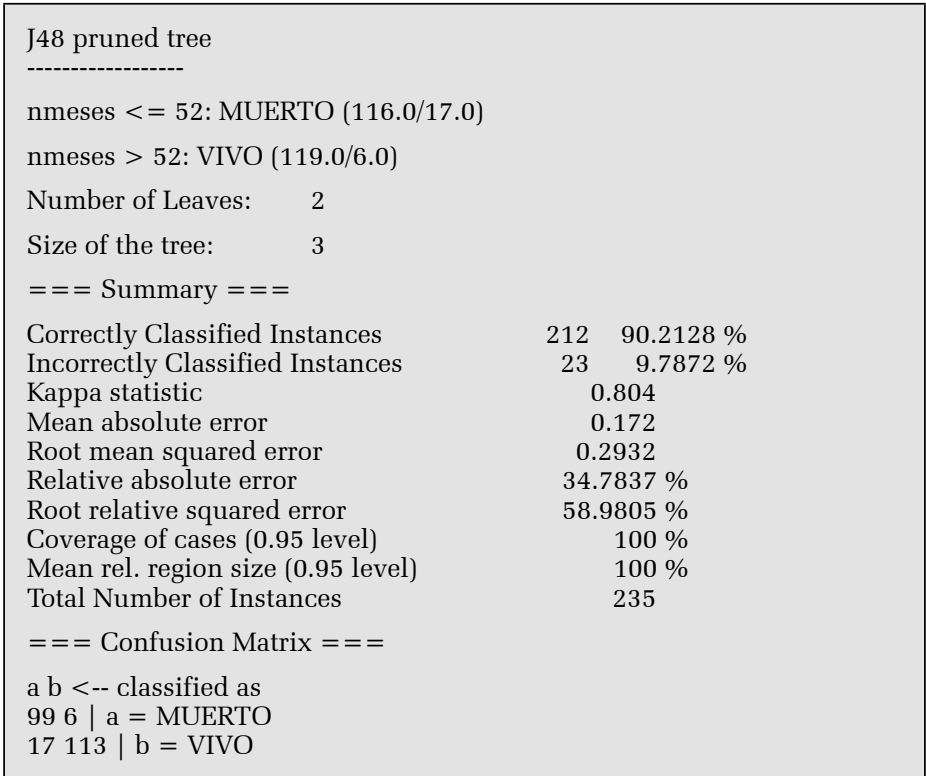


Figura 24. Mejores resultados para los factores socioeconómicos del repositorio T235A15ECON.

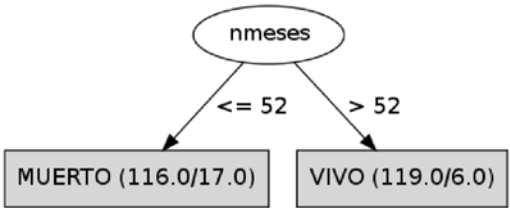


Figura 25. Mejor árbol para los factores socioeconómicos del repositorio T235A15ECON.

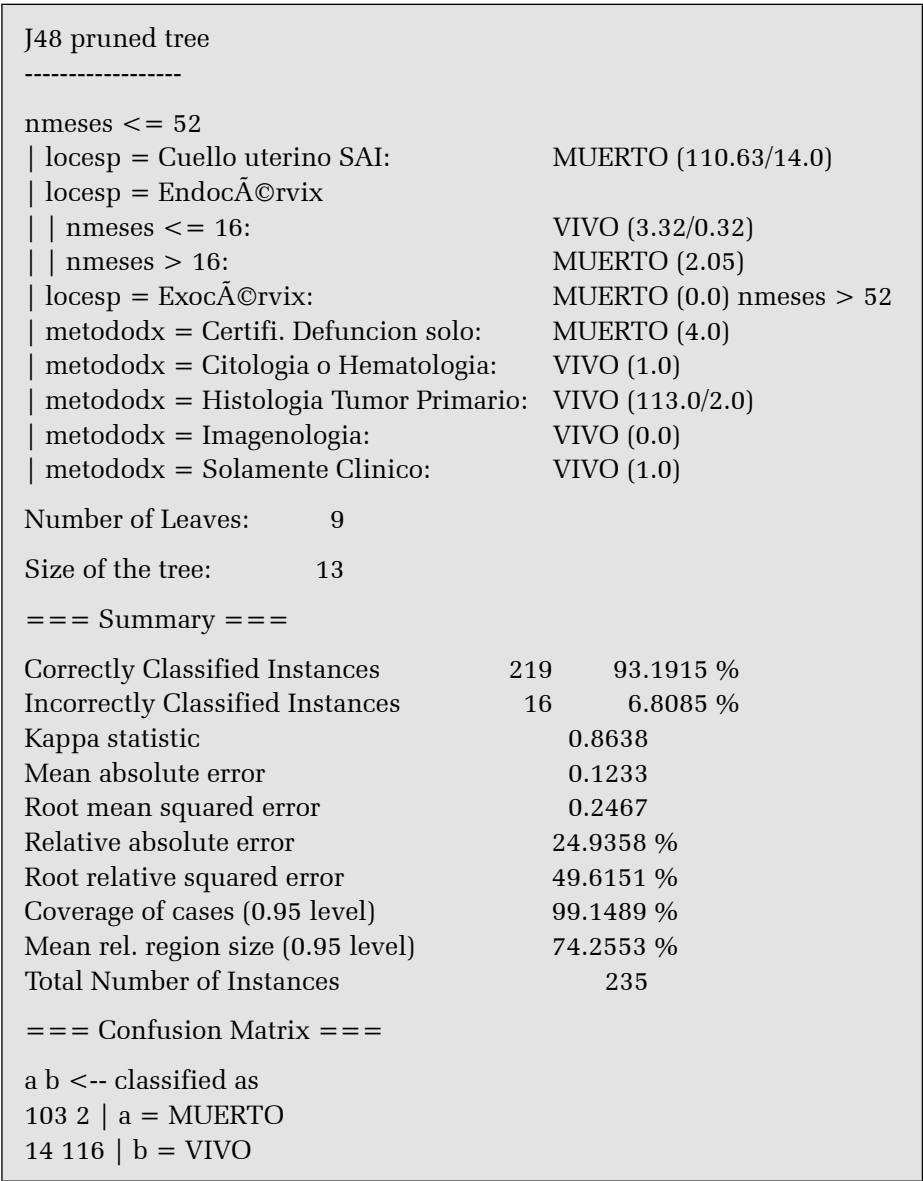


Figura 26. Resultados para los factores clínicos del repositorio T235A-09CLINICO con $M=2$ y $C=0.2$.

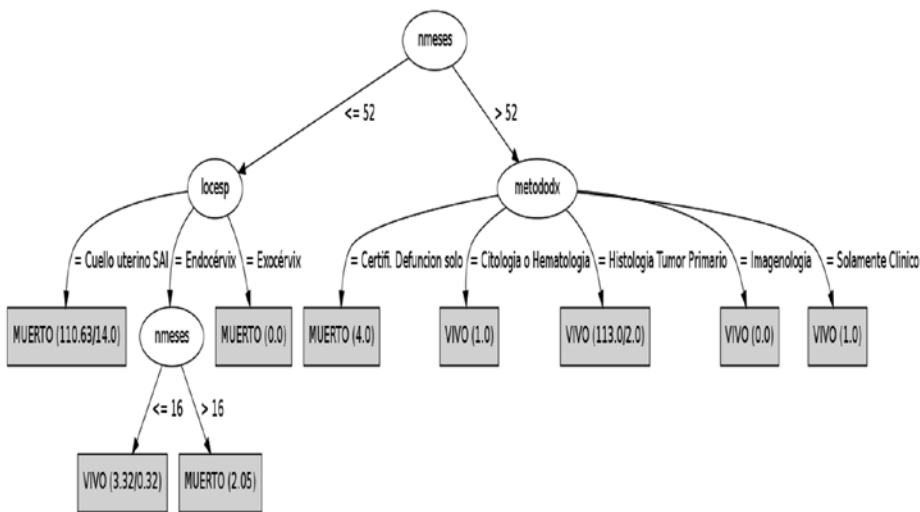


Figura 27. Árbol para los factores clínicos del repositorio T235A09CLINICO con $M=2$ y $C=0.2$.

Igual que sucedió con los repositorios T507A15ECON y T507A09CLINICO, los modelos de clasificación obtenidos para los factores socioeconómicos y clínicos con los conjuntos de datos T235A15ECON y T235A09CLINICO, guardan una relación directa con el atributo nmeses, por lo tanto, se planteó realizar otras pruebas quitando este atributo. Aplicando los mismos criterios que la prueba 4, el árbol construido con los parámetros $M=2$ y $C=0.4$ para los factores socioeconómicos, fue el que mejor resultados obtuvo con un porcentaje de 78.2979 % de instancias correctamente clasificadas correspondiente a 184 instancias de 235. Estos resultados se presentan en la figura 28.

Para los factores clínicos, el árbol construido para el repositorio T235A09CLINICO con los parámetros $M=2$ y $C=0.5$, fue el que mejor resultados obtuvo con un porcentaje de 68.9362 % de instancias correctamente clasificadas correspondiente a 162 instancias de 235. Estos resultados se presentan en la figura 29.

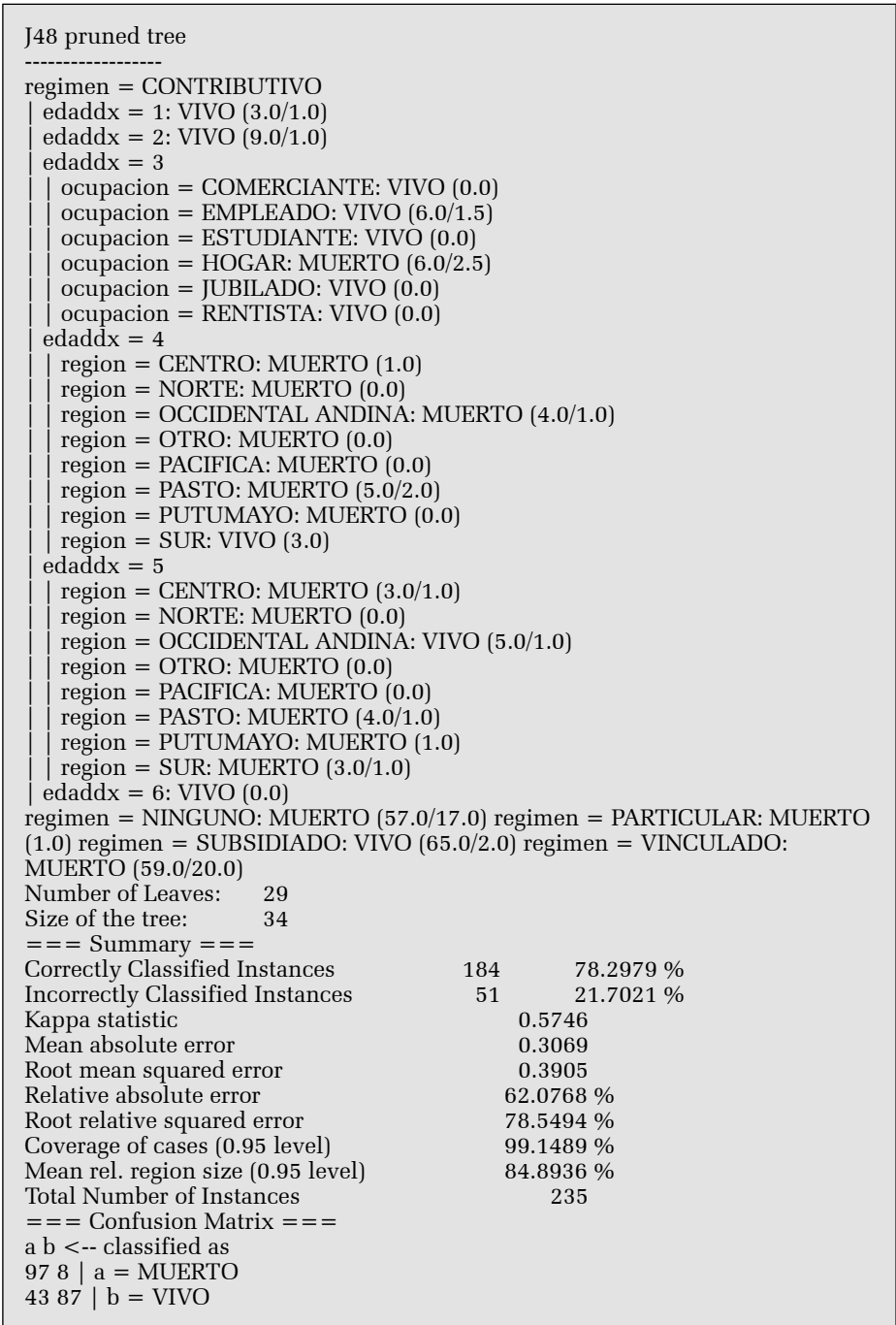


Figura 28. Resultados factores socioeconómicos con el repositorio T235A15ECON sin el atributo nmeses y con los parámetros $M=2$ y $C=0.4$.

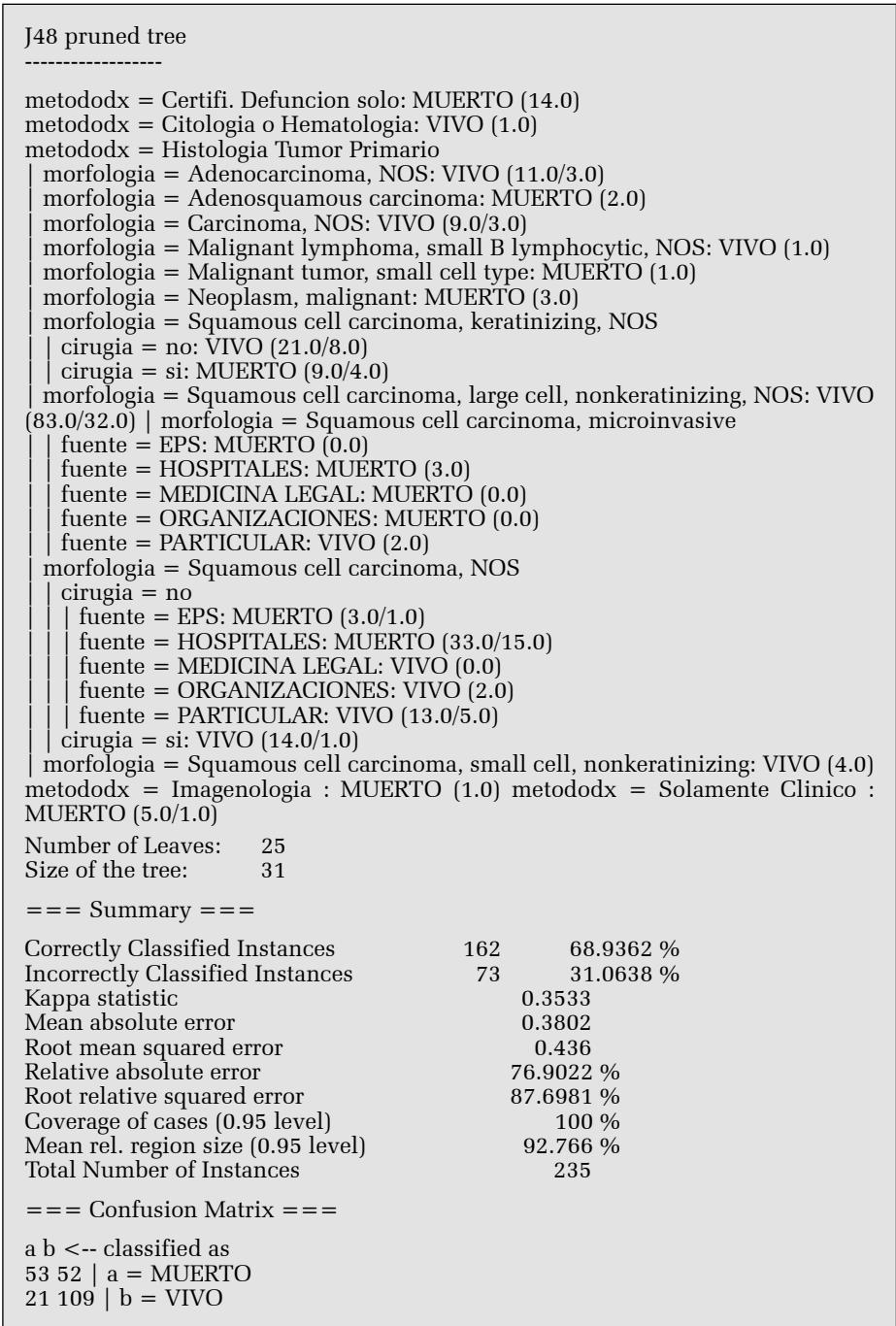


Figura 29. Resultados factores clínicos con el repositorio T235A09CLINICO sin el atributo nmeses y con los parámetros $M=2$ y $C=0.5$.

4.5.2 Tarea de Asociación

De igual manera como en las pruebas anteriores realizadas para la cohorte 1998-2007, con el fin de obtener reglas de asociación interesantes y significativas con una alta precisión, se utilizó el repositorio T235A22ALL descrito en la tabla 15, para descubrir patrones que relacionen los factores socioeconómicos y clínicos de todas las mujeres que sobrevivieron al cáncer invasivo de cuello uterino, que fueron diagnosticadas entre los años 1998 y 2002 con un ventana de observación hasta el 2007. Se fijó como mínima confianza el 80% (0.8), un soporte mínimo superior de 1.0, un soporte mínimo inferior de 0.1, un Δ de 0.5 y un numero 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). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 30 y las mejores 25 reglas generadas con una confianza del 100%, en la figura 31.

```
##
## Apriori
## =====##
## Minimum support: 0.1 (24 instances)
## Minimum metric <confidence>: 0.8
## Number of cycles performed: 18
##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 48
##
## Size of set of large itemsets L(2): 319
##
## Size of set of large itemsets L(3): 966
##
## Size of set of large itemsets L(4): 1676
##
## Size of set of large itemsets L(5): 1840
##
## Size of set of large itemsets L(6): 1326
##
## Size of set of large itemsets L(7): 628
##
## Size of set of large itemsets L(8): 191
##
## Size of set of large itemsets L(9): 35
##
## Size of set of large itemsets L(10): 3
```

Figura 30. Parámetros de ejecución del algoritmo Apriori para el repositorio T235A22ALL.

```

## Best rules found:
## ## 1. cabezafamilia=0 nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 2. edaddx=3 metododx=Histologia Tumor Primario
regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 3. edaddx=3 biopsia=si regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 4. estadocivil=CASADO discapacidad=NINGUNA
fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 5. estadocivil=CASADO metododx=Histologia Tumor Primario
fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 6. estadocivil=CASADO biopsia=si fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 7. discapacidad=NINGUNA cabezafamilia=0 nmeses_2007=(84,96]
24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 8. metododx=Histologia Tumor Primario regimen=SUBSIDIADO
region=PASTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 9. metododx=Histologia Tumor Primario cabezafamilia=0
nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 10. fuente=HOSPITALES locesp=Cuello uterino SAI
nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 11. morfologia=Squamous cell carcinoma, large cell,
nonkeratinizing, NOS cirugia=no regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 12. radio=si cabezafamilia=0 nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 13. biopsia=si cabezafamilia=0 nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 14. tipovivienda=CASA O APARTEMENTO
fuenteagua=ACUEDUCTO cabezafamilia=0 24conf:(1)
## ==> vivo_muerto_2007=VIVO 24

```

```

## 15. edaddx=3 metododx=Histologia Tumor Primario radio=si
regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 16. edaddx=3 radio=si biopsia=si regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 17. escolaridad=PRIMARIA metododx=Histologia Tumor Primario
locesp=Cuello uterino SAI nivelsisben=1 24
## ==> vivo_muerto_2007=VIVO 24 nf:(1)
## 18. escolaridad=PRIMARIA metododx=Histologia Tumor Primario
biopsia=si nivelsisben=1 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 19. escolaridad=PRIMARIA metododx=Histologia Tumor Primario
nivelsisben=1 regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 20. escolaridad=PRIMARIA locesp=Cuello uterino SAI
tipovivienda=CASA O APARTEMENTO regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1) ##
21. escolaridad=PRIMARIA biopsia=si nivelsisben=1
regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 22. estadocivil=CASADO discapacidad=NINGUNA radio=si
fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 23. estadocivil=CASADO discapacidad=NINGUNA
fuenteagua=ACUEDUCTO regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 24. estadocivil=CASADO metododx=Histologia Tumor Primario
radio=si fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 25. estadocivil=CASADO metododx=Histologia Tumor Primario
fuenteagua=ACUEDUCTO regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)

```

Figura 31. Mejores reglas generadas con Apriori con el conjunto de datos T235A22ALL.

También se obtuvieron reglas de asociación que relacionen solo los factores socioeconómicos con la supervivencia de las mujeres con cáncer de cuello uterino, diagnosticadas entre los años 1998

y 2002 con una ventana de observación hasta el 2007. Estas se obtuvieron utilizando el conjunto de datos T235A15ECON y con los mismos parámetros de la prueba anterior. Se encontraron las mejores reglas con una confianza del 80% y un soporte mínimo del 10 % (0.1). Los parámetros de ejecución del algoritmo Apriori se presentan en la figura 32 y las 18 reglas generadas en la figura 33.

```
##
## Apriori
## =====
##
## Minimum support: 0.1 (24 instances)
## Minimum metric <confidence>: 0.8
## Number of cycles performed: 18
##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 30
##
## Size of set of large itemsets L(2): 81
##
## Size of set of large itemsets L(3): 67
##
## Size of set of large itemsets L(4): 25
##
## Size of set of large itemsets L(5): 3
##
```

Figura 32. Parámetros de ejecución del algoritmo Apriori para los factores socioeconómicos del repositorio T235A15ECON.

De igual manera, se obtuvieron reglas de asociación que relacionen solo los factores clínicos con la supervivencia de las mujeres con cáncer de cuello uterino. Estas se obtuvieron utilizando el conjunto de datos T235A09CLINICO y con los mismos parámetros de las pruebas. De esta manera se lograron encontrar reglas con una confianza mínima del 80% (0.8) y un soporte mínimo del 10 % (0.1). Los parámetros de ejecución del algoritmo Apriori y las 8 reglas generadas, con una confianza del 53%, se presentan en la figura 34.

```

##
## Best rules found:
##
## 1. cabezafamilia=0 nmeses_2007=[84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 2. estadocivil=CASADO discapacidad=NINGUNA
fuenteagua=ACUEDUCTO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 3. discapacidad=NINGUNA cabezafamilia=0 nmeses_2007=[84,96]
24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 4. tipovivienda=CASA O APARTAMENTO
fuenteagua=ACUEDUCTO cabezafamilia=0 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 5. estadocivil=CASADO discapacidad=NINGUNA
fuenteagua=ACUEDUCTO regimen=SUBSIDIADO 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 6. ocupacion=HOGAR discapacidad=NINGUNA
fuenteagua=ACUEDUCTO cabezafamilia=0 24 ## ==> vivo_
muerto_2007=VIVO 24 conf:(1)
## 7. discapacidad=NINGUNA tipovivienda=CASA O
APARTAMENTO fuenteagua=ACUEDUCTO cabezafamilia=0 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 8. tipovivienda=CASA O APARTAMENTO
fuenteagua=ACUEDUCTO regimen=SUBSIDIADO cabezafamilia=0 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 9. ocupacion=HOGAR discapacidad=NINGUNA
fuenteagua=ACUEDUCTO regimen=SUBSIDIADO cabezafamilia=0 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 10. discapacidad=NINGUNA tipovivienda=CASA O
APARTAMENTO fuenteagua=ACUEDUCTO regimen=SUBSIDIADO
cabezafamilia=0 24 ## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 11. region=PASTO regimen=SUBSIDIADO 25
## ==> vivo_muerto_2007=VIVO 24 conf:(0.96)
## 12. ocupacion=HOGAR fuenteagua=ACUEDUCTO cabezafamilia=0
25
## ==> vivo_muerto_2007=VIVO 24 conf:(0.96)
## 13. ocupacion=HOGAR fuenteagua=ACUEDUCTO
regimen=SUBSIDIADO cabezafamilia=0 25
## ==> vivo_muerto_2007=VIVO 24 conf:(0.96)
## 14. cabezafamilia=1 26

```

```
## ==> vivo_muerto_2007=VIVO 24 conf:(0.92)
## 15. estadocivil=CASADO tipovivienda=CASA O APARTAMENTO
26
## ==> vivo_muerto_2007=VIVO 24 conf:(0.92)
## 16. regimen=NINGUNO nmeses_2007=[0,12] 26
## ==> vivo_muerto_2007=MUERTO 24 conf:(0.92)
## 17. regimen=NINGUNO cabezafamilia=0 nmeses_2007=[0,12] 26
## ==> vivo_muerto_2007=MUERTO 24 conf:(0.92)
## 18. ocupacion=HOGAR nivelsisben=1 28
## ==> vivo_muerto_2007=VIVO 24 conf: (0.86)
```

Figura 33. Parámetros de ejecución del algoritmo Apriori para los factores socioeconómicos del repositorio T235A15ECON.

```
## Apriori
## =====
##
## Minimum support: 0.1 (24 instances)
## Minimum metric <confidence>: 0.8
## Number of cycles performed: 18
##
## Generated sets of large itemsets:
##
## Size of set of large itemsets L(1): 21
##
## Size of set of large itemsets L(2): 76 ##
## Size of set of large itemsets L(3): 138
##
## Size of set of large itemsets L(4): 139
##
## Size of set of large itemsets L(5): 77
##
## Size of set of large itemsets L(6): 21
##
## Size of set of large itemsets L(7): 2
##
## Best rules found:
##
## 1. fuente=HOSPITALES locesp=Cuello uterino SAI nmeses_2007 =
(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 2. metododx=Histologia Tumor Primario fuente=HOSPITALES
locesp=Cuello uterino SAI nmeses_2007=(84,96] 24 ## ==> vivo_
muerto_2007=VIVO 24 conf:(1)
```

```

## 3. fuente=HOSP## ==> vivo_muerto_2007=VIVO 24 conf:(1)
ITALEs locesp=Cuello uterino SAI radio=si nmeses_2007=(84,96] 24
## 4. fuente=HOSPITALEs locesp=Cuello uterino SAI biopsia=si
nmeses_2007=(84,96] 24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 5. metododx=Histologia Tumor Primario fuente=HOSPITALEs
locesp=Cuello uterino SAI radio=si nmeses_2007=(84,96] 24 ## ==>
vivo_muerto_2007=VIVO 24 conf:(1)
## 6. metododx=Histologia Tumor Primario fuente=HOSPITALEs
locesp=Cuello uterino SAI biopsia=si nmeses_2007=(84,96] 24
## ==## 7. fuente=HOSPITALEs locesp=Cuello uterino SAI radio=si
biopsia=si nmeses_2007=(84,96] 24 => vivo_muerto_2007=VIVO 24
conf:(1)
## ==> vivo_muerto_2007=VIVO 24 conf:(1)
## 8. metododx=Histologia Tumor Primario fuente=HOSPITALEs
locesp=Cuello uterino SAI radio=si biopsia=si nmeses_2007=(84,96]
24
## ==> vivo_muerto_2007=VIVO 24 conf:(1)

```

Figura 34. Parámetros y reglas generadas con Apriori para los factores clínicos con el conjunto de datos T235A09CLINICO.

4.5.3 Tarea de Clustering

Se aplicó la tarea de clustering, de igual manera como se hizo en las tareas anteriores de clasificación y asociación, con los casos de las mujeres que sobrevivieron al cáncer invasivo de cuello uterino, diagnosticadas entre los años 1998 y 2002 con una ventana de observación hasta el 2007. Se utilizó el repositorio T235A22ALL descrito en la tabla 15, para encontrar grupos similares que relacionen los atributos socioeconómicos y clínicos de estas pacientes. Como las anteriores pruebas realizadas para la cohorte 1998-2007, lo importante es encontrar clusters donde como mínimo haya, un grupo de vivos y otro de muertos. Se configuró el número de clusters en 2, con una semilla por defecto de 10. Para evaluar los resultados del agrupamiento se utilizó el propio conjunto de entrenamiento Use tranining set igual al 66%. Como se puede observar en la tabla 21, el algoritmo K-means clasificó en el cluster 0 a 112 vivos y en el cluster 1 a 123 muertos. Los atributos que los determinan corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

Tabla 21. Clusters resultantes con K-means con el repositorio T235A22ALL

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_Primary	Histologia_Tumor_Primary
fuelle	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
tipooivienda	CASA O APARTAMENTO	CASA_O_ APARTAMENTO	CASA_O_ APARTAMENTO
fuelleagua	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_2007	[0,12]	[60,72]	[0,12]
vivo_muerto_2007	VIVO	VIVO	MUERTO

Para encontrar grupos que relacionen los factores socioeconómicos de las pacientes, diagnosticadas en el periodo 1998-2002 con una ventana de observación hasta el 2007, con cáncer invasivo de cuello uterino, se utilizó el conjunto de datos T235A15ECON descrito en la tabla 15. Se configuró el número de clusters en 2, 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%. En la tabla 22 se pueden observar los resultados con el algoritmo K-means clasificando en el clúster 0 a 105 vivos y en el cluster 1 a 130 muertos. De igual manera que el caso anterior, los atributos que los determinan corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

Tabla 22. Clusters resultantes con K-means con el repositorio T235A15E-CON

ATRIBUTO	FULL DATA (235)	CLUSTER 0 (105)	CLUSTER 1 (130)
edaddx	3	3	4
escolaridad	PRIMARIA	PRIMARIA	PRIMARIA
estadocivil	CASADO	CASADO	CASADO
ocupacion	HOGAR	HOGAR	HOGAR
discapacidad	NINGUNA	NINGUNA	NINGUNA
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_2007	[0,12]	[60,72]	[0,12]
vivo_muerto_ 2007	VIVO	VIVO	MUERTO

Para encontrar grupos que relacionen los factores clínicos de las pacientes, diagnosticadas en el periodo 1998-2002 con una ventana de observación hasta el 2007, con cáncer invasivo de cuello uterino, se utilizó el conjunto de datos T235A09CLINICO descrito en la tabla 15. Se encontró grupos de vivos y uno de muertos cuando se configuró el número de clusters en 3, con una semilla por defecto de 10. Para evaluar los resultados del

agrupamiento se utilizó el propio conjunto de entrenamiento Use tranining set igual al 66%. En la tabla 23 se pueden observar los resultados con el algoritmo K-means clasificando en el cluster 0 a 132 muertos, en el cluster 1 a 77 vivos y en el cluster 2 a 26 vivos. Como en los casos anteriores, los atributos que determinan a los grupos corresponden a las características del centroide de cada grupo, así, la media para los atributos numéricos y la moda para los atributos nominales (Hernández, Ramirez & Ferri, 2005).

Tabla 23. Clusters formados con K-means con el repositorio T235A09CLINICO

ATRIBUTO	FULL DATA (235)	CLUSTER 0 (132)
metododx	Histologia_Tumor_Primario	Histologia_Tumor_Primario
fuelle	HOSPITALES	HOSPITALES
morfologia	Squamous_cell_carcinoma, large_cell, nonkeratinizing,NOS	Squamous_cell_carcinoma,large cell,nonkeratinizing,NOS
locesp	Cuello_uterino_SAI	Cuello_uterino_SAI
cirugia	no	no
radio	si	si
biopsia	si	si
nmeses 2007	[0,12]	[0,12]
vivo_muerto_2007	VIVO	MUERTO
ATRIBUTO	CLUSTER 1 (77)	CLUSTER 2 (26)
metododx	Histologia_Tumor_Primario	Histologia_Tumor_Primario
fuelle	HOSPITALES	HOSPITALES
morfologia	Squamous_cell_carcinoma,NOS	Squamous_cell_carcinoma, large_cell, nonkeratinizing,NOS
locesp	Cuello_uterino_SAI	Cuello_uterino_SAI
cirugia	no	si
radio	si	si
biopsia	si	si
nmeses_2007	[84,96]	[60,72]
vivo_muerto_2007	VIVO	VIVO

4.6 ETAPA DE EVALUACIÓN E INTERPRETACIÓN DE RESULTADOS

En esta etapa se mide la calidad de los modelos obtenidos y 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.

Las técnicas de minería de datos producen modelos que explican de manera general los datos y su comprensibilidad al ser un factor subjetivo, depende en gran modo de la experiencia y conocimiento de los usuarios de los modelos (Hernández, Ramírez & Ferri, 2005). Los sistemas de reglas, como los utilizados en este estudio, son considerados como una de las representaciones que permiten comprender más fácilmente el comportamiento de un modelo, sin embargo, por el número de reglas, el número de atributos y las métricas de confiabilidad utilizadas, se hace necesaria su interpretación para facilitarles la comprensibilidad a los usuarios. Por otra parte, los diferentes modelos pueden producir patrones similares si se trabajan con repositorios de datos comunes, pero esto no significa que son redundantes ya que los diferencia la técnica que se aplica, el conjunto de datos que se utiliza, las métricas como la confianza y el soporte que se utiliza.

4.6.1 Evaluación e Interpretación de resultados para la cohorte 1998-2007

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 2007, almacenados en los conjuntos de datos T507A22ALL, T507A15ECON y T507A09CLINICO, aplicando las tareas de clasificación, asociación y clustering.

4.6.1.1 Modelos de Clasificación

Como se observó en la fase de minería de datos, evaluando el mejor modelo de clasificación que se obtuvo, utilizando el conjunto de datos T507A22ALL, fue el árbol de decisión de la prueba

4 (ver figura 8), construido con un factor de confianza $C=0.4$ y un mínimo de instancias por nodo $M=2$, utilizando el método de evaluación mediante validación cruzada, el cual clasificó 440 instancias correctamente, que corresponde a un porcentaje de precisión del 86.785% y 67 instancias incorrectamente, correspondiente a un porcentaje de error del 13,215%.

Ahora, si se observa la matriz de confusión del árbol, este clasifica correctamente a 139 casos de las pacientes que murieron por cáncer y 301 casos de las sobrevivientes. Además, clasifica incorrectamente a 34 casos de las pacientes que murieron y 33 casos de las pacientes que sobrevivieron. Teniendo en cuenta la distribución de los valores del atributo clase vivomuerto del repositorio T507A22ALL que es de 334 vivos y 173 muertos, esto significa que el modelo clasifica correctamente al 80.35% de las pacientes que murieron y el 90.12% de las pacientes que sobreviven.

Para medir la concordancia entre dos observadores en sus correspondientes clasificaciones según una serie de posibilidades (categorías) mutuamente excluyentes, se utiliza el coeficiente de kappa (Molinero, 2001), (Cerdeira & Villaroel, 2008). Este coeficiente está comprendido entre 0 y 1. 0. Así pues, en caso de concordancia perfecta el valor de kappa es 1; si la concordancia observada es igual a la esperada kappa vale 0; y en el caso de que el acuerdo observado sea inferior al esperado la índice kappa es menor que cero. Landis y Koch (1977) propusieron unos límites para valorar el grado de acuerdo en función del índice kappa. En la tabla 24 se muestran estos márgenes.

Tabla 24. Valoración del coeficiente Kappa

KAPPA	GRADO DE ACUERDO
<0	Sin acuerdo
0 - 0.2	Insignificante
0.2 - 0.4	Bajo
0.4 - 0.6	Moderado
0.6 - 0.8	Bueno
0.8 - 1.0	Muy bueno

Fuente: Landis y Koch (1977)

En el caso de la clasificación por árboles de decisión, este coeficiente mide la coincidencia de la predicción entre los resultados de las clasificaciones con el conjunto de entrenamiento y con el conjunto de prueba, con la clase real.

Teniendo en cuenta lo anterior, el modelo de clasificación que se obtuvo con el repositorio de datos T507A22ALL tiene un coeficiente kappa de 0.7056, que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera bueno.

Los porcentajes de instancias correctamente clasificadas presentados en el árbol como en la matriz de confusión y su grado de concordancia indican que el modelo cumple con los parámetros de confianza, establecidos en este estudio, para predecir nuevos casos y por lo tanto sus patrones son confiables.

Por otra parte, con el fin de mejorar la comprensibilidad del modelo de clasificación obtenido en la etapa de minería de datos (ver figura 9), se presenta la interpretación de los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios. El soporte es el porcentaje de casos que se clasifican de la manera como muestra el patrón, con relación al total de casos considerados en el estudio y la confianza es el porcentaje de casos que cumplen el patrón, con relación únicamente al total de casos de la clase, a la que el patrón clasifica. Es necesario aclarar que los patrones que se describen a continuación, son los que el algoritmo de clasificación (J48) obtuvo, basado en la ganancia de información de los atributos que forman parte de los respectivos patrones, teniendo en cuenta los datos del repositorio T507A22ALL y por lo tanto no se pueden considerar redundantes. En la discusión de resultados se presenta el análisis de estos patrones.

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 sobrepasa los 37 meses y el método de diagnóstico fue una Histología de Tumor Primario entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 40.04% de los

507 casos de mujeres con cáncer invasivo de cuello uterino consideradas en este estudio, se clasifican de esta manera y el 60.78% de las 334 mujeres sobrevivientes al cáncer cumplen con este patrón.

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses, pertenece al nivel 1 del SISBEN y el régimen de salud es subsidiado entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 12.23 % de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 18.56% de las 334 mujeres vivas cumplen con este patrón.
- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses y pertenece al nivel 2 del SISBEN entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 9.27 % de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 14.07% de las 334 sobrevivientes cumplen con este patrón.
- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses, no está clasificada en el sistema SISBEN, la localización específica del cáncer fue Cuello uterino SAI y fue diagnosticada en un hospital entonces la mujer tiene mayor probabilidad de morir en un periodo no mayor que cinco años. El 22.29 % de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasificaron de esta manera y el 65.32 % de las 173 mujeres que no sobrevivieron al cáncer cumplen con este patrón.

Como se observó en la fase de minería de datos, el modelo de clasificación que se obtuvo para los factores socioeconómicos, utilizando el conjunto de datos T507A15ECON de la figura 10, utilizando el método de evaluación mediante validación cruzada, muestra que el modelo tiene un porcentaje de precisión del

84.8126% correspondiente a 430 instancias correctamente clasificadas, de un total de 507 instancias, y un porcentaje de error de 15.1874% correspondiente a 77 instancias incorrectamente clasificadas.

En la matriz de confusión se puede observar que 136 de 173 casos de mujeres que no sobrevivieron al cáncer son correctamente clasificadas y 294 de 334 casos de mujeres que sobrevivieron al cáncer también están de igual manera clasificadas. Por otra parte, 37 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 40 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 78.61% de las pacientes que no sobrevivieron al cáncer y al 88.02% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.6636, que de acuerdo al grado de concordancia de la clasificación de la tabla 24 se considera moderado.

De igual manera, con el fin de mejorar la comprensibilidad del modelo de clasificación obtenido en la etapa de minería de datos, para los factores socioeconómicos asociados a la supervivencia en mujeres con cáncer invasivo de cuello uterino (ver figura 11), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 sobrepasa los 37 meses entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 41.81% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 63.47% de las 334 mujeres que sobreviven al cáncer cumplen con este patrón.
- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses, es cabeza de familia y no tiene SISBEN entonces la mujer tiene mayor probabilidad de morir en un periodo no mayor que cinco años. El 32.54% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se

clasifican de esta manera y el 95.38% de las 173 mujeres que no sobreviven al cáncer cumplen con este patrón.

Como se observó en la fase de minería de datos, el modelo de clasificación que se obtuvo para los factores clínicos, utilizando el conjunto de datos T507A09CLINICO de la figura 12, utilizando el método de evaluación mediante validación cruzada, muestra que de un total de 507 instancias, 401 instancias son correctamente clasificadas, correspondiente a un porcentaje de precisión del 79.0927% y 106 instancias incorrectamente clasificadas, correspondiente a un porcentaje de error de 20.9073%.

En la matriz de confusión se puede observar que 136 de 173 casos de mujeres que no sobrevivieron al cáncer están correctamente clasificadas y 265 de 334 casos de mujeres que sobrevivieron al cáncer también lo están. Por otra parte, 37 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 69 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 78.61% de las pacientes que no sobrevivieron al cáncer y al 79.34% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.5548, que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera moderado.

De igual manera, con el fin de mejorar la comprensibilidad del modelo de clasificación obtenido en la etapa de minería de datos, para los factores clínicos relacionados con la supervivencia en mujeres con cáncer invasivo de cuello uterino (ver figura 13), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 sobrepasa los 37 meses y el método de diagnóstico fue una Histología de Tumor Primario entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 40.04% de los 507 casos de mujeres con cáncer invasivo de cuello uterino consideradas en este estudio, se clasifican de esta manera y el 60.78% de las 334 mujeres sobrevivientes al cáncer cum-

plen con este patrón. Este patrón es el mismo que se obtuvo con el modelo general.

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses y se le realizó una cirugía para extirpar el tumor entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 12.50 % de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 19.16% de las 334 mujeres que sobrevivieron al cáncer cumplen con este patrón.
- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el final del año 2007 no sobrepasa los 37 meses, no se le realizó cirugía como método de tratamiento, la fuente de diagnóstico fue un hospital y la localización específica del cáncer es en el Cuello Uterino SAI entonces la mujer tiene mayor probabilidad de morir en un periodo no mayor que cinco años. El 29.33% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera y el 85.95% de las 173 mujeres que no sobreviven al cáncer cumplen con este patrón.

Dado que los patrones socioeconómicos y clínicos obtenidos con los repositorios T507A15ECON y T507A09CLINICO respectivamente, son similares a los patrones generales resultantes con el repositorio T507A22ALL, se optó por no tener en cuenta el atributo nmeses en los dos repositorios y volver a generar estos patrones.

Evaluando el modelo construido en la fase de minería de datos, con el conjunto de datos T507A15ECON, para los factores socioeconómicos relacionados con la supervivencia de las mujeres diagnosticadas con cáncer de cuello uterino, sin tener en cuenta el atributo nmeses (ver figura 14), utilizando el método de evaluación mediante validación cruzada, se puede observar que este clasifica correctamente a 401 instancias de 507, con una precisión del 79.0927%. El modelo clasifica incorrectamente a 106 instancias que representa un porcentaje de error del 20.9073%.

En la matriz de confusión se puede observar que 114 de 173 casos de mujeres que no sobrevivieron al cáncer son clasificadas correctamente y 287 de 334 casos de mujeres que sobrevivieron al cáncer también están de igual manera clasificadas. Por otra parte, 59 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 47 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 65.907% de las pacientes que no sobrevivieron al cáncer y al 85.93% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.5271, que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera moderado.

Con el fin de mejorar la comprensibilidad del nuevo modelo de clasificación obtenido en la etapa de minería de datos, para los factores socioeconómicos relacionados con la supervivencia en mujeres con cáncer invasivo de cuello uterino, sin tener en cuenta el atributo nmeses, (ver figura 15), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si una paciente pertenece a un régimen de salud subsidiado entonces ella tiene mayor probabilidad de sobrevivir al cáncer invasivo de cuello uterino por lo menos cinco años. El 45.76 % de los 507 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 69.46% de las 334 mujeres que sobrevivieron al cáncer cumplen con este patrón.
- Si una paciente pertenece a un régimen de salud contributivo entonces ella tiene mayor probabilidad de sobrevivir al cáncer invasivo de cuello uterino por lo menos cinco años. El 20.32 % de los 507 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 30.84% de las 334 mujeres que sobrevivieron al cáncer cumplen con este patrón.
- Si una paciente no pertenece a ningún régimen de salud entonces ella tiene mayor probabilidad de morir de cáncer

invasivo de cuello uterino en un periodo no mayor que cinco años. El 17.36 % de los 507 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 50.87% de las 173 mujeres que no sobrevivieron al cáncer cumplen con este patrón.

De igual manera, evaluando el modelo construido en la fase de minería de datos con el conjunto de datos T507A09CLINICO, para los factores clínicos relacionados con la supervivencia de las mujeres diagnosticadas con cáncer de cuello uterino, sin tener en cuenta el atributo nmeses (ver figura 16), utilizando el método de evaluación mediante validación cruzada, se puede observar que este clasifica correctamente a 366 instancias de 507, con una precisión del 72.1893%. El modelo clasifica incorrectamente a 141 instancias que representa un porcentaje de error del 27.8107%.

En la matriz de confusión se puede observar que 34 de 173 casos de mujeres que no sobrevivieron al cáncer son clasificadas correctamente y 332 de 334 casos de mujeres que sobrevivieron al cáncer también están de igual manera clasificadas. Por otra parte, 139 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 2 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 19.65% de las pacientes que no sobrevivieron al cáncer y al 99.40% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.2355, que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera bajo.

Con el fin de mejorar la comprensibilidad del nuevo modelo de clasificación obtenido en la etapa de minería de datos, para los factores clínicos relacionados con la supervivencia en mujeres con cáncer invasivo de cuello uterino, sin tener en cuenta el atributo nmeses, (ver figura 16), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si una paciente fue diagnosticada con cáncer de cuello uterino mediante una histología de tumor primario y la morfología del tumor era Squamous cell carcinoma, large cell, nonke-

ratinizing, NOS, entonces ella tiene mayor probabilidad de sobrevivir a este tipo de cáncer, por lo menos cinco años. El 34.12 % de los 507 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 51.80% de las 334 mujeres que sobrevivieron al cáncer cumplen con este patrón.

- Si una paciente fue diagnosticada con cáncer de cuello uterino mediante una histología de tumor primario y la morfología del tumor era Squamous cell carcinoma, NOS, entonces ella tiene mayor probabilidad de sobrevivir a este tipo de cáncer, por lo menos cinco años. El 33.14 % de los 507 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 50.30% de las 334 mujeres que sobrevivieron al cáncer cumplen con este patrón.

4.6.1.2 Modelos de Asociación

Se utilizó el conjunto de datos T507A22ALL para obtener reglas de asociación que relacionen los factores socioeconómicos y clínicos de 507 mujeres, diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998 al 2007, con la supervivencia de estas.

Las 25 reglas de asociación generadas (ver figura 26) 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 5 más representativas están:

Regla 5. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, tienen una edad entre 15 y 28 años, no tienen ninguna discapacidad, se les practicó una biopsia de la lesión y poseen servicio público de acueducto. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 8. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, tienen una edad

entre 15 y 28 años, con un cáncer localizado en el Cuello Uterino SAI, poseen servicio público de acueducto 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 9. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, son de un nivel de escolaridad primaria, de ocupación hogar, no tienen ninguna discapacidad y fueron diagnosticadas en un hospital. 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 al cáncer invasivo de cuello uterino por lo menos cinco años, son de un nivel de escolaridad primaria, no tienen ninguna discapacidad, pertenecen al régimen de salud subsidiado y son cabeza de familia. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 14. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, no tienen ninguna discapacidad, su cáncer tiene una morfología de tipo Squamous cell carcinoma, NOS, localizado en el Cuello uterino SAI y poseen servicio público de acueducto. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 16. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, fueron diagnosticadas mediante una Histología de Tumor Primario, su cáncer tiene una morfología de tipo Squamous cell carcinoma, large cell, nonkeratinizing NOS, se les practicó una biopsia de la lesión y viven en una casa o apartamento. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Con respecto a las reglas de asociación que relacionan solo los factores socioeconómicos de los 507 casos de mujeres, diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998

al 2007, con la supervivencia de estas, se utilizó el conjunto de datos T507A15ECON.

Se generaron 11 reglas de asociación (ver figura 28) con una confianza mínima del 80% y un soporte mínimo del 10%. De las 11 reglas encontradas, 8 tienen una confianza del 100% y 3 una confianza mayor que la confianza mínima. Por cumplir con los parámetros establecidos, todas estas reglas son reglas fuertes (strong rules) y por lo tanto interesantes, significativas y con una alta precisión.

Entre las reglas de asociación de tamaño 5 más representativas que asocian los factores socioeconómicos están:

Regla 4. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, trabajan en el hogar, poseen servicio público de acueducto, son del nivel 1 del SISBEN 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 5. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, son de un nivel de escolaridad primaria, no tienen ninguna discapacidad, poseen servicio público de acueducto y son del nivel 1 del SISBEN. 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 al cáncer invasivo de cuello uterino por lo menos cinco años, viven en una casa o apartamento, poseen servicio público de acueducto, pertenecen a un régimen de salud subsidiado y no son cabeza de familia. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 8. El 100% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, son de un nivel de escolaridad primaria, no tienen ninguna discapacidad, poseen servicio público de acueducto, son de nivel 1 del SISBEN y pertenecen a un régimen de salud subsidiado. El 10% de todas

las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

En cuanto a las reglas de asociación que relacionan solo los factores clínicos de los 507 casos de mujeres, diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998 al 2007, con la supervivencia de estas, se utilizó el conjunto de datos T507A-09CLINICO.

Únicamente 2 reglas de asociación (ver figura 29) cumplen con los requerimientos mínimos de confianza del 50% (se tuvo que bajar del 80% al 50% para poder encontrar reglas) y de soporte del 10%. Aunque la confianza no es buena, estas dos reglas son reglas fuertes (strong rules) y por lo tanto pueden ser interesantes. Estas son:

Regla 1. El 53% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, fueron diagnosticadas mediante una Histología de Tumor Primario, en un hospital, con cáncer localizado en el Cuello uterino SAI, no se les practicó cirugía, se les realizó un tratamiento con radioterapia y tienen un tiempo de diagnosticadas hasta de 12 meses con relación a la ventana del 31 de diciembre de 2007. El 19.13% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 2. El 52% de las mujeres que sobreviven al cáncer invasivo de cuello uterino por lo menos cinco años, fueron diagnosticadas mediante una Histología de Tumor Primario, en un hospital, con cáncer localizado en el Cuello uterino SAI, no se les practicó cirugía y fueron diagnosticadas de esta enfermedad entre 0 y 12 meses antes de 31 de diciembre de 2007, fecha límite de la ventana de observación del estudio. El 19.33% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

4.6.1.3 Modelos de Clustering

Se utilizó el conjunto de datos T507A22ALL, para aplicarle la técnica de clustering, con el fin de formar grupos homogéneos

entre todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2007, y que relacionen los factores socioeconómicos y clínicos de estas mujeres.

Como muestran los resultados de la tabla 18, se formaron dos clusters (parámetro $K=2$). En el cluster 0 se agruparon a 304 mujeres que posiblemente sobrevivieron al cáncer de cuello uterino por lo menos cinco años y en el cluster 1 se agruparon a 203 mujeres que supuestamente no sobrevivieron en un periodo no mayor a cinco años. En el repositorio T507A22ALL hay realmente 334 mujeres sobrevivientes al cáncer y 173 que no sobrevivieron. Por esa razón, los resultados indican que 30 pacientes, que realmente sobrevivieron, están en el grupo de las que murieron, porque sus características se asemejan más a las de ese grupo. Midiendo la precisión del modelo, se puede decir que este agrupa correctamente al 91% de las mujeres sobrevivientes al cáncer, con una tasa de error general del 5.9% 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 60% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2007 sobreviven al cáncer por lo menos cinco años y son aquellas cuya edad está entre 42 y 54 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 y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 24 y 36 meses.

Cluster 1. El 40% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2007 no sobreviven al cáncer en un periodo no mayor a cinco años y son aquellas cuya edad está entre 55 y 67 años, su cáncer tiene una morfología de tipo Squamous cell carcinoma, NOS, no están en el sistema SISBEN, no pertenecen a ningún régimen de salud y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 0 y 12 meses.

Para encontrar grupos similares que relacionen los factores socioeconómicos de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2007, se utilizó el conjunto de datos T507A15ECON. Como muestran los resultados de la tabla 19, se formaron dos clusters (parámetro $K=2$), en los cuales, en el cluster 0 se encuentran 298 mujeres que supuestamente sobrevivieron al cáncer y en el cluster 1 se agruparon a 209 mujeres que posiblemente no sobrevivieron. Teniendo en cuenta que se conoce realmente el número de mujeres sobrevivientes y las que no, los resultados indican que únicamente 36 pacientes de las 334 encontradas como vivas no tienen esa similitud y por sus características se asemejan más a las que murieron. Midiendo la precisión del modelo para los factores socioeconómicos, se puede decir que este agrupa correctamente al 89.2 % de las mujeres sobrevivientes al cáncer, con una tasa de error general del 7.1% que indica que el modelo es bueno.

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

Cluster 0. El 58.8% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2007 sobreviven al cáncer por lo menos cinco años y son aquellas cuya edad está entre 42 y 54 años, son de nivel 1 del SISBEN y pertenecen a un régimen de salud subsidiado.

Cluster 1. El 41.2% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2007 no sobreviven al cáncer en un periodo no mayor a cinco años y son aquellas cuya edad está entre 55 y 67 años, no están en el sistema SISBEN y pertenecen a un régimen de salud contributivo.

Para encontrar grupos similares que relacionen los factores clínicos de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2007, se utilizó el conjunto de datos T507A09CLINICO. Como muestran los resultados de la tabla 20, se formaron siete clusters (parámetro $K=7$), en los cuales, en el cluster 0 están 121 mujeres que sobrevivieron al cáncer, en el

cluster 1 están 11 sobrevivientes, en el cluster 2 se encuentran 68 sobrevivientes, en el cluster 3 se encuentran 146 sobrevivientes, en el cluster 4, están 117 sobrevivientes, en el cluster 5 están 11 sobrevivientes y en el cluster 6 se agruparon a 33 mujeres que no sobrevivieron.

El modelo agrupó en 6 clusters a las mujeres sobrevivientes, que suman 474 y en un cluster a las que no (33). La diferencia entre el número de mujeres que realmente sobrevivieron y el número de las agrupadas en los clusters de sobrevivientes es de 140. Esto indica que a este grupo llegaron 140 mujeres que realmente no sobrevivieron al cáncer invasivo de cuello uterino, pero que por sus similitudes se agrupan en los clusters de sobrevivientes. Midiendo la precisión de este modelo, se puede decir que este agrupa correctamente al 19% de las mujeres que no sobrevivieron al cáncer, con una tasa de error general del 27.6% que indica que el modelo no es bueno.

Teniendo en cuenta las características o atributos clínicos que diferencian a estos siete clusters de la moda que se encuentra en el grupo denominado full data y en el cluster 4 (ver tabla 20) y que el número de casos o soporte en cada cluster sea superior al 10%, los patrones más representativos son:

Cluster 0. El 23.9% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2007 sobreviven al cáncer por lo menos cinco años y son aquellas que se les practicó cirugía como tratamiento del cáncer y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 24 y 36 meses.

Cluster 2. El 13.4% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino, en el periodo 1998-2007 sobreviven al cáncer por lo menos cinco años y son aquellas cuyo número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 96 y 108 meses.

Cluster 3. El 28.8% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino, en el periodo 1998-2007 sobreviven al cáncer por lo menos cinco años y son aquellas que su

cáncer tiene una morfología de tipo Squamous cell carcinoma, NOS y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 12 y 24 meses.

4.6.2 Evaluación e Interpretación de resultados en la cohorte 1998-2002

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 los conjuntos de datos T235A22ALL, T235A15ECON y T235A09CLINICO, aplicando las tareas de clasificación, asociación y clustering.

4.6.2.1 Modelos de Clasificación

Evaluando los resultados obtenidos en la fase de minería de datos para la prueba de clasificación realizada con el conjunto de datos T235A22ALL, 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, utilizando el método de evaluación mediante validación cruzada, se puede observar que el árbol de decisión resultante (ver figura 22), clasifica 219 instancias correctamente, que corresponde a un porcentaje de precisión del 93.1915%, y 16 instancias incorrectamente, correspondiente a un porcentaje de error del 6.8085%.

Teniendo en cuenta la distribución de los valores del atributo clase vivomuerto del repositorio T235A22ALL 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. Por otra parte, 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. El estadístico Kappa es de 0.8638,

que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera muy bueno.

Los porcentajes de instancias correctamente clasificadas presentados en el árbol como en la matriz de confusión y su grado de concordancia indican que el modelo cumple con los parámetros de confianza, establecidos en este estudio, para predecir nuevos casos y por lo tanto sus patrones son confiables.

Por otra parte, con el fin de mejorar la comprensibilidad del modelo de clasificación obtenido en la etapa de minería de datos (ver figura 23), se presenta la interpretación de los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- 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 tiene mayor probabilidad de sobrevivir por lo menos cinco años. 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 52 y la localización específica del cáncer fue “Cuello uterino SAI” entonces la mujer tiene mayor probabilidad de morir en un periodo no mayor que cinco años. 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.

De igual manera, si se observan los resultados obtenidos en la fase de minería de datos, para los factores socioeconómicos,

utilizando el conjunto de datos T235A15ECON de la figura 24, utilizando el método de evaluación mediante validación cruzada, muestran que, de un total de 235 instancias, 212 fueron clasificadas correctamente para un porcentaje de precisión del 90.2128% y 23 fueron clasificadas incorrectamente para un porcentaje de error de 9.7872%.

En la matriz de confusión se puede observar que 99 de 105 casos de mujeres que no sobrevivieron al cáncer son correctamente clasificadas y 113 de 130 casos de mujeres que sobrevivieron al cáncer también están correctamente clasificadas. Por otra parte, 6 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 17 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 94.3% de las pacientes que no sobrevivieron al cáncer y al 86.9% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.804, que de acuerdo al grado de concordancia de la clasificación de la tabla 24 se considera muy bueno.

Como en los casos anteriores, con el fin de mejorar la comprensibilidad del modelo de clasificación obtenido en la etapa de minería de datos, para los factores socioeconómicos asociados a la supervivencia en mujeres con cáncer invasivo de cuello uterino (ver figura 25), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- 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 entonces la mujer tiene mayor probabilidad de sobrevivir por lo menos cinco años. El 48.1% de los 235 casos de mujeres consideradas en este estudio, se clasifican de esta manera y el 86.9% 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 que 52 entonces la mujer tiene mayor probabilidad de morir en un periodo no mayor que cinco años. El 42.1% de los 235 casos de mujeres consideradas en este estudio, se clasifican de esta manera y el 94.3% de las 105 mujeres no sobrevivientes al cáncer cumplen con este patrón.

Por otra parte, si se observan los resultados obtenidos en la etapa de minería de datos para los factores clínicos, utilizando el conjunto de datos T235A09CLINICO de la figura 26, utilizando el método de evaluación mediante validación cruzada, muestran que de un total de 235 instancias, 219 fueron clasificadas correctamente para un porcentaje de precisión del 93.1915% y 16 fueron clasificadas incorrectamente para un porcentaje de error de 6.8085%.

En la matriz de confusión se puede observar que 103 de 105 casos de mujeres que no sobrevivieron al cáncer están correctamente clasificadas y 116 de 130 casos de mujeres que sobrevivieron al cáncer también lo están. Por otra parte, 2 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 14 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 98.1% de las pacientes que no sobrevivieron al cáncer y al 89.2% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.8638, que de acuerdo al grado de concordancia de la clasificación de la tabla 24 se considera muy bueno.

El modelo que se construyó con el repositorio T235A09CLINICO que se muestra en la figura 27, es igual al modelo general construido con el conjunto de datos T235A22ALL, presentado en la figura 22, por lo tanto, los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino descubiertos y que relacionan los factores clínicos son los mismos que los patrones generales.

Como en la cohorte 1998-2007, los patrones socioeconómicos y clínicos obtenidos con los repositorios T235A15ECON y T235A09CLINICO respectivamente, son similares a los patrones generales resultantes con el repositorio T235A22ALL, se optó por

no tener en cuenta el atributo nmeses en los dos repositorios y volver a generar estos patrones.

Evaluando el modelo construido en la etapa de minería de datos, con el conjunto de datos T235A15ECON para los factores socioeconómicos, que inciden en la supervivencia de las mujeres diagnosticadas con cáncer de cuello uterino, sin tener en cuenta el atributo nmeses y que se presenta en la figura 28, se puede observar que este clasifica correctamente a 184 instancias de 235, con una precisión del 78.2979%. El modelo clasifica incorrectamente a 51 instancias que representa un porcentaje de error del 21.7021%.

En la matriz de confusión se puede observar que 97 de 105 casos de mujeres que no sobrevivieron al cáncer son clasificadas correctamente y 87 de 130 casos de mujeres que sobrevivieron al cáncer también están de igual manera clasificadas. Por otra parte, 8 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 43 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 92.4% de las pacientes que no sobrevivieron al cáncer y al 66.9% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.5746, que de acuerdo al grado de concordancia de la clasificación de la tabla 24 se considera moderado.

Con el fin de mejorar la comprensibilidad del nuevo modelo de clasificación obtenido en la etapa de minería de datos, para los factores socioeconómicos relacionados con la supervivencia en mujeres con cáncer invasivo de cuello uterino, sin tener en cuenta el atributo nmeses, (ver figura 28), se presenta la interpretación de los patrones más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si una paciente pertenece a un régimen de salud subsidiado entonces ella tiene mayor probabilidad de sobrevivir al cáncer invasivo de cuello uterino por lo menos cinco años. El 26.8 % de los 235 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 48.5%

de las 130 mujeres que sobrevivieron al cáncer cumplen con este patrón.

- Si una paciente no pertenece a ningún régimen de salud entonces ella tiene mayor probabilidad de morir de cáncer invasivo de cuello uterino en un periodo no mayor que cinco años. El 24.3 % de los 235 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 54.3% de las 105 mujeres que no sobrevivieron al cáncer cumplen con este patrón.

De igual manera, evaluando el modelo construido en la etapa de minería de datos, con el conjunto de datos T235A09CLINICO, para los factores clínicos que inciden en la supervivencia de las mujeres diagnosticadas con cáncer de cuello uterino, sin tener en cuenta el atributo nmeses y que se puede ver en la figura 29, utilizando el método de evaluación mediante validación cruzada, se puede observar que este clasifica correctamente a 162 instancias de 235, con una precisión del 68.9362%. El modelo clasifica incorrectamente a 73 instancias que representa un porcentaje de error del 31.0638%.

En la matriz de confusión se puede observar que 53 de 105 casos de mujeres que no sobrevivieron al cáncer son clasificadas correctamente y 109 de 130 casos de mujeres que sobrevivieron al cáncer también están de igual manera clasificadas. Por otra parte, 52 mujeres que no sobrevivieron están clasificadas incorrectamente y de igual manera 21 mujeres que sobrevivieron. Esto significa que el modelo clasifica correctamente al 50.5% de las pacientes que no sobrevivieron al cáncer y al 83.8% de las pacientes que sobrevivieron. El estadístico Kappa es de 0.3533, que según el grado de concordancia de la clasificación descrito en la tabla 24, se considera bajo.

Con el fin de mejorar la comprensibilidad del nuevo modelo de clasificación obtenido en la etapa de minería de datos, para los factores clínicos relacionados con la supervivencia en mujeres con cáncer invasivo de cuello uterino, sin tener en cuenta el atributo nmeses, (ver figura 29), se presenta la interpretación de los patro-

nes más representativos descubiertos, de acuerdo a su confianza y soporte y de esta manera sean entendibles por los usuarios:

- Si una paciente fue diagnosticada con cáncer de cuello uterino mediante una histología de tumor primario y la morfología del tumor era Squamous cell carcinoma, large cell, nonkeratinizing, NOS, entonces ella tiene mayor probabilidad de sobrevivir a este tipo de cáncer, por lo menos cinco años. El 35.3 % de los 235 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 63.80% de las 130 mujeres que sobrevivieron al cáncer cumplen con este patrón.
- Si una paciente fue diagnosticada con cáncer de cuello uterino mediante una histología de tumor primario y la morfología del tumor era Squamous cell carcinoma, NOS, no se le practica cirugía para extirpar el tumor y fue atendida en un hospital entonces la paciente tiene mayor probabilidad de morir con este tipo de cáncer, en un periodo no mayor que cinco años. El 14.04 % de los 235 casos de mujeres diagnosticadas con este tipo de cáncer se clasifican de esta manera y el 31.43% de las 105 mujeres que no sobrevivieron al cáncer cumplen con este patrón.

4.6.2.2 Modelos 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 31, 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 4 más representativas están:

Regla 2. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, 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 al cáncer de cuello uterino por lo menos cinco años, 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 al cáncer de cuello uterino por lo menos cinco años, 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 al cáncer de cuello uterino por lo menos cinco años, 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 al cáncer de cuello uterino por lo menos cinco años, 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 al cáncer de cuello uterino por lo menos cinco años, 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.

Se utilizó el conjunto de datos T507A15ECON para generar reglas de asociación que relacionan solo los factores socioeconómicos de los 235 casos de 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.

Se generaron 18 reglas de asociación que se muestran en la figura 33, con una confianza mínima del 80% y un soporte mínimo del 10%. Diez reglas tienen una confianza del 100% y ocho reglas, una confianza mayor que 80%. Por cumplir con los parámetros establecidos, todas estas reglas son fuertes (strong rules) y por lo tanto interesantes, significativas y con una alta precisión.

Entre las reglas de asociación de tamaño 5 más representativas que asocian los factores socioeconómicos están:

Regla 5. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, son casadas, no tienen ninguna discapacidad, poseen servicio público de acueducto y pertenecen al régimen subsidiado de salud. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 6. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, trabajan en el hogar, no tienen ninguna discapacidad, poseen servicio público de acueducto y son cabeza de familia. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 8. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, viven en una casa o apartamento, poseen servicio público de acueducto, pertenecen a un régimen de salud subsidiado y son cabeza de familia. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Entre las reglas cuya confianza es mayor que 80% pero menor que 100% están:

Regla 13. El 96% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, trabajan en el hogar, poseen servicio público de acueducto, pertenecen a un régimen de salud subsidiado y son cabeza de familia. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Regla 17. El 92% de las mujeres que no sobreviven al cáncer de cuello uterino por un periodo no mayor a cinco años, no pertenecen a un régimen de salud, son cabeza de familia y el número de meses de vida contados a partir de la fecha de diagnóstico está entre 0 y 12 meses. El 10% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino cumplen con este patrón.

Se utilizó el conjunto de datos T235A09CLINICO para generar reglas de asociación que relacionan solo los factores socioeconómicos de los 235 casos de 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 8 reglas de asociación generadas y que se muestran en la figura 34, 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 4 más representativas que asocian los factores clínicos están:

Regla 2. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, fueron diagnosticadas mediante el método de Histología de Tumor Primario, 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 8. El 100% de las mujeres que sobreviven al cáncer de cuello uterino por lo menos cinco años, fueron diagnosticadas mediante una Histología de Tumor Primario, en un hospital, con cáncer localizado en el Cuello uterino SAI, se les realizó un tratamiento con radioterapia, se les practicó una biopsia como procedimiento diagnóstico 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.

4.6.2.3 Modelos de Clustering

Se utilizó el conjunto de datos T235A22ALL, para aplicarle la técnica de clustering, con el fin de encontrar grupos homogéneos 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 21, 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 agruparon 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 al cáncer por lo menos cinco

años 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 en un periodo no mayor que cinco años 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.

Para encontrar grupos similares que relacionen los factores socioeconómicos de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2003, observadas hasta 2007, se utilizó el conjunto de datos T235A15ECON. Como muestran los resultados de la tabla 22, se formaron dos clusters (parámetro $K=2$), en los cuales, en el cluster 0 se encuentran 105 mujeres que por su similitud posiblemente sobrevivieron al cáncer y en el cluster 1 se agruparon a 130 mujeres que supuestamente no sobrevivieron. En el repositorio T235A15ECON hay 130 sobrevivientes y 105 que no lo son. Estas cifras muestran que 25 pacientes de las 130 realmente vivas no tienen esa similitud y por sus características se asemejan más a las que murieron. Midiendo la precisión del modelo, se puede decir que este agrupa correctamente al 80.8% de las mujeres sobrevivientes al cáncer, con una tasa de error general del 10.6% que indica que el modelo es aceptable.

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

Cluster 0. El 45% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002 y observadas hasta el 2007, sobreviven al cáncer por lo menos cinco años y son aquellas cuya edad está entre 42 y 54 años, 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 55% 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 en un periodo no mayor que cinco años y son aquellas cuya edad está entre 55 y 67 años, 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.

Para encontrar grupos similares que relacionen los factores clínicos de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino entre los años 1998 y 2002, con una ventana de observación hasta el 2007, se utilizó el conjunto de datos T235A09CLINICO. Como muestran los resultados de la tabla 23, se formaron tres clusters (parámetro $K=3$), en los cuales, en el cluster 0 están 132 mujeres supuestamente que no sobrevivieron al cáncer, en el cluster 1 están 77 consideradas sobrevivientes y en el cluster 2 se encuentran 68 supuestamente sobrevivientes. El modelo agrupó en 2 clusters a las mujeres sobrevivientes, que sumándolas da un resultado de 103, y en un cluster a las que no. Midiendo la precisión del modelo, se puede decir que este agrupa correctamente al 79.2% de las mujeres sobrevivientes al cáncer, con una tasa de error general del 7.2% que indica que el modelo es aceptable.

Teniendo en cuenta las características o atributos clínicos que diferencian al cluster 0, que agrupa a las pacientes que se consideran que no sobrevivieron al cáncer, con las del cluster 1 y cluster 2 que agrupan a las mujeres que se consideran sobrevivientes, los patrones más representativos son:

Cluster 0. El 56 % de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino en el periodo 1998-2002, observadas hasta el año 2007, no sobreviven al cáncer en un periodo no mayor que cinco años y son aquellas que su cáncer tiene una morfología de tipo Squamous cell carcinoma, large cell, nonkeratinizing NOS y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 0 y 12 meses.

Cluster 1. El 33% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino, en el periodo 1998-2002, observadas hasta el año 2007, sobreviven al cáncer por lo menos cinco años y son aquellas que su cáncer tiene una morfología de tipo Squamous cell carcinoma, NOS y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 84 y 96 meses.

Cluster 2. El 11% de todas las mujeres diagnosticadas con cáncer invasivo de cuello uterino, en el periodo 1998-2002, observadas hasta el año 2007, sobreviven al cáncer por lo menos cinco años y son aquellas a las cuales se les practicó cirugía y el número de meses de vida, contados a partir de la fecha de diagnóstico, está entre 60 y 72 meses.

CAPÍTULO V

DISCUSIÓN DE RESULTADOS

El objetivo de esta investigación fue detectar patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino con técnicas de minería de datos a partir de la información almacenada en el RPCMP en el periodo 1998-2007. Para cumplir este objetivo se escogieron las tareas de clasificación, asociación y clustering. Se analizaron 507 casos de mujeres con cáncer invasivo de cuello uterino. Se obtuvieron diferentes tipos de patrones de supervivencia dependiendo de la técnica de minería de datos aplicada. Los patrones resultantes de clasificación con árboles de decisión, permiten predecir las características de las mujeres que sobreviven, para los nuevos casos de cáncer invasivo de cuello uterino diagnosticados en el municipio de Pasto. Las reglas de asociación descubiertas determinan cuales factores o atributos, que inciden en la supervivencia de las mujeres con cáncer, se están presentando juntos y los clusters obtenidos, permiten conocer las características de las mujeres con cáncer invasivo de cuello uterino, que forman los distintos grupos, basado en sus similitudes.

Cabe recordar que el tiempo de supervivencia se define como el tiempo transcurrido desde el acontecimiento o estado inicial

que puede ser el diagnóstico de la enfermedad hasta el estado final que puede ser la muerte o una ventana de observación. En las enfermedades crónicas, tales como el cáncer, la supervivencia se mide como una probabilidad de permanecer vivo durante una determinada cantidad de tiempo. La supervivencia al año o a los 5 años es a menudo expresada como indicador de la severidad de una enfermedad y como pronóstico. Típicamente, el pronóstico del cáncer se valora determinando el porcentaje de pacientes que sobrevive por lo menos cinco años después del diagnóstico. En este estudio, la supervivencia de estas mujeres se tomó a cinco años, teniendo en cuenta que el RPCMP depura la información de acuerdo a la metodología internacionalmente establecida para estimar la incidencia, mortalidad y prevalencia de cáncer por quinquenios.

Con clasificación, se ha descubierto un patrón general que predice que una mujer tiene un tiempo de supervivencia de por lo menos cinco años, si su tiempo de vida sobrepasa los 37 meses después de haber sido diagnosticada con cáncer invasivo de cuello uterino. El 63.5% de las mujeres que sobreviven sobrepasan este umbral. Al no existir estudios en el Departamento de Nariño y en general en Colombia, que determinen patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino con minería de datos, que permitan confrontar el conocimiento descubierto con este patrón general, se realizaron las mismas pruebas de clasificación para obtener estos patrones, en la cohorte 1998-2002, donde se analizaron 235 casos. El patrón obtenido en esta cohorte fue que una mujer tiene un tiempo de supervivencia de por lo menos cinco años, si su tiempo de vida sobrepasa los 52 meses después de haber sido diagnosticada con cáncer invasivo de cuello uterino. El 86.9% de las mujeres que sobreviven, sobrepasan este umbral.

Analizando estos dos patrones, la disminución del umbral de 52 a 37 meses de vida, para considerar que una mujer con cáncer invasivo de cuello uterino tiene mayores probabilidades de tener un tiempo de supervivencia de por lo menos cinco años, en el periodo 1998-2007, se debe a que la información de estos casos en el RPCMP se encontraba mejor depurada y de mayor calidad. Además, se aumentó la cobertura de fuentes de información que

inicialmente en la cohorte 1998-2002 no se contaban y se incorporaron algunas modificaciones relacionadas con las localizaciones y diagnósticos a presentar, utilizando la metodología propuesta por la Asociación Internacional de Registros de Cáncer (IARC), con el fin de ser comparables con los datos publicados por otros registros en el mundo (Yépez et al., 2012a). Por otra parte, el tener cohortes más grandes aumenta la potencia del estudio e identifican otros factores de pronóstico.

Este patrón general del umbral para considerar la probabilidad de tener un tiempo de supervivencia de por lo menos cinco años va acompañado del método de diagnóstico a través de una Histología de Tumor Primario. Este patrón determina que un estudio de las células a partir de tejido de tumor primario, está asociado a una supervivencia más larga de la paciente. De este hallazgo se puede deducir la importancia del diagnóstico temprano por medio de este método y se convierte en conocimiento muy útil para los especialistas médicos en la toma de decisiones respecto al tratamiento óptimo para una paciente con este tipo de cáncer. Al respecto, la Fundación Contra el Cáncer (2012) determina que esta exploración confirma el tipo de cáncer y permite la verificación y precisión de la primera estadificación de la enfermedad, para saber hasta qué punto ha avanzado. Esta información adicional es necesaria para decidir si se necesita más tratamiento y qué tipo de tratamiento. En América Latina y el Caribe, la supervivencia de las mujeres con cáncer de cuello uterino es más corta porque a menudo solicitan atención cuando la enfermedad ya está avanzada. La baja supervivencia también se relaciona con una atención paliativa inadecuada y un tratamiento incompleto (Lewis, 2004).

Sin tener en cuenta el tiempo de supervivencia al que aduce el patrón general, otro patrón relevante que se relaciona con la supervivencia de las mujeres y que asocia factores socioeconómicos es el que predice que una mujer con cáncer invasivo de cuello uterino del municipio de Pasto, tiene mayor probabilidad de tener una supervivencia de por lo menos cinco años, si cuenta con un régimen de salud bien sea subsidiado o contributivo. El contar con un régimen subsidiado, que es un patrón común en las dos cohortes estudiadas, se explica debido a que en el municipio

de Pasto, el mayor porcentaje de mujeres con cáncer invasivo de cuello uterino tienen como régimen de aseguramiento el 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 y que de alguna forma permite, en el caso estudiado, que las mujeres diagnosticadas con cáncer invasivo de cuello uterino, reciban tratamiento y por lo tanto tengan una probabilidad de supervivencia mayor a las que no lo tienen. El segundo patrón, que se da en la cohorte 1998-2007, se explica porque al régimen contributivo, según el Ministerio de Protección Social (Minsalud, 2004), están afiliadas las personas que tienen una vinculación laboral, es decir, con capacidad de pago como los trabajadores formales e independientes, los pensionados y sus familias que les da derecho a recibir una amplia gama de servicios incluidos en el Plan Obligatorio de Salud –POS tales como protección integral para la salud en caso de enfermedad general o maternidad, incluyendo promoción de la salud, prevención de enfermedades, diagnóstico, tratamiento y rehabilitación para todas las enfermedades, independientemente de lo costosa o compleja que sea la atención requerida; beneficios que en el caso del tratamiento del cáncer invasivo de cuello uterino van a influir en la supervivencia de las mujeres con esta enfermedad.

De igual manera, si no se tiene en cuenta el tiempo de supervivencia al que se refiere el patrón general obtenido en la cohorte 1998-2007, otro patrón relevante que se relaciona con la supervivencia de las mujeres y que asocia factores clínicos es el que predice que una mujer tiene mayor probabilidad de tener una supervivencia de por lo menos cinco años, al cáncer invasivo de cuello uterino, si además del método de diagnóstico mediante una Histología de Tumor Primario, la morfología del tumor es un Squamous cell carcinoma, large cell, nonkeratinizing, (carcinoma de células escamosas, de células grandes, no queratinizante) o Squamous cell carcinoma (carcinoma de células escamosas). Este patrón es común a las dos cohortes, lo que indica que es un patrón que se mantiene sin importar el número de años o la cantidad de casos registrados. La explicación de este patrón se debe a que

estas dos formas de cáncer son las más comunes en las mujeres con cáncer de cuello uterino y además tienen menor capacidad de invasión que otras formas de cáncer cervical. Esta explicación se soporta en estudios que indican que en los países en desarrollo, entre el 90% y 95% de los cánceres invasores que se originan en el cuello uterino corresponden histológicamente a carcinomas de células escamosas y entre el 2% y 8% son adenocarcinomas (IARC,s.f), (ACS,2014). Además, en varios estudios retrospectivos con base en la población, se observa un desenlace más precario en las pacientes de adenocarcinoma, en quienes se observa un aumento de metástasis a distancia en comparación con aquellas con características histológicas escamosas (NCI, s.f). Se ha mostrado que el adenocarcinoma incide en una menor supervivencia a 5 años para cada una de las etapas del cáncer cervical, diferencia que se hace más significativa conforme avanzan los estadíos de este cáncer (Eifel et al., 1994).

Analizando los resultados de Asociación, la mayoría de las reglas tiene factores socioeconómicos y clínicos que se repiten, se presentan juntos y están asociados con la supervivencia, por lo menos de cinco años, de las mujeres con cáncer invasivo de cuello uterino del municipio de Pasto. Además, estos son comunes en las reglas de las cohortes 1998-2007 y 1998-2002, lo que indica que la asociación se mantiene sin importar el número de años o la cantidad de casos registrados.

Entre los factores socioeconómicos asociados a la supervivencia por lo menos de cinco años de las mujeres con cáncer invasivo de cuello uterino del municipio de Pasto, están: el tener una edad entre 15 y 28 años al momento del diagnóstico, no tener ninguna discapacidad, poseer servicio público de acueducto, pertenecer a un sistema de salud subsidiado y tener un nivel de escolaridad de primaria. Estos hechos reafirman los hallazgos realizados en otros estudios en los cuales factores socioeconómicos se asocian con la supervivencia de cáncer de cuello uterino (Lewis, 2004). En el caso del rango de edad, se ha observado que la supervivencia disminuye conforme aumenta la edad de las mujeres, notándose que las pacientes con edades superiores a los 65 años presentan las probabilidades más bajas de supervivencia al cabo de 5 años.

Sin embargo, este hallazgo puede estar asociado a la edad en la cual se hace el diagnóstico ya que, la proporción de carcinomas invasores se hace mayor, siendo también una posible causa de la menor supervivencia (Peto et al., 2004), (Yépez et al, 2009). En algunos estudios longitudinales se observó que, en pacientes de cáncer de cuello uterino in situ que no se trataron, entre 30 y 70% presentarán un carcinoma invasivo en un período de 10 a 12 años. Sin embargo, en alrededor de 10% de las pacientes, las lesiones pueden evolucionar de in situ a invasivas en un período menor de 1 año (NCI, 2014). Por lo tanto, se podría inferir que las probabilidades de sobrevivir son mayores en aquellas pacientes a quienes se les realiza el diagnóstico de cáncer cervical en edades tempranas, como es el caso del patrón descubierto. Este conocimiento debe ser incluido dentro de los programas de prevención contra el cáncer de cuello uterino que se implementen en el municipio de Pasto.

El no tener ninguna discapacidad como factor asociado a la supervivencia se debe a que, en este estudio, el 85% de las pacientes analizadas no tienen ninguna discapacidad y por lo tanto, esta variable por su soporte alto, estará asociada a otras que pasan el mínimo soporte, establecido en reglas de Asociación, como es el caso de la supervivencia.

El poseer servicio público de acueducto y pertenecer a un sistema de salud subsidiado como factores asociados con la supervivencia de por lo menos cinco años, se debe a las mismas razones descritas en los patrones de clasificación, en el sentido que, en el municipio de Pasto, el mayor porcentaje de mujeres con cáncer invasivo de cuello uterino tienen como fuente de abastecimiento de agua, el servicio de acueducto, que les permite tener una mejor calidad de vida que las que no lo tienen y como régimen de aseguramiento el régimen subsidiado, que les permite acceder a un tratamiento adecuado a su enfermedad y por lo tanto tener mejor probabilidad de supervivencia. Además, en la tarea de Asociación, por el soporte alto de estas dos variables, estarán relacionadas con la variable supervivencia.

El tener un nivel de escolaridad de primaria como factor asociado a la supervivencia, de por lo menos cinco años, de las mujeres con cáncer cervical en el municipio de Pasto, se explica porque la gran mayoría de mujeres analizadas tiene este nivel de escolaridad. Este hallazgo se podría inferir como contrario a los diferentes reportes que se hacen en otros estudios. En esos estudios, esta variable ha sido importante por cuanto el nivel educativo es un factor que incide en la adopción de conductas preventivas y de estilos de vida que procuran disminuir el riesgo de llegar a etapas avanzadas de la enfermedad, que aumenta la probabilidad de morir antes de los cinco años de supervivencia que es un tiempo esperado (Ortiz & Manrique-Leal, 2005). Sin embargo, en estudios realizados por Yépez et al. (2011) y Arellano (s.f.) se encontraron datos similares a los reportados en esta investigación y la interpretación dada fue por el bajo número de mujeres con alta escolaridad que se analizaron en la cohorte.

Entre los factores clínicos, asociados a la supervivencia de por lo menos cinco años, de las mujeres con cáncer invasivo de cuello uterino en el municipio de Pasto, están: la práctica de una biopsia para determinar si tiene cáncer o no, la realización de una Histología de Tumor Primario como método de diagnóstico para determinar el tipo de cáncer de cuello uterino, la morfología del cáncer de tipo Squamous cell carcinoma, NOS, la localización del cáncer en el Cuello uterino SAI (el término SAI, de acuerdo a la CIE-O (2000), significa sin otra indicación o en general) y un hospital o clínica como fuente de diagnóstico. Estos resultados indican que, con adecuados procedimientos clínicos de diagnóstico del cáncer cervical, realizados en entidades de salud que poseen la infraestructura necesaria para este tipo de procedimientos, las pacientes con cáncer de cuello uterino escamoso en general tienen mayores probabilidades de tener una supervivencia más larga. Confirman los estudios que expresan que la mayoría de los casos de cáncer de cuello uterino se pueden prevenir con exámenes de detección de rutina y con el tratamiento de las lesiones precancerosas. Como consecuencia, la mayoría de los casos de cáncer invasivo de cuello uterino se diagnostican en mujeres que viven en regiones desprovistas de protocolos de detección adecuados

y bajo situación de pobreza. Muchas mujeres con bajos ingresos no tienen acceso fácil a servicios adecuados de atención a la salud. Esto significa que es posible que no se hagan las pruebas de detección ni reciban tratamiento para precánceres y cánceres de cuello uterino (NCI, 2014), (ACS, 2014).

Con la tarea de agrupación, se formaron dos grupos homogéneos de mujeres con cáncer invasivo de cuello uterino diagnosticadas en el municipio de Pasto en la cohorte 1998-2007, con características socioeconómicas y clínicas similares en cada grupo. Los valores de los atributos que los determinan corresponden a la media en el caso de que el atributo sea numérico y a la moda en los atributos nominales. Los atributos cuyos valores son diferentes para cada grupo son los relevantes y los que caracterizan a cada grupo. En el primer grupo están las mujeres que sobrevivieron por lo menos cinco años (el mayoritario) al cáncer invasivo de cuello uterino y en el segundo las que no.

Teniendo en cuenta únicamente los factores socioeconómicos relevantes relacionados con la supervivencia, el grupo de mujeres que sobrevivieron por lo menos cinco años al cáncer invasivo de cuello uterino en el municipio de Pasto está formado por aquellas, cuyas edades al momento de ser diagnosticadas estaban entre 42 y 54 años, clasificadas en el nivel 1 del SISBEN, con un régimen de salud subsidiado y con un tiempo de vida transcurrido desde el momento del diagnóstico hasta el año 2007 entre 24 y 36 meses.

Las características de este grupo indican que, a pesar de que en varios estudios, se acepta que las mujeres de más bajo nivel social, cultural y educativo, tienen mayor incidencia de cáncer del cuello uterino (Ortiz et al., 2004), el contar con un régimen subsidiado de salud le da a estas mujeres, de bajos recursos económicos (nivel 1 del SISBEN) mayor probabilidad de sobrevivir un tiempo más largo (cinco o más años). Hecho que confirma los hallazgos encontrados también en los resultados de las tareas de clasificación y asociación.

Por otra parte, el tener una edad entre 42 y 54 años al momento del diagnóstico del cáncer invasivo, da a este grupo de mujeres

del municipio de Pasto mayor probabilidad de sobrevivir por lo menos cinco años. La edad del diagnóstico concuerda con el estudio realizado por el Dr. Lewis (2004) en el sentido de que la mayor carga absoluta del cáncer cervical recae sobre las mujeres de mediana edad. La mayoría de los casos se diagnostican entre los 35 y los 50 años, con un máximo entre los 40 y los 45. Hay un número significativo de casos desde los 30 años. Por otra parte, un análisis de los datos sobre mortalidad por cáncer cervical, notificados por los países de América Latina y el Caribe, indica que las mujeres entre los 35 y 54 años de edad conforman reiteradamente la mayor proporción de las defunciones anuales (Lewis, 2004). De acuerdo con este estudio, se puede inferir que, a este grupo de mujeres del municipio de Pasto, a pesar que tienen mayor probabilidad de sobrevivir por lo menos cinco años, por el tiempo transcurrido desde el momento del diagnóstico (entre 2 y 3 años) hasta el año de corte de este estudio, su supervivencia a 5 años puede disminuir.

Teniendo en cuenta solamente los factores clínicos relevantes relacionados con la supervivencia, el grupo de mujeres que sobrevivieron, por lo menos cinco años, al cáncer invasivo de cuello uterino, en el municipio de Pasto está formado por aquellas, en las cuales su cáncer tiene una morfología de tipo Squamous cell carcinoma, large cell, nonkeratinizing NOS. Este factor de supervivencia ya ha sido discutido y explicado en los resultados de las tareas de clasificación y asociación.

Los resultados encontrados en este estudio muestran la importancia de analizar los diferentes factores asociados a la supervivencia de mujeres con cáncer invasivo de cuello uterino como un fundamento para tomar decisiones en programas de prevención y tratamiento así como en la calidad de la prestación de los servicios.

CAPÍTULO VI

CONCLUSIONES

En esta investigación se descubrieron patrones de supervivencia de las mujeres con diagnóstico de cáncer invasivo de cuello uterino utilizando las técnicas de minería de datos clasificación por árboles de decisión, asociación y agrupación o clustering, a partir de los datos reportados, depurados y almacenados en el Registro Poblacional de Cáncer del Municipio de Pasto entre los años 1998 y 2007, siguiendo cada una de las etapas del proceso de descubrimiento de conocimiento en bases de datos.

Se construyó, limpió y transformó un repositorio de datos con la información de las mujeres del municipio de Pasto que fueron diagnosticadas con cáncer de cuello uterino en ese periodo de 10 años, complementada con datos de bases de datos externas principalmente 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. Se seleccionaron también otras fuentes externas como las bases de datos de clínicas privadas, hospitales públicos y privados, empresas sociales del estado, empresas prestadoras de servicios de salud, laboratorios de patología, instituciones de servicios

especializados de salud y de la Registraduría Nacional del Estado Civil. Para confrontar los resultados obtenidos en la cohorte 1998-2007 se aplicaron las mismas técnicas de minería de datos para la cohorte 1998-2003.

Como resultado de esta investigación se ha obtenido un patrón general que predice un tiempo de supervivencia, de por lo menos cinco años, de las mujeres con cáncer invasivo de cuello uterino 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 siempre y cuando como método de diagnóstico se le haya hecho un estudio de las células a partir de tejido de tumor primario.

Los resultados del estudio muestran que existe un patrón asociado a la supervivencia (de cinco o más años) de las mujeres con cáncer invasivo de cuello uterino del municipio de Pasto con sus condiciones socioeconómicas como es el contar con un régimen de salud (subsidiado o contributivo) y con un diagnóstico de esta enfermedad en edades tempranas, reafirmando los hallazgos realizados en otros estudios en los cuales este tipo de cáncer se asocia con la clase social.

Por otra parte, se obtuvo un patrón que asocia los factores clínicos para que las mujeres del municipio de Pasto con cáncer invasivo de cuello uterino tengan un tiempo de supervivencia de por lo menos cinco años, como es el contar con adecuados procedimientos clínicos de diagnóstico del cáncer cervical, realizados en entidades de salud que poseen la infraestructura necesaria para este tipo de procedimientos y para el tratamiento del cáncer de cuello uterino escamoso en general.

Por los resultados obtenidos en este proyecto de investigación, se debe tener en cuenta que un programa eficaz de prevención y control del cáncer de cuello uterino debe abordar varios aspectos, entre ellos la cobertura y la calidad de los servicios de salud y de detección temprana, así como la disponibilidad de un diagnóstico, tratamiento y atención subsiguiente rápidos, fiables y asequibles. Así mismo, debe propender, facilitar y apoyar este tipo de inves-

tigaciones que brindan herramientas para que los especialistas en el área de la salud puedan tomar decisiones sobre el tratamiento óptimo de esta enfermedad y los factores de riesgo asociados a su desarrollo y supervivencia.

Es importante que el conocimiento descubierto en este estudio, se incorpore al existente, sobre la supervivencia de mujeres con cáncer de cuello uterino y una vez sea socializado con los organismos gubernamentales y privados del sector salud del municipio de Pasto, ayude a la toma de decisiones en lo relacionado con el planteamiento de programas de prevención y control de las mujeres con este tipo de cáncer.

Como trabajos futuros está el continuar esta investigación, una vez el RPCMP haya depurado la información del quinquenio 2007-2012, en una segunda etapa, analizando las cohortes 1998-2012 y 2003-2007.

REFERENCIAS BIBLIOGRÁFICAS

ACS (2014). Cáncer de Cuello Uterino (Cervical). American Cancer Society. Last Medical Review: 10/30/2014. Disponible en: <http://www.cancer.org/acs/groups/cid/documents/webcontent/002288-pdf.pdf>.

AGRAWAL, R., GHOSH S., IMIELINSKI, T., IYER, B. & SWAMI, A. (1992). An Interval Classifier for Database Mining Applications. In: Proceedings VLDB Conference, Vancouver, Canada.

AGRAWAL, R., IMIELINSKI, T. & SWAMI, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD. Washington DC, USA.

AGRAWAL, R., MEHTA, M., SAFER J. & SRIKANT, R. (1996). The Quest Data Mining System. In: 2º Conference KDD y Data Mining. Portland, Oregon, USA.

AGRAWAL, R. & SRIKANT, R. (1994). Fast Algorithms for Mining Association Rules. In: VLDB Conference, Santiago de Chile, Chile.

AGRAWAL, R. & SRIKANT, R. (1995). Mining Sequential Patterns. In: The 11th International Conference on Data Engineering ICDE, pp 3-14. Taipei, Taiwan.

ARENDS, M., BUCKLEY, C. & WELLS, M. (1998). Aetiology, pathogenesis, and pathology of cervical neoplasia. *J ClinPathol*, 51:96-103.

ARIAS, S. (2009). Inequidad y cáncer: Una revisión conceptual. En: revista Facultad Nacional de Salud Pública Vol. 27, No 3, Universidad de Antioquia. Medellín, Colombia.

ARELLANO, R. (s.f.). Sobrevivencia en cáncer. Servicio de Salud Hospital Metropolitano. Escuela de Salud Pública. Facultad de Medicina. Universidad de Chile.

ASPORT, S. & RIVERO, T. (2004). Plan Nacional de Control de Cáncer de Cuello Uterino 2004-2008. Ministerio de Salud y Deportes. Bolivia.

ASHFORD, L. & COLLYMORE, Y. (2005). La Prevención del Cáncer Cervical a Nivel Mundial. Hoja informativa. Population Reference Bureau. Washington, USA.

BELLAACHIA, A. & GUVEN, E. (2005). Predicting Breast Cancer Survivability Using Data Mining Techniques. The George Washington University. Washington D.C., USA.

BOLAÑOS, H., HIDALGO, A. & YÉPEZ, M.C. (2007). Incidencia de cáncer en el Municipio de Pasto periodo 1998-2002. Editorial Universitaria. San Juan de Pasto, Colombia.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J. & ZANASI A. (1997). *Discovering Data Mining from Concept to Implementation*, Prentice Hall PTR.

CANNISTRA, S. & NILOFF, J. (1996). Medical progress: Cancer of the uterine cervix. *New England Journal of Medicine*. 334:16, 1030-1038.

CASADO. A., VIDART J. & CORONADO. M. (2004). Cáncer de Cérvix. Capítulo 6. Disponible en: <http://www.todocancer.com/NR/rdonlyres/7AC45FDA-5CF3-465C-A9722F5B7C25DA5D/0/Capitulo6.pdf>.

CASTRO, M., VERA, L. & POSSO, H. (2006). Epidemiología del cáncer de cuello uterino: estado del arte. *Revista Colombiana de Obstetricia y Ginecología*, vol.57, No.3 Bogotá, Colombia.

CERDA, J. & VILLAROEL, L. (2008). Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. *Revista Chilena de Pediatría*. Vol. 79, No. 1, pp. 54-58. Disponible en <http://www.scielo.cl/pdf/rcp/v79n1/art08.pdf>.

CHEN, M., HAN, J. & YU, P. (1996). Data Mining: An Overview from Database Perspective. In: *IEEE Transactions on Knowledge and Data Engineering*.

CHICHAREON, S., HERRERO, R., MUNOZ, N., BOSCH, F., JACOBS, M. & DEACON, J. (1998). Risk factors for cervical cancer in Thailand: a case-control study. *J Natl. Cancer Inst.*, 90:50-7.

CIENCIA HOY (2006). Cáncer del cuello uterino. *Revista Ciencia Hoy*. Vol. 16, No.95, ISSN 1666-5171. Buenos Aires, Argentina. Disponible en: <http://www.cienciahoy.org.ar/ch/ln/hoy95/cancer.htm>.

CIE-O (2003). *Clasificación Internacional de Enfermedades para Oncología*. Organización Panamericana de la Salud, Oficina Regional de la Organización Mundial de la Salud. Tercera Edición. ISBN:92-4-154534-8. Washington D.C., USA.

CORRAL, F, CUEVA, P, YÉPEZ, J. & MONTES, E. (1996). La baja escolaridad como factor de riesgo en el cáncer de cuello de útero. *Bol of Sanit. Panam*, pp 121:511-7.

DAEDALUS (2009). *Aplicación de minería de datos para el diagnóstico de accidentes cerebrovasculares agudos*. Madrid, España. Disponible en: <http://sistemaminergescon.blogspot.com/>.

DAVIES, J., GIBBONS, J., HARRIS, S. & CRICHTON, C. (2012). *The Cancer Grid Experience: Metadata-Based Model-Driven Engineering for Clinical Trials*.

In: Science of Computer Programming. Departament of Computer Science, University of Oxford. Disponible en: <http://www.cs.ox.ac.uk/people/jeremy.gibbons/publications/cancergrid.pdf>.

EIFEL, P.J., MORRIS, M. & WHARTON, J.T. (1994). The influence of size tumor and morphology on the outcome of patients with FIGO stage IB squamous cell carcinoma of the uterine cervix. *Int J Radiol Oncol Biol Phys* 29: 9-16.

ELDER, J. & PREGIBON, D. (1996). A Statistical Perspective on Knowledge Discovery in Databases. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Pres/ The MIT Press.

ESMO (2012). Cáncer de cuello uterino: una guía para pacientes – Basada en la Guía de Práctica Clínica de la ESMO – v. 2012.1. Fundación Contra el Cáncer. European Society for Medical Oncology - ESMO. Disponible en: <http://www.esmo.org/content/download/6673/115548/file/ESMO-ACF-Cancer-de-CuelloUterino-Guia-para-Pacientes.pdf>.

FAYYAD, U., PIATETSKY-SHAPIO, G. & SMYTH, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Communications of the ACM*, Vol 39, No. 11.

FERLAY, J., BRAY, F., PISANI, P. & PARKIN, D.M. (2004). Cancer incidence, mortality and prevalence worldwide, version 2.0. In: *IARC Cancer Base N. ° 5*. Lyon, France: IARC Press.

FERNÁNDEZ, G. (2009). Extracción de Información de la Web usando Técnicas de Minería de datos. Disponible en: <http://www.tdg-seville.info/Download.ashx?id=48>.

FRANCO, E., DUARTE-FRANCO, E. & FERENCZY, A. (2001). Cervical cancer: Epidemiology, prevention and the role of human papillomavirus infection. *CMAJ*, 164:101725.

GARCÍA, D. (s.f.). Manual de Weka. Disponible en: <http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf>.

GARCÍA, M. & ÁLVAREZ, A. (2010). Análisis de Datos en WEKA – Pruebas de Selectividad. Disponible en: <http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf>.

GILL, H. & RAO P. (1996). Data warehousing: La integración de información para la mejor toma de decisiones. Ed. Prentice-Hall.

GLOBOCAN (2012). Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. Disponible en: http://globocan.iarc.fr/Pages/bar_site_sel.aspx.

HALL, M., FRANK, E. & WITTEN, I. (2011). Practical Data Mining: Tutorials. University of Waikato. Disponible en: [http:// www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf](http://www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf).

HAN, J. & KAMBER, M. (2012). Data Mining Concepts and Techniques. Ed. Morgan Kaufmann Publishers, San Francisco, USA.

HEARD, I., PALEFSKY, J.M. & KAZATCHKINE, M.D. (2004). The impact of HIV antiviral therapy on human papillomavirus (HPV) infections and HPV-related diseases. In: International Medical Press, Antiviral Therapy. ISSN: 1359-6535, 9: 13-22.

HERNÁNDEZ, H (2014). Aplicación de minería de datos a información de pacientes prediabéticos. En: Revista Iberoamericana de Producción Académica y Gestión Educativa, vol. 1, num. 1. ISSN: 2007-8412.

HERNÁNDEZ, E. & LORENTE, R. (2009). Minería de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III, Madrid, España. Disponible en: <http://www.it.uc3m.es/jvillena/irc/practicass/08-09/14.pdf>.

HERNÁNDEZ, J., RAMÍREZ, M.J & FERRI, C. (2005). Introducción a la Minería de datos. Editorial Pearson Prentice Hall, ISBN 84-205-4091-9. Madrid, España.

HERNÁNDEZ, J. & FERRI, C. (2006). Introducción al Weka. Curso de Doctorado en Extracción Automática de Conocimiento en Bases de Datos e Ingeniería de Software. Universidad Politécnica de Valencia. Valencia, España.

HERRERO, R., BRINTON, L., REEVES, W., BRENES, M., TENORIO, F. & DE BRITTON, R. (1990). Sexual behavior, venereal diseases, hygiene practices, and invasive cervical cancer in a high-risk population. In: Cancer, pp. 65: 380-6.

HURTADO, V., SILVA, V., SALAS, R. & LOBOS, A (2008). 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, Chile.

INC (2012). Instituto Nacional de Cáncer. Cáncer en cifras. Bogotá, Colombia. Disponible en: <http://www.cancer.gov.co/~incancer/instituto/content/estadisticas>.

IARC (s.f.). Introducción al cáncer invasor del cuello uterino. International Agency for Research on Cancer. World Health Organization. IARC Screening Group. Disponible en: <http://screening.iarc.fr/colpochap.php?lang=3&chap=3>.

JAIN, V. & GODORA, S. (2017). Comparative Study of Data Mining Classification Methods in Brain Tumor Disease Detection. In: IJCSC, vol. 8, issue 2, pp. 12-17. ISSN: 0973-7391.

KUMARI M. & SUNILA G (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. In: IJCST, vol. 2, issue 2. ISSN:0976-8491 (On line).

LANDIS, J. & KOCH, G. (1977). The measurement of observer agreement for categorical data. *Revista Biometrics*, Vol. 33, pp. 159-74.

LÓPEZ, R.A. (2011). Descubrir el Comportamiento de los Factores de Riesgo de Cáncer de Cuello Uterino utilizando Minería de datos. Tesis de Maestría. En: Biblioteca Digital Repositorio Institucional Universidad Nacional de Colombia. Bogotá, Colombia.

LEWIS, M. J. (2004). Análisis de la situación del cáncer cervicouterino en América Latina y el Caribe. Washington D.C., OPS. ISBN: 92-75- 32531- 6.

MADELEINE, M., DALING, J., SCHWARTZ, S., SHERA, K., MCKNIGHT, B. & CARTER, J. (2001). Human papillomavirus and long-term oral contraceptive use increase the risk of adenocarcinoma in situ of the cervix. In: *Cancer Epidemiol Biomarkers Prev*, 10:171-7.

MOLERO, G., CÉSPEDES, Y. & CAMPAÑA, M (2013). Caracterización y análisis de la base de datos de cáncer de mama SEER-DB. En: *Memorias de IX Congreso Internacional Informática en Salud*.

MINSALUD (2004). Sistema de Seguridad Social en Salud: Régimen Contributivo. Bogotá, Colombia. Disponible en: <http://www.minsalud.gov.co/Documentos%20y%20Publicaciones/GUIA%20INFORMATIV A%20DEL%20REGIMEN%20CONTRIBUTIVO.pdf>.

MINSALUD (2012). Plan Decenal para el Control del Cáncer en Colombia, 2012-2021. Ministerio de Salud y Protección Social. Instituto Nacional de Cancerología, ESE. Bogotá, Colombia. ISBN: 978-958-57632-2-7.

MERLE, J. L. (2004). Análisis de la Situación del Cáncer Cérvico Uterino en América Latina y el Caribe. OPS. Washington, USA.

MOLINERO, L. M. (2001). Medidas de concordancia para variables cualitativas. Asociación de la Sociedad Española de Hipertensión, Liga Española para la lucha contra la Hipertensión Arterial. Disponible en <http://www.seh-lelha.org/pdf/concor2.pdf>.

MOMJIAN, B. (2001). PostgreSQL: Introduction and Concepts. Editorial Addison-Wesley, 455 p, ISBN: 0-201-70331-9. New York, USA.

MORA, R. (2009). El papel de la Minería de datos en la Detección y Diagnóstico de Cáncer. Universidad de Salamanca. Salamanca, España. Disponible en: <http://sistemaminergescon.blogspot.com/>

MUÑOZ, N, BOSCH, F.X & DE SANJOSE, S. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer. In: New England Journal of Medicine, 348:518-527.

NCI (s.f.). Diccionario de Cáncer. Instituto Nacional del Cáncer de los Institutos Nacionales de la Salud de EE.UU. Disponible en: <http://www.cancer.gov/diccionario>.

NCI (2014). Cáncer de cuello uterino: Tratamiento. Instituto Nacional del Cáncer de los Institutos Nacionales de la Salud de EE.UU. Disponible en: http://www.cancer.gov/espanol/pdq/tratamiento/cuellouterino/HealthProfessional/page1#_391_toc. Actualización: diciembre 2014.

NG, R. & HAN, J. (1994). Efficient and Effective Clustering Method for Spatial Data Mining. In: VLDB Conference, Santiago de Chile.

ORTIZ, A. & MANRIQUE-LEAL, M. (2005). Sobrevida en pacientes con cáncer de cérvix. Acta Médica Costarricense (On-line versión). Volumen 47. No. 4, pp. 192-195. San José de Costa Rica. ISSN: 0001- 6012.

ORTIZ, R., URIBE, C., DÍAZ, L. & DANGOND, Y. (2004). Factores de Riesgo para Cáncer de Cuello Uterino. Revista Colombiana de Obstetricia y Ginecología, Volumen 55, No. 2, pp. 146-160. Bogotá, Colombia. ISSN: 0034-7434.

OPS (s.f.). Cáncer Cervicouterino. Organización Panamericana de la Salud. Disponible en: <http://www.paho.org/hq/index.php>.

OVIEDO, E., OVIEDO, A. & VÉLEZ, G. (2015). Minería de datos: Aportes y tendencias en el servicio de salud de ciudades inteligentes. En: Revista Politécnica, vol. 11, número 20, pp. 111-120. ISSN: 2256-5353 (En línea).

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

PÉREZ, A. (2012). Aplicación de la red de probabilidad neuronal y escala de framingham para predicción de la hipertension arterial. En: Memorias Convención Internacional de Salud Pública. La Habana (Cuba).

PÉREZ, M. (2007). Investigación multidisciplinar para el diagnóstico temprano de cáncer de páncreas. Revista Bit 163, las Tics en la sanidad, pp. 51-54.

PÉREZ, C. & SANTÍN, D. (2006). Data Mining: Soluciones con Enterprise Miner. Editorial Alfaomega Ra-Ma. ISBN: 84-7897-695-7. México D.F (México).

PETO, J., GILHAM, C., FLETCHER, O. & MATTHEWS, F. (2004). The cervical cancer epidemic that screening has prevented in the UK. The Lancet 2004; Volume 364, No. 9430, pp. 249-256. DOI: 10.1016. ISSN:0140-6736

PIATETSKY-SHAPIRO, G., BRACHMAN, R. & KHABAZA, T. (1996). An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. AAAI Pres/ The MIT Press.

PIÑEROS, M., MURILLO, R. & PORRAS, R. (2011). Guía para el análisis de la situación del cáncer. Ministerio de Salud y Protección Social. Instituto Nacional de Cancerología E.S.E, Bogotá, Colombia.

PROKOPCZYK, B., COX, J., HOFFMANN, D. & WAGGONER, S. (1997). Identification of tobacco-specific carcinogen in the cervical mucus of smokers and nonsmokers. In: J Natl Cancer Inst, 89:868-73.

QUINLAN, J.R. (1986). Induction of Decision Trees. In: Machine Learning Journal, Vol. 1, No. 1, Kluwer Academic Publishers, pp. 81-106. Boston, USA.

QUINLAN, J.R. (1993). C4.5: Programs for Machine Learning. Ed. Morgan Kaufmann Publishers.

REPARAZ, D., MERLINO, H., RANCAN, C., RODRÍGUEZ, D., BRITOS, P. & GARCÍAMARTÍNEZ, R. (2008). Determinación de la Eficacia de la Braquiterapia en Tratamiento de Cáncer basada en Minería de datos. En: Proceedings WIIC 2008.

ROWE, L.A. & STONEBRAKER, M. (1987). The POSTGRES Data Model. En: Proceedings of the 13 th VLDB Conference, Brighton, England.

SHAIKH, T. (2014). A Prototype of Parkinson's and Primary Tumor Diseases Prediction Using Data Mining Techniques. In: International Journal of Engineering Science Invention, 3(4).

SAIZ, J. & LAU, R. (2009). Análisis de Datos Médicos de Cáncer Pediátrico por medio de Minería de datos. Disponible en: <http://www.cca.ipn.mx/.../03MinDatoslineasdeInvestigacion-v2.pdf>.

SANKARANARAYANAN, R., BUDUKH, A. & RAJKUMAR, R. (2001). Effective screening programmes for cervical cancer in low- and middle-income developing countries. Bull World Health Organ, pp. 79:954-62.

SATTLER, K. & DUNEMANN, O. (2001). SQL Database Primitives for Decision Tree Classifiers. In: CIKM, Atlanta, Georgia, USA.

SHARMA, N., BAJPAI, A. & LITORIYA, R. (2012). Comparison the various clustering algorithms of Weka tools. In: International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, ISSN: 2250-2459.

SOLARTE, R & CASTRO, Y.V. (2012). Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial. En: Tecnura, 35-52.

SRIKANT R. & AGRAWAL, R. (1996). Mining Quantitative Association Rules in Large Relational Tables, ACM SIGMOD. Montreal, Canada.

STONEBRAKER, M. & ROWE, L., A. (1986). The Design of POSTGRES. In: Proceedings of the ACM-SIGMOD Conference, Washington D.C., USA.

STUVER, S. & ADAMI, H. (2002). Cervical cancer. In: Adami HO, Hunter DJ, Trichoponlos D, eds. Textbook of Cancer Epidemiology. Oxford: Oxford University Press, pp. 340-58.

TEJEDA M., SERRANO, F. & GUMEZ-PASTRANA, N. (2007). Cáncer de cuello uterino. Estado actual de las vacunas frente al virus del papiloma humano (VPH). *Oncología*, 2007; 30 (2):42-59.

TIMARÁN R. & MILLÁN M. (2006). 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.

TIMARÁN, R. (2009). 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, pp 39-58, junio 2009, Manizales, Colombia.

TIMARÁN, R & YÉPEZ, M. (2012). La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. En: Revista Universidad y Salud, pp. 117-129.

TIMARÁN, S.R., HERNÁNDEZ, I., CAICEDO, J., HIDALGO, A. & ALVARADO, J (2016). Descubrimiento de Patrones de Desempeño Académico con árboles de decisión en las competencias genéricas de la formación profesional. Ediciones Universidad Cooperativa de Colombia. Bogotá, Colombia. ISBN:978-958-760-049-0. DOI: <http://dx.doi.org/10.16925/9789587600490>.

VILLA, LL (2003). Vaccines against papilloma virus infections and disease. *Salud Pública, Mexico*, 45, suppl 3: S443-8.

WALBOOMERS, J.M., JACOBS, M.V. & MANOS, M.M. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. In: *Journal of Pathology*, 189: 1219.

WANG, M., IYER, B. & SCOTT, V.J. (1998). Scalable Mining for Classification Rules in Relational Databases, In: International Database Engineering and Application Symposium - IDEAS, p.p. 58-67. Cardiff, Wales, U.K.

WITTEN, I. & FRANK, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Ed. Morgan Kaufmann Publishers, 365 p, ISBN: 155860-552-5. San Francisco, CA, USA.

YÉPEZ, M.C., CERÓN, E. & CERÓN, C. (2009). Percepción de los factores culturales atribuidos a la supervivencia en mujeres diagnosticadas con cáncer invasivo de cuello uterino. En: Revista Universidad y Salud, Año 9, Vol. 1, No. 11, pp. 7-23. Pasto, Colombia.

YÉPEZ, M.C., CERÓN, E., HIDALGO, A. & CERÓN, C. (2011). Supervivencia de mujeres con cáncer de cuello uterino, Municipio de Pasto. En: Revista Universidad y Salud, Año 11, Vol. 2 No. 14 2011, pp. 7-18. Pasto, Colombia.

YÉPEZ, M.C., BRAVO, L., HIDALGO, A., JURADO, D. & BRAVO, M. (2012a). Incidencia y Mortalidad por Cáncer Municipio de Pasto 2003-2007. Editorial Universitaria, Universidad de Nariño, 48 p. ISBN: 978-958-8609-49-2.

YÉPEZ, M.C., BRAVO, L., HIDALGO, A., JURADO, D. & BRAVO, M. (2012b). Cancer incidence and mortality in the municipality of Pasto, 1998-2007. En: Revista Colombia Médica, Vol. 43, No. 4 (Oct-Dec), pp. 256:266.

YÉPEZ, M.C., JURADO, D.M., BRAVO, L.M. & BRAVO, L.E. (2018). Trends in cancer incidence, mortality and survival in Pasto, Colombia. En: Revista Colombia Médica, 49(1): pp. 42-54, doi: 10.25100/cm.v49i1.3616.

ZAMARRON, C., GARCÍA, V., CALVO, U., PICHEL, F. & RODRÍGUEZ, J. (2006). Aplicación de la Minería de datos al Estudio de las Alteraciones Respiratorias durante el Sueño. Pneuma, pp. 6: 156-166.

ZHANG, T., RAMAKRISHNAN, R. & LIVNY, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. In: ACM SIGMOD International Conference on Management of Data, pp. 103-114. Montreal, Canadá.



Editorial
Universidad de **Nariño**

Este libro se terminó de imprimir
en el mes de marzo de 2021,
en Graficolor Pasto sas
Calle 18 No. 29-67
Tels. 7310652 - 7311833
graficolorpasto@hotmail.com

La mortalidad debida al cáncer de cuello uterino es uno de los grandes problemas de Salud Pública que confrontan las mujeres en todo el mundo, especialmente en los países menos desarrollados, donde ocurre el 80% de las muertes relacionadas a este tipo de cáncer.

En este libro se presentan los resultados del proyecto de investigación que tuvo como objetivo detectar patrones de supervivencia de las mujeres diagnosticadas con cáncer invasivo de cuello uterino utilizando técnicas de Minería de Datos, a partir de la información reportada en el Registro Poblacional de Cáncer del municipio de Pasto (Colombia). Aplicando como metodología las diferentes etapas del proceso de descubrimiento de conocimiento en bases de datos, 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 2007. Se utilizaron las tareas de minería de datos clasificación, asociación y agrupación para detectar diferentes patrones de supervivencia de estas mujeres a cinco años. Se encontró un patrón general que asocia el tiempo de supervivencia a cinco años con el número de meses de vida de la mujer transcurridos a partir del diagnóstico del cáncer invasivo de cuello uterino. Se detectaron además patrones que asocian factores socioeconómicos y clínicos a la supervivencia de este grupo poblacional. Se espera que el conocimiento descubierto se incorpore al existente sobre la supervivencia de mujeres con cáncer de cuello uterino en el municipio de Pasto y ayude a la toma de decisiones de los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de programas de prevención y control de las mujeres con este tipo de cáncer.



Universidad de **Nariño**

