Tesis defendida por

Cesar Carlo López Portillo

y aprobada por el siguiente Comité

Dr. Hugo Homero Hidalgo Silva *Codirector*	Dra. Ana Isabel Martínez García *Codirectora*
Dr. Armando Paniaga Michel *Miembro del comité*	Dr. Jesús Favela Vara *Miembro del comité*
Dr. Pedro Negrete Regagnon *Miembro del comité*	
Dr. José Antonio García Macías *Coordinador del programa de posgrado en Ciencias de la Computación*	Dr. David Hilario Covarrubias Rosales *Director de estudios de posgrado*

12 de febrero de 2013

CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR

DE ENSENADA

Programa de posgrado en Ciencias

En Ciencias de la Computación

Minería de datos como soporte en el diagnóstico y tratamiento del cáncer de mama.

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en ciencias

Presenta:
César Carlo López Portillo

Ensenada, Baja California, México
2013

Resumen de la tesis de César Carlo López Portillo, presentada como requisito parcial para la obtención del grado de Maestro en Ciencias de la Computación. Ensenada, Baja California, México.

Minería de datos como soporte en el diagnóstico y tratamiento del cáncer de mama

Resumen aprobado por:

_____ _____

Dra. Ana Isabel Martínez García Dr. Hugo Homero Hidalgo Silva

El cáncer de mama es el cáncer de la mujer más común en el mundo, y su incidencia aumenta en los países en vías de desarrollo. El caso de México no es la excepción, donde además es el cáncer con tasa de mortalidad más alta. En este trabajo se presentan resultados de estudios realizados a datos recopilados de las áreas de Radiología y Oncología de una clínica hospital de salud pública en la ciudad de Ensenada, México. Los datos se procesaron y analizaron usando algoritmos de Minería de Datos. De los reportes de Radiología fue posible clasificar las observaciones de tumores en benignos y malignos con una exactitud de 96% a 98% y producto sensibilidad-especificidad de 0.65 a 0.91. Los mejores resultados se obtuvieron con los métodos Bayesianos y el algoritmo re regresión Elastic Net. Los datos referentes a la descripción morfológica de las observaciones fueron los más relevantes para la clasificación, de acuerdo al análisis de reducción de dimensionalidad y a los nodos resultantes en los árboles de decisión inducidos. El cuadrante de la mama y la lateralidad resultaron irrelevantes para los modelos. De los reportes de Oncología se realizó un análisis de predicción de recurrencia de la enfermedad, obteniendo los mejores resultados por métodos de imputación Bayesiana y predicción por regresión logística, con exactitud de 72% a 95% y producto sensibilidad-especificidad de 0.5 a 0.94. En estos datos, cuando existe sobreexpresión del gen Her2/Neu el peso de la persona resulta como atributo relevante. Para los datos denominados LUMINAL, el atributo correspondiente a la aplicación de hormonoterapia aparece como relevante para prevenir la recurrencia. El análisis del conjunto denominado Triple Negativo tuvo los más pobres resultados, sin embargo es posible apreciar que la administración de capecitabina aparece como atributo relevante para prevenir recurrencia.

Palabras Clave: Cáncer **de mama, predicción de datos, recurrencia, minería de datos.**

Abstract of the thesis presented by César Carlo López Portillo as a partial requirement to obtain the Master in Science degree in Computer Science. Ensenada, Baja California, México.

Data mining as an aid in breast cancer detection and treatment.

Abstract approved by:

_____ _____

Dra. Ana Isabel Martínez García Dr. Hugo Homero Hidalgo Silva

Breast cancer is the most extended type of cancer in the world among women, with an increasing incidence on Mexico, where it has attained the largest mortality rates for cancer. Several data mining studies were conducted with results of Radiology and Oncology reports gathered from a public hospital in Ensenada, México. From mammogram screening reports of Radiology, several classifiers were built and tested, with the goal to predict malignity from mammogram results. Regularized regression data imputation was applied to some missing data about tumor size. Then, a cluster analysis applied to completed data revealed some of the data categorization by BIRADS. Also Histo-pathological diagnostic was revealed on the clusters observed on the first principal components. The completed data were applied to a dimensionality reduction algorithm and then classifiers were built, considering algorithms SVM, Bayesian networks, naïve Bayes, logistic, multilayer perceptron, nearest neighbors, decision trees (C4.5, random trees and CART) and regularized regression. Best classifiers were obtained with Bayesian methods and regularized regression with Elastic Net algorithm, attaining an accuracy of 96% to 98% and sensitivity-specificity product of 0.65-0.91. Morphological description data were the most relevant for classification, according to a dimensionality reduction study and of the developed nodes in the decision trees.

In the second part, data mining was applied to data of cancer diagnosed patients, in an effort to predict disease recurrence. The prediction study was realized on the Oncology reports data. The best results were attained by applying Bayesian imputation methods for some missing data retrieval and then logistic regression applied to the completed data. An accuracy of 72% to 95% was achieved, with sensitivity-specificity products of 0.5-0.94. When an overexpression of the gene Her2/Neu exists, the weight of the person appears as a relevant attribute. For the subset named LUMINAL, the attribute hormone-therapy appears as a relevant attribute for preventing recurrence. The analysis for the subset called Triple Negative had the worst results, but is possible to foresee that the administration of capecitabine appears as a relevant attribute for preventing recurrence.

Key Words: **Breast cancer, data prediction, regression, data mining.**

Dedicatorias.

Dedico este trabajo a mis padres,

Nora y Marco Tulio,

por todo el amor y el apoyo que me brindaron.

Agradecimientos.

Quiero agradecer a los buenos amigos que hice durante mi estancia en el CICESE, a Valeria, Maythe, Gio, Roberto, Ivan y a tantos que me apoyaron y me aguantaron durante el tiempo que estuvimos juntos.

A la Dra. Ana Isabel, por su paciencia, su ayuda, su comprensión y sus consejos.

Al Dr. Hugo Hidalgo, por su ayuda, su paciencia y por compartir sus conocimientos.

Al Dr. Armando Paniagua, por su valiosa ayuda al permitirme conocer los procesos hospitalarios y por compartir su conocimiento y experiencia.

Al personal de la clínica, por su gran ayuda, a la Dra. Xochitl Valenzuela, a la Dra. Janet Castelo, el Dr. Ricardo Salido, a Lorena, Franz y a las chicas de expediente, en especial a Gris.

A Dolores García, por haberme hecho sentir parte de su familia en estos 2 años.

A Arturo Cárdenas, por todo su apoyo y comprensión.

Al Consejo Nacional de Ciencia y Tecnología, por las facilidades y apoyos que me dieron para obtener este grado.

Y a todas las personas que formaron parte de este proceso, y que me ayudaron a crecer y convertirme en una mejor persona.

CONTENIDO

Resumen en español……i

Resumen en inglés……ii

Dedicatorias……iii

Agradecimientos……iv

Contenido……v

Lista de figuras……viii

Lista de tablas……xi

Introducción. 1

I.1 Antecedentes 1

I.2 Trabajo previo. 4

I.3 Planteamiento del problema. 6

I.4 Objetivo. 7

I.4.1 Objetivos específicos. 7

I.5 Metodología de la investigación. 8

I.5.1 Revisión de la literatura. 8

I.5.2 Caso de estudio. 8

I.5.3 Minería de datos (Proceso KDD) 9

I.5.4 Integración o difusión del conocimiento descubierto. 10

I.5 Contenido de la tesis. 10

Marco teórico. 12

II.1 Proceso de descubrimiento de conocimiento en datos. 12

II.2 Minería de datos. 15

II.2.1 Reconocimiento de patrones. 16
II.2.2 Aprendizaje supervisado y aprendizaje no supervisado. 17
II.2.3 Clasificadores para aprendizaje supervisado. 18
II.2.4 Aprendizaje no supervisado. 28
II.2.5 Evaluación del conocimiento. 29
II.3 Resumen. 31
Contexto médico y caso de estudio. 33
III.1 Introducción. 33
III.2 El cáncer de mama. 33
II.3.1 Diagnóstico. 36
II.3.2 Tratamiento. 38
III.3 Estudio del proceso de diagnóstico y tratamiento. 40
II.3.3 Descripción textual del proceso. 40
II.3.4 Modelado del proceso 48
III.4 Resumen. 63
Proceso de descubrimiento de conocimiento sobre registros de mastografías. 64
III.1 Antecedentes. 64
III.2 Herramientas y materiales. 65
III.3 Extracción de datos. 65
III.4 Transformación y codificación de datos. 68
III.4.1 Reposición de datos faltantes. 69
III.4.2 Análisis de clústeres: BI-RADS 73
III.5 Resumen. 88

Proceso de descubrimiento de conocimiento sobre repositorios de pacientes en tratamiento para cáncer de mama 91

IV.1 Antecedentes. 91

IV.2 Materiales y métodos 92

IV.3 Extracción de datos. 92

IV.4 Transformación y codificación. 94

IV.5 Pérdida y recuperación de datos. 96

IV.6 Resultados. 99

IV.6.1 Primera parte: Repositorio incompleto 100

IV.6.2 Segunda parte: Imputación de datos y reducción en la dimensionalidad. 101

IV.6.3 Comparación de resultados. 107

IV.7 Modelos resultantes. 110

IV.8 Resumen del capítulo. 112

Conclusiones 114

Referencias bibliográficas…….117

LISTA DE FIGURAS

Figura 1. Metodología de la investigación. 9
Figura 2. Proceso de descubrimiento de conocimiento en datos. (Orallo et al., 2004) 15
Figura 3. Ejemplo de árbol de decisión. 18
Figura 4. Modelo de regresión logística. 21
Figura 5. Ejemplo de hiperplano separador 26
Figura 6. Descripción morfológica de tumores en mastografías. 38
Figura 7. Proceso general de diagnóstico y tratamiento del cáncer de mama 51
Figura 8. Proceso de diagnóstico del cáncer de mama cuando la paciente acude por detectar una lesión. 52
Figura 9. Proceso de diagnóstico del cáncer de mama. Proceso que ocurre cuando la paciente es diagnosticada en una clínica particular. 53
Figura 10. Proceso de diagnóstico del cáncer de mama. Actividades seguidas en el subproceso de ginecología. 54
Figura 11. Proceso de tratamiento del cáncer de mama. Proceso de acudir a consulta para revisión. 55
Figura 12. Representacion de una actividad en IDEF 56
Figura 13. Diagrame general del proceso de detección del cáncer de mama. 58
Figura 14. Diagrama del proceso de diagnóstico del cáncer de mama. 59
Figura 15. Vista del proceso de medicina familiar 60
Figura 16. Vista del proceso de ginecología 61
Figura 17. Vista del proceso de cirugía (biopsia) 62
Figura 18. Ejemplo de registro mastográfico 66
Figura 19. Ejemplo de registro de reporte histopatológico. 66
Figura 20. Gráfica del tamaño del tumor contra el error en la predicción 71
Figura 21 Gráfica de la raíz del error cuadrado medio contra log lambda. 71
Figura 22 Gráfica de los coeficientes Vs. log lambda 73
Figura 23 Gráfica de los primeros dos componentes principales, con la categoría BIRADS como clase. 74

Figura 24 Imagen resultante del SOM aplicado a la base de datos de mastografías, con categoría BIRADS como clase. 76
Figura 25 Resultado obtenido con WEKA utilizando el algoritmo EM con n = -1...77
Figura 26 Resultado obtenido con WEKA utilizando el algoritmo EM con n = 4....77
Figura 27 Gráfica de los dos primeros componentes principales de la base de mastografías, con el diagnóstico patológico como clase. 78
Figura 28 Imagen resultante del SOM aplicado a la base de datos de mastografías con diagnóstico patológico como clase.....79
Figura 29. Relevancia de atributos según el método de ganancia en la tasa de la información. 81
Figura 30. Gráfica de la exactitud de los algoritmos 82
Figura 31. Gráfica de la evaluación de los algoritmos. 82
Figura 32. Árbol de decisión inducido con el algoritmo CART 84
Figura 33. Árbol de decisión inducido con el algoritmo C4.5 84
Figura 34. Árbol de decisión inducido por el algoritmo Random Tree....85
Figura 35. Ejemplo de registro de paciente. 93
Figura 36. Gráfica de exactitud....100
Figura 37. Medidas de evaluación. 100
Figura 38. Gráfica de exactitud (EM) 102
Figura 39. Gráfica evaluación (EM) 102
Figura 40. Gráfica exactitud EM-Infogain.....103
Figura 41. Gráfica evaluación EM-Infogain....103
Figura 42. Gráfica exactitud ElasticNet....104
Figura 43. Gráfica evaluación ElasticNet....104
Figura 44. Gráfica exactitud ElasticNet-Infogain....106
Figura 45. Gráfica evaluación ElasticNet-Infogain 106
Figura 46. Resultados BUGS....107
Figura 47. Gráfica de comparación exactitud 108
Figura 48. Gráfica de comparación de evaluación.....108
Figura 49. Gráfica comparación de exactitud entre algoritmos....109

Figura 50. Gráfica comparación de evaluación entre algoritmos 109
Figura 51. Gráfica comparativa de la exactitud entre métodos imputación.....110
Figura 52. Gráfica comparativa de la evaluación de los métodos de imputación.110

LISTA DE TABLAS.

Tabla 1. Matriz de confusión para las clases A y B.....30
Tabla 2. Categorías BI-RADS.....37
Tabla 3. Simbología de una gráfica rica.....48
Tabla 4. Datos del repositorio de mastografías.....67
Tabla 5. Clasificaciones BI-RADS.....67
Tabla 6. Diccionario de datos de la base BI-RADS.....68
Tabla 7. Comparación de algoritmos de regresión 70
Tabla 8. Coeficientes obtenidos mediante Elastic Net.....72
Tabla 9. Resultados de exactitud y evaluación de los algoritmos. 82
Tabla 10. Coeficientes resultantes del algoritmo Elastic Net. 87
Tabla 11. Descripción de los campos que componen al repositorio. 93
Tabla 12. Descripción de los datos codificados. 95
Tabla 13. Estadísticas descriptivas de los subconjuntos encontrados.....96
Tabla 14. Resultados al aplicar los algoritmos sobre repositorio incompleto.....100
Tabla 15. Resultados de repositorio imputado con algoritmo EM.....101
Tabla 16. Resultados de conjunto EM reducido con Infogain 103
Tabla 17. Resultados de imputación con Elastic Net.....104
Tabla 18. Resultados ElasticNet - Infogain 105
Tabla 19. Resultados de repositorio recuperado con BUGS 106
Tabla 20. Comparación de métodos de reducción de dimensionalidad.....108
Tabla 21. Comparación de algoritmos de minería de datos.....109
Tabla 22. Comparación entre métodos de imputación.....109
Tabla 23. Modelos resultantes para los tres subconjuntos de datos.....111

Capítulo I. Introducción

I.1 Antecedentes

El cáncer de mama es el cáncer de la mujer más común en el mundo, y su incidencia aumenta en los países en vías de desarrollo (OMS, 2011). En el caso de México no es la excepción, donde además es el cáncer con tasa de mortalidad más alta (INEGI, 2009). Cada año son diagnosticados con cáncer miles de pacientes y un gran porcentaje de estos casos son detectados en etapas avanzadas. En estos casos, el tratamiento solamente puede ser paliativo, es decir, se pueden aliviar síntomas y mejorar la condición de vida, pero no se puede alcanzar la curación, a diferencia de cuando la condición es detectada en etapas tempranas donde a través de terapias específicas es posible alcanzar la curación y además mejorar el nivel de vida del paciente.

Actualmente en México existe el programa nacional de detección oportuna del cáncer de mama (DOCMA), en el que personal capacitado hace actividades de promoción en las salas de espera de los hospitales, para educar a las mujeres en edad de riesgo a auto-explorarse y a detectar cambios en su anatomía, además de campañas de concientización en medios masivos de comunicación. A pesar de ello, la mortalidad por causa de esta enfermedad va en aumento.

El diagnóstico del cáncer de mama en sí, es un producto o la culminación de un proceso de toma de decisión, y como tal, es posible modelarlo para analizarlo y mejorarlo. En el presente trabajo de tesis, se expone el análisis efectuado sobre dos repositorios de datos de casos reales obtenidos de la clínica VIII del Instituto Mexicano del Seguro Social en la ciudad de Ensenada Baja California. El análisis mencionado consiste en descubrir conocimiento oculto, a través de un proceso llamado minería de datos, con la finalidad de integrar éste conocimiento y dar soporte al proceso de diagnóstico y cáncer de mama en sus diferentes etapas.

El proceso de minería de datos (KDD por sus siglas en inglés: Knowledge Data Discovery) se ha aplicado exitosamente en bases de datos grandes y ha tenido mucho éxito para descubrir conocimiento en repositorios de instituciones financieras y comerciales. El análisis de datos clínicos no es una excepción para aplicar el proceso KDD además de que ya se ha investigado extensamente el la minería de datos en registros clínicos, como se puede ver en la sección de trabajo previo.

El proceso KDD es el siguiente: (Orallo et al., 2004)

1. Integración y recopilación de datos**.** En esta fase se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas. A continuación se transforman todos los datos a un formato común.
2. Selección, limpieza y transformación. Esta fase consiste en eliminar o corregir datos incorrectos. Además se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes.
3. Minería de datos. Aquí se decide cuál es la tarea a realizar (clasificar, agrupar, regresión, etc.) y se elige el método que se va a utilizar.
4. Evaluación e interpretación. En esta fase se evalúan los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración.
5. Difusión y uso. En esta fase se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios.

Para el punto 3, que corresponde al de minería de datos, los algoritmos más comúnmente utilizados son los siguientes (Berner, 2006), (Gupta et al., 2011)

- Máquina vector de soporte. Las máquinas vector de soporte pertenecen a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones kernel) con maximización del margen.

- Árboles de decisión. Este modelo consiste en un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.
- Redes neuronales. Las redes neuronales artificiales son sistemas conexionistas dentro del campo de la inteligencia artificial, las cuales, dependiendo del tipo de arquitectura neuronal, pueden tener diferentes aplicaciones. Una de estas aplicaciones es la minería de datos.
- Métodos Bayesianos. Son métodos basados en la probabilidad y en la teoría de Bayes que se utilizan para realizar inferencias a partir de los datos, induciendo modelos probabilísticos que después serán usados para razonar (formular hipótesis) sobre nuevos valores observados.
- Métodos relacionales y estructurales. En este método, se utiliza un lenguaje de representación relacional, y además, los métodos relacionales permiten descubrir patrones más complejos, haciendo uso de la estructura de los propios datos y las relaciones entre ellos.
- Reglas de asociación y dependencia. En este método, se buscan reglas que expresen patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos. En concreto, estas reglas expresan las combinaciones de valores de los atributos que suceden más frecuentemente.
- Extracción de conocimiento con algoritmos evolutivos y reglas difusas. En el proceso de extracción de conocimiento en bases de datos y, en concreto, en el proceso de minería de datos, existen distintas tareas o problemas que se pueden enfocar y resolver como problemas de optimización y búsqueda. Los algoritmos evolutivos imitan los principios de la evolución natural para formar procedimientos de búsqueda y optimización global y son aplicables para el desarrollo de algoritmos de minería de datos propiamente dichos, como algoritmos de pre o pos-procesamiento o como herramientas para la optimización de los parámetros de otros algoritmos.

- Métodos basados en casos y en vecindad. Este tipo de métodos de aprendizaje tratan de resolver un problema a partir de información extraída de un conjunto de ejemplos existentes previamente. Los ejemplos son los que aportarán la información necesaria para poder predecir el comportamiento de un nuevo dato no perteneciente al conjunto de ejemplos.

El interés del área de la medicina en la aplicación de las técnicas de minería de datos, se debe al aumento en la utilización de sistemas de expediente clínico electrónico y a la existencia de grandes volúmenes de información en sus bases de datos. La minería de datos permite la interpretación de una gran variedad de información clínica procedente de fuentes tales como radiografías, tomografía computarizada, imágenes de ultrasonido, bases de datos epidemiológicas y otra información procedente de otros estudios clínicos (Sboner, 2005)

I.2 Trabajo previo.

Existen varios estudios en donde se aplica la minería de datos para el descubrimiento de conocimiento en repositorios clínicos. En algunos de ellos se tratan problemas como la evaluación del riesgo de padecer alguna enfermedad, o para asisitir en el tratamiento y pronóstico de otras enfermedades. En esta sección se presentan algunos trabajos encontrados en la literatura que tratan sobre este mismo tema y los resultados que se obtuvieron.

En el trabajo de Bellaachia y Guven (Bellaachia & Guven, 2006) se presenta un análisis sobre la tasa de supervivencia de pacientes con cáncer de mama utilizando minería de datos. Se utilizó una base de datos SEER y se aplicaron tres algoritmos de clasificación: Naïve Bayes, red neuronal con retropropagación y el árbol de decisión C4.5 siendo éste último el que obtuvo mejores resultados que los otros dos algoritmos. En este trabajo, el análisis se hizo para determinar la tasa de supervivencia de pacientes que ya padecieron el cáncer de mama y no se hace un estudio sobre la prevención del padecimiento.

En el estudio de (Delen et al., 2005), se hace una comparativa entre tres algoritmos de minería de datos, red neuronal, árbol de decisión C5 y regresión logística, para predecir la sobrevivencia en pacientes que han padecido cáncer de mama y se encuentran en estado de remisión. Los algoritmos se aplicaron sobre una base de datos de más de 200,000 casos reales y se utilizó validación cruzada de 10-fold. De los resultados obtenidos, el algoritmo C5 fue el que obtuvo los mejores resultados, con una precisión de 93.6% (siendo la mejor obtenida que cualquiera reportada en la literatura existente) seguido por la red neuronal con una precisión del 91.2% y por último la regresión logística con una precisión de 89.2% En este estudio, al igual que en el anterior, el análisis se hace en la predicción de la sobrevivencia de pacientes, es decir, en cuanto tiempo queda de vida una vez que han padecido el cáncer de mama y no sobre el riesgo que tiene una persona de padecerlo, o sobre aspectos preventivos del mismo.

En (Kharbat et al., 2007) se hizo un estudio sobre una base de datos de un hospital británico, Frenchay Breast Cancer (FBC) la cual contiene datos histopatológicos de miles de pacientes con cáncer de mama. El objetivo del estudio consistió en comparar cuatro algoritmos de minería de datos determinar qué tan correctamente clasificaban los datos según si pertenecían a un paciente con cáncer o no. Los algoritmos utilizados fueron: Naïve Bayes, Máquina vector de soporte, C4.5 y XCS (Wilson, 1995). Se obtuvieron resultados más precisos con XCS (80.1%) seguido de C4.5 (77.4%). En el estudio se validaron las reglas generadas por los algoritmos C4.5 y XCS por un experto en el área y se reporta que las reglas generadas por XCS fueron cualitativamente más descriptivas y útiles que aquellas generadas por C4.5 a pesar de que este último arrojó reglas más compactas. De las limitaciones del estudio, se puede decir que el análisis se hizo sobre datos provenientes exclusivamente de tejido tumoral por lo que es necesario realizar una biopsia a una paciente para determinar su riesgo de padecer cáncer y no se toman en cuenta otros factores clínicos que pudieran alertar del mismo riesgo.

En Tan et al (2002) se hace uso de programación genética para analizar datos provenientes de pacientes con enfermedades relacionadas con el dolor de pecho. Se obtuvieron reglas para clasificar 12 enfermedades relacionadas con el dolor de pecho, entre ellas se encuentran la angina de pecho, infarto, embolia pulmonar y úlcera péptica.

En el trabajo de (Bellazzi & Zupan, 2008) se hace una revisión del proceso de minería de datos en medicina y se propone un framework para hacer frente a los problemas de construir, evaluar y explotar modelos de minería de datos en medicina, por ejemplo, en genómica, espectros de masa, y predicción del diagnóstico del paciente. El trabajo es muy completo y trata con el aspecto de la construcción e implementación de un sistema de soporte a la toma de decisión clínica, aunque en el artículo no se trata con algún caso de estudio para algún padecimiento en particular.

En (Williams & Williamson, 2006) se utiliza una red bayesiana para predecir el pronóstico y la recurrencia local del cáncer de mama en pacientes en tratamiento para dicha enfermedad. La base de datos que se utiliza para éste estudio fue recopilada a finales de la década de los años de 1980 y no se tomaba en cuenta el diagnóstico inmunohistoquímico o de receptores hormonales que se encuentran en la superficie de las células tumorales y que le confieren propiedades y características particulares a cada caso de cáncer de mama.

I.3 Planteamiento del problema.

Como ya se mencionó anteriormente, el cáncer de mama es el cáncer más común de las mujeres y en el caso de México, ocupa el primer lugar en mortalidad entre mujeres, esto se debe a que es diagnosticado en etapas avanzadas. En estos casos, el tratamiento de la enfermedad únicamente tiene el propósito de aliviar síntomas pero no para alcanzar la curación.

Los trabajos de investigación sobre éste padecimiento, se basan en datos obtenidos mediante análisis histopatológicos (análisis microscópico de tejido para

estudiar la manifestación de una enfermedad) para predecir el transcurso de la enfermedad y para pronosticar el tiempo de vida, por otra parte, no se encontraron trabajos que traten de determinar el riesgo de padecer el cáncer de mama ni trabajos que contemplen los datos y mediciones que se realizan en la actualidad mediante protocolos modernos de medicina oncológica.

Dado lo anterior, surge la necesidad de realizar un trabajo de investigación para resolver el problema del diagnóstico y tratamiento del cáncer de mama utilizando minería de datos. Para el diagnóstico oportuno, es necesario clasificar los datos de mastografías en categorías de riesgo de padecer cáncer de mama. Para el tratamiento, es necesario analizar y clasificar los datos de pacientes diagnosticadas para estudiar su evolución y respuesta al tratamiento.

I.4 Objetivo.

El objetivo del presente trabajo de tesis es aplicar el proceso de minería de datos en repositorios relacionados con el cáncer de mama con la finalidad de descubrir conocimiento útil que apoye al proceso de diagnóstico y tratamiento del cáncer de mama en sus diferentes etapas.

I.4.1 Objetivos específicos.

- Modelar el proceso de diagnóstico y tratamiento del cáncer de mama.
- Aplicar el proceso de minería de datos (KDD) sobre los repositorios de mastografías y de pacientes en tratamiento con la finalidad de descubrir conocimiento.
- Elaborar reglas a partir del conocimiento descubierto.
- Validar las reglas elaboradas con los especialistas.
- Integrar las reglas en los procesos de diagnóstico y tratamiento del cáncer de mama.

I.5 Metodología de la investigación.

En la figura 1 se muestra un esquema de la metodología utilizada en el presente trabajo de investigación y subsecuentemente se describe a detalle cada una de las partes del esquema.

I.5.1 Revisión de la literatura.

La revisión de la literatura consiste en revisar material bibliográfico y artículos científicos relacionados con el tema de clasificación, regresión, estadística, tratamiento de datos perdidos y su aplicación en la investigación de datos clínicos.

I.5.2 Caso de estudio.

Se realizó un caso de estudio en una clínica con la finalidad de estudiar el proceso del diagnóstico y tratamiento del cáncer de mama. Se realizaron entrevistas con personal de epidemiología, médicos familiares, especialistas en oncología, patología y radiología, además se realizó un estudio sombra durante la consulta de pacientes diagnosticadas con cáncer de mama, con la finalidad de entender el proceso de toma de decisión que conlleva el diagnóstico y el tratamiento de dicha enfermedad. Otro de los propósitos de realizar el mencionado caso de estudio, fue el de localizar las fuentes de datos apropiadas para aplicar el proceso de minería de datos.

Al final de la realización del caso de estudio, se elaboró la descripción textual del proceso y se modeló con las gráficas correspondientes, además de contar también con los datos a los que se les aplicó el proceso de minería de datos.

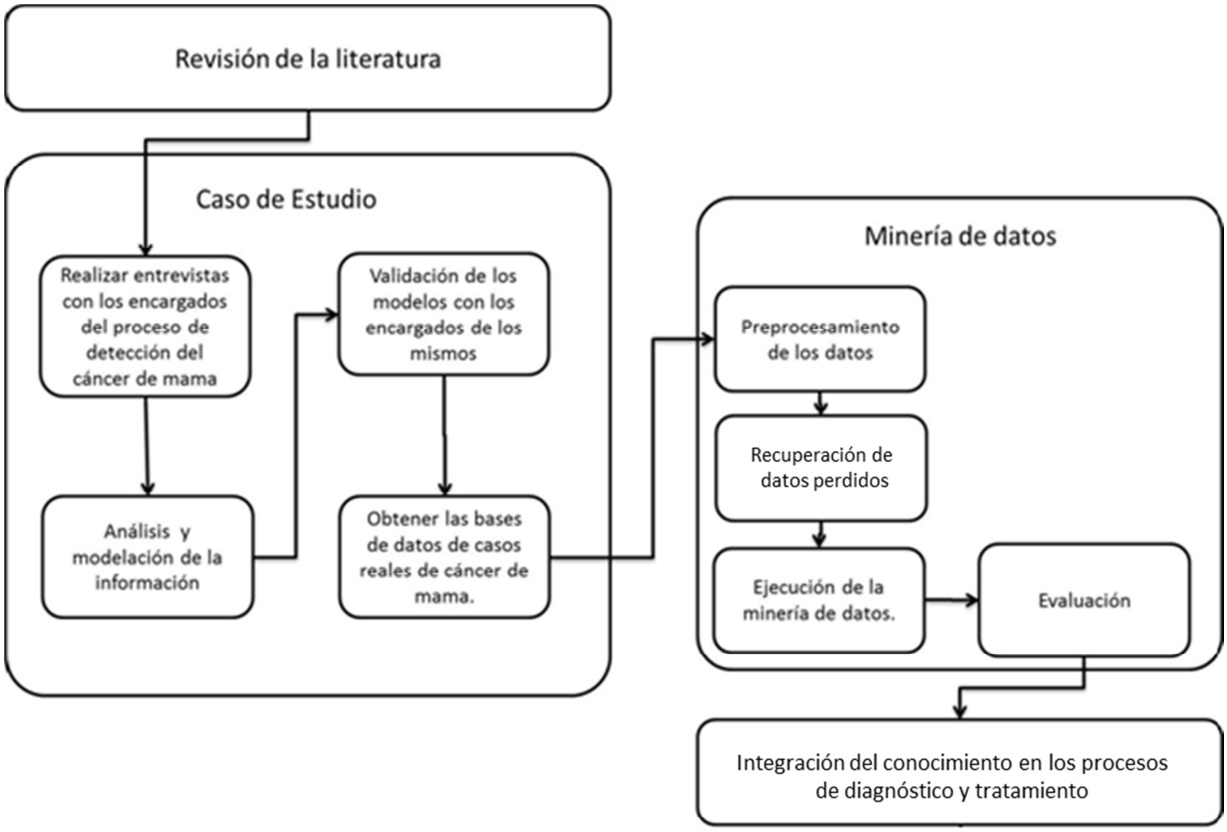


Figura 1. Metodología de la investigación.

I.5.3 Minería de datos (Proceso KDD)

Se aplicó el proceso de minería de datos sobre los repositorios obtenidos en el paso anterior. Uno de los respositorios corresponde a registros de mastografías y el otro a un registro de pacientes diagnosticadas y en tratamiento para el cáncer de mama. Dado que los registros se encontraban guardados a manera de redacción, se procedió a codificar y transformar los datos hasta obtener una vista minable que pudiera ser procesada por los algoritmos de clasificación y regresión. Antes de aplicar los algoritmos mencionados, se procedió a efectuar un paso

previo de recuperación de datos perdidos, debido a la existencia de varios casos que no podían ser desechados y además por la naturaleza de la variable o dato, tampoco podía ser descartada del análisis.

Por último se procedió a aplicar los algoritmos de clasificación y regresión sobre los repositorios transformados y recuperados. El conocimiento descubierto fue sometido a un proceso de evaluación con la finalidad de valorar su exactitud, sensibilidad y especificidad.

I.5.4 Integración o difusión del conocimiento descubierto.

Del conocimiento y reglas descubiertas en el proceso KDD, se procede a integrarlas y darlas a conocer dentro de los procesos de toma de decisión clínica. El conocimiento se difunde a manera de reglas y recomendaciones para los especialistas del proceso.

I.5 Contenido de la tesis.

En esta sección se describe de manera general el contenido y la organización del presente trabajo de tesis. En el primer capítulo se presenta la introducción, se enuncia el planteamiento del problema, los objetivos y sus objetivos específicos así como la metodología para lograr el objetivo.

En el capítulo II se aborda la teoría referente al proceso de descubrimiento de conocimiento en datos, de los algoritmos de minería de datos utilizados para la clasificación y para la identificación de cúmulos.

El capítulo III introduce al lector al contexto médico del problema que se aborda en este trabajo de tesis. Se da a conocer la terminología utilizada en los capítulos sucsecuentes y se presenta un caso de estudio sobre el proceso de diagnóstico y tratamiento del cáncer de mama. Se presentan los modelos gráficos del proceso y se introducen las fuentes de datos que se analizaron y cuyos resultados se presentan en los capítulos III y IV.

El capítulo IV explica el proceso de minería de datos que se aplicó en la base de datos de mastografías que se obtuvo del caso de estudio descrito en el capítulo III. Al final del proceso se obtienen los modelos predictivos.

El capítulo V trata sobre el proceso de minería de datos pero se aplica esta vez sobre una base de datos de pacientes de cáncer en tratamiento.

Finalmente el capítulo VI describe las conclusiones del trabajo y se propone trabajos para realizar en un futuro.

Capítulo II. Marco teórico

En este capítulo se presentan conceptos teóricos sobre el proceso de descubrimiento de conocimiento en datos y en particular sobre la fase de minería de datos y en cada uno de los algoritmos utilizados.

II.1 Proceso de descubrimiento de conocimiento en datos

La mayoría de las organizaciones internacionales producen en unos cuantos días más información de la que una sola persona leerá en toda su vida. Cada día se producen gigabytes de información a un ritmo que no es posible analizar ni monitorear por personas ya que el crecimiento de la producción de datos es exponencial. La producción automatizada de datos requiere que se desarrollen nuevas estrategias mecanizadas para la selección, filtrado y análisis de éstos datos.

El proceso de descubrimiento de conocimiento en datos o KDD por sus siglas en inglés (Knowledge Data Discovery) surge como la intersección de varias disciplinas: Inteligencia artificial, bases de datos, estadística, aprendizaje de máquinas, visualización, computación paralela y otras. El propósito del KDD es el procesamiento de grandes bases de datos con la finalidad de descubrir conocimiento útil que pueda ser aplicado para la toma de decisiones.

El término minería de datos ha sido utilizado por muchos autores como un sinónimo del proceso de descubrimiento de conocimiento en datos, sin embargo actualmente se denomina minería de datos a un paso específico del proceso KDD el cual corresponde exclusivamente al paso de descubrimiento de conocimiento.

Formalizando, podemos definir al proceso KDD como al procedimiento no trivial de extraer conocimiento desconocido, implícito, previamente desconocido y además potencialmente útil, a partir de datos (Fayyad et al., 1996)

Ahora bien, definiremos los pasos del proceso y en las subsecciones siguientes se profundizará en cada uno de ellos.

El proceso KDD consta de cinco pasos que se ilustran en la figura 2.

El primer paso es la fase de integración y recopilación, que consiste en identificar la meta del proceso y entender el dominio de la aplicación. Después se seleccionan los datos relevantes al dominio y se guardan en un formato en común.

Los datos pueden provenir de diferentes fuentes y presentarse en diferentes formatos digitales, por lo tanto será necesario realizar cambios para lograr que todos los datos se ajusten a una sola vista.

El segundo paso corresponde al de la limpieza y transformación de los datos. Existen varios motivos por los que un repositorio pudiera estar contaminado, uno de los más comunes es la existencia de registros duplicados debido a errores en la escritura de uno o más registros. Otro motivo es por la falta de consistencia en la captura, en particular al escribir fechas. Este tipo de contaminación es la que causa más problemas a la hora de descubrir patrones y es mejor prevenir este tipo de problemas desde el momento de la captura de datos.

Debido a que muchos de los datos pueden ser muy detallados para ser utilizados en la manera en que se encuentran, es necesario transformarlos o codificarlos. Por ejemplo un campo que representa a un domicilio, se puede recodificar como un código regional, o la edad se puede codificar en conteos que corresponden a intervalos de cinco años. La tarea de codificar atributos requiere de creatividad y además podría ser necesario repetir este paso hasta obtener los mejores resultados.

En algunos casos, se obtienen mejores resultados cuando se discretizan valores continuos. El proceso de discretización consiste en transformar los valores que toma un atributo numérico continuo en intervalos disjuntos que cubren todo el rango de valores del atributo en cuestión.

Una vez que se cuenta con un repositorio limpio y transformado ya se puede seleccionar y aplicar el algoritmo de minería de datos que mejor se adapte a las necesidades. De este paso se obtiene un modelo, que pueden ser reglas, árboles de asociación, o incluso algún modelo matemático como una red neuronal, un modelo estadístico o probabilístico o máquina vector de soporte.

El conocimiento adquirido del paso de minería de datos debe ser evaluado para descartar un posible regreso a alguna de las fases anteriores, ya que el proceso completo es iterativo y éste termina una vez que se han logrado resultados satisfactorios.

Si los resultados en el paso de evaluación son satisfactorios, se procede a hacer difusión del conocimiento encontrado, típicamente se utilizan reportes o se consolida a través de un sistema que hace uso del conocimiento.

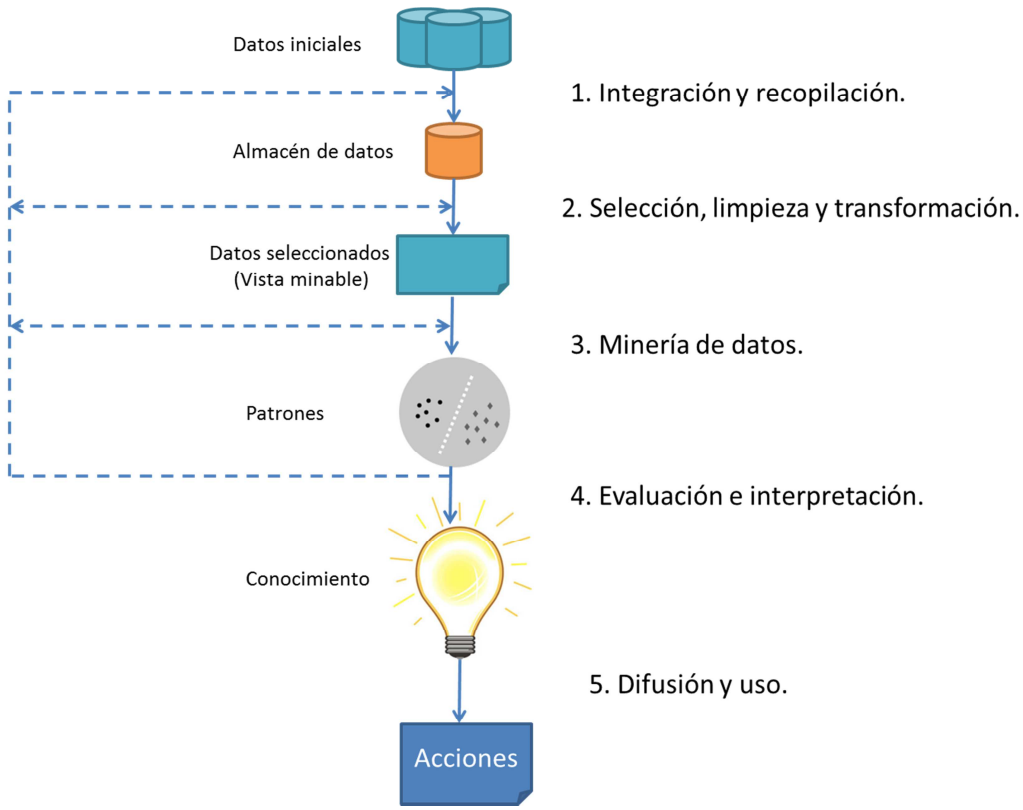


Figura 2. Proceso de descubrimiento de conocimiento en datos. (Orallo et al., 2004) p. 20

II.2 Minería de datos

La minería de datos es el proceso de descubrir patrones y relaciones en grandes bases de datos. Bajo este contexto la minería de datos abarca muchos campos, incluyendo el del reconocimiento de patrones, la estadística, las ciencias de la computación y la administración de bases de datos.

A grandes rasgos, el propósito de la minería de datos es transformar a los datos en información útil. Dicho de otra manera, trata de descubrir patrones nuevos para los usuarios. Los patrones pueden tener dos propósitos, la descripción y la predicción. La primera se enfoca en encontrar patrones y presentarlos a los usuarios de una manera interpretable. La predicción involucra identificar atributos

en la base de datos que puedan ser útiles para predecir valores o comportamientos futuros.

En el campo de la salud, la minería de datos no es una excepción. Día a día se generan grandes volúmenes de datos que pudieran explotarse con el objetivo de mejorar las prácticas médicas, mejorar el tratamiento de enfermedades y para optimizar recursos hospitalarios.

II.2.1 Reconocimiento de patrones

El reconocimiento de patrones es un área de la minería de datos. Es una ciencia que busca modelos analíticos con el propósito de describir o clasificar datos y mediciones. El reconocimiento de patrones tiene como objetivo inferir a partir de una colección de datos, mecanismos que faciliten los procesos de toma de decisión (Duda et al., 2001) (Friedman et al., 2008)

Uno de los enfoques que tiene el reconocimiento de patrones es el estadístico, en éste se utiliza la estadística para obtener modelos a partir de los datos o mediciones. La meta es seleccionar atributos que permitan a un patrón ser clasificado en uno o más grupos o clases.

Existen varios casos de éxito en donde se ha aplicado la minería de datos en medicina utilizando reconocimiento de patrones. En (Burroni, 2004) se hace un análisis de imágenes digitalizadas de lesiones en piel con el propósito de asisitir en el diagnóstico de melanoma. En (Bertani et al., 1999) se hace un análisis sobre el andar de un paciente con pie plano, para diagnosticar y monitorizar la recuperación de paciente después de cirugía. En el trabajo de (Leonard et al., 2002) se propone minar la información de la base de datos de Medline para obtener referencias sobre genes y proteínas y asistir en la relevancia en los resultados.

II.2.2 Aprendizaje supervisado y aprendizaje no supervisado

La minería de datos y la obtención de modelos predictivos se pueden concebir como aprendizaje a partir de datos (o evidencia). El aprendizaje puede ser supervisado y no supervisado.

II.2.2.1 Aprendizaje supervisado

En el aprendizaje supervisado se supone que el usuario conoce con certeza las categorías de cada registro con el que se cuenta.

Al proceso de transferir el conocimiento al modelo se le conoce como entrenamiento. Los datos utilizados en este proceso se llaman conjunto de entrenamiento. Los ejemplares del conjunto se componen de variables dependientes u objetivos y las variables independientes o predictores. El sistema se ajusta según cada ejemplar de entrenamiento y la señal de error que éste produce (típicamente la diferencia entre la respuesta esperada y la respuesta obtenida). Con una cantidad suficiente de ejemplares de entrenamiento, se puede minimizar el error en la predicción y por lo tanto obtener un modelo predictivo más preciso.

El objetivo de éste enfoque es establecer un modelo predictivo entre la variable dependiente y los predictores. La naturaleza de la variable dependiente determina el tipo de modelo predictivo, si la variable es del tipo discreta, el modelo será de clasificación; si la variable dependiente es continua, el modelo será de regresión.

II.2.2.2 Aprendizaje no supervisado.

En el aprendizaje no supervisado se cuenta con un conjunto de datos pero no existe información sobre cómo agrupar los datos en conglomerados. Basándose en similitudes entre los datos, se empiezan a desarrollar conglomerados o clústeres entre los datos hasta que comienzan a aparecer un conjunto de patrones diferenciables. Aquí no hay diferenciación entre variables independientes o predictores y variables dependientes.

II.2.3 Clasificadores para aprendizaje supervisado

En el aprendizaje supervisado, se pretende clasificar a los registros de datos en alguna de las categorías predefinidas. En el desarrollo de un modelo predictivo, la meta es crear un clasificador que pueda predecir la categoría de un registro basándose en los datos con los que cuenta.

II.2.3.1 Árboles de decisión

Los árboles de decisión son uno de los modelos más fáciles de entender y más utilizados de la categoría de aprendizaje supervisado. En la figura 3 se muestra un ejemplo de árbol de decisión. Para llegar a una decisión, se recorre el árbol de arriba abajo, y en cada uno de sus nodos se comprueba a cada atributo hasta llegar a una conclusión que siempre corresponde a una de las hojas del árbol. Por lo tanto, un árbol puede representarse como un conjunto de reglas del tipo Si – Entonces, que son fáciles de entender y de interpretar por un usuario.

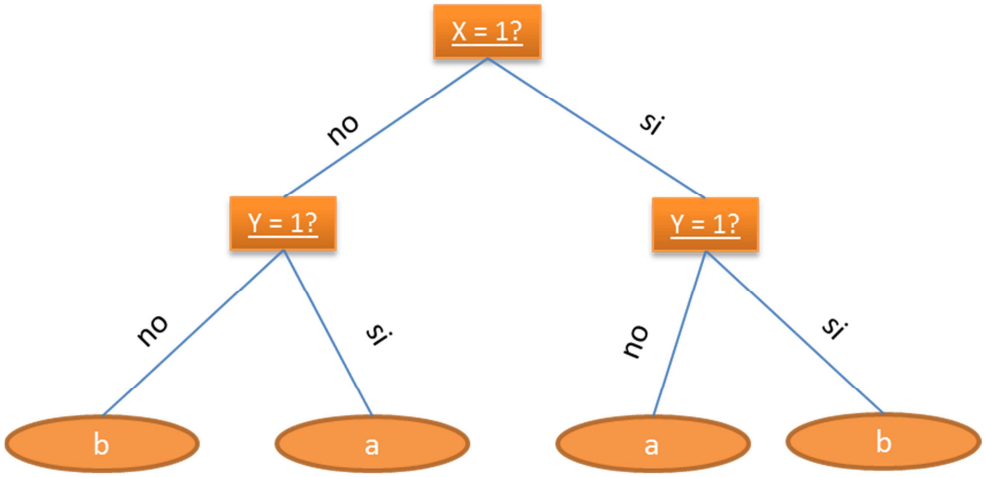


Figura 3. Ejemplo de árbol de decisión.

Otra de las ventajas en el uso de los árboles de decisión es su resistencia a la pérdida de datos en algunos de los atributos, además de que el algoritmo es robusto en cuanto a la presencia de ruido en los datos. (Quinlan, 1986).

Cada nodo del árbol corresponde a un atributo del repositorio. Éstos se van seleccionando recursivamente basándose en lo que denominan una medida de impureza de nodo. Para el caso de la inducción de un árbol de decisión con el algoritmo ID3, la medida de impureza de nodo que se utiliza es la entropía, que está basada en el trabajo de (Shannon & Weaver, 1949) sobre la teoría de la información.

Si expresamos formalmente a un árbol de decisión como un conjunto de experiencias donde los elementos se denominan:

$$A^j , B^j , C^j , \dots , Clase_p^j \tag{1}$$

En donde $A^j , B^j , C^j , \dots$ son los valores de los atributos A, B y C en el elemento j del conjunto de experiencias. La $Clase^j$ corresponde a la clase asignada, para el elemento *j*.

Sea un nodo en el que existen un conjunto de experiencias a clasificar y en el que existen los atributos A, B y C con todos sus valores. La entropía del nodo se define como:

$$H = -\sum_{i=1}^{n} P_i log_2 \ P_i \tag{2}$$

En donde

$$P_i = \frac{n \ elementos \ en \ el \ conjunto \ de \ experiencias \ en \ clase \ i}{n \ total \ de \ elementos \ en \ el \ conjunto \ de \ experiencias} \tag{3}$$

Entonces *H* mide la cantidad de información contenida en la clase *i* y ésta será cero si la clase *i* tiene probabilidad de uno.

Si se clasifica según el atributo A, se producirán h subconjuntos de experiencia asociados a cada nodo en cada modelo, para calcular la entropía parcial, denotada *HP*. Entonces si se estima la nueva entropía:

$$H = \sum Prob(A_i)HP_i \tag{4}$$

El incremento de información aportado por la clasificación según A se evalúa por la disminución de la entropía producida respecto de la inicial, es decir:

$$\Delta I = H - H(A) \tag{5}$$

El criterio para hacer la separación de los datos es seleccionar al atributo con mayor ΔI y similarmente obtener las siguientes ΔI para cada atributo y seleccionar la que tiene mayor valor.

II.2.3.2 Regresión logística

La regresión logística se usa para modelar datos en los que la variable dependiente toma valores binarios, en otras palabras, toma el valor de 1 con una probabilidad de éxito p, o toma un valor 0 con una probabilidad de fracaso 1 – p.

El objetivo es desarrollar unn modelo de regresión que relacione a la variable dependiente con los predictores. En lugar de predecir los valores de la variable dependiente, la regresión logística estima la probabilidad de que una variable dependiente tome un valor dado. Por ejemplo, en vez de predecir si un paciente sufre de una determinada enfermedad, la regresión logística trata de estimar la probabilidad de que el paciente padezca la enfermedad. Si la probabilidad estimada es mayor a 0.5, entonces existe una probabilidad más alta de que el paciente padezca la enfermedad, en lugar de que no la padezca. La función que relaciona a las probabilidades a los predictores no es lineal y se representa con la siguiente ecuación:

$$p(y) = \frac{1}{1 + e^{(-\beta_0 - \beta x)}} \tag{6}$$

En donde *p(y)* representa la probabilidad de que la variable dependiente ocurra según *x*, el valor del predictor. En la figura 4 se muestra un ejemplo de una curva que representa a un modelo de regresión logística. La curva está matemáticamente limitada a permanecer en el rango 0.0 a 1.0 del eje vertical.

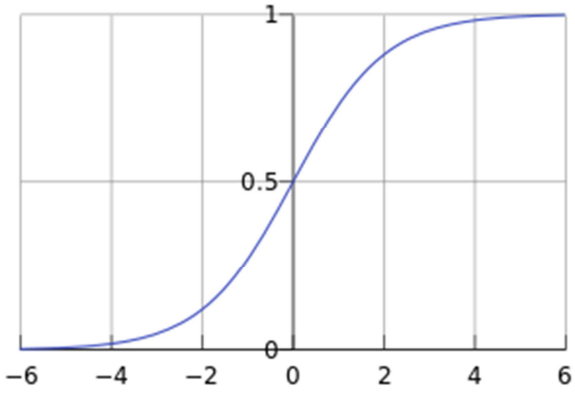


Figura 4. Modelo de regresión logística.

La construcción del modelo se realiza seleccionando los parámetros β que mejor representen al conjunto de datos de entrenamiento.

II.2.3.3 Métodos Bayesianos

Los métodos Bayesianos corresponden a los clasificadores del tipo estadístico y cumplen con una doble función: descripción y predicción. Como modelo descriptivo, los algoritmos de aprendizaje de las redes Bayesianas se centran en el descubrimiento de relaciones de independencia y relevancia entre sus variables, por lo que el modelo resultante refleja de forma explícita numerosas relaciones de interés. Como modelo predictivo, permite calcular de forma explícita la probabilidad asociada a cada una de las hipótesis posibles, lo cual le confiere una ventaja sobre los otros modelos.

Los métodos Bayesianos se centran fundamentalmente en el teorema de Bayes. El teorema de Bayes permite actualizar la creencia que se tiene en un suceso o conjunto de sucesos en base a nuevas observaciones o datos. En otras palabras, se pueda pasar de la probabilidad a priori P(suceso) a una probabilidad a posteriori P(suceso | observaciones).

La probabilidad a priori se puede ver como la probabilidad inicial. L probabilidad a posteriori se obtiene tras conocer información o evidencia. En base a los conceptos anteriores, el teorema de Bayes se representa por la siguiente expresión:

$$P(h|O) = \frac{P(O|h)P(h)}{P(O)} \tag{7}$$

En donde *P(h)* es la probabilidad a priori de la hipótesis *h* y de las observaciones *O* y a *P(O|h)* se le conoce como la verosimilitud de que la hipótesis *h* haya producido el conjunto de observaciones *O*.

Considerando al problema de clasificación con variable clase *C* y un conjunto de variables predictoras $\{A_1, \dots, A_n\}$ el teorema de Bayes queda de la siguiente forma:

$$P(C|A_1, \dots, A_n) = \frac{P(A_1, \dots, A_n|C)P(C)}{P(A_1, \dots, A_n)} \tag{8}$$

En donde la clase a devolver será:

$$\arg max_{c \in \Omega} = P(A_1, \dots, A_n|C)P(C) \tag{9}$$

El problema con el teorema de Bayes es su complejidad computacional debido a que se necesita trabajar con distribuciones de probabilidad que involucran muchas

variables. Con el uso del clasificador Bayes Simple se puede mejorar el problema de la complejidad computacional suponiendo independencia entre las variables.

Dado un conjunto de datos con muchos atributos, el costo computacional de calcular $P(C_j|X)$ sería muy alto. Para reducir el costo computacional, el clasificador Bayes Simple supone la independencia condicional de la clase. Por lo tanto nos quedaría:

$$P(C_j|X) = \prod_{k=1}^{n} P(x_k|C_i) \tag{10}$$

En el caso de las redes bayesianas, éstas representan al conocimiento cualitativo mediante un grafo dirigido acíclico, el modelo se articula en la definición de relaciones de dependencia/independencia entre las variables que componen al modelo. Estas relaciones abarcan desde una independencia completa hasta una dependencia funcional entre las variables del modelo. El hecho de que se utilice una representación gráfica del modelo, hace que las redes bayesianas sean una herramienta muy atractiva para representar al conocimiento.

Según (Witten & Frank, 2005) una red bayesiana para un conjunto de variables aleatorias $X = (X_1, X_2, X_3, \dots, X_n)$ es un par $B = (G, \Theta)$, donde *G* es un grafo acíclico dirigido, cuyos nodos están en correspondencia uno a uno con las variables X, Θ es un conjunto de funciones de probabilidad $P(X_i|Pa(X_i))$ desde *i* hasta *n.* Se usa $Pa(X_i)$ para denotar a los padres del nodo X_i en el grafo *G.*

La codificación de las relaciones de independencia condicional expresadas en una red bayesiana hace que la distribución de probabilidad se pueda almacenar mucho más eficiente y local a cada una de las variables del modelo.

$$P(X) = P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{11}$$

Las funciones de probabilidad local P son las funciones de probabilidad correspondiente a los términos de la ecuación (11).

Una vez construida, la red Bayesiana constituye un dispositivo potente para el razonamiento probabilístico, sin embargo queda la tarea de la construcción del modelo. Por lo tanto, el problema de aprendizaje de redes Bayesianas se puede definir como sigue: dado un conjunto de datos *D*, encontrar el grafo acíclico dirigido G que mejor represente al conjunto de dependencias/independencias presentes en los datos. Para solucionar este problema desde el punto de vista bayesiano, es necesario calcular la probabilidad posterior de una red bayesiana en particular, dado el conjunto de datos conocidos, esto es *P(G|D)*. Una vez que se sabe cómo calcular esa probabilidad, se tiene la medida de evaluación de cada grafo a los datos de partida y por consiguiente podremos comparar para quedarnos con la mejor entre las distintas redes bayesianas disponibles. Está demostrado que el problema del aprendizaje, tal y como se ha definido, es un problema NP-Duro (Chickering, 1996).

Para resolver el problema, se utiliza una métrica para medir la optimalidad de una red Bayesiana basado en el Criterio de información Bayesiana.

*El criterio n*o requiere la evaluación de la función de probabilidad a priori. Consecuentemente se trata de un criterio práctico usado en circunstancias apropiadas. Cuando se aplica a redes con distribución multinomial para las variables de este criterio, se tiene la siguiente expresión:

$$f_i(X_i|Pa(X_i):D) = \sum_{j=1}^{q}\sum_{k=1}^{r} N_{ijk} log \frac{N_{ijk}}{N_{jk}} - \frac{1}{2} Dim(B) logN \quad (12)$$

Dim(B) representa una función para calcular el número de parámetros libres que se necesitan especificar en la red Bayesiana, esto es $Dim(B) = \sum_{i=1}^{n}(r_i - 1)q_i$, que es el número de entradas independientes en las tablas de probabilidad.

Una vez especificado el problema de aprendizaje, se puede identificar que es un problema de optimización. Pero como este problema de optimización no tiene solución exacta, por consiguiente, se han planteado en la literatura específica del tema de algoritmos de búsqueda específicos o adaptación de algoritmos metaheurísticos para la resolución aproximada del aprendizaje estructural.

Los algoritmos de búsqueda más utilizados para el aprendizaje, son los basados en métodos voraces y locales (Cooper & Herskovits, 1992), (Buntine, 1994).

II.2.3.4 Máquina Vector de Soporte (SVM)

Con el propósito de introducir el concepto de máquina vector de soporte, comenzaremos a explicar el caso más sencillo, el clasificador lineal de dos clases.

Sea un conjunto S con n datos, $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ en donde cada x_i corresponde a un vector en un espacio de entrada X de dimensión D, y cada $y_i \in \{-1, 1\}$ indica la clase a la que pertence x_i; ahora supongamos que queremos separar y clasificar cada uno de éstos puntos en dichas clases utilizando un hiperplano separador. A esta forma de separar se le llama clasificación lineal. También es necesario seleccionar a aquel hiperplano que maximice el espacio entre los puntos más cercanos entre ambas clases.

El hiperplano separador de las muestras del conjunto S, se puede representar por:

$$f(x) = \langle w.x \rangle + b \tag{13}$$

En donde *x* corresponde a un vector de entrada, *w* es un vector de pesos que corresponde al gradiente o dirección del hiperplano *f(x); b* es un escalar que corresponde a la distancia del hiperplano *f(x)* al origen.

Para que los datos estén bien separados, se deben cumplir las siguientes propiedades:

$$\langle w.x\rangle + b \geq 1 \quad si \quad y_i = 1 \tag{14}$$

$$\langle w.x\rangle + b \leq -1 \quad si \quad y_i = -1 \tag{15}$$

En la figura 5 se muestra un ejemplo de hiperplano separador basado en máquina vector de soporte.

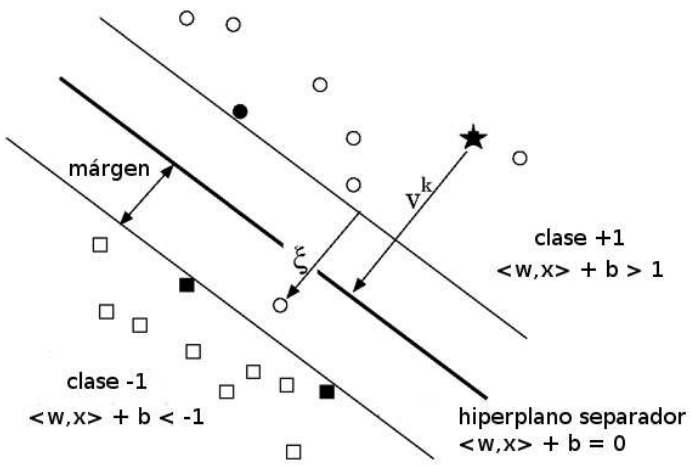


Figura 5. Ejemplo de hiperplano separador

La distancia de un vector *x* a un hiperplano *f*, se define por la expresión: $dist(f,x) = \frac{|f(x)|}{\|w\|}$, en donde $\|w\|$ es la norma asociada al producto escalar $\langle w.w\rangle$. Dado que existen un número infinito de soluciones que solo difieren en la escala de w, se puede reescalar w de tal manera que la distancia esté definida como:

$$dist(f,x) = \frac{1}{\|w\|} \tag{16}$$

Entonces se tiene la formulación más común de la SVM lineal con margen máximo en su forma primal:

$$Minimizar\ \frac{1}{2}\langle w.w\rangle \qquad (17)$$
$$Sujeto\ a\colon y_i(\langle w.x_i\rangle + b) \geq 1 \quad i = 1,2,\dots,n$$

El cual es un problema de optimización convexa, consistente en minimizar una función cuadrática con restricciones lineales. Para resolver este problema, se puede utilizar el método de los multiplicadores de Lagrange para encontrar la solución óptima, entonces el problema original con restricciones se puede formular como un problema de optimización sin restricciones:

$$L(w,b,\lambda) = \frac{1}{2}\langle w.w\rangle + \sum_{i=1}^{n}\lambda_i(1 - y_i(\langle w.x_i\rangle + b)) \qquad (18)$$

Derivando con respecto a las variables primales,

$$\frac{\partial L(w,b,\lambda)}{\partial w} = 0 \Rightarrow w - \sum_{i=1}^{N} y_i\lambda_i x_i = 0 \Rightarrow w = \sum_{i=1}^{N} y_i\lambda_i x_i \qquad (19)$$

$$\frac{\partial L(w,b,\lambda)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} y_i\lambda_i = 0 \qquad (20)$$

Entonces podemos reescribir la función lagrangiana generalizada en términos de las variables λ_i:

$$L_{dual}(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1} y_i y_j \, \lambda_i \lambda_j < x_i, x_j > \tag{21}$$

La ecuación anterior es la formulación dual del problema de optimización, para resolverlo, es necesario maximizar con respecto a los parámetros λ_i y las restricciones:

$$\sum_{i=1}^{N} y_i \lambda_i = 0 \tag{22}$$

El teorema de Karush-Kuhn-Tucker dice que λ_i^* es diferente de cero si las restricciones correspondientes en la siguiente ecuación son iguales:

$$\lambda_i^* [y_i (w . x_i + b) - 1] \tag{23}$$

Entonces los x_i para los cuales $\lambda_i^* > 0$, serán los vectores de soporte.

La formulación dual del problema de optimización tiene dos principales ventajas, una de ellas es que el problema de optimización es más fácil de resolverse debido a que no se tiene que lidiar con las desigualdades en las restricciones de su formulación primal, la otra ventaja, es que el problema queda en términos del producto interior de los vectores de entrada x.

II.2.4 Aprendizaje no supervisado

La clasificación no supervisada se utiliza en situaciones en donde el objetivo es clasificar una colección diversa de datos sin categoría en diferentes grupos basándose en las características propias del conjunto de datos. La clasificación no supervisada, también conocida como análisis de cúmulos o clustering, es un

término general para describir metodologías que han sido diseñadas para encontrar agrupamientos basados en similitudes entre los elementos del repositorio. El análisis de cúmulos ofrece una descripción generalizada de cada cúmulo, resultando en un entendimiento más profundo de las características del repositorio en general y provee de un comienzo para explorar futuras relaciones entre ellos.

Las técnicas de clustering son muy útiles en la minería de datos debido a la velocidad, confiabilidad y consistencia con la que organizan una gran cantidad de datos en diferentes grupos. A pesar de la disponibilidad de una gran cantidad de algoritmos de clustering en la literatura, éstos se basan en dos enfoques diferentes: clustering jerárquico y el clustering no jerarquico. En el primero, los datos se organizan en una secuencia anidada de grupos, el cual se puede desplegar a manera de árbol o dendograma.

Existen problemas asociados con el clustering. Uno de ellos es que los datos se agrupan en cúmulos con diferentes formas y tamaños. Otro problema es la resolución o granularidad. Este problema se vuelve más obvio cuando se trata de delinear una región que contiene una alta densidad de patrones comparados con el fondo. Por lo tanto, algunos autores definen a un cúmulo como aquel que consiste de una alta densidad relativa de puntos separados de otros cúmulos por una baja en la densidad de puntos. La selección de una medida apropiada de similitud para definir cúmulos permanece como un desafío en el área de análisis de cúmulos (Hardin & Chhieng, 2007).

II.2.5 Evaluación del conocimiento

Una vez construidos los modelos de predicción, es necesario evaluar la calidad de éstos. En las subsecciones siguientes se discutirán algunos de los métodos más utilizados

II.2.5.1 Evaluación mediante validación cruzada

El método de validación cruzada consiste en dividir el conjunto de entrenamiento en k subconjuntos disjuntos de similar cardinalidad. Entonces se entrena un modelo con la unión de k-1 subconjuntos y el subconjunto restante se utiliza para evaluar la exactitud. Se repite el procedimiento k veces procurando siempre no repetir los elementos de los k subconjuntos. Al final el error de la muestra se calcula como la media aritmética de los k errores de muestra parcial.Comúnmente se utiliza un valor de k = 10 (Witten et al., 2005).

II.2.5.2 Medidas de evaluación de un clasificador

Existen varias maneras de evaluar el desempeño del clasificador, siendo la tasa de éxito o exactitud la más común. La tasa de éxito es la relación entre las clasificaciones correctas y el total de clasificaciones. La tasa de errores es uno menos la tasa de éxitos.

Otra herramienta muy utilizada es el análisis de la matriz de confusión. Dadas *m* clases, una matriz de confusión es una matriz de tamaño *m x m*. Para una celda determinada CM_{ij} indica el número de ejemplares de la clase *i* que han sido clasificados en la clase *j*. Para que un clasificador tenga buena exactitud, la mayoría de los ejemplares deben estar representados en la diagonal de la matriz de confusión con el resto de las celdas quedando cercanas a cero.

En el caso de dos clases, se utilizan los términos de verdaderos positivos para referirse a los ejemplares que han sido clasificados como positivos correctamente, y los verdaderos negativos serán los ejemplares que han sido clasificados como negativos correctamente, como se muestra en la tabla 1.

Tabla 1. Matriz de confusión para las clases A y B

	A	**B**
A	Verdadero positivo (VP)	Falso negativo (FN)
B	Falso positivo (FP)	Verdadero negativo (VN)

Debido a que un clasificador pudiera estar reconociendo solamente los verdaderos positivos o los verdaderos negativos, entonces hace falta evaluar la manera en que un clasificador reconoce dichas clases, se utilizan las medidas conocidas como sensibilidad y especificidad.

La sensibilidad mide la proporción de verdaderos positivos del clasificador, mientras el la especificidad mide la proporción de verdaderos negativos.

Para calcular dichas medidas, se utilizan las siguientes expresiones:

$$Sensibilidad = \frac{VP}{VP + FP} \tag{24}$$

$$Especificidad = \frac{VN}{VP + FN} \tag{25}$$

También se acostrumbra a reportar el producto sensibilidad x especificidad (SE) (Witten et al., 2005).

II.3 Resumen

En el presente capítulo se presentó brevemente la teoría que conforma al proceso de descubrimiento de conocimiento en datos, los pasos que lo componen y después se profundiza más sobre el proceso de minería de datos y el reconocimiento de patrones.

Al final del capítulo se muestran las medidas de evaluación de los modelos predictivos obtenidos mediante los algoritmos de minería de datos.

En el siguiente capítulo se describe la teoría y los fundamentos relacionados con la enfermedad del cáncer de mama y se describe un caso de estudio que se realizó en una clínica con la finalidad de entender el proceso e identificar las fuentes de datos.

Si se quiere obtener más información sobre el tema del proceso de descubrimiento de conocimiento en datos y sobre los algoritmos de minería de datos, se recomienda revisar la bibliografía en la sección de referencias bibliográficas.

Capítulo III. Contexto médico y caso de estudio

III.1 Introducción

Para comprender mejor el presente trabajo de tesis, es necesario tener una idea general de la enfermedad del cáncer de mama e introducir algunos conceptos claves que serán útiles para comprender los estudios y análisis realizados y así poder interpretar los resultados.

En el presente capítulo, se presenta información referente al contexto médico del cáncer de mama, su definición, la forma de diagnosticarse y su tratamiento. La información que se reporta, está basada en las guías técnicas del IMSS para el cáncer de mama y en la guía del cáncer de mama de la American Cancer Society.(“Breast Cancer Detailed Guide,” n.d.) Después se describe el proceso de diagnóstico y tratamiento que se utiliza en una clínica de salud pública, de la ciudad de Ensenada Baja California, según el caso de estudio realizado para el presente trabajo de investigación.

III.2 El cáncer de mama

El cáncer es el crecimiento anormal e incontrolado de células, localizado en alguna parte específica del cuerpo. Este tipo de degeneración puede darse en varias partes del organismo. El cáncer comienza en las células, que constituyen los ladrillos del cuerpo. Normalmente, el cuerpo forma células nuevas a medida que se necesitan para reemplazar a las células envejecidas que mueren.

Algunas veces, este proceso no resulta ser el esperado; crecen células nuevas que no son necesarias y las células envejecidas no mueren cuando deberían. Estas células adicionales pueden formar una masa llamada tumor.

Los tumores pueden ser benignos o malignos. Los tumores benignos no son cáncer, mientras que los malignos sí lo son. Las células de los tumores malignos pueden invadir los tejidos cercanos; también pueden desprenderse y diseminarse

a otras partes del cuerpo.

La mayoría de los cánceres se nombran por su lugar de origen. Por ejemplo, el cáncer de pulmón comienza en el pulmón y el cáncer de seno comienza en el seno. La diseminación de un cáncer desde una parte del cuerpo a otra se denomina metástasis. Los síntomas y el tratamiento dependen del tipo de cáncer y de lo avanzada que esté la enfermedad. El tratamiento puede incluir la cirugía, radiación y/o quimioterapia.

El cáncer de mama es un tumor maligno que se origina en la glándula mamaria. Ésta enfermedad es más común en la mujer, aunque también puede ocurrir en hombres.

Existen varios tipos de cáncer de mama, algunos más comunes que otros y algunos tumores resultan ser una combinación de varios tipos.

- **Carcinoma ductal insitu**. Es el tipo más común de cáncer no invasor en la glándula mamaria. Eso quiere decir que las células están confinadas a los ductos pero no han atravesado sus paredes hacia el tejido mamario circundante.
- **Carcinoma ductal infiltrante**. Es la forma más común de cáncer de mama. En ésta forma de cáncer, el tumor se origina dentro de los ductos de leche de la glándula mamaria, el tumor rompe las paredes de los ductos e invade al tejido circundante. En este punto, puede invadir (metástasis) a través del sistema linfático y sanguíneo a otros órganos distantes. Aproximadamente cada 8 de 10 cánceres resultan ser carcinoma ductal infiltrante.
- **Carcinoma lobulillar infiltrante**. Este tipo comienza en los lobulillos, que son las glándulas que producen la leche materna. Al igual que el carcinoma ductal infiltrante, éste tipo puede pasar a otros órganos distantes a través del sistema linfático o sanguíneo. También es más difícil de detectar mediante mastografía.

- **Cáncer de mama inflamatorio**. Es un tipo muy raro de cáncer de mama que representa aproximadamente entre un 1% y un 3% de los reportes de cáncer de mama. Dado que no comienza por un tumor, es más difícil de detectar tempranamente y por lo tanto tiene más oportunidades de provocar una enfermedad metastásica.
- **Tumor phyllodes**. Este tipo de cáncer de mama ocurre raramente y comienza en el tejido conectivo del seno, a diferencia de los anteriores que ocurren dentro de la glándula mamaria.
- **Angiosarcoma.** Este tipo de cáncer comienza en las células que revisten a los vasos sanguíneos y los ductos linfáticos. Tiende a ser agresivo y a invadir rápidamente a otros tejidos.

Un factor de riesgo es cualquier condición que afecta las posibilidades de tener alguna enfermedad, como puede ser el cáncer. Los factores de riesgo para el cáncer de mama los podemos clasificar como intrínsecos y extrínsecos. Los primeros son aquellos que no se pueden alterar modificando el estilo de vida, como la edad y la genética, mientras que los factores extrínsecos pueden ser alterados cambiando el estilo de vida que lleva una persona.

Los siguientes son factores de riesgo de padecer cáncer de mama.

- Género. El simple hecho de ser mujer aumenta las posibilidades de que una persona padezca cáncer de mama. Esta enfermedad es 100 veces más común en la mujer que en el hombre.
- Edad. Las posibilidades de tener cáncer de mama aumentan con la edad.
- Genética. Tales como mutaciones en los genes BRCA1, BRCA2 y un historial familiar de cáncer de mama aumentan las posibilidades de padecer la enfermedad.
- Haber padecido cáncer de mama en el pasado.

- Tejido mamario denso. Las mujeres con un tejido mamario denso tienen más posibilidades de tener cáncer de mama. Además la alta densidad hace más difícil realizar un diagnóstico mediante mastografía.
- Condiciones benignas. Tales como la enfermedad fibroquística, ectasia ductal, infecciones, etcétera.
- Ciclos menstruales. Las mujeres que han tenido más ciclos menstruales, ya sea porque comenzaron a menstruar a una edad temprana y/o tuvieron su menopausia a una edad tardía, tienen un riesgo ligeramente más alto de padecer el cáncer de mama.
- Haber recibido radiación en el pecho.
- No haber tenido hijos. Las mujeres que no tuvieron hijos tienen un riesgo ligeramente más alto de padecer ésta enfermedad.
- Uso de anticonceptivos.
- Uso de remplazo hormonal después de la menopausia.
- Amamantar. Diversos estudios sugieren que una mujer que amamanta a sus hijos tiene un riesgo ligeramente menor de padecer cáncer de mama.
- Beber alcohol.
- Sobrepeso. Tener sobrepeso u obesidad aumenta el riesgo de padecer cáncer de mama, especialmente después de la menopausia.
- Actividad física. Existe evidencia que indica que la actividad física en forma de ejercicio reduce el riesgo de padecer la enfermedad.

III.1.1 Diagnóstico

Según la norma oficial mexicana vigente para el diagnóstico y tratamiento del cáncer de mama NOM-041-SSA2-2011, y la organización mundial de la salud (OMS), la mastografía y el ultrasonido permanecen como la regla de oro para realizar un tamizaje de las pacientes mayores de 40 años.

Desde 1992, el American College of Radiology (ACR) desarrolló un sistema llamado Breast Imagin Reporting and Data System (BI-RADS) (American College

of Radiology (ACR), 2003) con el objetivo de estandarizar la terminología y sistemática del informe mastográfico, además estandariza las lesiones estableciendo un grado de sospecha y asigna a cada una de ellas una recomendación sobre qué acción tomar en cada uno de los grados . En la actualidad se utiliza la 4ª edición del sistema BI-RADS que fue publicada en el año 2003 y que incluye secciones para ultrasonido y resonancia magnética.

El informe BI-RADS se reporta de la siguiente manera:

Tabla 2. Categorías BI-RADS.

CATEGORÍA BI-RADS	INTERPRETACIÓN	RECOMENDACIÓN
0	Insuficiente	Se requiere complementar con ultrasonido.
1	Normal	Control mastográfico de 1 a 2 años.
2	Hallazgo benigno	Control mastográfico a 1 año.
3	Hallazgo probablemente benigno	Se sugiere control mastográfico a 6 meses durante 2 años o biopsia.
4	Hallazgos sospechosos de malignidad en lesión.	Se sugiere biopsia.
5	Altamente sugestivo de malignidad	Se sugiere biopsia, acción inmediata.
6	Maligno	Confirmado previamente mediante estudio histopatológico.

Para el caso de hallazgos de nódulos o tumores, la clasificación BI-RADS se hace en base a la descripción morfológica descrita en la figura 6.

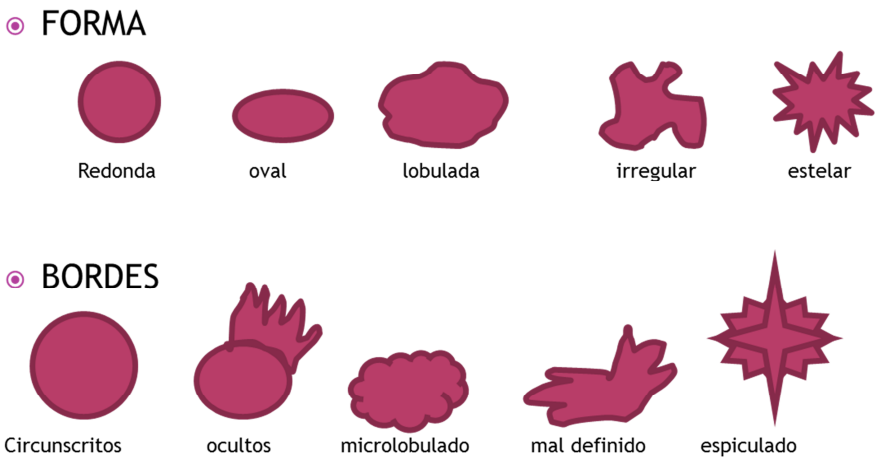


Figura 6. Descripción morfológica de tumores en mastografías.

En aquellas mastografías reportadas en categoría BI-RADS 4 y BI-RADS 5, la norma oficial vigente recomienda someter a la paciente a un procedimiento de biopsia escisional para determinar la naturaleza del tejido mediante un estudio histopatológico. El resultado del estudio revela, entre otras características de la muestra de tejido, si la lesión es benigna o maligna.

III.1.2 Tratamiento

Según la norma oficial mexicana vigente: NOM-041-SSA2-2011 (2011); la cirugía y la radioterapia permanecen como el estándar para el tratamiento del cáncer de mama, aunado a éstos, se encuentran varios y diversos agentes de quimioterapia, como aquellos basados en platino, las antraciclinas, los taxanos, etc. La hormonoterapia con Tamoxifeno y los inhibidores de aromatasa, y la inmunoterapia con Herceptin para los casos de cáncer de mama con el gen Her2/Neu amplificado. Con la cirugía es posible eliminar el tumor y ganglios linfáticos afectados por células cancerosas, después de la cirugía y en caso de requerirse se aplica la radioterapia para eliminar rastros microscópicos del tumor

dejados en el tejido sano, y en los ganglios linfáticos para evitar el crecimiento de nuevas tumoraciones en el sitio.

El análisis histopatológico del tejido extraído mediante cirugía, revela el sitio de la glándula mamaria en donde se originó el tumor, el grado de malignidad que se califica en escala del 1 al 3 siendo éste último el menos favorable, la dimensión del tumor en su diámetro mayor, el número de ganglios afectados, y además mediante otras técnicas del tipo inmunohistoquímicas, se analiza en la superficie de las células del tumor, la presencia de ciertas proteínas que pueden ser receptores hormonales para estrógeno y progesterona, o la sobreexpresión del gen Her2/Neu. La presencia o ausencia de dichas proteínas, le dan características particulares a cada caso, por ejemplo, a la presencia de receptores de estrógeno y/o progesterona se le conoce como subtipo **Luminal** e indica la existencia de un tumor que depende de la señal de dichas hormonas para reproducir sus células y crecer, por lo que las pacientes con este tipo de tumor se ven beneficiadas con terapia hormonal. Un diagnóstico inmunohistoquímico de **Her2/Neu positivo**, indica una amplificación en la expresión de dicho gen. Antes de la existencia de la inmunoterapia con Herceptin, un diagnóstico del tipo Her2/Neu positivo era desfavorable para la paciente. Por último, a la ausencia de todo receptor de estrógeno, progesterona y un diagnóstico negativo para Her2/Neu se le conoce como **Triple Negativo**. Para este subtipo de cáncer de mama no existe una terapia específica, aunque las pacientes se ven beneficiadas por la aplicación de quimioterapia.

La clasificación en los subtipos de cáncer mencionados anteriormente también ayuda a pronosticar el comportamiento del tumor, el transcurso de la enfermedad y a planear un tratamiento adecuado y personalizado para cada caso particular. Una meta importante en la terapia contra el cáncer de mama es diseñar un tratamiento personal que evite el retorno de la enfermedad y que minimice los efectos secundarios de los medicamentos.

En base a lo anterior, podemos definir a la recurrencia del cáncer de mama como la recurrencia de actividad tumoral en el sitio original de la cirugía, después de haber recibido radioterapia en caso de haberse requerido, terapia hormonal, inmunoterapia y/o quimioterapia en un plazo menor a cinco años de su diagnóstico inicial.

III.3 Estudio del proceso de diagnóstico y tratamiento

Se realizó un estudio del proceso del diagnóstico y tratamiento del cáncer de mama en una clínica de salud pública. El propósito del estudio fue para entender el proceso de toma de decisiones, identificar a los actores involucrados en el proceso y principalmente identificar las fuentes de datos que serían utilizadas en el proceso de minería de datos.

Para el estudio y modelado del proceso, se realizaron 8 entrevistas con médicos familiares, 2 entrevistas con el especialista en radiología, 2 entrevistas con el especialista en ginecología, 1 entrevista con un especialista en patología y 2 entrevistas con el oncólogo. Además se estudiaron los censos para cáncer de mama que se realizan en el área de epidemiología de dicha clínica, se revisaron expedientes de pacientes diagnosticadas con cáncer de mama y se realizó un estudio sombra en dicha clínica durante dos semanas para entender la consulta de pacientes en la especialidad de oncología.

III.1.3 Descripción textual del proceso

Existen varios escenarios en los cuales una paciente puede ser diagnosticada y tratada de cáncer de mama, dos de los más típicos son cuando la paciente detecta una lesión en su glándula mamaria mediante auto-exploración, el otro es a través del programa de detección oportuna del cáncer de mama (DOCMA) en el cual, personal entrenado hace actividades de promoción en la sala de espera a la población en general y educa a las mujeres en edad reproductiva a auto explorarse y a detectar cambios en su cuerpo.

Un tercer escenario es aquel en el que la paciente ha sido diagnosticada en lo privado, y acude la clínica para recibir tratamiento. En este caso, al igual que en todos, se inicia con una consulta en la unidad médico familiar, la diferencia es que la paciente es referida directamente al servicio de oncología para recibir el tratamiento adecuado.

III.1.3.1 Escenario 1. Paciente detecta lesión mediante autoexploración.

Empezaremos por describir el caso en el que la paciente ha detectado una lesión mediante auto-exploración y acude a la unidad médico familiar.

III.1.3.1.1Consulta con médico familiar

Durante la primera cita, el médico familiar escucha los síntomas que le comenta la paciente, revisa su expediente electrónico y hace una exploración de la lesión que le reporta la paciente y le hace preguntas sobre antecedentes familiares de cáncer de mama y de factores de riesgo asociados a éste.

Paso siguiente, se emite una orden de estudios radiológicos para mastografía si la paciente tiene más de 40 años, si tiene menos de 40, la solicitud de estudios radiológicos será para ultrasonido. En caso de que la paciente tenga antecedentes familiares directos para cáncer de mama, está indicada la mastografía de tamizaje 10 años antes de la edad en que apareció la enfermedad en dicho familiar.

En la solicitud de estudios radiológicos se le indica a la paciente la fecha y el lugar donde se efectuará el estudio.

III.1.3.1.2Consulta con radiólogo

El médico radiólogo efectúa la mastografía o el ultrasonido en las pacientes que fueron remitidas por las clínicas. Los resultados de los hallazgos se escriben en la solicitud de estudios radiológicos y se clasifican según la clasificación BIRADS de la siguiente manera:

- BI-RADS 0 Estudio insuficiente, se necesita reevaluar con otro estudio. No es posible determinar ninguna patología ni la ausencia.

- BI-RADS 1 Mastografía negativa a malignidad, sin ganglios o calcificaciones. 0% de posibilidades de cáncer.
- BI-RADS 2 Mamografía negativa a malignidad, pero con hallazgos benignos (ganglios intramamarios, calcificaciones mamarias, etc.)
- BI-RADS 3 Resultado con posible benignidad pero se requiere control a seis meses. Puede presentar nódulos circunscritos o algún grupo pequeño de calcificaciones puntiformes y redondeadas. 2.24% de posibilidades de cáncer.
- BI-RADS 4 Resultado dudoso de malignidad. Requiere una confirmación histopatológica. Consta de 3 grados de acuerdo con su porcentaje de malignidad que van del 3 al 94%
    - Baja sospecha de malignidad. 3 a 49%
    - Sospecha media de malignidad. 50 a 89%
    - Sospecha intermedia de malignidad. 90 a 94%
- BI-RADS 5 Alta sospecha de malignidad. Requiere biopsia para confirmar diagnóstico. Más de 95% de posibilidades de malignidad.
- BI-RADS 6 Malignidad confirmada mediante biopsia

Si el estudio resulta clasificado como BI-RADS 3 en adelante, los resultados se envían de regreso a la clínica para que la paciente sea contactada a la brevedad posible. El trabajador social es el responsable de contactar y dar aviso a la paciente sobre los resultados y su próxima cita.

Por otro lado, si el resultado es clasificado como BI-RADS 1 ó BI-RADS 2, la paciente acude con el médico familiar para valoración y se le indica que se efectúe otro estudio dentro de seis meses.

A las pacientes con resultado clasificado como BI-RADS 3 en adelante, se les agenda una cita a especialidad con el ginecólogo, en donde hará una valoración inicial sobre los resultados radiológicos y programará una biopsia para confirmar

los hallazgos. Para la cita, se emite el documento de cita a especialidades y se genera para archivar en el expediente una hoja de referencia – contrarreferencia.

Todos los hallazgos, así como el resumen clínico de cada paciente, se anotarán en el expediente electrónico, en el apartado de notas médicas.

III.1.3.1.3Consulta con ginecólogo.

En una consulta con el ginecólogo, el paciente llega con resultados de análisis de laboratorio y resultados radiológicos que indican una clasificación de BIRADS 3 en adelante. En algunos casos el ginecólogo vuelve a solicitar estudios radiológicos para confirmar resultados dudosos.

El especialista en ginecología tiene la capacidad de efectuar una biopsia por aspiración en el mismo consultorio ya sea en la primera consulta o en alguna subsecuente. La muestra de la biopsia se envía a patología junto con su respectiva solicitud de estudios histopatológicos.

Existen casos en los que la biopsia por aspiración no arroja resultados concluyentes, en este caso es necesario efectuar una biopsia abierta. Cuando el ginecólogo solicita una biopsia, ésta es programada como una cirugía general, en donde es necesario efectuar análisis preoperatorios y consulta pre-anestésica. El médico emitirá una solicitud de exámenes de laboratorio y hará uso del documento de notas médicas y prescripción para hacer un resumen clínico de la paciente y el tratamiento farmacológico que pudiera prescribir. Al final se le entrega al paciente la solicitud de exámenes de laboratorio, una receta en caso de requerirse y la cita a consulta pre-anestésica

III.1.3.1.4Consulta pre-anestésica.

La paciente acude a la consulta pre-anestésica con el especialista en anestesiología. En dicha consulta se hace una exploración física del paciente y se anotan los hallazgos en el formato de Consulta pre-anestésica y consentimiento informado. Después se hacen las siguientes preguntas sobre antecedentes:

- Alergias
- Tabaquismo
- Hemorragias
- Cirugías previas
- Anestesias
- Asma
- Toxicomanías
- Cardiopatía
- Hipertensión arterial
- Diabetes
- Hepatitis
- Tiroides
- Obesidad
- Enfermedad renal

Los resultados de los análisis preoperatorios son anotados en el formato mencionado anteriormente, así como los resultados de la valoración cardiológica, prueba de Goldman de riesgo anestésico y la técnica anestésica a utilizar.

Al final del formato se incluye una carta de consentimiento, en donde se explica al paciente los riesgos de someterse al procedimiento anestésico y se solicita su firma de consentimiento para efectuarlos.

El especialista en anestesiología hace uso del documento de notas médicas y prescripción para enterarse de la situación clínica de la paciente, así como para escribir sus observaciones y resultados de los análisis preoperatorios y además anota los resultados de las pruebas anestésicas.

III.1.3.1.5Procedimiento quirúrgico de biopsia

Previo al procedimiento, se le explica al paciente los riesgos que conllevan la cirugía y se le solicita firmar la carta de consentimiento para realizar intervenciones quirúrgicas y procedimientos invasivos y se recibe la solicitud de programación quirúrgica. El procedimiento se hace en el quirófano, se retira tejido de la mama y se envía a patología junto con la solicitud de exámenes de patología quirúrgica y citológica para hacer un estudio histopatológico. La paciente es mantenida en observación, se monitorean sus signos vitales, dieta, líquidos y todo es registrado en la hoja de registros clínicos de enfermería.

III.1.3.1.6Trabajo social

En la mayoría de los expedientes revisados, la paciente fue referida a trabajo social inmediatamente después de haber sido intervenida quirúrgicamente. El deber del trabajador social es dar apoyo psicológico y social al paciente, en las hojas de trabajo social se apuntan las observaciones que se obtienen de la entrevista con la paciente, sus hábitos alimenticios, higiénicos, historia personal, familiares y se discuten cambios de hábitos que se deben de hacer para tener una mejor recuperación.

III.1.3.1.7Estudios Histopatológicos

Los estudios histopatológicos son efectuados en la clínica principal. Para ello se envía la muestra de tejido o la glándula mamaria sumergida en formol al laboratorio de patología. Junto con la muestra se envía una solicitud de exámenes de patología quirúrgica y citología, en el caso de que procedan de un paciente con sospecha de malignidad, la solicitud deberá venir debidamente clasificada como tal para asignarle la prioridad más alta en la atención.

La muestra de tejido debe permanecer como mínimo 24 horas en la solución de formol, para luego ser rebanada en láminas para poder ser observadas bajo el microscopio. En promedio se procesan 25 muestras al día.

Los hallazgos y anotaciones son escritos en la solicitud de exámenes patológicos y citológicos y son enviados al archivo.

Los resultados de la muestra pueden ser positivos, en caso de que se determine presencia de cáncer en la muestra, negativos en caso contrario. En caso de que se haya confirmado la presencia de cáncer (malignidad) se envía la misma muestra a un laboratorio externo (aprovechando laminillas y tejido extraído mediante biopsia) para determinar con exactitud el tipo de cáncer del que se trata. Al procedimiento anterior se le denomina estudio inmunohistoquímico y sirve para determinar cuál será el mejor tratamiento para la paciente. Los resultados del estudio son comunicados al servicio de patología por medio de correo electrónico.

Existen casos en los que un paciente no haya ido a recoger los resultados de sus estudios y resulta que éstos indican presencia de cáncer. En estos casos se turna el expediente al trabajador social, quien contactará por cualquier medio posible al paciente y enviará una ambulancia por él en caso de que no le sea posible trasladarse por sus propios medios. El trabajador social hará del conocimiento del paciente los resultados de los estudios y las opciones con las que cuenta.

III.1.3.1.8Consulta con oncólogo

La paciente llega a consulta con el especialista en oncología una vez que se ha determinado que la lesión ha sido provocada por cáncer. Se deberán tener los resultados histopatológicos, en la solicitud de exámenes de patología quirúrgica y citología e inmunohistoquímicos, diagnóstico patológico y por correo electrónico, y el especialista decidirá de acuerdo a éstos, cuál será el tratamiento adecuado para el paciente. El también podrá escribir en las notas médicas y prescripción sus observaciones y el tratamiento que asignará a la paciente.

En ocasiones el especialista en oncología solicitará hacer estudios adicionales (Ultrasonido, gama-gramma óseo, etc.) para determinar si ha ocurrido metástasis. Dichos estudios podrán ser subrogados a otros laboratorios si no se cuenta con el equipo para elaborarlos, para ello se llenará el formato de solicitud de servicios

subrogados y los resultados se recibirán por escrito por medio de la carta de diagnóstico.

Durante la evolución del cuadro clínico de la paciente, es el especialista en oncología quien estará al pendiente de los cambios que ocurran en la salud de la paciente y anotará sus observaciones, evolución, tratamiento y cambios durante el tratamiento en las notas médicas y prescripción.

III.1.3.2 Escenario 2. Se detecta lesión durante consulta con el médico.

En este escenario, la paciente ha acudido a consulta por un motivo diferente a aquel donde se detecta una lesión por autoexploración. La consulta puede ser con el médico familiar o con el especialista en ginecología, o enfermera en si se trata del programa DOCMA, quien hará una revisión de rutina en la glándula mamaria de la paciente. Si se detecta una lesión en la glándula mamaria, entonces se lo hará saber a la paciente y le solicitará estudios radiológicos de mastografía o ultrasonido según la edad de la paciente y a continuación se seguirá el mismo procedimiento que en el escenario anterior a partir de que se obtienen los resultados y la clasificación BI-RADS del estudio radiológico.

III.1.3.3 Escenario 3. La paciente fue diagnosticada en lo privado.

En este caso, la paciente fue diagnosticada en algún hospital o consultorio privado, llega a la clínica con los resultados histopatológicos e inmunohistoquímicos para recibir tratamiento. El médico familiar revisa los resultados y el diagnóstico y la deriva al servicio de oncología para recibir el tratamiento adecuado a su padecimiento.

Según lo visto en los expedientes médicos, el escenario 3 implica el menor tiempo para que una paciente reciba tratamiento una vez que ésta ha detectado lesiones, generalmente tres semanas. En los demás escenarios el tiempo promedio ha sido de tres meses.

III.1.4 Modelado del proceso

Una manera fácil de representar un proceso para comprenderlo y darlo a conocer es a través de diagramas. Existen varias técnicas para diagramar procesos aunque los que se utilizaron en éste caso son la gráfica rica, desarrollada por Peter Checkland como parte de la metodología Soft Systems (Monk & Howard, 1998), el diagrama de Rol Actividad (RAD) (Ould, 1995) y el diagrama Integration Definition for Function Modeling (IDEF0) (PUBS, 1993).

La gráfica rica le da al lector una visión global e instantánea de un proceso, además de que se puede definir a diferentes niveles de detalle. A los responsables de los procesos se les representa como un ícono de rol, y se puede indicar su interacción con otros roles, su propósito en el proceso, conflictos, miedos, deseos e inquietudes. En la tabla 3 se muestra la simbología de una gráfica rica.

Tabla 3. Simbología de una gráfica rica

	Acciones o preocupaciones
	Rol
	Actividad
	Interacciones
	Entradas y salidas
	Artefactos

Acciones o preocupaciones. Se representa como una nube de pensamiento. Sirve para representar o capturar la motivación o preocupación que lleva al agente a formar parte del sistema.

Roles. Representan a los agentes que participan en el proceso.

Actividad. Son representaciones de tareas o acciones de y entre roles. Por medio de ellas se establecen las interacciones.

Interacciones. Se representan mediante líneas simples. Indican alguna interacción entre un rol o varios roles y alguna actividad.

Entradas y salidas. Se representan por flechas. Indican si un artefacto es entrada o salida de alguna actividad, o ambos.

Artefactos. Entrada o salida de una actividad.

Ahora bien, si ponemos en práctica la simbología y las normas para modelar un proceso de acuerdo a la gráfica rica, nos quedan los diagramas del proceso como se muestran en las figuras 7, 8, 9, 10, 11.

En la figura 7 se muestra la gráfica rica general del escenario 1, que es en el que una paciente acude a la unidad médica familiar porque ha detectado una lesión y en la figura 8 se muestra una gráfica rica del escenario cuando la paciente es diagnosticada en lo particular y acude al sector público a recibir tratamiento. Para ambos escenarios las fuentes de datos las podemos ubicar en la actividad de radiología, como un artefacto de salida que es el radiodiagnóstico. Y en la actividad de “Realizar estudio histopatológico” se obtiene el diagnóstico patológico final, que se puede cruzar con los datos obtenidos de radiología. De esta manera se puede saber cuáles registros clasificados como BI-RADS 3 a BI-RADS 5 tienen diagnóstico histopatológico.

En la figura 9 se muestran las actividades de ginecología con respecto al proceso de diagnóstico y tratamiento del cáncer de mama.

En las figuras 10 y 11 se modelan las actividades que se efectúan en el área de especialidad. La figura 10 muestra el proceso de consulta y la figura 11 detalla el proceso de una paciente acudiendo a recibir tratamiento. En la parte o actividad llamada “acudir a consulta con oncólogo” se tiene como artefacto de entrada y salida al cuadro: “Notas médicas y prescripción”. En dicho artefacto se detalla la

información necesaria para realizar el análisis sobre recurrencia del cáncer de mama.

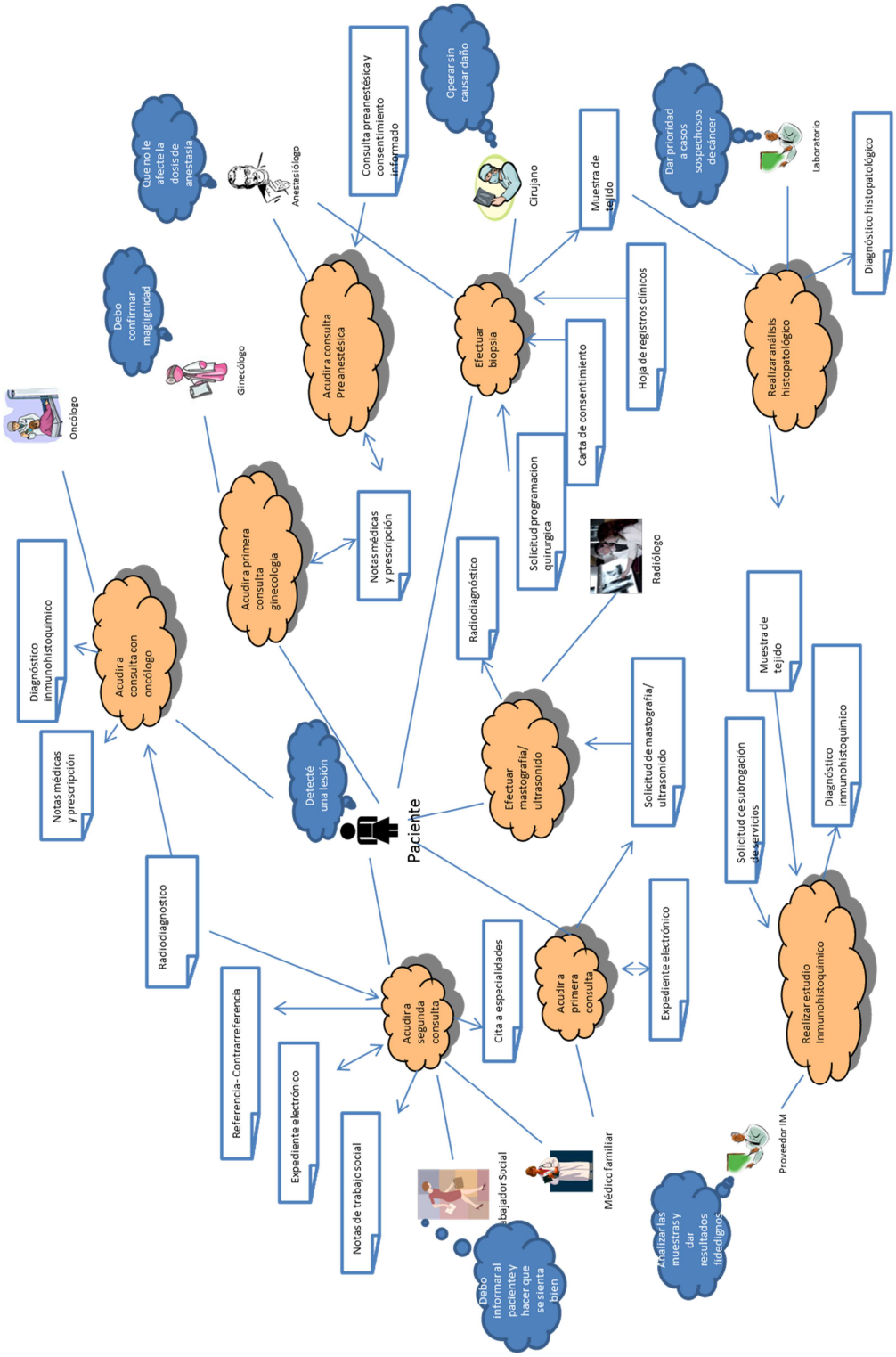


Figura 7. Proceso general de diagnóstico y tratamiento del cáncer de mama

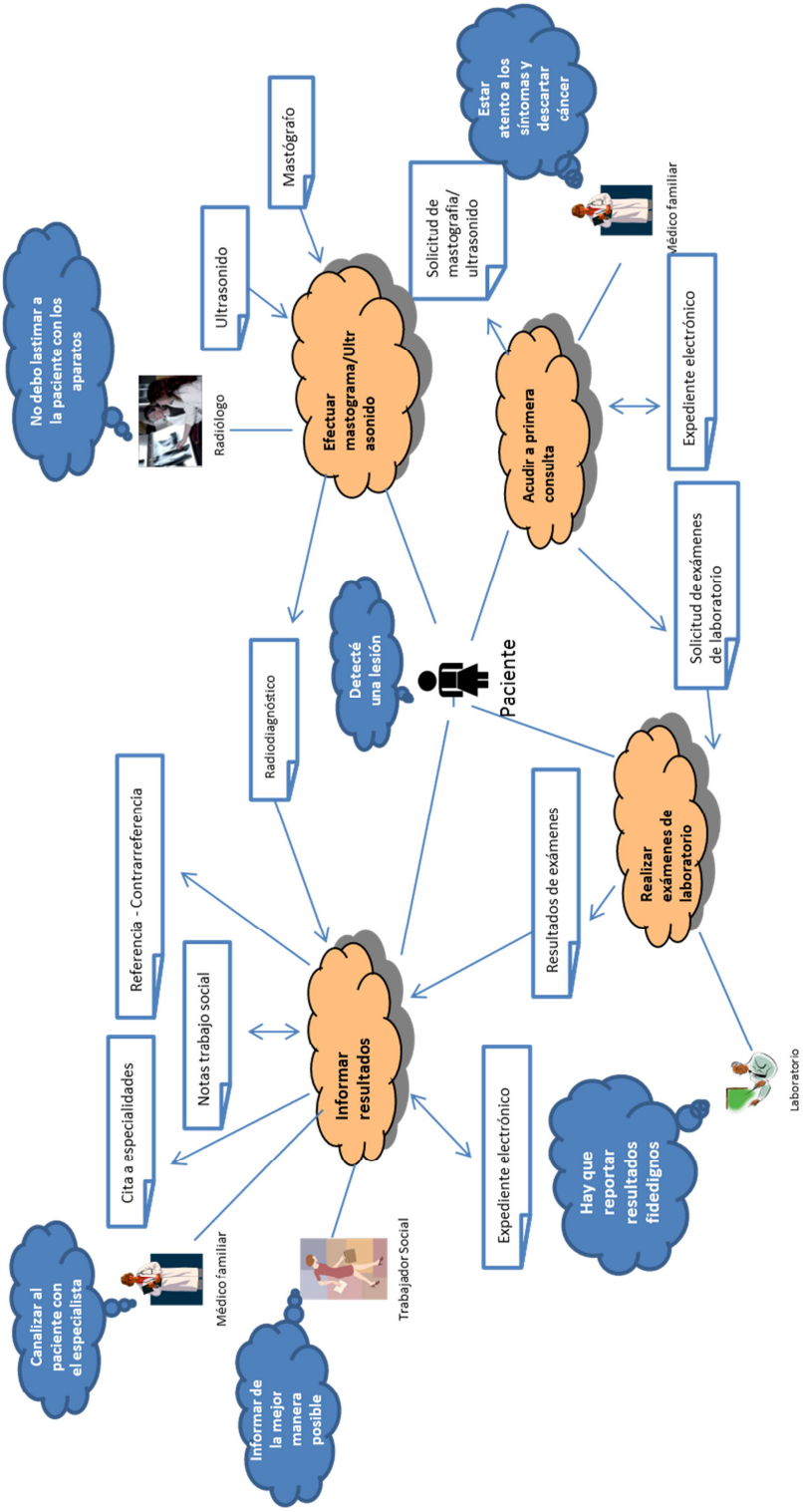


Figura 8. Proceso de diagnóstico del cáncer de mama cuando la paciente acude por detectar una lesión.

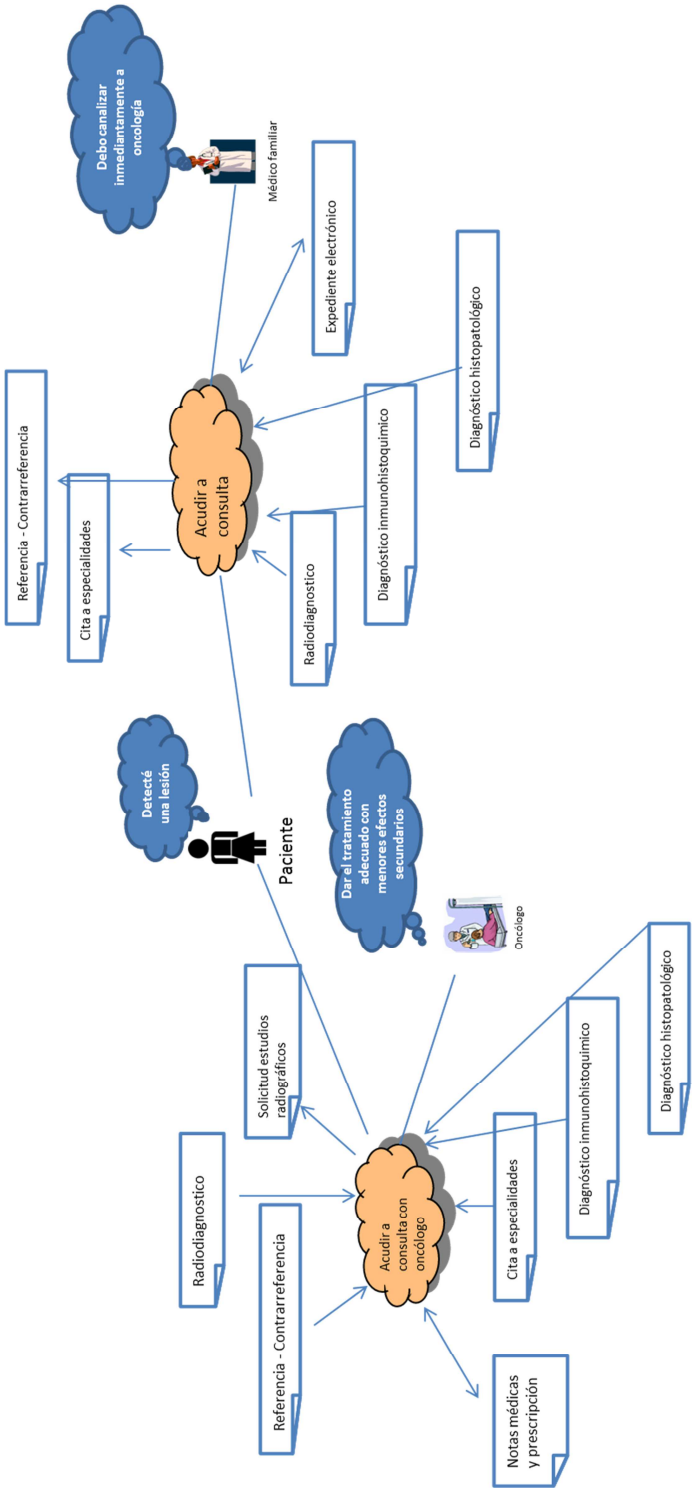


Figura 9. Proceso de diagnóstico del cáncer de mama. Proceso que ocurre cuando la paciente es diagnosticada en una clínica particular.

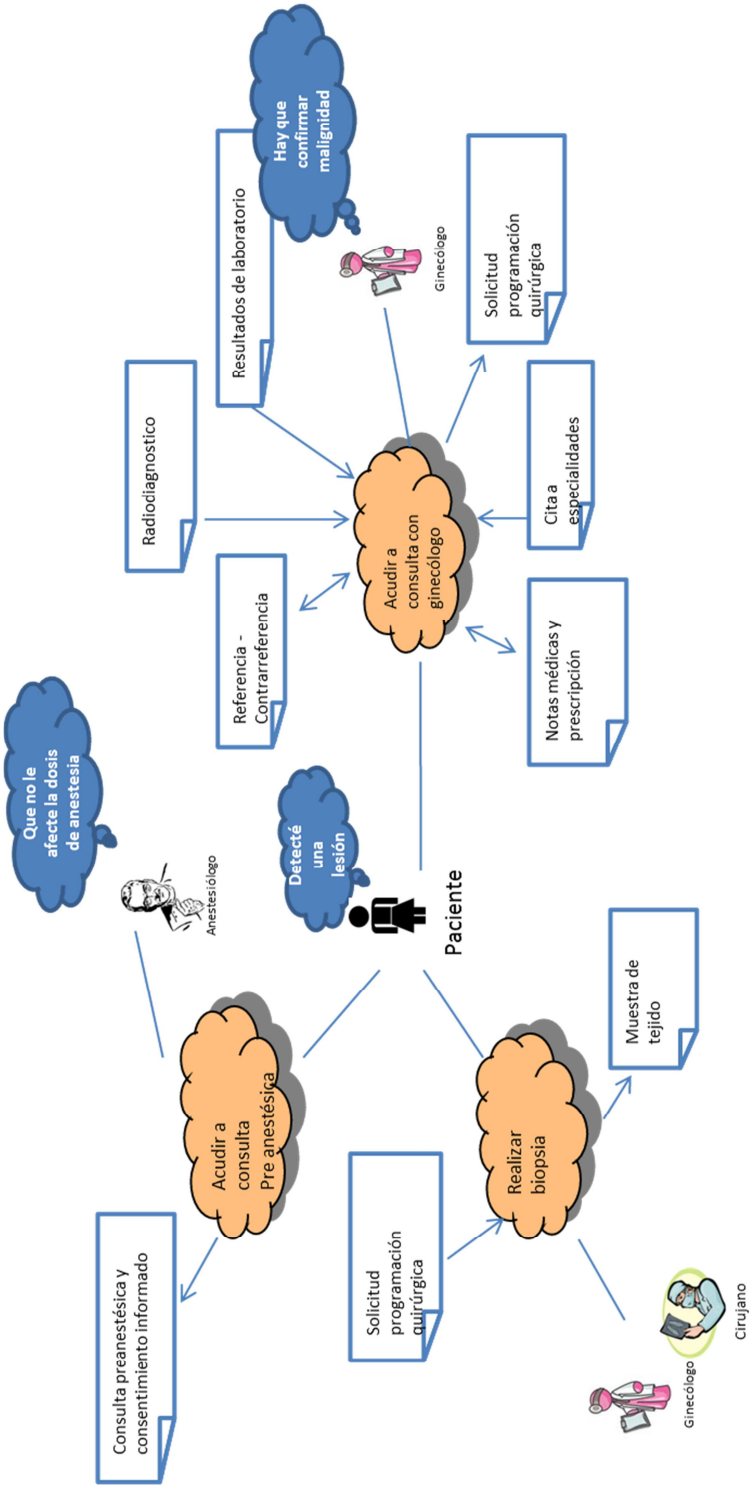


Figura 10. Proceso de diagnóstico del cáncer de mama. Actividades seguidas en el subproceso de ginecología.

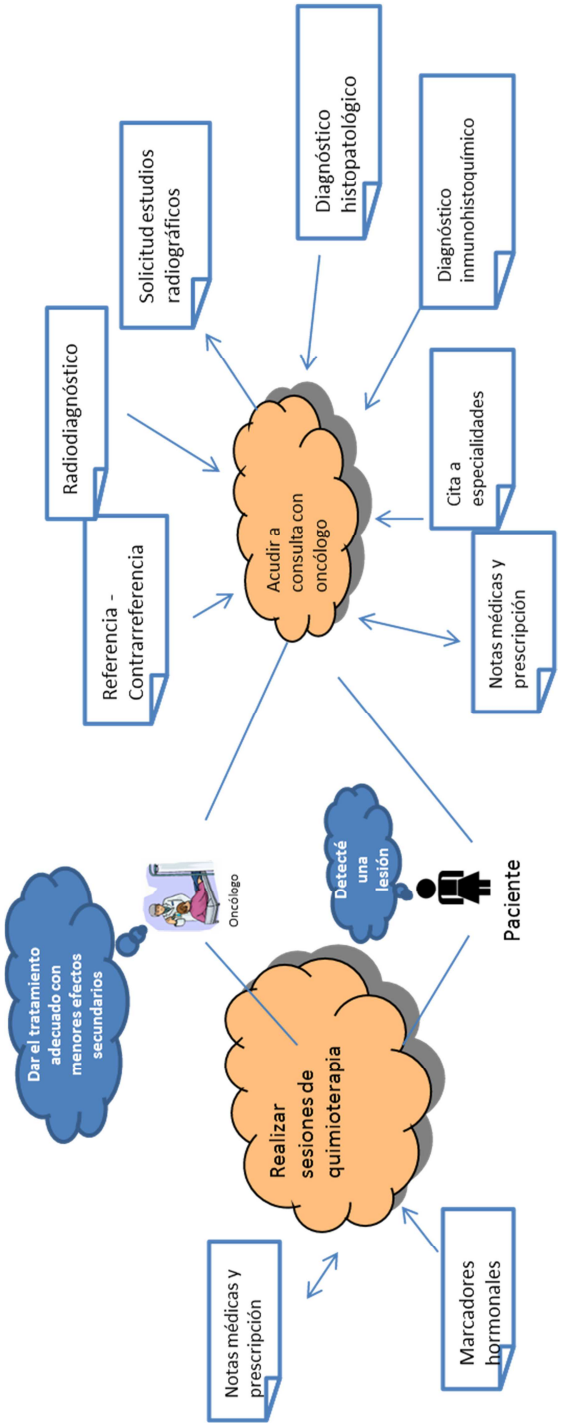


Figura 11. Proceso de tratamiento del cáncer de mama. Proceso de acudir a consulta para revisión.

III.1.4.1 Diagrama de Integración Definida (IDEF)

Los Diagramas de Integración Definida, IDEF por sus siglas en inglés (PUBS, 1993) representan de manera estructurada y jerárquica las actividades que conforman un proceso y los objetivos o datos que apoyan la interacción de esas actividades.

Permiten modelar gráficamente procesos a cualquier nivel de detalle y está basado en un estándar con especificaciones precisas y rigurosas.

Mediante estos diagramas se obtiene una vista estratégica de un proceso, y de esta forma se facilita el análisis del proceso para identificar puntos de mejora.

En la figura 12 se muestra la forma de representar o modelar un proceso mediante IDEF.

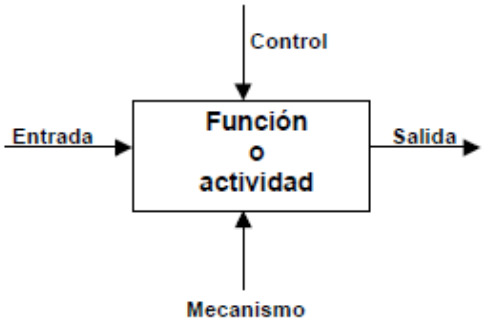


Figura 12. Representacion de una actividad en IDEF

Entradas:

- Material o información consumida o transformada por una actividad para producir “salidas”

- Asociadas al lado izquierdo de la actividad (caja)

- Una actividad puede no tener entradas

Salidas:

- Objetos producidos por la actividad o proceso

- Asociadas al lado derecho de la actividad

Control:

- Objetos que gobiernan o regulan cómo, cuándo y si una actividad se ejecuta o no. Están asociados al lado superior de la actividad.

Mecanismos:

- Recursos necesarios para ejecutar un proceso.
- Asociados al lado inferior de la actividad.
- Una actividad puede no tener mecanismos.

El diagrama IDEF agrupa las actividades jerárquicamente de manera que se puede hacer un análisis tan específico como sea requerido.

En la figura 13, vemos el diagrama general del proceso de detección y tratamiento del cáncer de mama, con sus entradas, salidas y controles, como las normas oficiales y guías técnicas. En la figura 14 podemos apreciar a una granularidad menor el proceso de detección, mostrando los principales subprocesos involucrados, que son medicina familiar, ginecología y oncología.

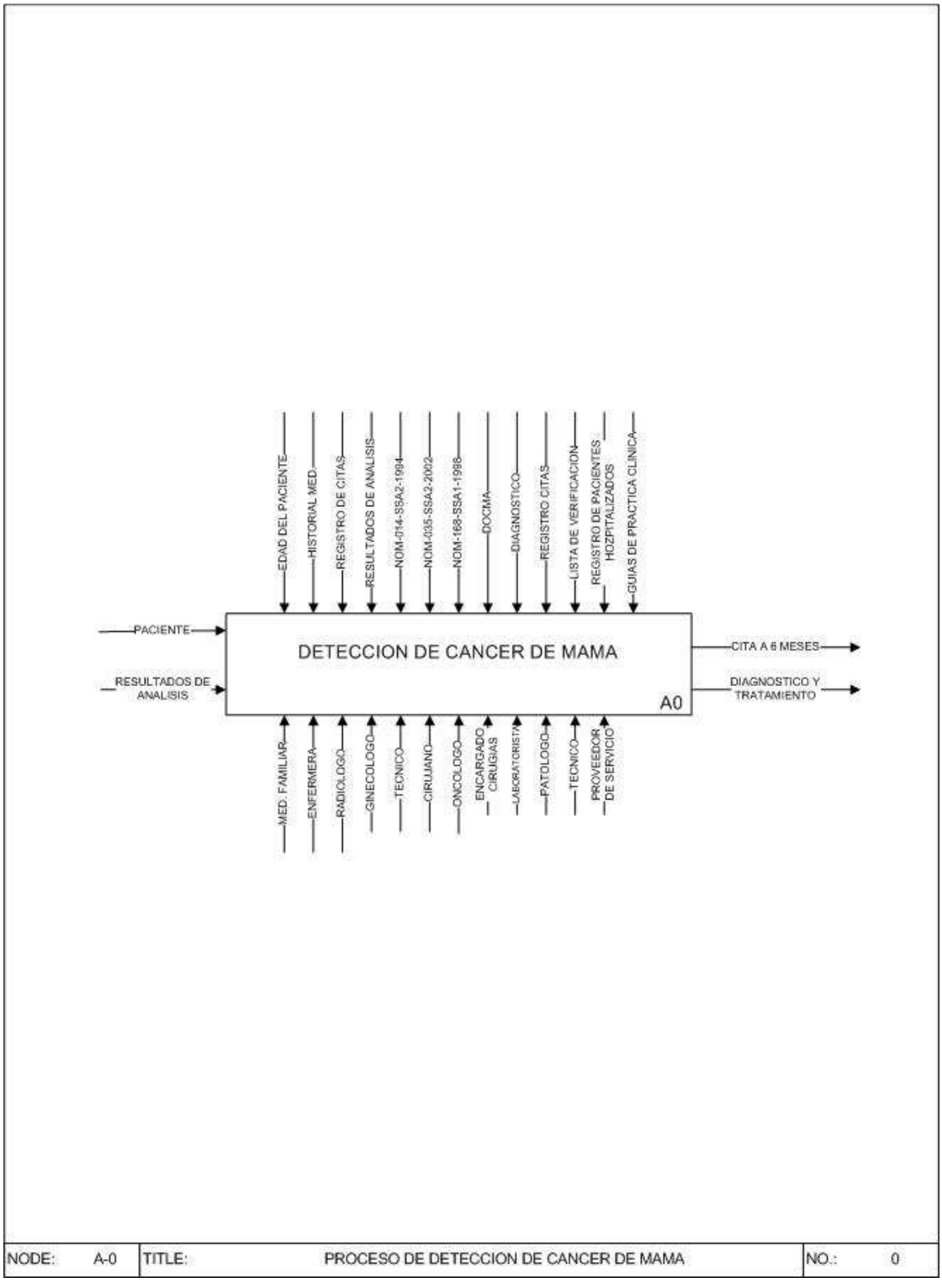


Figura 13. Diagrame general del proceso de detección del cáncer de mama.

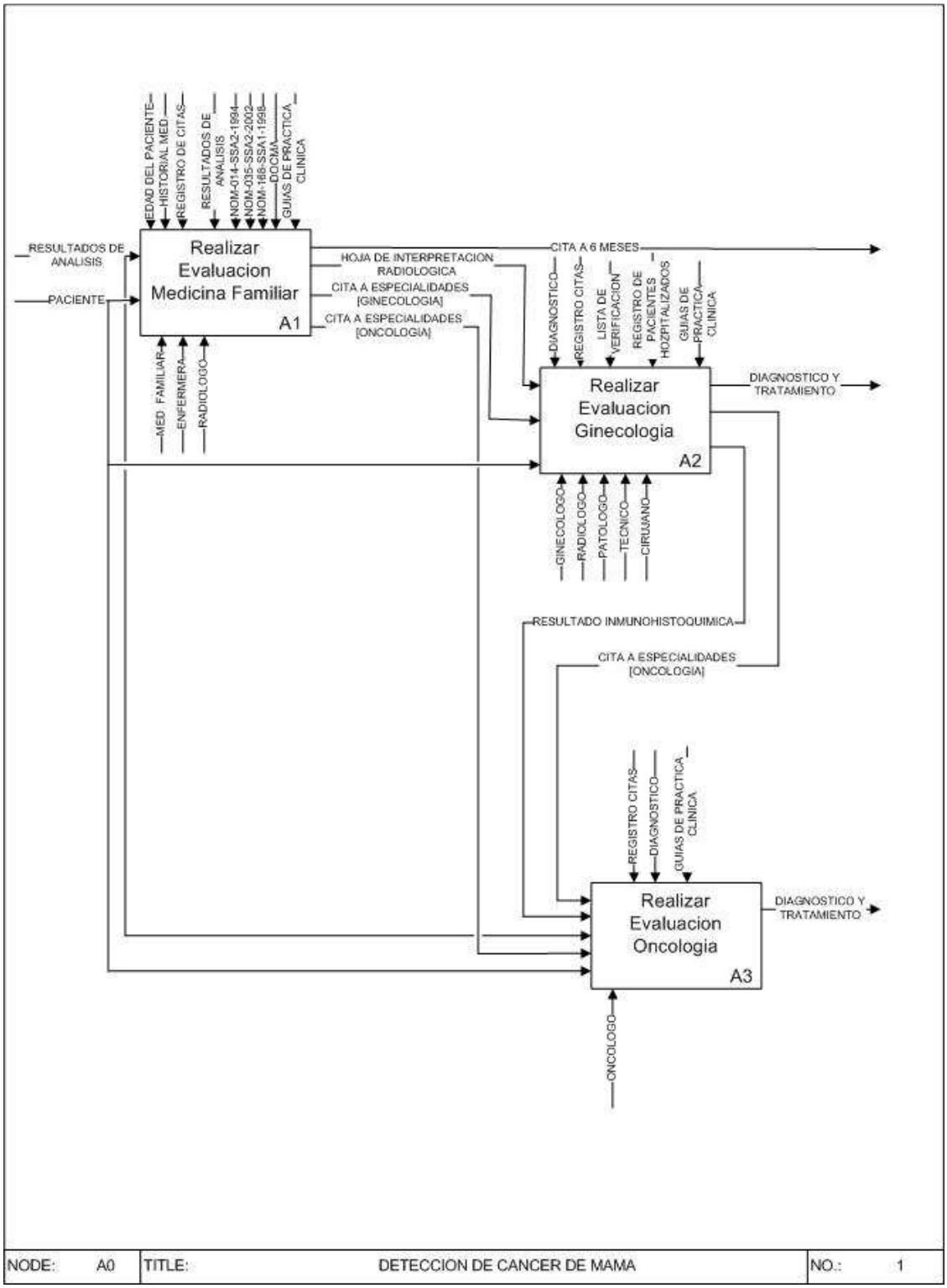


Figura 14. Diagrama del proceso de diagnóstico del cáncer de mama.

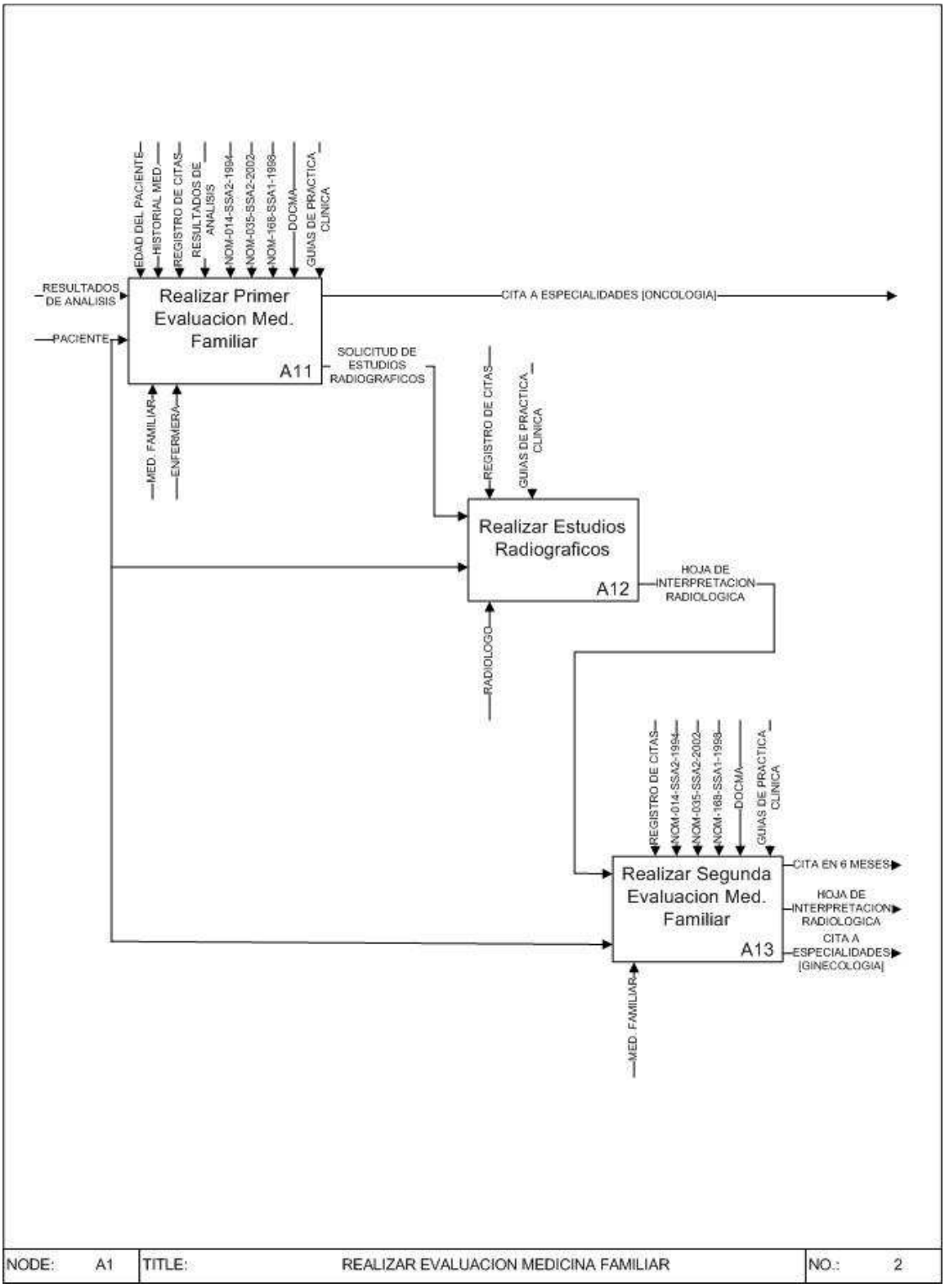


Figura 15. Vista del proceso de medicina familiar

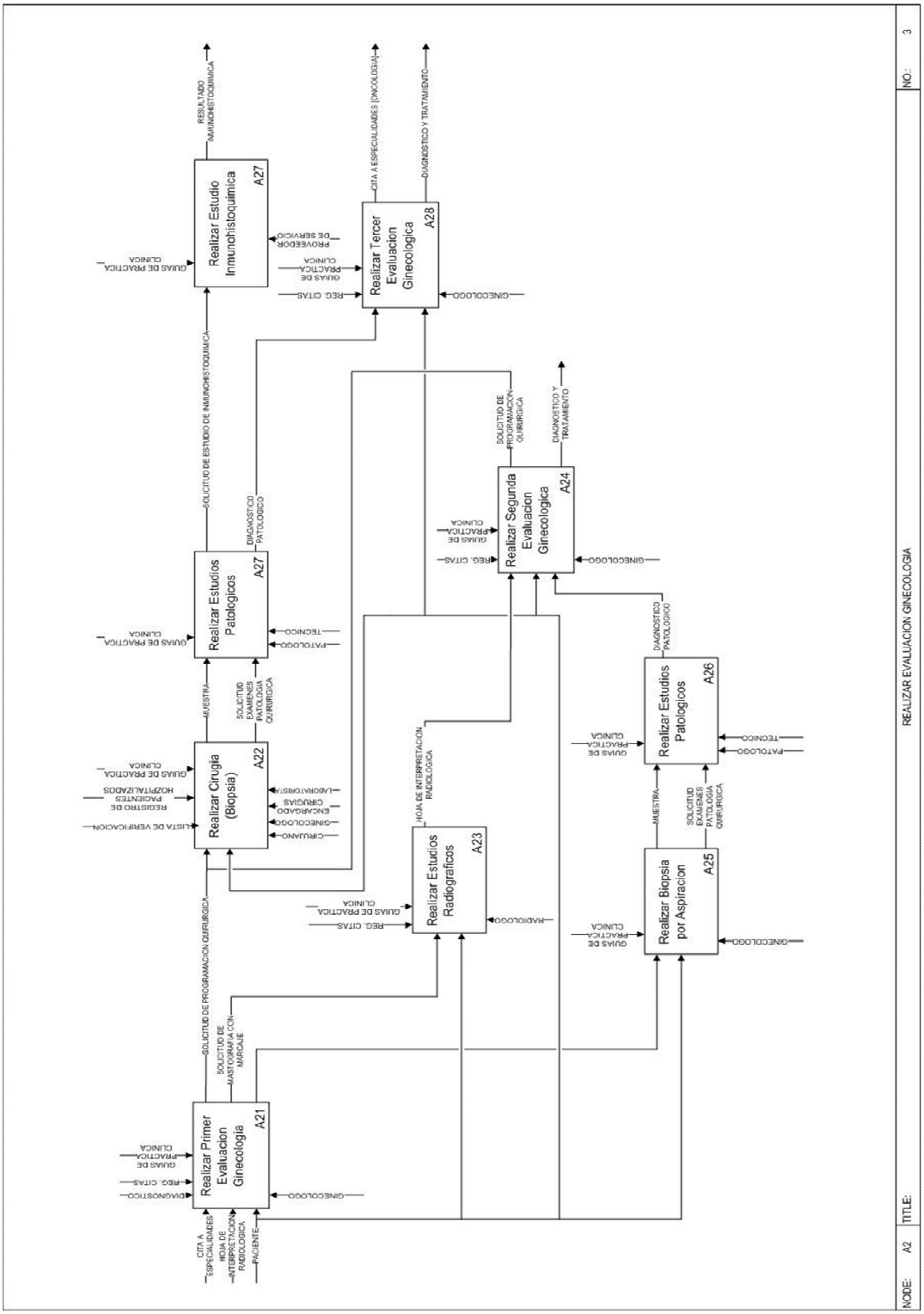


Figura 16. Vista del proceso de ginecología

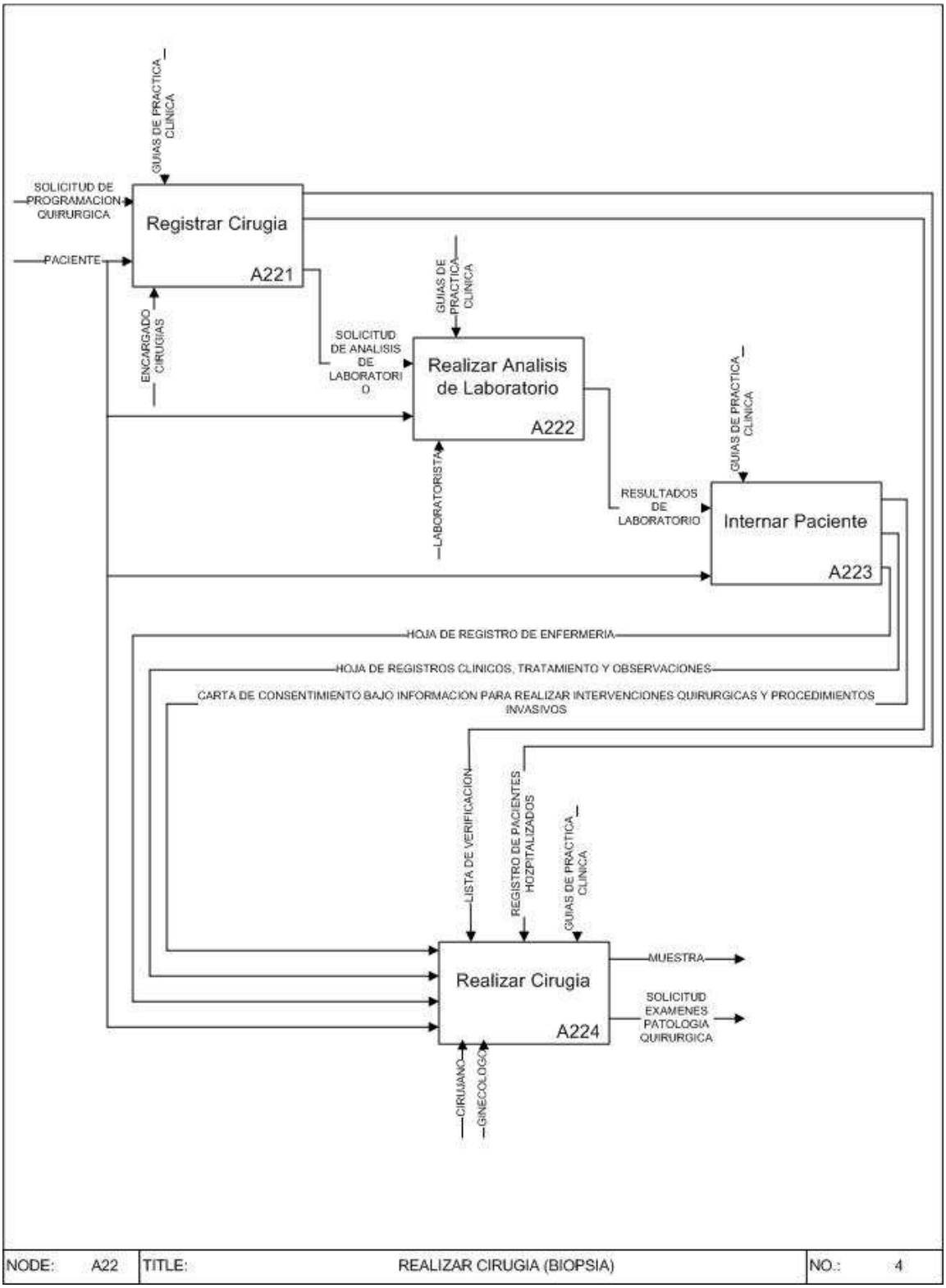


Figura 17. Vista del proceso de cirugía (biopsia)

III.4 Resumen

En este capítulo se presenta información relevante de la enfermedad del cáncer de mama. En primer lugar se mencionan la teoría, normas y el contexto médico que rodea a la enfermedad, su descripción, los factores de riesgo, el diagnóstico, el tratamiento y algunas definiciones que serán útiles para comprender más profundamente el trabajo de investigación que se realizó en el presente trabajo de tesis.

En la segunda parte se describe el proceso en particular que se sigue en la clínica VIII del Instituto Mexicano del Seguro Social de la ciudad de Ensenada Baja California, según el caso de estudio que se reporta en éste capítulo. Se presenta una descripción textual del proceso, se describe a los participantes del proceso, los documentos y datos que se utilizan y por último se presenta un modelado gráfico de dicho proceso, cuyo propósito es la comprensión del proceso de toma de decisión clínica y la identificación de fuentes de datos que se utilizaron para el proceso de minería de datos que se discute en los capítulos subsecuentes.

Los modelos obtenidos de éste paso de la investigación, también serán útiles para difundir o poner en práctica los conocimientos descubiertos del proceso de minería de datos, en las distintas etapas del proceso de toma de decisión clínica.

Capítulo IV. Proceso de descubrimiento de conocimiento sobre registros de mastografías

En el presente capítulo se presenta el proceso de minería de datos aplicado a uno de los repositorios de datos obtenidos del caso de estudio presentado en el capítulo III. El proceso comienza con la extracción de la información relevante a partir de datos crudos obtenidos en su mayoría a manera de redacción. Después sigue la transformación de los datos y la codificación en valores enteros de aquellas variables del tipo nominal. A continuación se aplicaron y se comparan dos métodos para reducción de variables con el fin de obtener un modelo más preciso. Por último se hizo un análisis exploratorio con los algoritmos de minería de datos y se evaluaron los resultados de los modelos obtenidos.

IV.1 Antecedentes

De acuerdo al estudio realizado en hospitales mexicanos, (Bright et al., 2011) un diagnóstico tardío es la principal causa de muerte por la enfermedad. Las causas de un diagnóstico tardío se deben principalmente a dos motivos, por un lado están los motivos socioculturales que provocan que una mujer acuda al médico mucho tiempo después de que aparecen los primeros síntomas del cáncer, después se encuentran aquellos que por falta de sistemas informáticos y de equipamiento médico, retrasan el diagnóstico a pesar de que la paciente ya ha acudido con su médico y de que existen indicios o sospechas de la existencia de la enfermedad.

En aquellas mastografías reportadas en categoría BI-RADS 3, BI-RADS 4 y BI-RADS 5, la norma oficial vigente recomienda someter a la paciente a un procedimiento de biopsia escisional para determinar si el tumor es benigno o maligno.

En base a lo explicado anteriormente, en el presente trabajo de investigación se reporta el estudio de minería de datos aplicado sobre la información registrada en el área de radiología con la finalidad de clasificarla en maligno o benigno, según el resto de los datos con los que se cuenta.

IV.2 Herramientas y materiales

Para la implementación de los algoritmos de minería de datos, se utilizó el programa Weka (Hall et al., 2009), que consiste en una colección de algoritmos de máquinas de aprendizaje y herramientas para el procesamiento de datos.

Para el análisis de componentes principales se utilizó el software estadístico Minitab, y para la aplicación del algoritmo Elastic Net (Friedman, 2010) se utilizó la implementación llamada Glmnet, para R y Matlab.

Los datos crudos se obtuvieron del caso de estudio presentado en el capítulo III. Éstos se generan en dos etapas separadas del proceso. En la primera etapa se obtienen los datos que se capturan en radiología, en éstos se incluyen: edad, forma del tumor, tipo de bordes del tumor, dimensiones, presencia de micro calcificaciones, presencia de macro calcificaciones, cuadrante de la mama en donde se reporta el hallazgo, lateralidad de la mama, densidad, clasificación BI-RADS de la mastografía. En la segunda etapa se reporta el diagnóstico histopatológico realizado sobre tumores que fueron clasificados como BI-RADS 3 a BI-RADS 5.

IV.3 Extracción de datos

Los datos obtenidos de radiología venían sin procesar, de manera cruda y en forma de redacción. En las figuras 18 y 19 se muestran unos ejemplos de cómo venían los datos en los registros. Para el caso de mastografías se tenían archivos de Word con los reportes a manera de redacción. Los reportes de patología no eran muy diferentes, aunque la redacción se efectuaba en registros de Access, no tenían ningún tipo de estructura normalizada. Entonces se procedió a reescribir los registros en una cuadrícula de Excel, a partir de los registros redactados con la ayuda del encargado de epidemiología y de radiología de la clínica en donde se realizó el caso de estudio. El propósito es separar a manera de matriz todos los datos que se utilizarían para realizar el resto del proceso de minería de datos.

Al final se obtuvo una matriz o cuadrícula con los datos organizados en columas para las variables y renglones para cada registro, en la tabla 4 se muestran las variables o las columnas que conforman a la matriz resultante.

EDAD: 75 AÑOS

MASTOGRAFIA

REPORTE:

Se realizó mastografía bilateral con técnica de baja radiación, tomando las proyecciones habituales, visualizando lo siguiente:

Simetría mamaria en forma, volumen y densidad
La estructura mamaria es de tipo mixto, conformada predominantemente por tejido fibroglandular hacia cuadrantes superiores y externos, el resto presenta tejido de sustitución grasa.
Se observan calcificaciones puntiformes aisladas, en "cielo estrellado", de forma bilatera, algunas agrupadas en acinosl de predominio derecho.
Se observan calcificaciones gruesa en "doble riel" de tipo vascular hacia cuadrantes superiores e inferiores en ambas mamas.
No se visualizan zonas de distorsión de la arquitectura.
Las regiones axilares sin evidencia de alteraciones.
Nota: No se cuenta con estudios previos para hacer analisis comparativo de los hallazgos.

HALLAZGOS COMPATIBLES CON:

1. MAMAS CON TEJIDO FIBROGLANDULAR DISPERSO.
2. CALCIFICACIONES ABUNDANTES PUNTIFORMES AGRUPADAS DE PREDOMINO EN MAMA DERECHA, QUE REQUIEREN VALORACION HISTOPATOLOGICA.
3. BIRADS 4

Figura 18. Ejemplo de registro mastográfico

EDAD: 57

ESPECIMEN:
MAMA IZQUIERDA

DESCRIPCIÓN MICROSCÓPICA:
EN CORTES DE SECCION SE APRECIAN DUCTOS DILATADOS REVESTIDOS POR EPITELOCUBICA, RODEADOS POR ELEMENTOS INFLAMATORIOS, EL ESTOMA CON FIBROSIS, PROLIFERACION DE FIBRAS COLAGENAS (REPARACION) MACROFAGO Y FIBROBLASTOS, ASI COMO ELEMENTOS INFLAMATRIOS Y CELULAS GIGANTES DE TIPO CUERPO EXTRAÑO. PIEL Y PEZON SIN ALTERACIONES, LOS GANGLIOS LINFATICOS EN NUMERO DE 15, UNO DE ELLAS CON INVACION TUMORAL POR ESTRUCTURAS DUCTALES REVESTIDAS POR EPITELO CUBICO CON CELULAS ANAPLASICAS.

DESCRIPCIÓN MACROSCÓPICA:
IDX: BIOPSIA EXCISIONAL DE TUMOR MAMARIO IZQUIERDO.
- CARCINOMA DUCTAL INFILTRANTE CON GRADO NUCLEAR DE 8 CON LA CLASIFICACION DE BLOOM- RICHARDSON, GRUPO III, MIDE 3 X 2.8 CM, ASOCIADO A ENFERMEDAD FIBROQUISTICA.
- CARCINOMA DUCTAL IN SITU EN EL TEJIDO ADYACENTE Y CALCIFICACIONES.

Figura 19. Ejemplo de registro de reporte histopatológico.

Tabla 4. Datos del repositorio de mastografías.

Característica	Tipo	Representación
Edad	Numérico	Numérico continuo [0 <= Edad <= 100]
Hallazgo	Nominal	{Masa, Masa y calcificaciones, Distorsión, Inflamatorio}
Forma	Nominal	{Redondo, Ovalado, Irregular, Estelar}
Bordes	Nominal	{Circunscrito, Microlobulado, Oculto, Irregular, Mal definido}
Dimensión	Numérico	Numérico continuo [0 <= Dimension <= 20]
Micro calcificaciones agrupadas	Binario	{Si, No}
Micro calcificaciones aisladas	Binario	{Si, No}
Macro calcificaciones agrupadas	Binario	{Si, No}
Macro calcificaciones aisladas	Binario	{Si, No}
Cuadrante	Nominal	{Superior externo, Superior interno, Inferior externo, Inferior interno, Retroareolar}
Mama	Binario	{Izquierda, Derecha}
Densidad	Nominal	{Fibroglandular disperso, Fibroglandular residual, Heterogeneamente denso}
BI-RADS	Numérico	{1,2,3,4,5}
Diagnóstico	Binario	{Benigno, Maligno}

En total se procesaron 424 registros con 17 casos confirmados con cáncer de mama. Las clasificaciones BI-RADS se distribuyen de la forma que se muestra en la tabla 5:

Tabla 5. Clasificaciones BI-RADS

Categoría	Cantidad	Malignos	Porcentaje
BI-RADS 2	379	0	0 %
BI-RADS 3	13	1	7.7 %
BI-RADS 4	20	7	35 %
BI-RADS 5	11	9	81 %

IV.4 Transformación y codificación de datos

En base a los datos extraídos que se mencionaron de la sección anterior, se procedió a codificar y transformar los datos, también con la ayuda del jefe de epidemiología del hospital.

En la tabla 6 se muestra un diccionario de los datos limpios y las claves correspondientes tras el proceso de limpieza y transformación.

Tabla 6. Diccionario de datos de la base BI-RADS

Variable	Descripción	Código
Edad	15 <= Edad <= 100	Numérico
Hallazgo	Masa	1
	Masa + calcificacion	2
	Calcificaciones	3
	Distorsión arquitectura	4
	Inflamatorio	5
Forma	Redonda	1
	Ovalada	2
	Irregular	3
	Estelar	4
Bordes	Circunscrito	1
	Microlobulado	2
	Oculto	3
	Irregular	4
	Especulado	5
	Mal definido	6
Mic-agr	Micro-calcificación agrupada	{0,1}
Mic-ais	Micro-calcificación aislada	{0,1}
Mac-agr	Macro-calcificación agrupada	{0,1}
Mac-ais	Macro-calcificación aislada	{0,1}

Cuadrantes	Superior externo	1
Variable	**Descripción**	**Código**
Cuadrantes	Superior interno	2
	Inferior externo	3
	Inferior interno	4
	Retroareolar	5
Mama	Derecha	1
	Izquierda	2
Densidad	Fibroglandular disperso	1
	Fibroglandular residual	2
	Heterogeneamente densas	3
BI-RADS	Clasificación BI-RADS	2,3,4,5
Diagnóstico	Benigno	1
	Maligno	2

IV.4.1 Reposición de datos faltantes

Según el National Research Council en su reporte sobre el tratamiento de datos faltantes del 2010, menciona que se puede utilizar la imputación simple mediante regresión para tratar con éste problema. En nuestro caso era necesario reponer las dimensiones de 21 tumores en el registro, para los cuales no se había reportado el dato del tamaño; por lo tanto se aplicó la técnica de regresión para recuperar a aquellos datos faltantes a partir de la información existente de 191 registros de tumores.

En la base de datos se cuenta con información de 191 tumores y se dividió de tal manera que 75% de sus registros seleccionados aleatoriamente serían utilizados para entrenamiento, el 25% restante para prueba. Se hicieron pruebas con el algoritmo de regresión de soporte vectorial (SVR) con diferentes núcleos, los cuales se reportan en la tabla 7. Para las pruebas con SVR se utilizó el paquete WEKA con el plugin LibSVM.

Con los mismos datos de entrenamiento y prueba utilizados en el experimento anterior, se realizó otro experimento con el algoritmo Elastic Net (Friedman et al., 2008). Para determinar el parámetro óptimo de lambda, se utilizó validación cruzada con 10-fold; y se obtuvo un valor óptimo con lambda = 0.50, tal como se ve en la figura 21. El resultado de la comparación entre SVR y Elastic Net se muestra en la tabla 7 en términos del error RMSE.

En la figura 20 se muestra la gráfica del error en la predicción contra la dimensión del tumor. En el eje horizontal se denotan las dimensiones de los tumores, ordenadas de menor a mayor y en el eje vertical se muestra la raíz del error cuadrado medio para cada dimensión. De la figura se observa que el algoritmo Elastic Net posee mayor variabilidad en sus predicciones. Además se observa que todos presentan errores grandes, mayores a 10, cuando la dimensión del tumor supera los 16 milímetros.

La figura 21 muestra la gráfica del parámetro log lambda con respecto del error. La gráfica muestra el error de cada parámetro lambda y la línea vertical punteada indica el valor óptimo de lambda. El eje horizontal inferior representa al valor del logaritmo de lambda, el horizontal superior representa la cantidad de coeficientes diferentes de cero que resultan con dicho valor de lambda. El eje vertical representa al error cuadrado medio, y en la gráfica se ven marcados los límites superior e inferior del error cuadrado medio para cada valor de lambda.

Tabla 7. Comparación de algoritmos de regresión

	SVR			**Elastic Net**
	Base radial	Polinomial grado 2	Polinomial grado 3	Lambda 0.50
RMSE[1]	6.88	6.85	6.66	6.29

[1] Raíz del error cuadrado medio.

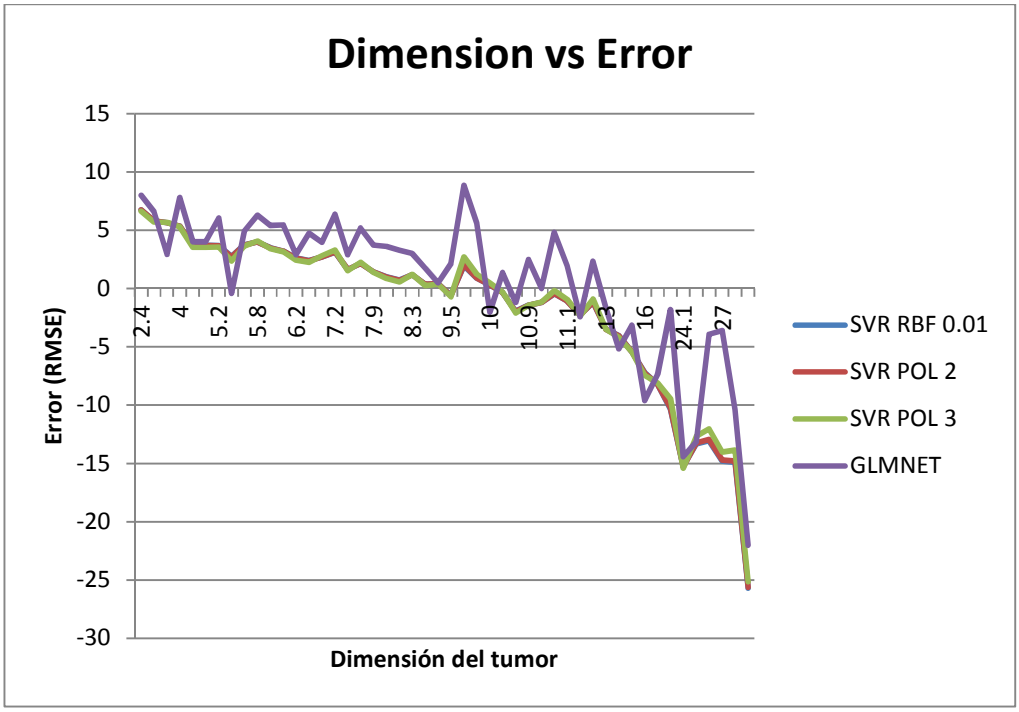


Figura 20. Gráfica del tamaño del tumor contra el error en la predicción

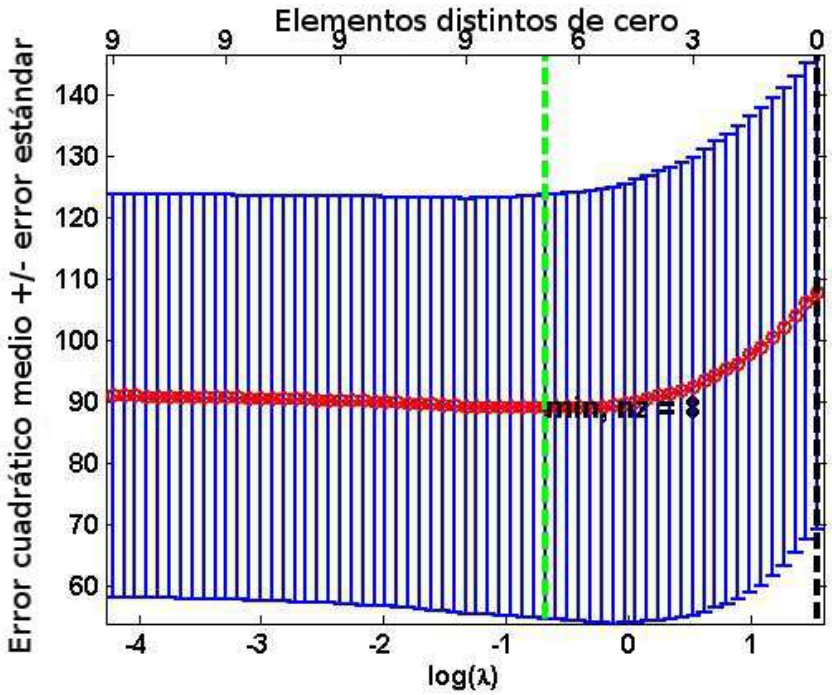


Figura 21 Gráfica de la raíz del error cuadrado medio contra log lambda.

Una de las características de Elastic Net es que impone una penalización en el tamaño de los coeficientes de regresión, llegando a cero en algunos de ellos y por lo tanto se disminuye la dimensionalidad del conjunto de datos inicial. La reducción de dimensionalidad permite obtener un modelo más simple y con menor error en la predicción.

En el experimento realizado, se redujo la dimensionalidad al eliminar dos de los atributos de la base de datos, que corresponden a la forma del borde del tumor y la presencia de micro calcificaciones agrupadas. Los resultados se pueden apreciar en la tabla 8. Es interesante ver que el atributo de la edad recibió un coeficiente cercano a cero.

BIAS	EDAD	FORMA	BORDES	MIC-AGR	MIC-AIS	MAC-AIS	MAC-AGR	CUADRANTE	DENSIDAD	BIRADS
0.91	-0.03	6.49	0	0	-0.18	-1.95	-1.12	-1.17	-1.06	1.84

Tabla 8. Coeficientes obtenidos mediante Elastic Net

La figura 22 corresponde a una gráfica del tamaño de los coeficientes de regresión contra log lambda. Se puede apreciar que los coeficientes de regresión se reducen conforme el parámetro de penalización log lambda aumenta, como resultado de la penalización en el tamaño de los coeficientes.

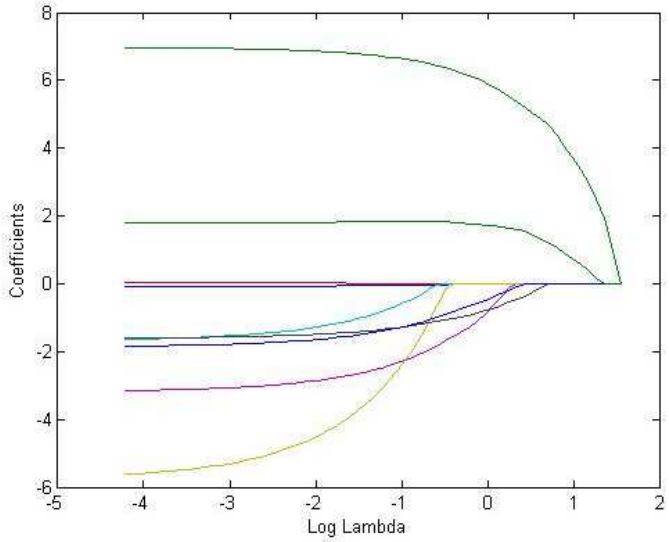


Figura 22 Gráfica de los coeficientes Vs. log lambda

IV.4.2 Análisis de clústeres: BI-RADS

Una vez que se recuperaron las dimensiones faltantes de los tumores, se aplicaron tres técnicas para analizar los datos de registro BI-RADS, el análisis de componentes principales, el mapa auto organizado y el algoritmo EM.

IV.4.2.1 Análisis de componentes principales (PCA)

La imagen de la figura 23 corresponde a la gráfica de puntos de los primeros dos componentes principales de la base BI-RADS. En la figura, podemos observar la formación de dos cúmulos principales, los cuales corresponden con los hallazgos de calcificaciones en la parte de arriba, y con masas, en el cúmulo de abajo. Además se puede apreciar que para cada caso, hay correspondencia de los cúmulos encontrados, con los grupos de riesgo de la clasificación BIRADS.

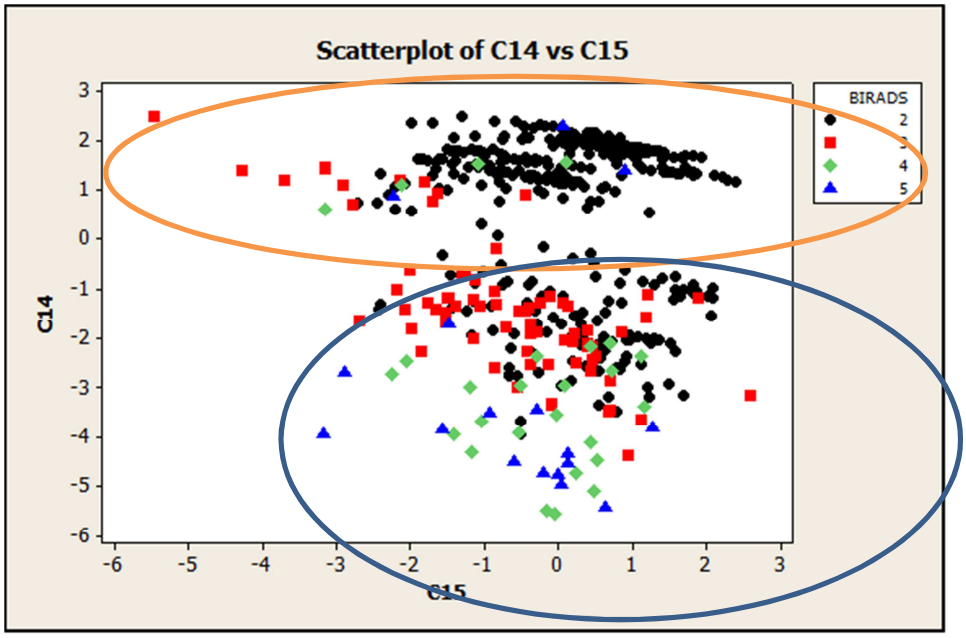


Figura 23 Gráfica de los primeros dos componentes principales, con la categoría BIRADS como clase.

IV.4.2.2 Red neuronal de Kohonen (SOM)

Se aplicó el algoritmo SOM sobre los datos para identificar la formación de cúmulos, se utilizó el software Spice SOM para el análisis de los datos. Como función de vecindad se utilizó la función gaussiana, topología hexagonal y tasa de aprendizaje del 0.1 con un máximo de 100 iteraciones. La imagen resultante se presenta en la figura 24.

Como se puede apreciar en la figura 24, existe la formación de cúmulos diferenciados para cada categoría BI-RADS, con la clase BI-RADS 3, empalmándose en tres de ellos, y las categorías BI-RADS 4 Y BI-RADS 5 se diferencian del resto acumulándose en la esquina superior izquierda.

IV.4.2.3 Expectation –maximization (EM)

Se hicieron pruebas con el algoritmo EM para estimar los parámetros de los modelos de los datos. Se utilizó WEKA para hacer las pruebas e inicialmente se dejó el parámetro del número de clústeres en -1; para permitir que el algoritmo

encontrase los cúmulos mediante validación cruzada; luego se fijó el parámetro en 4, que es el número de grupos BIRADS presentes en el conjunto de datos.

Para el caso de n = -1, en donde por medio de validación cruzada se busca automáticamente el número de clústeres, se obtuvieron los siguientes resultados:

Se encontraron 3 clústeres, equivalentes a las clases BI-RADS 4, BI-RADS 3 Y BI-RADS 2; la clase BI-RADS 5, tal como se vio en el análisis de componentes principales, aparece empalmada junto con la clase BIRADS 4, por lo tanto el algoritmo EM no la identificó como un clúster aparte.

Los resultados obtenidos con Weka se muestran en la figura 25.

Para el caso de n = 4; se presentan los siguientes resultados:

Se aprecia un empalme entre las clases BI-RADS 4 Y BI-RADS 5; la clase BI-RADS 2 se encuentra diferenciada del resto; la clase BI-RADS 3 se encuentra empalmada con la clase BI-RADS 2.

Los resultados obtenidos con Weka se muestran en la figura 26.

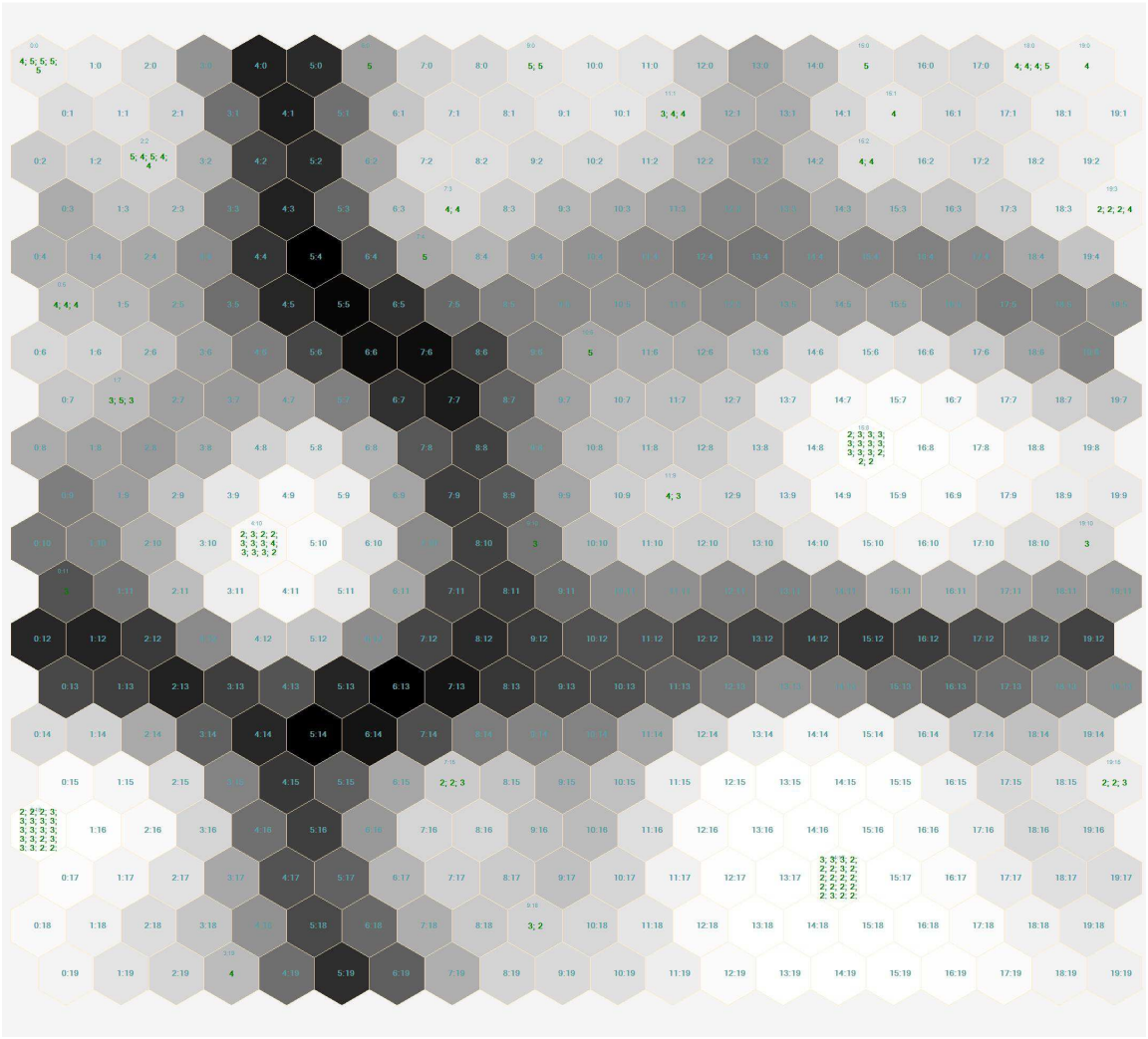


Figura 24 Imagen resultante del SOM aplicado a la base de datos de mastografías, con categoría BIRADS como clase.

```
Class attribute: BIRADS
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  3 18 92 | 2
  3 26 33 | 3
 19  2  1 | 4
 12  1  1 | 5

Cluster 0 <-- 4
Cluster 1 <-- 3
Cluster 2 <-- 2

Incorrectly clustered instances :   74.0   35.0711 %
```

Figura 25 Resultado obtenido con WEKA utilizando el algoritmo EM con n = -1

```
Class attribute: BIRADS
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 18 92  3  0 | 2
 26 33  2  1 | 3
  2  1 12  7 | 4
  1  1  6  6 | 5

Cluster 0 <-- 3
Cluster 1 <-- 2
Cluster 2 <-- 4
Cluster 3 <-- 5

Incorrectly clustered instances :   75.0   35.545  %
```

Figura 26 Resultado obtenido con WEKA utilizando el algoritmo EM con n = 4

IV.4.2.4 Análisis de clústeres: Diagnóstico histopatológico.

Se aplicó el análisis de componentes principales y el del mapa auto organizado (SOM) sobre los mismos datos, pero esta vez se utilizó como criterio de clase al diagnóstico histopatológico.

IV.4.2.4.1 Análisis de componentes principales.

La imagen de la figura 27 corresponde a la gráfica de los dos primeros componentes principales de la base BI-RADS hecha en base a la información de masas, con el diagnóstico histopatológico como clase. Los puntos rojos

corresponden a la clase MALIGNO, mientras que los negros corresponden a la clase BENIGNO.

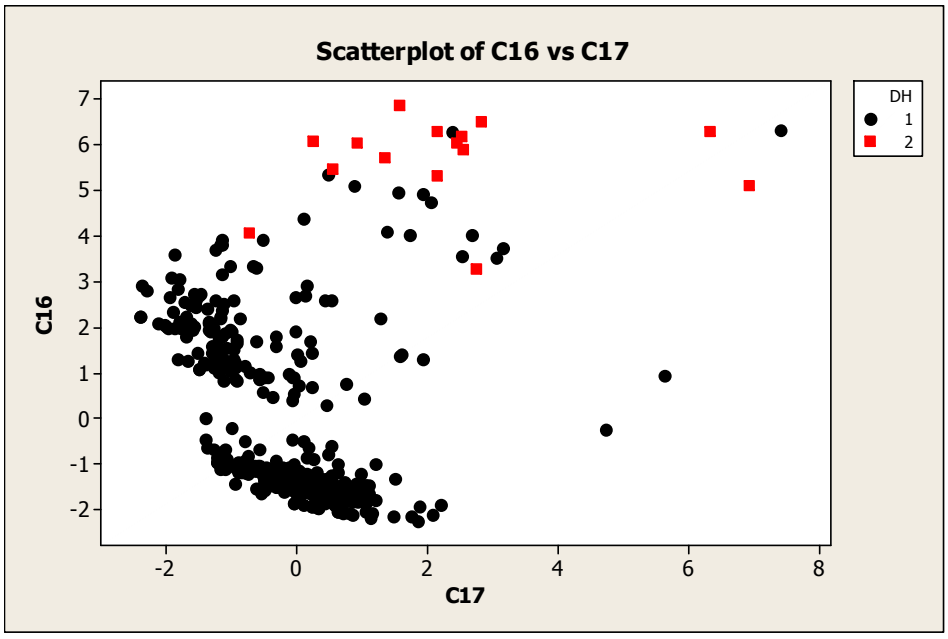


Figura 27 Gráfica de los dos primeros componentes principales de la base de mastografías, con el diagnóstico patológico como clase.

IV.4.2.4.2 Red neuronal de Kohonen (SOM)

Al igual que en el análisis de cúmulos BI-RADS, se utilizó el paquete Spice SOM. Los mejores resultados se obtuvieron con una función de vecindad gaussiana, topología hexagonal y tasa de aprendizaje de 0.1 con un máximo de 100 iteraciones. La imagen resultante se presenta en la figura 28, de donde marcamos con un óvalo rojo al cúmulo de casos correspondientes a un diagnóstico de malignidad.

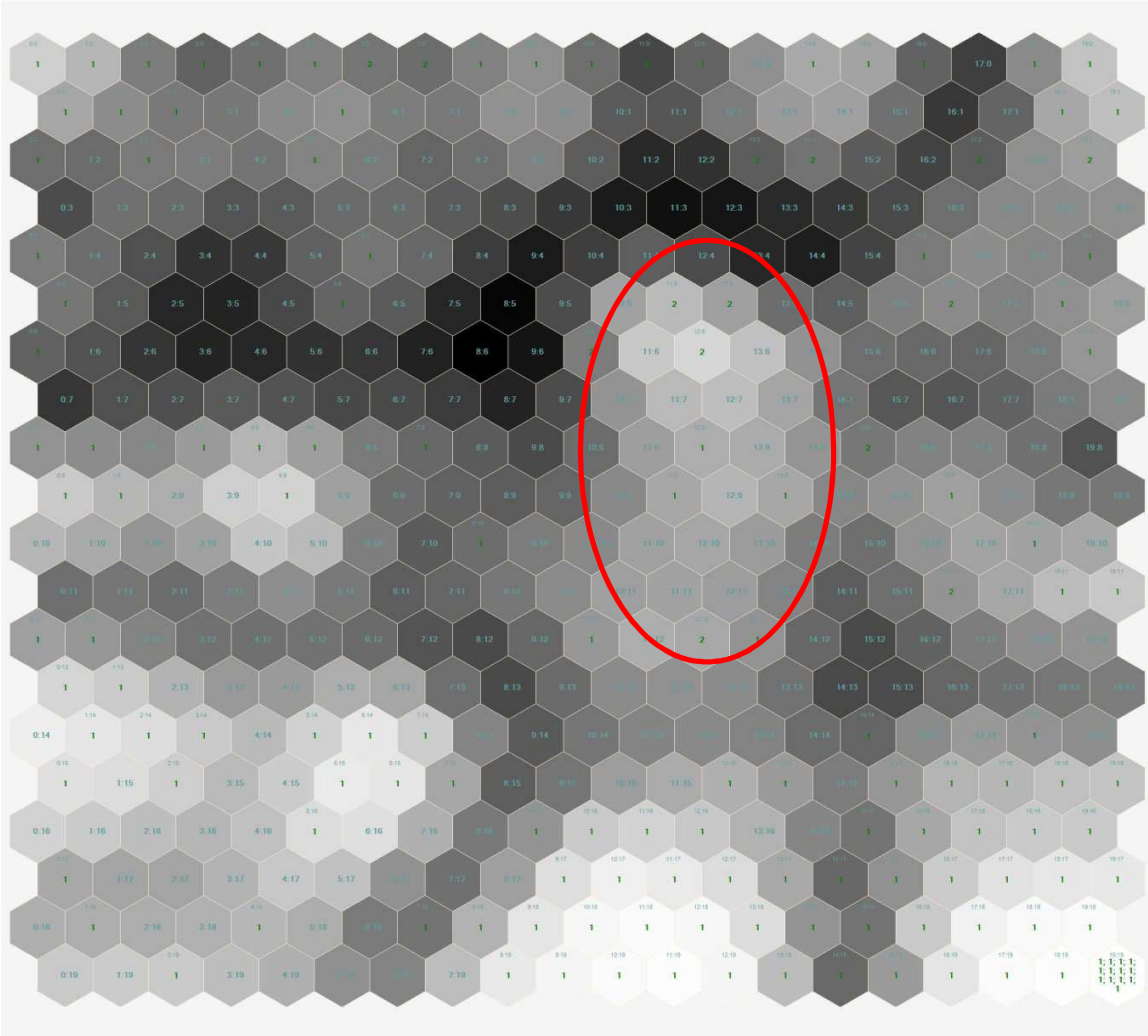


Figura 28 Imagen resultante del SOM aplicado a la base de datos de mastografías con diagnóstico patológico como clase.

IV.4.2.5 Reducción de la dimensionalidad.

Con la finalidad de mejorar la exactitud y la calidad del modelo de clasificación se aplicó la reducción de dimensionalidad sobre la vista minable obtenida del paso anterior. Éste método permite expresar el modelo en función de menos variables y se eliminan atributos irrelevantes para la clasificación.

Se aplicaron dos técnicas para reducir la dimensionalidad: la ganancia en la tasa de la información y el método de Elastic Net, que como vimos en la sección de recuperación de datos, reduce los coeficientes de regresión, llegando a convertir en cero a algunos de estos coeficientes.

Para el método de ganancia en la tasa de la información, se aplicaron una serie de pasos para valorar la relevancia que tienen las columnas de la vista minable en la clasificación. El primer paso consiste en aplicar el método sobre todos los atributos de la vista minable, se elimina al atributo menos relevante para el modelo y se aplica un algoritmo de clasificación, en este caso hemos utilizado el algoritmo de Bayes Simple. Se aplica nuevamente el método de ganancia en la tasa de la información y se elimina al atributo menos relevante y volvemos a evaluar con el mismo algoritmo de clasificación. Se continúa de esta manera hasta obtener un resultado deseado en la exactitud en la clasificación.

En la figura 29 se muestra los atributos ordenados según su relevancia para la clasificación utilizando el método de ganancia en la tasa de la información. Como explicamos anteriormente, se utilizó el algoritmo de Bayes Simple para medir la exactitud en la clasificación y se obtuvo un 96.19% de registros clasificados correctamente y además el modelo se puede expresar ahora en función de seis variables: Bordes, clasificación BI-RADS, dimensión, presencia de micro calcificaciones agrupadas, forma del tumor y el hallazgo.

Figura 29. Relevancia de atributos según el método de ganancia en la tasa de la información.

Para el método de reducción con Elastic Net, el modelo se puede expresar en función de ocho variables: Hallazgo, bordes, presencia de micro calcificaciones agrupadas, presencia de micro calcificaciones aisladas, presencia de macro calcificaciones aisladas y densidad. Con éste último modelo se obtiene una exactitud de 98.57% de registros clasificados correctamente.

IV.4.2.6 Clasificación.

En esta sección reportaremos los resultados obtenidos al aplicar algoritmos de clasificación en el repositorio de mastografías. Para esta tarea se utilizó el programa Weka para minería de datos y se utilizó validación cruzada con 10-fold para evaluar los resultados. Los algoritmos de clasificación que se utilizaron y evaluaron fueron los siguientes: Máquina vector de soporte (SVM), regresión logística, perceptrón multicapa (MLP), k-vecinos cercanos (KNN), Bayes simple, red Bayesiana (BAYES NET) , árbol de decisión C4.5, árbol de decisión CART, árbol aleatorio y Elastic Net.

Los resultados de la clasificación se reportan en la tabla 9 y las gráficas de la exactitud y la evaluación de los modelos se presentan en las gráficas de las figuras 30 y 31.

ALGORITMO	EXACTITUD	AUC	SENSIBILIDAD	ESPECIFICIDAD	SE
SVM	97.62	0.859	0.733	0.985	0.722
LOGISTIC	97.86	0.984	0.660	0.990	0.653
MLP	97.86	0.950	0.660	0.990	0.653
KNN	97.62	0.983	0.666	0.980	0.653
BAYES	96.19	0.990	0.933	0.980	0.914
C4.5	97.14	0.899	0.600	0.998	0.599
RANDOM TREE	97.86	0.826	0.660	0.990	0.653
CART	98.57	0.806	0.800	0.999	0.799
BAYES NET	96.19	0.984	0.866	0.965	0.836
GLMNET	98.57	0.970	0.733	0.990	0.726

Tabla 9. Resultados de exactitud y evaluación de los algoritmos.

Figura 30. Gráfica de la exactitud de los algoritmos

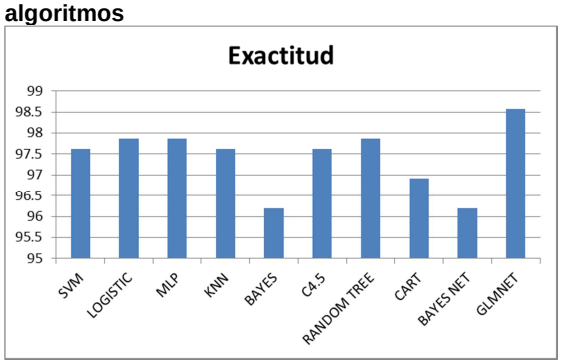
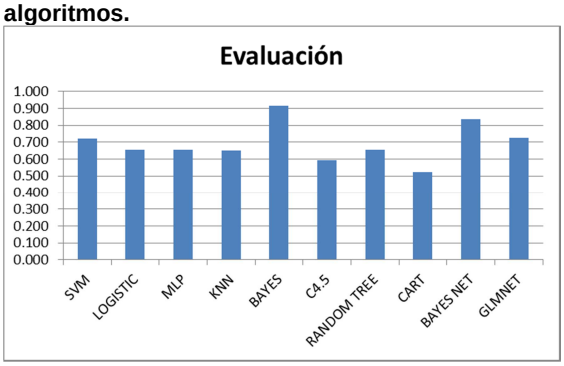


Figura 31. Gráfica de la evaluación de los algoritmos.



De los algoritmos analizados en esta etapa, nos enfocaremos especialmente en los resultados obtenidos a partir de los árboles de decisión y de la regresión logística, no solo por los buenos resultados obtenidos en cuanto a exactitud en la clasificación y evaluación, sino que éstos modelos tienen la particularidad de retroalimentar e informar sobre la relación que existe entre las variables de entrada y su clasificación.

IV.4.2.7 Modelos resultantes.

En esta sección describiremos los modelos predictivos obtenidos a partir de los algoritmos de árbol de decisión y regresión logística.

En la primera subsección se describirán los modelos basados en árbol de decisión y posteriormente los modelos de regresión logística.

IV.4.2.7.1 Modelos predictivos basados en árbol de decisión.

En las figuras 32, 33 y 34 mostramos los árboles de decisión inducidos por los algoritmos C4.5, CART y Random Tree respectivamente. Podemos apreciar que el árbol inducido por el algoritmo CART es el más simple y sencillo de interpretar, además de haber obtenido los mejores números en cuanto a exactitud en la precisión y sus medidas de sensibilidad y especificidad.

En la tabla 9 se encuentran las medidas de evaluación y desempeño para cada uno de los algoritmos de árbol de decisión.

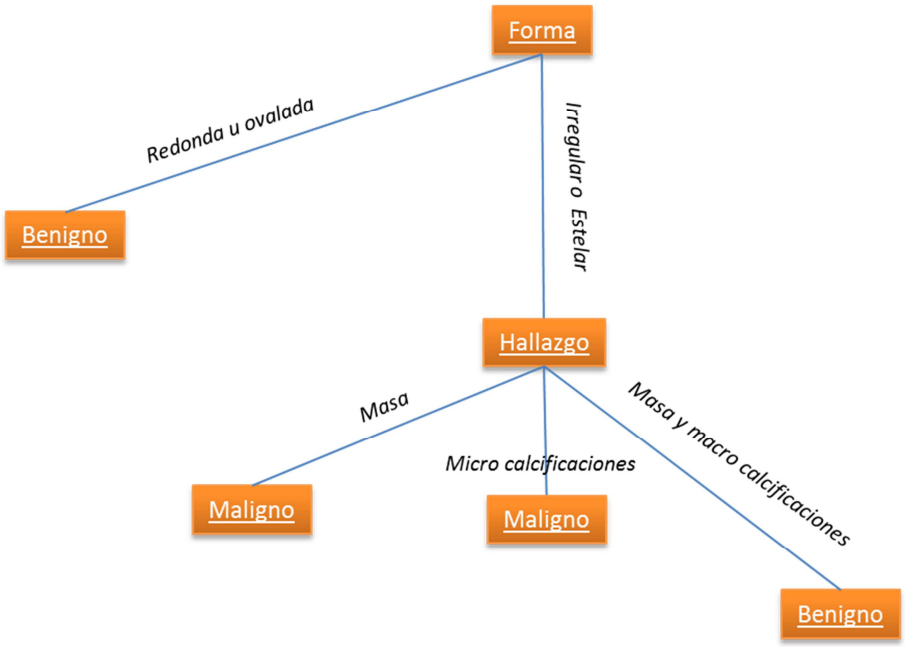


Figura 32. Árbol de decisión inducido con el algoritmo CART

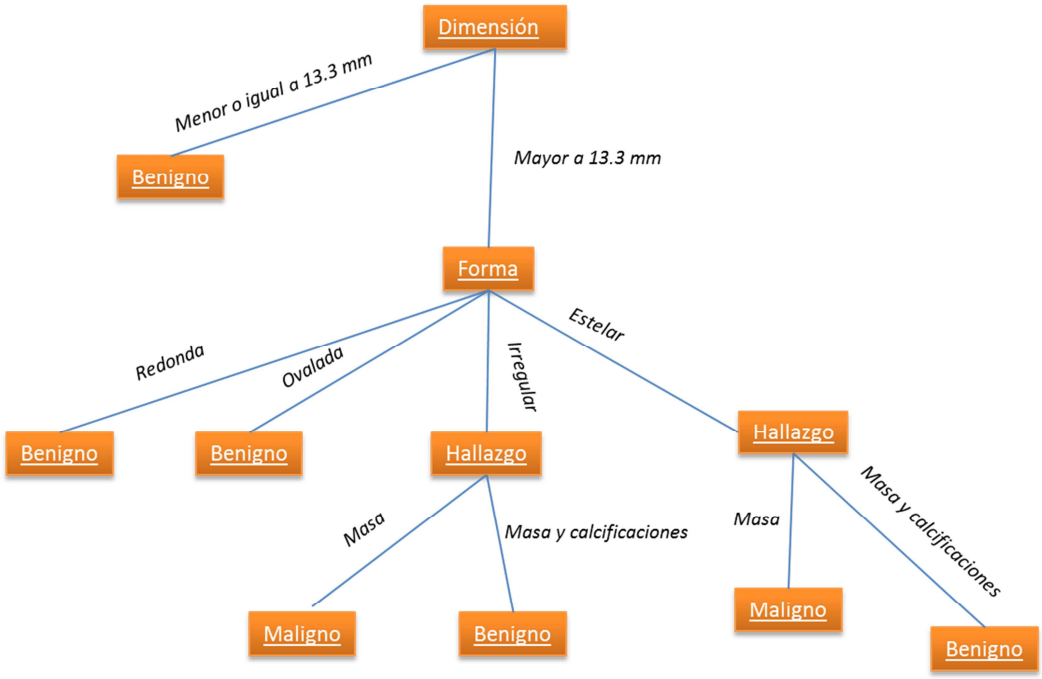


Figura 33. Árbol de decisión inducido con el algoritmo C4.5

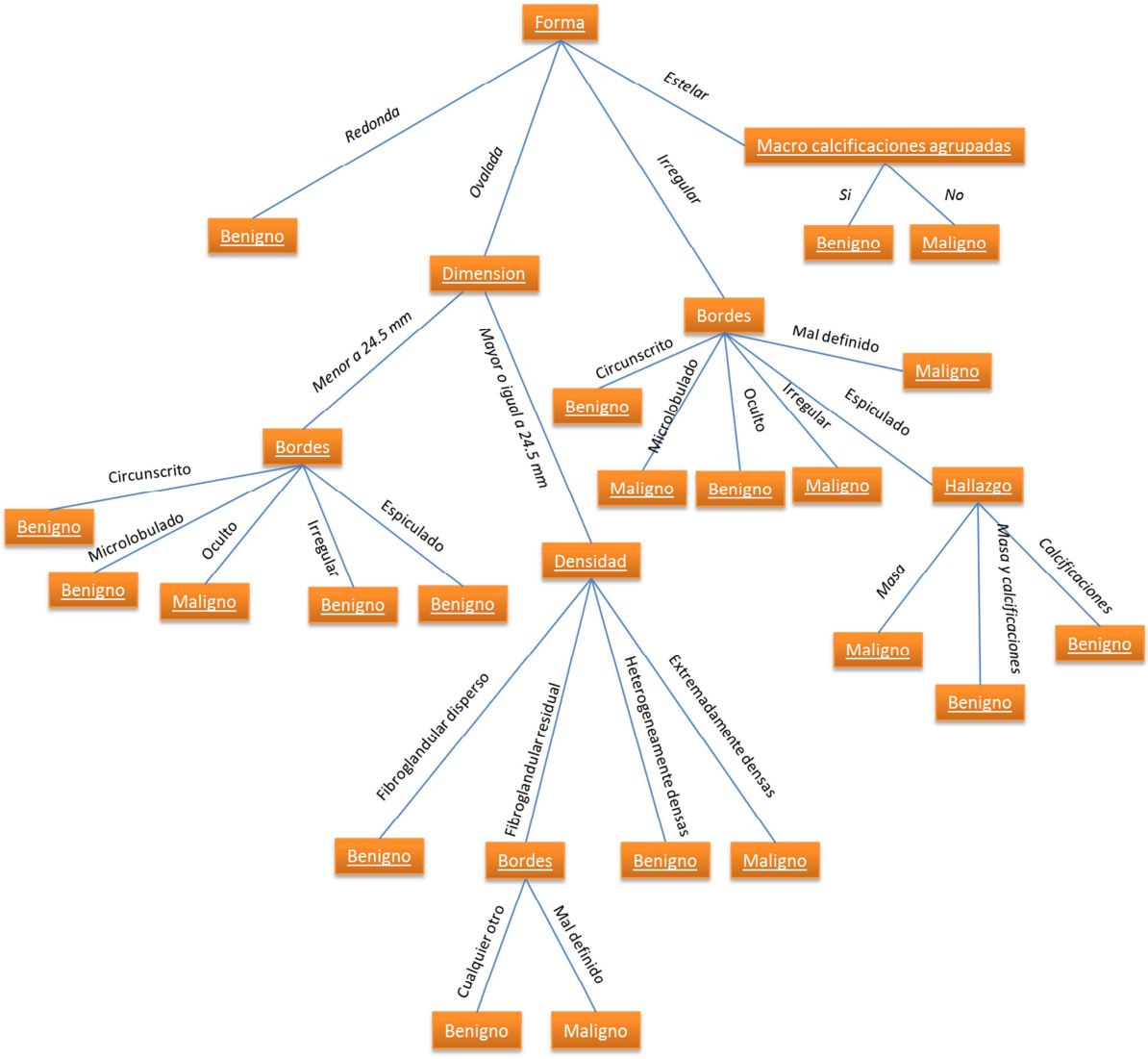


Figura 34. Árbol de decisión inducido por el algoritmo Random Tree

IV.4.2.7.1.1 Ejemplo de uso de un árbol de decisión.

Supongamos que una paciente de 45 años se presenta a realizarse una mastografía. De la imagen se obtuvieron los siguientes hallazgos: La estructura mamaria es de tipo mixto con tejido heterogéneamente denso hacia cuadrantes superiores y externos. No se visualizan calcificaciones ni zonas de distorsión de la

arquitectura. Se observa un nódulo ovalado de contornos circunscritos con dimensiones aproximadas de 15.2 x 4.6 x 6.8 milímetros en cuadrante superior externo de la mama izquierda.

Hallazgos compatibles con BI-RADS 3

Transformando la información anterior se obtiene un registro de la siguiente manera:

Edad	Hallazgo	Forma	Bordes	Dim.	Mic. Agr	Mic. Ais.	Mac. Agr	Mac. Ais	Cuadr.	Mama	Densidad	BIRADS
45	1	2	1	15.2	0	0	0	0	1	2	2	3

Para aplicar el modelo del árbol inducido por CART (de la figura 42) solamente es necesario contar con los atributos de forma y hallazgo, por lo tanto podemos construir la siguiente regla:

> **Si** forma = ovalada, **entonces** diagnóstico = benigno.
>
> Probando ahora con el árbol inducido por C4.5 tenemos:
>
> **Si** dimensión > 13.3 y forma = ovalada **entonces** diagnóstico = benigno.

Ahora aplicando el modelo inducido con el algoritmo Random Tree (figura 40) tenemos:

> **Si** forma = ovalada y dimensión < 24.5 y bordes = circunscrito, **entonces** diagnóstico = benigno.

IV.4.2.7.2 Modelos predictivos basados en regresión logística.

En la tabla 10 se muestra el modelo basado en regresión. Las medidas correspondientes al desempeño y evaluación se muestran en la tabla 9 en este mismo capítulo. Podemos apreciar que los coeficientes de la edad, la presencia de macro calcificaciones aisladas, el cuadrante y el lado se redujeron a cero, siendo éstos irrelevantes para el modelo. Los coeficientes correspondientes a la densidad, dimensión y hallazgo obtuvieron valores muy cercanos a cero.

INTERCEPT	HALLAZGO	EDAD	FORMA	BORDES	DIM.	MIC AIS	MIC AGR	MAC AGR	MAC AIS	CUADR.	LADO	DENS.
-7.0308	-0.0943	0	0.679	0.8327	0.0441	0.8206	-0.8266	-0.6571	0	0	0	0.089

Tabla 10. Coeficientes resultantes del algoritmo Elastic Net.

Ejemplo de uso de un modelo de regresión logística:

Supongamos que una paciente de 45 años se presenta a realizarse una mastografía. De la imagen se obtuvieron los siguientes hallazgos: La estructura mamaria es de tipo mixto con tejido heterogéneamente denso hacia cuadrantes superiores y externos. No se visualizan calcificaciones ni zonas de distorsión de la arquitectura. Se observa un nódulo ovalado de contornos circunscritos con dimensiones aproximadas de 15.2 x 4.6 x 6.8 milímetros en cuadrante superior externo de la mama izquierda.

Hallazgos compatibles con BI-RADS 3

Transformando la información anterior se obtiene un registro de la siguiente manera:

Edad	Hallazgo	Forma	Bordes	Dim.	Mic. Agr	Mic. Ais.	Mac. Agr	Mac. Ais	Cuadr.	Mama	Densidad	BIRADS
45	1	2	1	15.2	0	0	0	0	1	2	2	3

Ahora calculamos la probabilidad de la paciente de tener un diagnóstico clasificado como maligno utilizando la siguiente expresión:

$$\Pr(Diagnostico = benigno|x) = \frac{1}{1 + e^{\beta_0 + x^T\beta}}$$

En donde los valores de beta corresponden a los coeficientes de la tabla 8.

Si sustituimos los valores del registro transformado en x de la ecuación anterior tenemos:

$$\Pr(Diagnostico = benigno|x)$$
$$= \frac{1}{1 + e^{-7.0308+(-0.0943)+(0.679)(2)+(0.8324)+(0.0441)(15.2)+(0.8206)(0)+(0.8266)(0)+(0.6571)(0)+(0.089)(2)}}$$

$$\Pr(Diagnostico = benigno|x) = 0.9834$$

Y por lo tanto:

$$\Pr(Diagnostico = maligno|x) = 1 - \Pr(Diagnostico = benigno|x) = 0.0166$$

Entonces se tiene que la probabilidad de que el registro tenga un diagnóstico benigno es del 0.9834, contra el 0.0166 de probabilidad de tener un diagnóstico de malignidad.

IV.5 Resumen

En el presente estudio, se reporta el análisis de cúmulos y de clasificación realizado sobre un repositorio de registros de mastografías. Para el análisis de cúmulos, se realizó un estudio con el propósito de verificar la correspondencia de los grupos de riesgo BI-RADS con el diagnóstico histopatológico de benignidad y malignidad. El método de análisis de componentes principales y la red neuronal de

Kohonen permiten visualizar de manera gráfica a los cúmulos en un espacio bidimensional, sin embargo, los métodos de K-Medias y el algoritmo EM cuantifican la cercanía y la formación de dichos cúmulos.

En base al resultado obtenido, podemos concluir que es posible diferenciar cúmulos en los datos mastográficos y que éstos se corresponden con las categorías de riesgo, a pesar de existir un traslape entre algunos de los cúmulos. El mismo análisis de cúmulos se aplicó para el estudio de benignidad vs malignidad.

Con respecto a la clasificación del conjunto, el mejor resultado en la exactitud se obtuvo con Elastic Net y el menor con Bayes Simple, aunque en éste último se obtuvieron mejores resultados en la evaluación del algoritmo en cuanto a valores de sensibilidad y especificidad.

En cuanto a la reducción de dimensionalidad, encontramos como elementos en común entre el método de ganancia en la tasa de información y Elastic Net, la forma, los bordes, las dimensiones del tumor, la presencia de micro calcificaciones agrupadas y micro calcificaciones aisladas. Lo cual concuerda con los criterios utilizados por el ACR para evaluar el riesgo de padecer cáncer de mama a través de su sistema BI-RADS de clasificación.

Los modelos definitivos que se recomiendan, basándonos en los resultados obtenidos y en la literatura son los de los árboles de decisión y la regresión logística. Además, con los árboles de decisión es posible que sean interpretados a manera de reglas por un especialista y el modelo de regresión logística es sencillo de implementarse en una hoja de cálculo, además de que provee información sobre la relevancia de los atributos o predictores con respecto a la clase a la que pertenecen.

Desde un punto de la ingeniería de procesos, se recomienda formalizar un formato para capturar la información relevante para la predicción de la presencia del

cáncer de mama. Además se debe facilitar la integración de dicha información con el diagnóstico histopatológico obtenido del área de patología. Un sistema de expediente electrónico con especialidad podría fácilmente integrar información de diferentes áreas, incluso de estudios y análisis efectuados en otras clínicas e instituciones externas al Instituto, con la ventaja de mantener la integridad y validez de la información ingresada. La principal ventaja de contar con un sistema de expediente como el mencionado, es la posibilidad de integrar un repositorio de datos más extenso y realizar experimentos con más datos, con el fin de elaborar reglas de clasificación más robustas.

Capítulo V. Proceso de descubrimiento de conocimiento sobre repositorios de pacientes en tratamiento para cáncer de mama

En el presente capítulo, se presenta el proceso de descubrimiento en datos en uno de los repositorios obtenidos del caso de estudio presentado en el capítulo III. El repositorio en cuestión corresponde a casos de cáncer de mama que recibieron tratamiento o se encuentran recibiéndolo y el propósito del estudio es predecir la recurrencia de la enfermedad del cáncer de mama en base a los datos disponibles del repositorio.

En las siguientes subsecciones se detallan los datos trabajados, el problema de los datos faltantes y su tratamiento y finalmente el análisis y evaluación de los modelos obtenidos mediante el proceso de minería de datos.

V.1 Antecedentes

En el contexto del presente trabajo de investigación, cuando hablamos de la recurrencia del cáncer de mama, nos referimos a la facultad de predecir o a cuantificar la posibilidad de recurrencia de la enfermedad después de haber sido diagnosticada y tratada. En el trabajo de Williams y Williamson (Williams et al., 2006), se reporta un trabajo realizado sobre un repositorio[2] de pacientes de cáncer de mama. El propósito del trabajo era pronosticar utilizando redes bayesianas, la recurrencia de la enfermedad basándose en los datos recopilados, tales como la edad de la paciente, el cuadrante del tumor primario, la etapa del cáncer al momento del diagnóstico, etcétera. El repositorio se recopiló en la década de los 1980, en momentos cuando no se realizaban pruebas de inmunohistoquímica que permiten detectar marcadores en la superficie de las células del tumor y que además le confieren cualidades específicas en cuanto a tratamiento y pronóstico. En la actualidad tales pruebas se realizan rutinariamente en todas las muestras que han resultado positivas para malignidad. Por lo tanto, en el presente trabajo

[2] University Medical Centre, Institute of Oncology, Ljubljana, Slovenia

de investigación, se realiza un trabajo similar al de Williams y Williamson, pero esta vez añadiendo los marcadores tumorales en los registros de las pacientes.

V.2 Materiales y métodos

Para el análisis con los algoritmos de minería de datos se utilizó el software Weka y el paquete GLMNET que contiene la implementación del algoritmo Elastic Net en R. Además para el análisis de recuperación de datos perdidos se utilizó OpenBUGS[3] que consiste en una implementación del método de inferencia bayesiana utilizando el muestreo de Gibbs.

Los datos crudos se obtuvieron del caso de estudio presentado en el capítulo III y se recibieron como archivos de FileMaker[4].

V.3 Extracción de datos

Al igual que los datos del capítulo IV, éstos se presentaron sin ningún tipo de preprocesamiento, es decir, estaban a manera de redacción y en registros de FileMaker. Por lo que fue necesario filtrar a aquellos registros que no contenían los datos necesarios para el análisis y con la ayuda del especialista en oncología, separar y organizar en una hoja de cálculo los registros originales. En la figura 35 se muestra un ejemplo de cómo venían los registros. Al final de este proceso se obtuvo una hoja de cálculo con los datos organizados en filas y columnas.

[3] Disponible en: http://www.openbugs.info/w/
[4] http://www.filemaker.com/

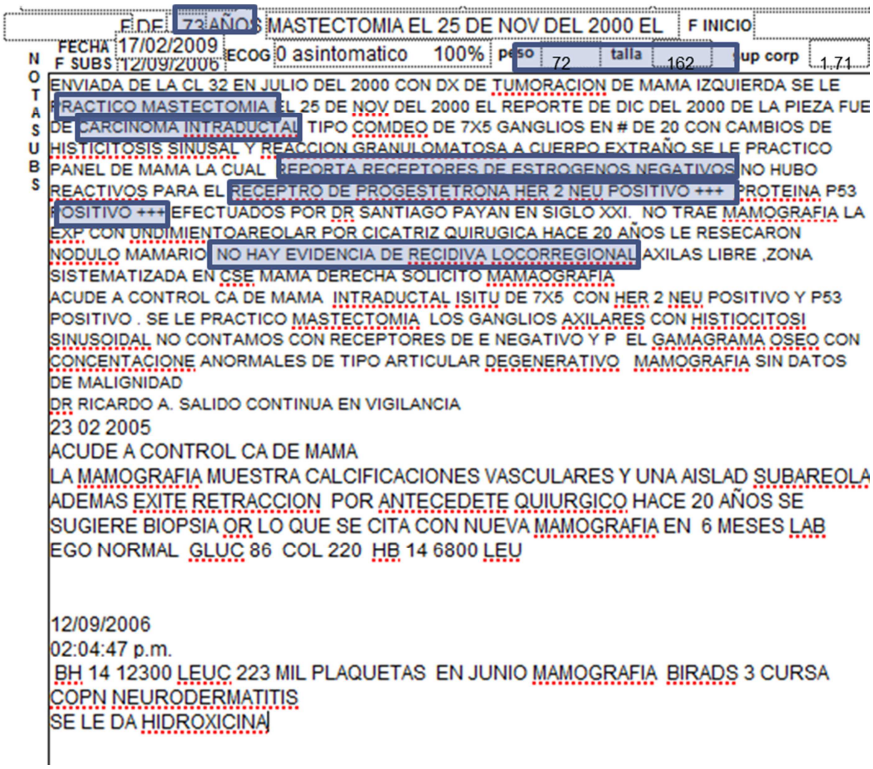


Figura 35. Ejemplo de registro de paciente.

Al final de la extracción de datos, se obtuvo un nuevo repositorio de 217 registros. En la tabla 11 se muestra una descripción de cada una de las variables que componen dicho repositorio.

Tabla 11. Descripción de los campos que componen al repositorio.

Campo	Tipo	Descripción
Edad	Numérico	Numérico [18 <= Edad <= 100]
Peso	Numérico	Numérico continuo [30 <= Peso <= 150]
Talla	Numérico	Numérico continuo [140 <= Talla <= 180]
Índice de masa corporal	Numérico	Numérico continuo [20 <= IMC <= 50]
Superficie corporal	Numérico	Numérico continuo [1.38 <= SC <= 2.30]
Diagnóstico	Nominal	{Carcinoma ductal infiltrante, Lobulillar infiltrante, Ductal insitu, Medular, Mixto, Inflamatorio, Mucinoso, Papilar, Cel. Gig., Angiosarcoma, Indiferenciado}
Grado	Nominal	{Bajo, Medio, Alto}
Dimensiones	Numérico	Numérico continuo [0 <= Dimensiones <= 9]

Subtipo	Nominal	{Luminal NS, Luminal A, Luminal B, Her2+++, Triple Negativo}
Estadio	Nominal	{0, 1, 2A, 2B, 3A, 3B, 4 }
Receptores estrógeno	Numérico	Numérico continuo [0 <= RE <= 100]
Receptores progesterona	Numérico	Numérico continuo [0 <= RP <= 100]
Her2/Neu	Nominal	{+, ++, +++}
Proteina Ki67	Numérico	Numérico continuo [0 <= KI67 <= 100]
Proteina P53	Numérico	Numérico continuo [0 <= P53 <= 100]
Quimioterapia	Binario	{Si, No}
Cirugia	Nominal	{Mastectomia, Cuadrantectomia, Lumpectomia, Ninguna}
Radioterapia	Numérico	{0, 25, 35
Hormonoterapia	Binario	{Si, No}
Inmunoterapia	Binario	{Si, No}
Quimioprofilaxis	Nominal	{Nada, Capecitabina, Vinorelbina}
Recurrencia	Binario	{Si, No}

V.4 Transformación y codificación

De los datos extraídos del paso anterior, se procedió a realizar una transformación y codificación de los datos. La codificación se hizo con ayuda de los especialistas en oncología y epidemiología.

En la tabla 12 se muestra un diccionario con los datos resultantes del proceso.

Tabla 12. Descripción de los datos codificados.

Variable	Valores	Código
Recurrencia	{Si, No}	{1,2}
Edad	15 <= Edad <= 100	
Peso	30 <= Peso <= 150	
Talla	140 <= Talla <= 180	
Índice de masa	20 <= IMC <= 50	
Superficie	1.38 <= SC <= 2.30	
Diagnóstico	Carcinoma ductal infiltrante	1
	Carcinoma lobulillar infiltrante	2
	Carcinoma ductal insitu	3
	Carcinoma medular infiltrante	4
	Carcinoma mixto	5
	Inflamatorio	6
	Mucinoso	7
	Carcinoma papilar	8
	Carcinoma de células gigantes	9
	Angiosarcoma de mama	10
Variable	**Valores**	**Código**
Grado:	Bajo	1
	Medio	2
	Alto	3
Dimension:	0 <= Dimension <= 9	
Subtipo	Luminal A	
	Luminal B	
	Her2+++	
	Triple Negativo	
Estadio:	0	1
	1	2
	02:00 a.m.	3
	2 B	4
	03:00 a.m.	5
	3 B	6
	4	7
Receptores	0 <= RE <= 100	
Receptores	0 <= RP <= 100	
Her2 Neu	+	1
	++	2
	+++	3
Proteína Ki-67	0 <= Ki67 <= 100	
Proteína P53	0 <= P53 <= 100	
Quimioterapia	Si	1
	No	0
Cirugía:	Mastectomia	1
	Cuadrantectomia	2
	Lumpectomia	3
	Ninguna	4
Radioterapia:	0,25,35	
Hormonoterapia:	Si	1
	No	0
Inmunoterapia:	Si	1
	No	0
Quimioprofilaxis	Si	1
	No	0

Se dividió al conjunto inicial de 217 registros en tres subconjuntos separados por subtipo de cáncer de mama: el conjunto Luminal, conjunto Her2 y el conjunto Triple Negativo; separado de esta manera de acuerdo a los marcadores tumorales reportados en los mismos atributos de los registros. En la tabla 13 se muestran algunos datos acerca de los subconjuntos resultantes.

Tabla 13. Estadísticas descriptivas de los subconjuntos encontrados.

	N	ATRIBUTOS	SIN RECURRENCIA	%	RECURRENCIA	%
HER2+++	44	22	31	70.5	13	29.5
LUMINAL	113	22	88	77.9	25	22.1
TRIPLE NEG	60	22	41	68.3	19	31.7
TOTALES	**217**		**160**		**57**	

V.5 Pérdida y recuperación de datos

Un problema común en el análisis de repositorios de datos clínicos es la existencia de datos faltantes, los cuales provocan problemas al momento de realizar análisis de clasificación e inferencias sobre los datos. Para comprender mejor el problema de la pérdida de datos y su solución, es necesario conocer primero los mecanismos por los cuales ocurre. Según el reporte del National Research Council para el tratamiento y prevención de datos faltantes (*The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials ; National Research*, 2010) y (Little & Donald, 2002), se documentan tres mecanismos de pérdida de datos:

- Datos perdidos completamente al azar (MCAR). En este mecanismo, la pérdida se da de manera completamente aleatoria, independiente de otros datos y del valor del dato perdido, como por ejemplo en el diseño intencional de pérdida de datos, las pruebas o mediciones resultan

costosas y solamente se pueden realizar a un número reducido de participantes elegidos al azar.

- Datos perdidos al azar (MAR). Aquí la distribución de los datos perdidos depende del valor de las otras variables o datos, pero no del valor del dato perdido en si, por ejemplo el valor de una variable que se elimina para registros o participantes que pertenezcan a un determinado rango de edad.
- Datos no perdidos al azar (MNAR). En éste mecanismo, la distribución de los datos perdidos depende del valor del dato perdido, como por ejemplo en la censura de datos que rebasen determinado valor.

En el mecanismo tipo MCAR, se asume que la distribución de los datos perdidos es igual a la de los datos existentes y la existencia de este mecanismo de pérdida no implica un sesgo en la inferencia o clasificación, por el contrario, con el mecanismo del tipo MAR se asume diferente la distribución y su existencia si implica sesgo a la hora de realizar inferencias, aunque se tiene la ventaja de que con este mecanismo los valores perdidos se pueden predecir a partir de los datos existentes.

La mayoría de los paquetes de software estadístico o de minería de datos, por default eliminan aquellos registros donde existen datos faltantes en alguna de sus columnas. Ésta solución podría producir intervalos de confianza más amplios y coeficientes más pequeños para el caso de regresión lineal y se tiene el potencial de introducir sesgo en la inferencia si el motivo de pérdida en determinada columna está en función del valor de otra columna en el mismo registro. En otros casos se tiene la opción de sustituir o imputar a aquel dato faltante por el promedio global de esa columna. La solución es simple y no afecta al promedio de la columna, pero ahora se tiene el problema de que el valor imputado no tiene relación con el resto de las variables en el registro. En el reporte del National Research Council (*The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials ; National Research*, 2010) sobre el tratamiento de datos faltantes en pruebas clínicas, se hace referencia al uso del

algoritmo EM, del análisis de regresión y a la imputación múltiple como métodos para la reposición de datos faltantes en pruebas clínicas.

La imputación múltiple consiste en rellenar para cada observación de dato faltante, con un conjunto S (por ejemplo S = 10) de valores imputados para cada observación, después se realiza un análisis de clasificación para el conjunto completado para cada repositorio completado de S. el resultado de cada repositorio de S es el promedio de cada una de las estimaciones resultantes del conjunto y la varianza es el promedio de las varianzas. Con este método se tiene la ventaja de que la varianza refleja la incertidumbre asociada al proceso de rellenar con valores desconocidos a través de imputar múltiples veces y se obtienen estimados más precisos que al utilizar imputación simple.

Con el método de imputación con el algoritmo EM, se obtiene un conjunto de datos casi igual al original con la diferencia de que el algoritmo ha introducido valores en los campos vacíos. La manera en que lo hace es suponiendo una distribución normal de cada una de las columnas, y tomando en cuenta a la matriz de covarianza del conjunto entero de datos, de tal manera que se sustituyen los campos vacíos por cantidades que el algoritmo calcula, en base a los valores no faltantes en el repositorio de datos. Al término de la ejecución del algoritmo EM, se obtiene un repositorio completo el cual es posible analizar con el resto de los algoritmos de minería de datos que requieren datos completos.

Además de los métodos de imputación, se tiene a los métodos bayesianos de clasificación y los algoritmos de clasificación por árbol de decisión que son resistentes o insensibles a la pérdida de información (Witten et al., 2005) (Duda et al., 2001) a diferencia del resto de los algoritmos de clasificación como K-NN, SVM, o perceptrón multicapa, que se ven beneficiados por los métodos de imputación (Farhangfar et al., 2008).

En el análisis que se presenta en éste reporte, hemos supuesto un mecanismo de pérdida del tipo MCAR y se han aplicado los métodos de imputación con el

algoritmo EM, Elastic Net, imputación múltiple utilizando OpenBUGS (Lunn & Spiegelhalter, 2009) y también se aplicó el análisis de clasificación sobre los repositorios sin imputar, es decir, con datos faltantes aplicando los algoritmos de clasificación por árbol de decisión y los métodos de Bayes Simple y Red Bayesiana.

V.6 Resultados

En esta sección se reportan los resultados en dos partes, en la primera se reporta el análisis con el repositorio con datos faltantes aplicando los algoritmos resistentes a la pérdida de datos. En la segunda parte se reportan los resultados del análisis efectuado sobre el repositorio con datos imputados y aplicando dos métodos de reducción de dimensionalidad, la reducción de coeficientes con Elastic Net y el método de ganancia en la tasa de la información.

Para los repositorios procesados por imputación EM y Elastic Net y aquellos repositorios donde no se aplicó imputación de ningún tipo, se utilizó validación cruzada de 10-fold para la ejecución de los algoritmos de minería de datos. Para el repositorio procesado por imputación múltiple con BUGS, los subconjuntos se dividieron aleatoriamente con una proporción de 75% para entrenamiento y 25% para prueba.

La evaluación de los resultados, se realizó analizando la matriz de confusión y calculando la sensibilidad del resultado, que es la proporción de verdaderos positivos, que en este análisis se consideró como positivo a aquellos registros pertenecientes a la clase recurrente y luego calculando la especificidad, que consiste en la proporción de verdaderos negativos en la matriz de confusión. En éste análisis, los verdaderos negativos son aquellos registros que pertenecen a la clase no-recurrente. Después de calcular los valores de sensibilidad y de especificidad, se calculó el producto de ambos valores y los resultados se reportaron en las tablas y gráficas de cada sección.

V.6.1 Primera parte: Repositorio incompleto

Para este análisis se utilizaron únicamente los algoritmos resistentes a la pérdida de información, además de considerar todos los atributos del repositorio. Se utilizó validación cruzada 10-fold para la ejecución de los algoritmos mencionados y para la evaluación. En la tabla 14 se muestran los resultados, la exactitud en la clasificación y la segunda columna es una medida de evaluación que es el producto de la sensibilidad y la especificidad. Las figuras 36 y 37 muestran las gráficas de la exactitud en la clasificación y las medidas de evaluación, respectivamente.

Tabla 14. Resultados al aplicar los algoritmos sobre repositorio incompleto

ALGORITMO	HER2+++	SE	LUMINAL	SE	TRIPLE NEG	SE
BAYES SIMPLE	81.81%	0.555	69.90%	0.258	56.60%	0.119
BAYES NET	86.30%	0.595	77.87%	0.115	68.30%	0.01
C4.5	79.54%	0.535	79.64%	0.2981	63.30%	0.092
CART	84.09%	0.521	79.64%	0.264	68.30%	0.01
RANDOM TREE	75%	0.401	76.90%	0.35	60%	0.087

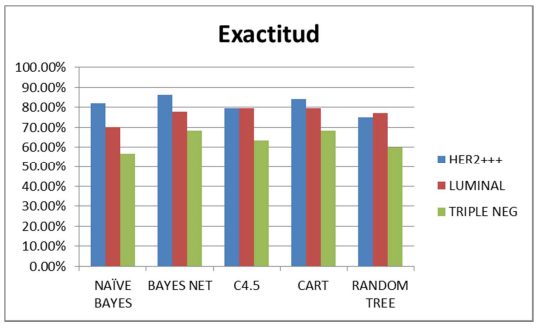


Figura 36. Gráfica de exactitud

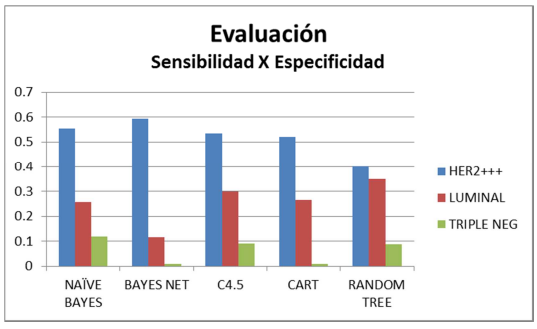


Figura 37. Medidas de evaluación.

Observando la exactitud, parece que los mejores resultados los provee Bayes Net, sin embargo, al obtener el producto SE, se concluye que el tratamiento del

repositorio incompleto no es satisfactorio. Un valor satisfactorio para el producto SE debería ser cercano a 1.

V.6.2 Segunda parte: Imputación de datos y reducción en la dimensionalidad

En esta segunda parte, se reportan resultados del análisis realizado sobre el conjunto de datos recuperado por los tres métodos mencionados con anterioridad, Imputación por EM, regresión con Elastic Net e Imputación Múltiple con Inferencia Bayesiana utilizando muestreo de Gibbs (BUGS por sus siglas en inglés). Además se evalúan dos métodos de reducción en la dimensionalidad, la reducción de coeficientes de Elastic Net y el método de reducción por ganancia en la tasa de la información.

V.6.2.1 Imputación con EM.

Para implementar el algoritmo EM como método de imputación, se utilizó la librería EM Imputation disponible para WEKA.

Al repositorio imputado con el algoritmo EM se le aplicaron dos métodos de reducción en dimensionalidad, Elastic Net y el método de ganancia en la tasa de la información. En la tabla 15 se muestran los resultados obtenidos con la reducción con Elastic Net. Los resultados se muestran separados para cada subconjunto, su exactitud en la clasificación y la medida de evaluación que corresponde al producto de sensibilidad y especificidad. Las gráficas de la tabla 15 se muestran en las figuras 38 y 39, que corresponden a las gráficas de exactitud y evaluación respectivamente.

Tabla 15. Resultados de repositorio imputado con algoritmo EM

	HER2	SE	LUMINAL	SE	TRIPLENEG	SE
NAÏVE BAYES	84.09%	0.63	82.30%	0.409	71.60%	0.199
LOGISTIC	88.63%	0.756	82.30%	0.409	71.60%	0.146
SVM	93.18%	0.806	82.30%	0.409	73.30%	0.285
MLP	88.63%	0.706	82.30%	0.409	71.60%	0.279
KNN	90.90%	0.738	82.30%	0.409	71.60%	0.24

C4.5	75.00%	0.4	82.30%	0.409	63.30%	0.091
CART	72.70%	0.33	80.50%	0.334	66.66%	0.01
RANDOM TREE	79.50%	0.53	82.30%	0.409	66.66%	0.295
GLMNET	87.00%	0.706	80.50%	0.334	69.00%	0.193

Se observa una mejoría notable en el producto SE, obteniendo los mejores resultados con SVM y KNN para el subconjunto HER2. Para Luminal y Triple Neg no se obtienen resultados favorables.

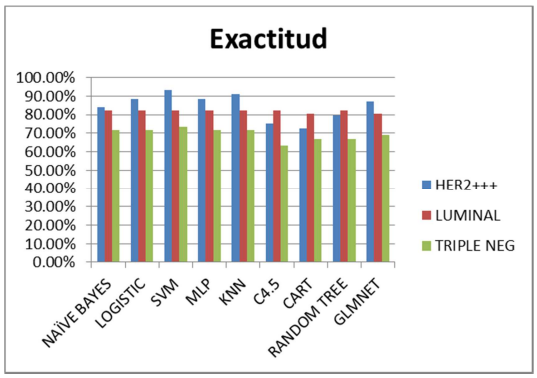


Figura 38. Gráfica de exactitud (EM)

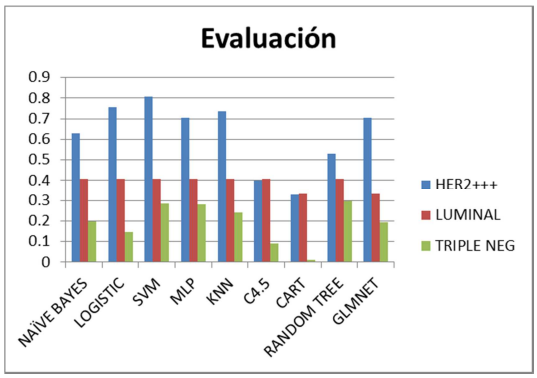


Figura 39. Gráfica evaluación (EM)

En la tabla 16 se muestran los resultados obtenidos de reducir el conjunto imputado por el algoritmo EM con el método de ganancia en la tasa de la información. Al igual que en la tabla 15, los resultados se dividen para cada subconjunto, su exactitud en la clasificación y la medida de evaluación, que es el producto de la sensibilidad y la especificidad. Las gráficas de la tabla 16 se muestran en las figuras 40 y 41.

Tabla 16. Resultados de conjunto EM reducido con Infogain

	HER2	SE	LUMINAL	SE	TRIPLENEG	SE
NAÏVE BAYES	86.36%	0.59	81.41%	0.372	70.00%	0.277
LOGISTIC	86.36%	0.59	81.41%	0.372	65.00%	0.318
SVM	88.63%	0.61	82.30%	0.409	66.66%	0.357
MLP	88.63%	0.61	81.41%	0.372	66.66%	0.225
KNN	93.18%	0.812	83.18%	0.415	73.30%	0.2503
C4.5	86.36%	0.59	82.30%	0.409	61.66%	0.131
CART	88.63%	0.664	79.60%	0.193	65.00%	0.001
RANDOM TREE	86.36%	0.645	74.30%	0.276	58.30%	0.3

De nuevo SVM y K-NN proveen los mejores resultados para HER2, y se siguen obteniendo bajos rendimientos para Luminal y Triple Negativo.

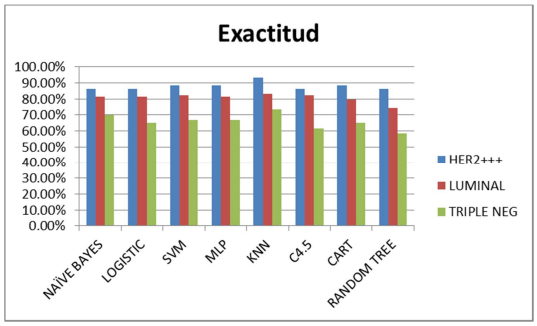


Figura 40. Gráfica exactitud EM-Infogain.

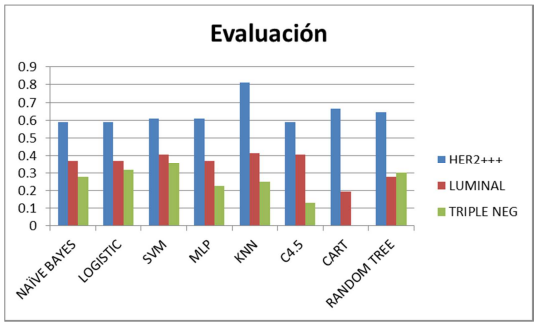


Figura 41. Gráfica evaluación EM-Infogain

V.6.2.2 Análisis sobre repositorio con datos imputados mediante Elastic Net.

En este apartado se detallan los resultados obtenidos al analizar el repositorio utilizando Elastic Net para realizar una imputación simple sobre los datos perdidos. Para la imputación, se tomaron en cuenta solamente a aquellas columnas en donde no había pérdidas, y con el fin de no propagar el error, las columnas con datos imputados no fueron utilizadas para imputar a las demás columnas con datos faltantes. Por último, con el repositorio resultante, se realizó un análisis con los algoritmos de minería de datos que requieren de completitud en los datos. Los

resultados se presentan a continuación reportando primero aquellos reducidos con el método de Elastic Net y después con el método de ganancia en la tasa de información.

En la tabla 17 se muestran los resultados de aplicar los algoritmos de minería de datos sobre el conjunto imputado mediante Elastic Net y que además se redujo la dimensionalidad con el algoritmo Elastic Net. Los resultados se presentan solamente para Her2 y Luminal, porque Elastic Net no desarrolló un modelo para Triple Neg, en términos de la exactitud en la clasificación y la medida de evaluación, en donde hemos utilizado el producto de sensibilidad y especificidad. Las gráficas de la tabla 17 se muestran en las figuras 42 y 43.

Tabla 17. Resultados de imputación con Elastic Net

	HER2	SE	LUMINAL	SE	TRIPLENEG	SE
NAÏVE BAYES	86.36%	0.649	81.41%	0.372		
LOGISTIC	88.63%	0.764	81.41%	0.372		
SVM	88.63%	0.719	82.30%	0.409		
MLP	88.63%	0.669	81.41%	0.372		
KNN	90.90%	0.744	78.70%	0.359		
C4.5	75.00%	0.401	82.30%	0.409		
CART	75.00%	0.287	79.60%	0.193		
RANDOM TREE	79.54%	0.536	74.30%	0.276		
GLMNET	82.00%	0.7	80.50%	0.381	68.00%	0.485

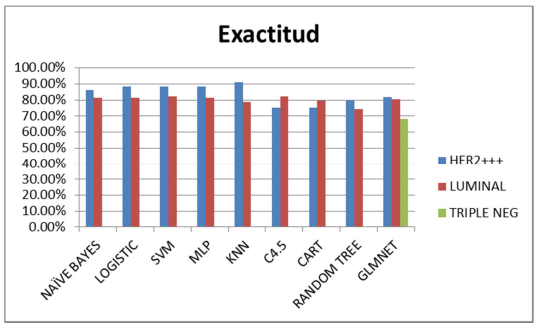


Figura 42. Gráfica exactitud ElasticNet

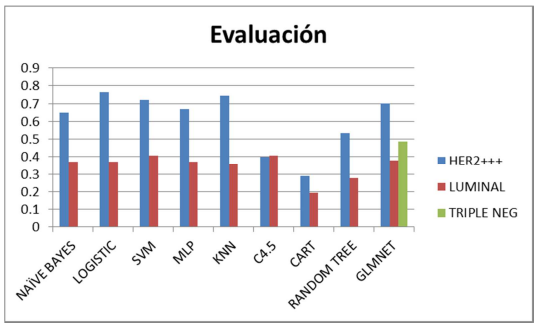


Figura 43. Gráfica evaluación ElasticNet

Se observa que los resultados son inferiores a los obtenidos anteriormente por EM como método de imputación.

En la tabla 18 se muestran los resultados de aplicar los algoritmos de minería de datos sobre el conjunto imputado mediante Elastic Net, en este caso se redujo la dimensionalidad con el método de ganancia en la tasa de la información. Los resultados se presentan por subconjunto de datos y para cada uno de ellos está la exactitud en la clasificación y la medida de evaluación, en donde hemos utilizado el producto de sensibilidad y especificidad. Las gráficas de la tabla 18 se muestran en las figuras 44 y 45.

Tabla 18. Resultados ElasticNet - Infogain

	HER2	SE	LUMINAL	SE	TRIPLENEG	SE
NAÏVE BAYES	84.09%	0.575	81.41%	0.372	63.33%	0.134
LOGISTIC	75.00%	0.496	81.41%	0.372	65.00%	0.179
SVM	86.36%	0.694	82.30%	0.409	71.66%	0.243
MLP	81.81%	0.629	81.41%	0.372	68.33%	0.269
KNN	88.60%	0.669	80.00%	0.359	70.00%	0.15
C4.5	84.09%	0.521	82.30%	0.409	61.66%	0.238
CART	84.09%	0.521	79.60%	0.193	63.33%	0.009
RANDOM TREE	88.60%	0.669	74.30%	0.276	53.33%	0.224

Se observa que los resultados obtenidos con este procedimiento son ligeramente inferiores a los obtenidos anteriormente.

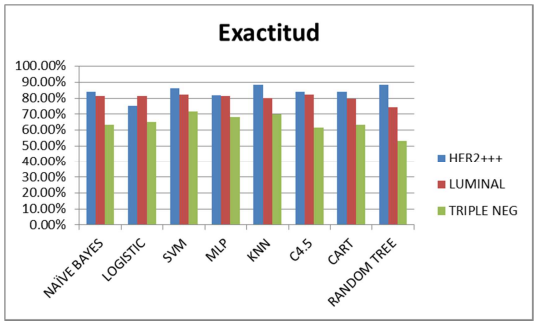


Figura 44. Gráfica exactitud ElasticNet-Infogain

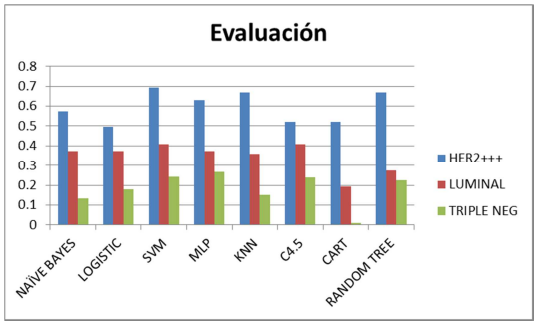


Figura 45. Gráfica evaluación ElasticNet-Infogain

V.6.2.3 Análisis sobre repositorios recuperados con imputación múltiple con BUGS

Del análisis efectuado con inferencia bayesiana utilizando muestreo de Gibbs, se obtiene un modelo de regresión logístico. En la tabla 19 se muestran los resultados de clasificación de los tres subconjuntos a partir del modelo generado por el análisis BUGS. En la figura 46 se muestra la gráfica de los resultados de la tabla 19.

Tabla 19. Resultados de repositorio recuperado con BUGS

	Exactitud	Sensibilidad	Especificidad	SE
HER2	95.4%	100.0%	94.0%	0.94
LUMINAL	81.4%	57.0%	89.0%	0.5
TRIPLE NEG	71.6%	100.0%	70.0%	.070

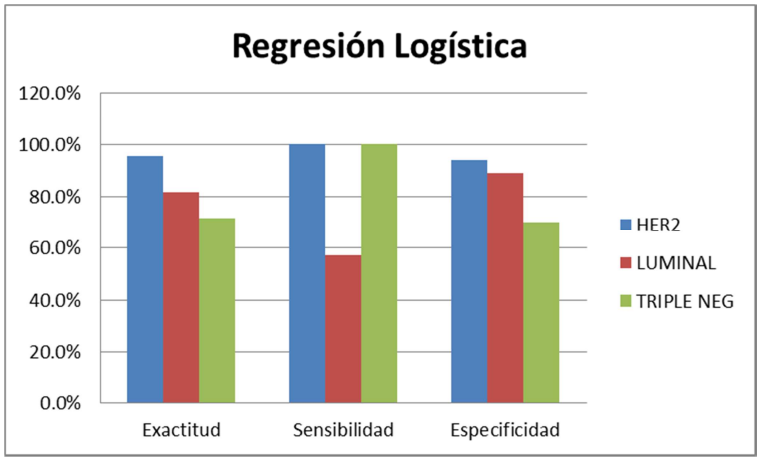


Figura 46. Resultados BUGS

En este caso, los peores resultados se obtuvieron para Luminal, con Triple Neg se obtuvo un SE = 0.70, mejor que todos los métodos anteriores

V.6.3 Comparación de resultados

En esta sección se presenta un reporte comparativo de resultados de los métodos de reducción de dimensionalidad, algoritmos de clasificación y de métodos de recuperación de datos. Para la comparación de los métodos de reducción de dimensionalidad, se reportan los mejores valores en exactitud de la clasificación sin importar el método de recuperación de datos perdidos ni el algoritmo de clasificación utilizado, así mismo, para la comparación de algoritmos de clasificación, se reportan los mejores resultados obtenidos sin importar el método de recuperación de datos ni el método de reducción de la dimensionalidad, por último, para la comparación de métodos de recuperación de datos perdidos, se reportan los mejores resultados independientemente del algoritmo de clasificación y del método de reducción de dimensionalidad utilizados.

Según la gráfica que se muestra en la figura 47, no hay diferencia entre utilizar el método de Elastic Net (o GLMNET según la gráfica) contra el de la ganancia en la tasa de la información, como métodos de reducción en la dimensionalidad, sin

embargo el uso de cualquiera de los anteriores demuestra un incremento en las medidas de exactitud y evaluación de los resultados.

En cuanto a los modelos de imputación, el uso del análisis BUGS, demuestra una mejora con respecto a los otros métodos cuando se utilizó el subconjunto Her2, según se aprecia en la tabla 22 y en las gráficas de la figura 51 y 52. Para el resto de los subconjuntos no se aprecia una mejoría notable.

Tabla 20. Comparación de métodos de reducción de dimensionalidad

COMPARATIVO DE METODOS DE REDUCCIÓN						
	HER2		LUMINAL		TRIPLE NEG	
	%	SE	%	SE	%	SE
ELASTIC NET	93.18%	0.806	82.30%	0.409	73.30%	0.485
INFOGAIN	93.18%	0.812	83.18%	0.415	73.30%	0.357
NINGUNO	86.30%	0.595	79.64%	0.35	68.30%	0.119

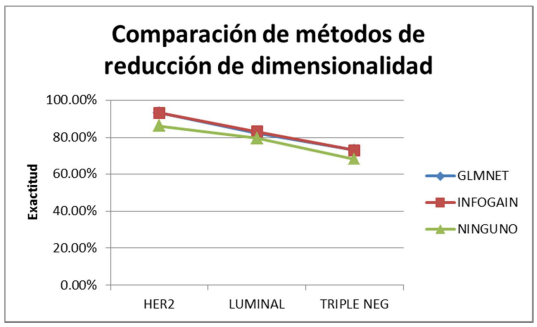


Figura 47. Gráfica de comparación exactitud

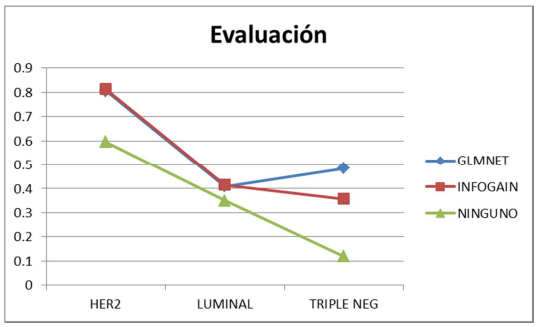


Figura 48. Gráfica de comparación de evaluación.

Tabla 21. Comparación de algoritmos de minería de datos.

COMPARATIVO DE ALGORITMOS						
	HER2	SE	LUMINAL	SE	TRIPLE NEG	SE
BAYES SIMPLE	86.36%	0.649	82.30%	0.409	71.60%	0.277
LOGISTIC	95.40%	0.94	82.30%	0.509	71.60%	0.7
SVM	93.18%	0.806	82.30%	0.409	73.30%	0.357
MLP	88.63%	0.706	82.30%	0.409	71.60%	0.279
KNN	93.18%	0.812	83.18%	0.415	73.30%	0.2503
C4.5	86.36%	0.59	82.30%	0.409	63.30%	0.238
CART	88.63%	0.664	80.50%	0.334	66.66%	0.0097
RANDOM TREE	88.60%	0.669	82.30%	0.409	66.66%	0.3
ELASTIC NET	87.00%	0.706	80.50%	0.381	69.00%	0.485

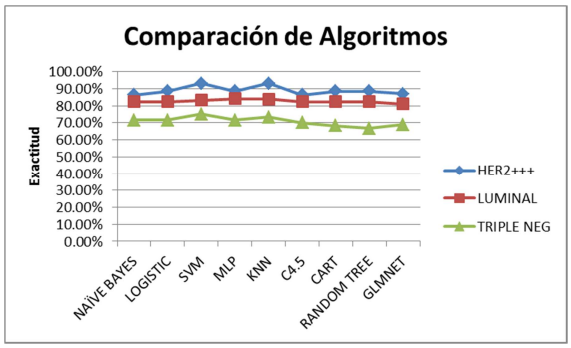


Figura 49. Gráfica comparación de exactitud entre algoritmos

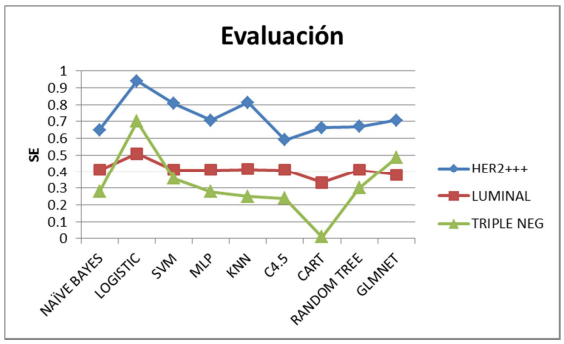


Figura 50. Gráfica comparación de evaluación entre algoritmos

Tabla 22. Comparación entre métodos de imputación.

COMPARATIVO DE METODOS DE IMPUTACION						
	HER2	SE	LUMINAL	SE	TRIPLE NEG	SE
EM	93.18%	0.812	83.18%	0.415	73.30%	0.357
ELASTIC NET	90.90%	0.764	82.30%	0.409	71.66%	0.485
BUGS	95.40%	0.94	81.40%	0.509	71.60%	0.7
INFERENCIA	86.30%	0.595	79.64%	0.35	68.30%	0.119

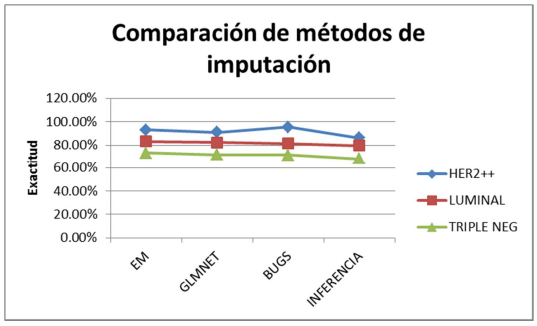


Figura 51. Gráfica comparativa de la exactitud entre métodos imputación.

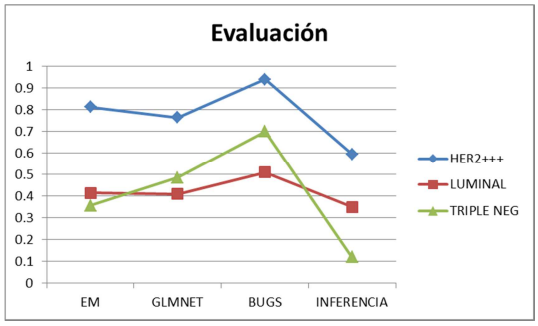


Figura 52. Gráfica comparativa de la evaluación de los métodos de imputación.

De la tabla 22 se observa que los mejores resultados se obtienen con inferencia Bayesiana mediante BUGS.

V.7 Modelos resultantes

En base a los resultados obtenidos, la facilidad de la implementación del modelo y además de la ventaja de la retroalimentación sobre los predictores, se eligió como modelo para la implementación al de regresión logística obtenida con BUGS.

En la tabla 23 se muestran los coeficientes resultantes de los modelos para los tres subconjuntos analizados: Her2, Luminal y Triple Negativo. La interpretación de los coeficientes para cada modelo se presenta a continuación.

Para el subconjunto Her2 podemos apreciar que los coeficientes correspondientes a los atributos de receptores de progesterona, proteína P53, receptores de estrógeno, índice de masa corporal, proteína Ki67 y superficie corporal, tienen valores cercanos a cero, lo cual se puede interpretar como atributos irrelevantes para el modelo de predicción para el subconjunto Her2.

En el caso de los coeficientes del modelo para el subconjunto Luminal, , los coeficientes menos relevantes para el modelo son: índice de masa corporal, el estadío, quimioprofilaxis, peso, radioterapia, superficie corporal, porcentaje de receptores de progesterona, proteína Ki67, proteína P53, edad, talla y el porcentaje de receptores de estrógeno.

Para el subconjunto Triple Negativo, los coeficientes del modelo se muestran en la tabla 23. Los atributos menos relevantes para este modelo son: superficie corporal, radioterapia, proteína P53, edad, proteína Ki67.

Tabla 23. Modelos resultantes para los tres subconjuntos de datos.

HER2			LUMINAL			TRIPLENEG	
COEFICIENTE	**VALOR**		**COEFICIENTE**	**VALOR**		**COEFICIENTE**	**VALOR**
INTERSECT	0.05332		INTERSECT	0.009258		INTERSECT	-0.06485
EDAD	-0.4418		EDAD	0.005932		EDAD	-0.00745
PESO	0.4791		PESO	-0.03099		PESO	0.07818
TALLA	-0.1588		TALLA	0.003477		TALLA	-0.07627
SUPCOR	-0.01398		SUPCOR	-0.02193		SUPCOR	-0.03001
IMC	0.01972		IMC	0.09108		IMC	-0.1836
GRADO	-0.5062		GRADO	0.2998		GRADO	1.113
DX	0.732		DX	-0.5869		DX	0.1041
MEDIDA	0.2895		MEDIDA	0.3767		MEDIDA	-0.1847
CIRUGIA	-0.2108		CIRUGIA	0.2711		CIRUGIA	0.9086
ESTADIO	1.213		ESTADIO	0.08382		ESTADIO	0.7287
SUBTIPO	0.4323		SUBTIPO	-0.6485		SUBTIPO	0.4126
ESTR	0.04086		ESTR	0.002014		ESTR	0.2767
PROG	-0.08747		PROG	-0.01826		PROG	0.05094
HER2	0.2571		HER2	-0.25		HER2	-0.07223
KI67	0.01598		KI67	-1.40E-02		KI67	0.001587
P53	0.04777		P53	-0.00934		P53	-0.00835
RADIO	-0.3462		RADIO	-0.02456		RADIO	-0.01833
QUIMIO	0.2179		QUIMIO	0.4814		QUIMIO	0.1687
HORMONO	-0.4475		HORMONO	-1.837		HORMONO	-0.4081
INMUNO	0.448		INMUNO	-0.6844		INMUNO	-0.06024
QUIMIOPROF	0.5518		QUIMIOPROF	0.03326		QUIMIOPROF	-0.6778

V.8 Resumen del capítulo

En el presente capítulo se presentó el análisis o el proceso de minería de datos aplicado sobre un conjunto de datos correspondientes a pacientes de cáncer de mama en tratamiento.

El primer paso fue extraer los datos necesarios a partir de los registros médicos. Los registros venían en forma de redacción sin una estructura de base de datos relacional ni como matriz, organizada en renglones y columnas. Con la ayuda del especialista en oncología, se procedió a leer los registros y a ir separando los datos en una hoja de cálculo.

Después de la extracción de los datos, se procedió a transformar y codificar los datos del paso anterior. A los atributos con valores nominales se les asignó como clave a un número entero, se discretizaron atributos continuos y se umbralizaron algunos atributos continuos. Cabe decir que no todos los atributos transformados llegan a formar parte del modelo final, ya que como hemos dicho anteriormente, el proceso de minería de datos es iterativo, y en algunos casos es necesario probar con diferentes transformaciones de los datos hasta obtener resultados satisfactorios.

Dada la proporción considerable de datos perdidos, se aplicaron tres métodos para la imputación de datos, el algoritmo EM, imputación por regresión e imputación múltiple con inferencia bayesiana utilizando muestreo de Gibbs. Por lo tanto se obtuvo al final de este paso tres nuevos repositorios con datos imputados.

El siguiente paso es la reducción de la dimensionalidad, o en otras palabras, la identificación de atributos relevantes para la clasificación. Se aplicaron y se compararon dos métodos para reducir y eliminar a los atributos menos relevantes para la clasificación, el método de ganancia en la tasa de la información y Elastic Net.

Por último, se analizaron algoritmos de minería de datos sobre los conjuntos resultantes de los pasos anteriores con la finalidad de seleccionar al modelo definitivo en base a los resultados de exactitud, sensibilidad y especificidad.

Capítulo VI. Conclusiones

Fue posible clasificar las observaciones de tumores como benignos y malignos en base a los datos observados y medidos en la clínica. Los datos referentes a la descripción morfológica de la observación fueron los más relevantes a la hora de hacer la clasificación, tal como se pudo constatar en la sección de reducción en la dimensionalidad en el capítulo IV y en los nodos de los árboles de decisión inducidos. El cuadrante de la mama y la lateralidad (izquierda, derecha) resultaron irrelevantes en los modelos.

El algoritmo Bayes Simple a pesar de haber tenido la menor proporción de registros clasificados correctamente, tuvo el mejor desempeño en cuanto a la evaluación de sensibilidad y especificidad fue el mejor de todos.

Para el caso del análisis de la recurrencia del cáncer de mama, del capítulo V, el peso aparece como un atributo con relevancia para la recurrencia del cáncer de mama cuando existe sobreexpresión del gen Her2/Neu, siendo el peso el único atributo que representa a un factor de riesgo extrínseco. En el análisis del subconjunto Luminal, el atributo correspondiente a la aplicación de hormonoterapia aparece como un atributo con relevancia para la recurrencia de la enfermedad. Según la revisión de expedientes, la ausencia de hormonoterapia se debe principalmente al abandono del tratamiento debido a efectos secundarios en el mismo. El análisis del subconjunto Triple Negativo tuvo resultados pobres en comparación de los otros dos subconjuntos pero a pesar de ello, es posible apreciar que la administración de capecitabina aparece como un atributo con relevancia para prevenir la recurrencia de la enfermedad.

Los modelos obtenidos por los algoritmos de árbol de decisión, además de haber obtenido buenos resultados, proveen un diagrama que puede ser interpretado directamente por un profesional de la salud. El modelo de regresión logística también tuvo muy buenos resultados, provee información sobre la relevancia de los atributos sobre la clasificación, es muy sencillo de implementarse en una hoja

de cálculo y además de clasificar, nos da como resultado a un componente probabilístico sobre la pertenencia a una u otra clase.

En cuanto a recuperación de datos perdidos, se obtuvieron resultados similares con los métodos de BUGS y algoritmo EM, pero con el método de BUGS se tiene más flexibilidad para definir las distribuciones estadísticas de los atributos. El algoritmo EM supone una distribución normal en todos los atributos.

Con el método de imputación simple se obtuvieron resultados inferiores que con los dos métodos anteriores.

Los algoritmos de minería de datos que son insensibles a la pérdida de datos arrojaron buenos resultados, sin embargo el desempeño en general aumentó con la recuperación de datos y con la reducción de la dimensionalidad.

Se obtuvieron resultados similares con los dos métodos de reducción de dimensionalidad que se probaron. Sin embargo, Elastic Net es más sencillo y rápido de utilizar. El parámetro lambda de Elastic Net permite calibrar la cantidad de predictores que se pretende obtener y el método es muy utilizado para la clasificación y reducción de datos biomoleculares en donde el número de dimensiones es mayor que la cantidad de registros u observaciones.

Se recomienda implementar un sistema informático de expediente médico con capacidad para soportar el almacenamiento de estudios de laboratorio e imagenología, además es necesario uniformar o normalizar la manera de reportar los resultados de los estudios, por ejemplo si éstos se informarán de manera porcentual como en el caso de los receptores para estrógeno, o umbralizados de manera positivo – negativo.

Para cuando no se cuenta con un sistema informático de expediente, se recomienda almacenar consistentemente por lo menos, los datos relevantes para el perfeccionamiento de los modelos.

Trabajo futuro

Es necesario repetir el análisis utilizando más datos, de preferencia integrarlos desde un sistema informático de expediente médico. Además se requiere realizar un muestreo estratificado para cada clase o categoría sobre la que se pretende hacer predicciones.

El análisis de los registros de mastografía se puede complementar con un estudio de procesamiento de imágenes sobre las mastografías directamente, con la finalidad de extraer los atributos automáticamente desde la imagen.

Se pueden repetir los análisis efectuados sobre los repositorios de pacientes en tratamiento, aumentando la cantidad de atributos, con la finalidad de mejorar la exactitud en la clasificación, especialmente para el subconjunto Triple Negativo.

Referencias bibliográficas

American College of Radiology (ACR). (2003). *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*. Reston, VA. Recuperado de: http://www.acr.org/Quality-Safety/Resources/BIRADS

Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, *58*(13), 10–110.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, *77*(2), 81–97.

Berner, E. (2010). *Clinical decision support systems: theory and practice*. Springer. 2a Edición. New York.

Bertani, A., Cappello, A., Benedetti, M. G., Simoncini, L., & Catani, F. (1999). Flat foot functional evaluation using pattern recognition of ground reaction data. *Clinical biomechanics*, *14*(7), 484–93.

Breast Cancer Detailed Guide. (2011). *American Cancer Society*. Recuperado de http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/index

Bright, K., Barghash, M., Donach, M., De la Barrera, M. G., Schneider, R. J., & Formenti, S. C. (2011). The role of health system factors in delaying final diagnosis and treatment of breast cancer in Mexico City, Mexico. *The Breast*, *20*, S54–9.

Buntine, W. (1994). Operations for learning with graphical models. *Journal of artificial intelligence research*, *2*(1), 159–225.

Burroni, M. (2004). Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study. *Clinical Cancer Research*, *10*(6), 1881–1886.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. *Networks*, *112*(1999), 121–130.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309–347.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113–27.

Duda, R. O., Hart, P. E., & Stork, D. (2001). *Pattern Classification.* Wiley-Interscience. 2a Edición. New York.

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*, 3692–3705.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. 17(3), 37–54.

Friedman, J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1–24.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer. 8a Edición. New York.

Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Journal of Computer Science*, *2*(2), 188–195.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, *11*(1), 10–18.

INEGI. (2009). Estadísticas a propósito del día mundial contra el cáncer. Recuperada de http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/contenidos/estadisticas/2009/cancer09.asp?s=inegi&c=2676&ep=8

Kharbat, F., Bull, L., & Odeh, M. (2007). Mining breast cancer data with XCS. *Proceedings of the 9th annual conference on Genetic and evolutionary computation - GECCO '07* (p. 2066). New York, ACM Press.

Leonard, J., Colombe, J., & Levy, J. (2002). Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, *18*(11), 1515–1522.

Little, R. J. A., & Donald, R. B. (2002). *Statistical Analysis with Missing Data* 2a Edición. Wiley-Interscience. New York

Lunn, D., & Spiegelhalter, D. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(June), 3049–3067.

Monk, A., & Howard, S. (1998). Methods & tools: the rich picture: a tool for reasoning about work context. *Interactions*, *5*(2), 21–30.

Orallo, J., Ramíirez, M., & Ferri, C. (2004). *Introducción a la minería de datos* (p. 680). Pearson Educación. Madrid

Ould, M. A. (1995). *Business Processes: Modelling and Analysis for Re-Engineering and Improvement* (p. 224). Wiley. Chichester

Federal Information Processing Standards (1993). Recuperado de: www.itl.nist.gov/fipspubs/idef02.doc

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 81–106.

Sboner, A. (2005). *User-tailored clinical decision support systems*. (Tesis de doctorado, Universidad de Trento, Italia).Recuperada de: http://static.digns.com/uploads/doctoral_school/documents/phd-thesis/XVII/sboner_andrea.pdf

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Recuperado de: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials ; National Research. (2010).*Social Sciences*. Washington, DC: The National Academies Press.

Organización Mundial de la Salud (2010). Recuperado de: http://www.who.int/cancer/detection/breastcancer/en/

Williams, M., & Williamson, J. O. N. (2006). Combining Argumentation and Bayesian Nets for Breast Cancer Prognosis. *Journal of Logic, Language and Information*, *15*(1), 155–178.

Wilson, S. W. (1995). Classifier Fitness Based on Accuracy. *Evolutionary Computation*, *3*(2), 149–175.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. 2a Edición. Morgan Kaufmann. New York