Trabajo de Investigación de Pregrado

# TAMIZAJE DE PACIENTES CON SOSPECHA DE CÁNCER DE MAMA MEDIANTE UN SISTEMA FUZZY USANDO VARIABLES CLÍNICAS.


Por:

**Olga Sofía Fragozo Díaz**

**Cristian Fernando Martínez Sierra**




Bucaramanga, mayo de 2010

# TAMIZAJE DE PACIENTES CON SOSPECHA DE CÁNCER DE MAMA MEDIANTE UN SISTEMA FUZZY USANDO VARIABLES CLÍNICAS.

**OLGA SOFÍA FRAGOZO DIÁZ**
**CRISTIAN FERNANDO MARTÍNEZ SIERRA**

**Trabajo de investigación desarrollado para optar al título de Ingeniero Electrónico**

**Director:**
**PhD. OSCAR GUALDRÓN GONZÁLEZ**

**Codirector:**
**MSC. EDWIN SANTIAGO ALFÉREZ BAQUERO**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER**
**FACULTAD DE INGENIERÍAS FISICO MECÁNICAS**
**ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES**
**BUCARAMANGA**
**2010**

## AGRADECIMIENTOS

*A Dios por darme la existencia y permitirme alcanzar los logros propuestos.*

*A mis padres por apoyarme durante el trascurso de mi formación.*

*A mis hermanos por aconsejarme oportunamente.*

*A mi director Oscar Gualdrón por guiar nuestro trabajo.*

*A Santiago Alférez por sus aportes y seguimiento en el desarrollo de este proyecto.*

*Al grupo de investigación CPS por acogernos y brindarnos las herramientas necesarias.*

*A todo el equipo médico por sus contribuciones y observaciones de acuerdo a su experiencia.*

*A Raquel Maldonado por su gestión para contactar a cada una de las pacientes que hacen parte de este estudio.*

*A mis amigos por acompañarme y transmitirme ánimos para la consecución de este logro.*

*Olga Sofía*

*Agradezco principalmente a DIOS que me dio la vida y me llenó de dones para alcanzar las metas propuestas.*

*A mis padres Pedro Rafael y Flor María quienes me brindaron apoyo durante todo este tiempo de formación.*

*A mi novia Diana Carolina que estuvo a mi lado y me apoyó en los momentos más difíciles.*

*A mis amigos que me apoyaron y contribuyeron en el desarrollo de este proyecto.*

*A Oscar Gualdrón quien dirigió este proyecto y nos guió en su desarrollo.*

*A Santiago Alférez quien desde el inicio de este proyecto fue nuestro principal apoyo y quien nos ayudo a encontrar las soluciones a los problemas que se presentaron.*

*Al grupo de investigación CPS que nos propuso este proyecto y que nos brindó las herramientas necesarias para su desarrollo.*

*Al grupo de investigación ONCOPAT del cual hacen parte el Doctor Alvaro Niño, la Doctora Olga Álvarez y el Doctor Luis Orozco quienes nos apoyaron aportando sus conocimientos y experiencia en este proyecto.*

*A Raquel Maldonado quien nos colaboró en la recolección de la información necesaria para realizar el estudio objeto de este trabajo.*

*A todas las personas que contribuyeron de forma directa o indirecta en el desarrollo de este trabajo y que olvido mencionar en este momento.*

*Cristian Fernando Martínez Sierra*

**Tabla de contenido**

## Lista de Figuras

## Lista de Tablas

**Lista de Anexos**

RESUMEN

TÍTULO: TAMIZAJE DE PACIENTES CON SOSPECHA DE CÁNCER DE MAMA MEDIANTE UN SISTEMA FUZZY EMPLEANDO VARIABLES CLÍNICAS[1]

AUTORES: Olga Sofía Fragozo Díaz; Cristian Fernando Martínez Sierra[2]

PALABRAS CLAVES: Cáncer de mama, Función de membresía, Factores de riesgo, Lógica Fuzzy, Clustering, Reconocimiento de patrones, Curva ROC.

DESCRIPCIÓN
El cáncer de mama es uno de los tipos de cáncer con mayor incidencia en la población femenina y con altas tasas de mortalidad. Con el objetivo de detectar tempranamente esta enfermedad se ha impulsado el uso de técnicas novedosas como la inteligencia artificial aplicada al diagnóstico médico. Este trabajo tiene como objetivo tamizar pacientes con sospecha de cáncer de mama mediante un sistema fuzzy usando variables clínicas. En primer lugar se estudia la relación de ciertas variables socio-demográficas, antecedentes familiares, hormonales y clínicas con el padecimiento de cáncer de mama, en base a la información recolectada por medio de encuestas a pacientes con sospecha de cáncer y su resultado de biopsia. De estas variables, se seleccionan las que presentan mayor correlación con el carcinoma. Se implementan varios sistemas de inteligencia artificial basados en lógica fuzzy, los cuales se fundamentan en clustering, reconocimiento de patrones y sistema de inferencia. Para la evaluación de estos sistemas se determinan los parámetros sensibilidad, especificidad y área bajo la curva ROC comparando el rendimiento entre estos. El mejor rendimiento lo obtiene el sistema de Inferencia fuzzy alcanzando un área bajo la curva ROC de 0.84. El reconocimiento de patrones obtiene el segundo mejor rendimiento con un área bajo la curva ROC de 0.82 y de último pero aún con un muy buen rendimiento se encuentra el clustering que alcanzó un área bajo la curva ROC de 0.8.

[1]Proyecto de grado
[2]Facultad de ingenierías físico mecánicas, Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones, Director: Ph.D Oscar Gualdrón González, Codirector: Msc. Edwin Santiago Alférez Baquero.

ABSTRACT

TITLE: SELECTIONS OF PATIENS WITH SOSPECTED BREAST CANCER USING A FUZZY SYSTEM AND CLINICAL VARIABLES [3]

AUTHORS: Olga Sofía Fragozo Díaz; Cristian Fernando Martínez Sierra[4]

KEYWORDS: Breast cancer, Membership function, Risk factors, Fuzzy logic, Clustering, Pattern recognition, ROC curve.

DESCRIPTION
The breast cancer is one of the most frequently occurring among other types of cancers affecting female population. If not diagnosed early, it has a great mortality rate. As the result of a search for an efficient method to detect the breast cancer on early stages of the disease development, the new technique of artificial intelligence in the area of medical diagnostics was introduced. This work concentrates on selecting the patients with the high possibility of breast cancer via fuzzy system using medical factors (variables) acquired from patient clinical data. The work started from studying the relationship among some variables and the breast cancer occurrences using the biopsy results and the information collected from the patients via questionary. Variables with most correlation were selected and used as an input to several fuzzy systems through clustering, pattern recognition, and an inference system. The validation of fuzzy systems was done by first finding parameters of their sensitivity and specificity, then calculating the area under the ROC. With these parameters it was the comparison between all systems. The best performance of the fuzzy systems was achieved with the area under the ROC of 0.84 by the inference system. Pattern recognition was the second best technique with the area under the ROC of 0.82. Finally clustering obtained a good performance with the area under the ROC of 0.8.

[3]Grade project
[4]Mechanical physics engineering faculty, School of Electrical Engineering, Electronics and Telecommunications, Director: Ph.D Oscar Gualdrón González, Codirector: Msc Edwin Santiago Alférez Baquero.

# INTRODUCCIÓN

El cáncer de seno es una de las enfermedades con mayor tasa de mortalidad en el mundo. Según la Organización Mundial de la Salud (OMS), el cáncer de mama contribuye con 548.999 muertes al año a nivel mundial [1]. En Colombia, el carcinoma de glándula mamaria ocupa el segundo lugar en incidencia [2], 33 de cada 100.000 mujeres padecen esta enfermedad y en Santander la tasa es de 36,42[5] [2]. En los países menos desarrollados la tasa varía desde 4 a 23,8; de manera que Colombia presenta mayor incidencia que este grupo. Estas alarmantes cifras muestran que la neoplasia de mama es un problema de salud pública, que requiere atención y prevención. Una de las razones que puede ser causante de este inconveniente es la asistencia tardía a centros médicos por parte de las pacientes y la inconstancia para el tratamiento de la enfermedad, al existir en Colombia ineficiencia en el sistema de salud pública y escaso cubrimiento a la población por parte de entidades promotoras de salud. Actualmente el método más utilizado para el diagnóstico de cáncer de seno es la mamografía, que ayuda a reducir la mortalidad hasta un 23 % en mujeres de 50 o más años de edad [3]. Es el método más confiable y preciso para la detección temprana del cáncer mamario [5]. Sin embargo, Colombia presenta un alto porcentaje de personas que no tienen acceso a dicho método debido a su alto costo. Otra alternativa para diagnosticar esta enfermedad es el examen clínico, que detecta el cáncer de seno en un 45% de los casos [6]. La efectividad de éste método, depende en gran parte de la experiencia de los médicos, lo cual genera incertidumbre.

En busca de una técnica que complemente el diagnóstico de la neoplasia de mama, se propone una metodología abordada a lo largo de este trabajo. Este libro se encuentra dividido en 7 capítulos. En el primer capítulo se relata la fundamentación teórica usada a lo largo del trabajo, iniciando por el concepto de factores de riesgo y factores pronósticos en la detección de cáncer de mama,

[5] Todas las tasas porcentuales están dadas por cada 100.000 habitantes por año.

pasando por las nociones médicas referentes a las regiones de incidencia de esta enfermedad, hasta finalizar con las bases de inteligencia artificial en el diagnóstico médico para la clasificación y reconocimiento. El segundo capítulo comenta la metodología seguida para el tamizaje de carcinoma de glándula mamaria usando las variables socio-demográficas, la historia familiar, los datos hormonales y la información clínica. El tercer capítulo describe el proceso de adquisición de la información, la cual está compuesta de tres partes fundamentales: un consentimiento informado, las encuestas que conllevan a la información socio-demográfica, hereditaria, hormonal y clínica, y el registro de la información en una base de datos. Aquí mismo, se explica la cantidad de variables recopiladas. Se detalla en el capítulo 4 la selección de variables, implementando tres técnicas: correlación lineal mediante los coeficientes de Pearson y Spearman, selección secuencial de variables y ranking a través de las pruebas T y Wilcoxon. Las variables seleccionadas, son usadas para implementar los algoritmos de clasificación y reconocimiento, como señala el capitulo 5. En este, se prueban cuatro metodologías: el agrupamiento con hard c-means, el agrupamiento con fuzzy c-means, el reconocimiento de patrones y los sistemas de inferencia fuzzy. En el transcurso de ésta sección, se validan cada uno de los algoritmos implementados, determinando los parámetros de sensibilidad, especificidad y área bajo la curva ROC para cada prueba. En el capítulo 6 se presentan las recomendaciones y conclusiones de toda la investigación y en el 7 se referencia toda la bibliografía que soporta el proyecto desarrollado.

# 1 MARCO TEÓRICO

Este trabajo de investigación, busca determinar las variables socio-demográficas, la historia familiar, los datos hormonales y la información clínica que permitan la mejor descripción a la hora de establecer una condición de anormalidad en las glándulas mamarias, especialmente en el posible diagnóstico de neoplasias malignas, además de, implementar un sistema de inteligencia artificial para el tamizaje de pacientes con sospecha de esta enfermedad. Por esto, antes de entrar en detalle con la metodología desarrollada, se tratarán conceptos fundamentales de interés iniciando con factores de riesgo y regiones de incidencia en la detección del carcinoma de glándula mamaria, hasta describir brevemente algunos métodos de inteligencia artificial aplicados al diagnóstico médico.

## 1.1 FACTORES DE RIESGO

Los factores de riesgo son características o atributos[6] que condicionan la probabilidad de presentar una enfermedad determinada. Dichos factores pueden estar presentes en población sana y aumentan el riesgo de tener la enfermedad. La identificación de los factores de riesgo es imprescindible para la prevención primaria. En el caso de los diferentes tipos de cáncer, cada uno tiene distintos factores de riesgo. Por ejemplo, la ventana estrogénica, entendida como el tiempo expuesto a hormonas, es un factor de riesgo para el cáncer de mama. Existen diferencias entre los factores de riesgo y los factores pronósticos, que son aquellos que predicen el curso clínico de un padecimiento una vez que la enfermedad está presente [7][8]. Un factor pronóstico puede suministrar información sobre la evolución que puede experimentar un enfermo en particular [9].

Se ha demostrado que existen ciertos factores de riesgo que pueden aumentar las

[6] Llamados a lo largo de este libro variables

posibilidades de contraer cáncer de mama, a pesar de que muchos médicos consideran esta enfermedad muy heterogénea, es decir, con manifestaciones clínicas variables [10]. En la tabla 1 se muestra una clasificación de algunos factores de riesgo para el cáncer de seno, de acuerdo con varias investigaciones.

| **Factor de riesgo** | | **Característica** |
|---|---|---|
| Socio-demográficos | Sexo | Ocurre unas 100 veces más en mujeres que en hombres [11]. |
| | Edad | Las tasas de incidencia aumentan enormemente en edades entre los 45 y 50 años [12] [12]. |
| | Raza | Las mujeres blancas son más propensas a padecer esta enfermedad que las de raza negra, aunque la mortalidad en éstas últimas es mayor, probablemente porque a ellas se les detecta en estadios más avanzados. Las que tienen menor riesgo de padecerlo son las mujeres asiáticas e hispanas [14]. |
| | Estado socioeconómico | Las mujeres de nivel socioeconómico alto tienen mayor riesgo de desarrollar cáncer de mama en comparación con las de estrato bajo [14]. |
| | Área de residencia | Aumento de la incidencia en mujeres de áreas urbanas comparadas con las que residen en áreas rurales [15][17]. |
| Hereditarios | Historia familiar de cáncer | Solo el 10% de mujeres diagnosticadas con esta lesión tienen antecedentes familiares positivos [17]. |
| | Historia genética | Los genes BRCA1 y BRCA2, según algunos estudios, muestran que entre el 50% y el 60% de mujeres que han heredado estos genes mutados pueden desarrollar el cáncer antes de los 70 años [14]. |
| Hormonales | Edad de la menarquía | La menarquía temprana (antes de los 12 años) ha sido asociada a un incremento del riesgo en un 10 a 20% [18][19]. |
| | Menopausia | Una menopausia tardía (después de los 54 años) incrementa el riesgo de desarrollar carcinoma mamario en un 3% por cada año que se tarde la menopausia [20][21]. |
| | Embarazo temprano | Mujeres con un embarazo a término y paridad aumentada tienen un riesgo disminuido a la mitad del presentado en las nulíparas [12]. |
| | Lactancia | Una lactancia prolongada ha demostrado ser un factor protector, disminuyendo el riesgo en un 3.4% por cada 12 meses de lactancia; adicionalmente, por cada parto el riesgo baja un 7.0% [22][23][24][26]. |

Tabla 1. Factores de riesgo en el cáncer de mama

## 1.2 INCIDENCIA DEL CÁNCER DE MAMA

En la Figura 1 se puede observar la subdivisión del seno en cuatro cuadrantes y la región pezón-areola. El cuadrante superior externo (CSE), se extiende desde el punto medio del pezón hasta la axila. El cuadrante superior interno (CSI), parte desde el punto medio del pezón hasta el encuentro con el esternón. Los otros dos cuadrantes (CIE y CII) parten desde el punto medio del pezón hasta la zona de curvatura del seno. Los cuadrantes CSE y CSI junto con la región pezón-areola es por donde mayor circulación arterial se da en la glándula mamaria, puesto que por esta área están ligadas arterias que nacen de la arteria axilar, mientras que la composición de los cuadrantes CIE y CII en la mayoría de los casos es tejido adiposo [27].

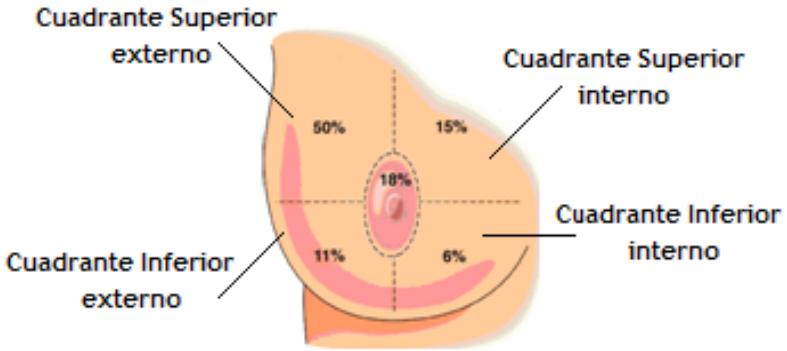



Figura 1. Regiones de la mama e incidencia del carcinoma de mama. Adaptada de [28].

Los valores porcentuales de zonas de incidencia conforme el cáncer de seno pueden verse en la Figura 1. El CSE tiene una incidencia del 50%, por encontrarse cerca de los ganglios axilares. La segunda zona de mayor incidencia, se da en el conjunto pezón-areola por encontrarse los ductos y vasos linfáticos. La tercera zona es la superior interna con un 15% y las dos zonas inferiores tienen una incidencia menor por la poca influencia de ganglios y ductos [28].

## 1.3 INTELIGENCIA ARTIFICIAL APLICADA AL DIAGNÓSTICO MÉDICO

En años recientes, los métodos de inteligencia artificial han sido ampliamente usados en diferentes áreas incluyendo aplicaciones médicas. Q. S. Ismail [29], hace referencia a un sistema experto, utilizando lógica fuzzy que toma como entradas variables tales como: PSA[7], edad y PV[8], y estima a la salida el riesgo de cáncer de próstata. El sistema diseñado resultó ser rápido, económico, sin riesgos y alta fiabilidad. Existen algunas publicaciones en el área de pronóstico de cáncer de mama con la ayuda de métodos de computación [30][31][32]. Ng et al. identifican variables obtenidas a partir de una encuesta realizada a cada paciente y del examen clínico al que fue sometido, denominándolas biodatos: edad del paciente, historia familiar de cáncer, terapia de reemplazo hormonal, edad de la menarquía del paciente, presencia de masas, dolor en el seno, edad de la menopausia, embarazo temprano, etc.

### 1.3.1 Clustering

Clustering o agrupamiento, es una técnica usada para determinar el número de clases en el que se puede dividir una variable teniendo una muestra de datos considerable. Hay dos tipos de agrupamiento de datos: “hard” (o crisp) para variables cuyos valores de membresía son 1 o 0, y “soft” (o fuzzy) para variables con valores de membresía entre 0 y 1. En ambos casos, como resultado del agrupamiento, se obtiene que los datos de un mismo grupo están más relacionados entre ellos que con los datos de otros grupos. Para determinar el grado en que dos datos están relacionados se usa una medida de similaridad que es la distancia euclidiana entre vectores en el espacio de características [34][35].

[7] Antígeno prostático específico.
[8] Volumen prostático.

Los métodos de agrupamiento, definen los grupos óptimos a través de un criterio denominado función objetivo, que mide el grado en el cual posibles grupos optimizan una suma ponderada de errores cuadrados entre datos y centros de grupos en el espacio de características [34]. Esta función busca minimizar la distancia euclidiana entre cada dato en un conjunto y su centro de grupo y maximizar la distancia euclidiana entre centros de grupo.

#### 1.3.1.1 Hard c-means (HCM)

Según Timothy Ross [34], este método de agrupamiento es usado para clasificar datos de modo que cada dato solo puede ser asignado a un grupo de datos. Por esta razón, estos grupos son llamados *particiones de datos*.

Considerando $\{A_1,\ A_2,\ A_3, \dots,\ A_C\}$ como $c$ particiones de $X$, se debe cumplir con las siguientes condiciones para obtener una partición "hard":

$$\bigcup_{i=1}^{C} A_i = X \qquad (1.1)$$

$$A_i \cap A_j = \emptyset \quad \text{para todo } i \neq j \qquad (1.2)$$

$$\emptyset \subset A_i \subset X \quad \text{para todo } i \qquad (1.3)$$

donde $c$ es el número de grupos y $X$ el universo de datos.

Usando la notación de teoría de funciones se pueden expresar las ecuaciones anteriores de la manera que sigue [34]:

$$\bigvee_{i=1}^{C} X_{A_i}(\boldsymbol{x}_k) = 1 \qquad (1.4)$$

$$X_{A_i}(\boldsymbol{x}_k) \wedge X_{A_j}(\boldsymbol{x}_k) = 0 \quad \text{para todo } i \neq j \qquad (1.5)$$

$$0 < \sum_{k=1}^{n} X_{A_i}(\boldsymbol{x}_k) < n \qquad (1.6)$$

donde la función característica $X_{A_i}(\boldsymbol{x}_k)$ es definida como:

$$X_{A_i}(\boldsymbol{x}_k) = \begin{cases} 1, & x_k \in A_i \\ 0, & x_k \notin A_i \end{cases} \qquad (1.7)$$

Por simplicidad en notación, la asignación de membresía del *j*-ésimo punto de dato en el *i*-ésimo grupo, es definido por $X_{ij} \equiv X_{A_i}(x_j)$. A continuación, se define una matriz U con c filas y n columnas compuesta por elementos $X_{ij}$ $(i = 1,2,..,c; j = 1,2,\dots,n)$. Entonces, se define un espacio de $c$ particiones para $X$ como el siguiente conjunto de matrices:

$$M_C = \left\{ U \middle| X_{ij} \in \{0,1\}, \sum_{i=1}^{C} X_{ik} = 1\,, 0 < \sum_{k=1}^{n} X_{ik} < n \right\} \qquad (1.8)$$

Así, cualquier matriz $U \in M_C$ es una *hard c-partition*. La cardinalidad, es decir, el número de posibles agrupaciones de los datos de cualquier *hard c-partition*, $M_C$, es [34]:

$$n_{M_C} = \left(\frac{1}{C!}\right)\left[\sum_{i=1}^{C} \binom{C}{i} (-1)^{C-i}\,.\, i^n\right] \qquad (1.9)$$

Seguidamente se selecciona la mejor partición, para esto se evalúa la función objetivo, denotada $J(U, v)$, donde $U$ es la matriz de partición y el parámetro $v$ es un vector de centro de grupo. Esta función objetivo está dada por:

$$J(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{C} X_{ik} (d_{ik})^2 \qquad (1.10)$$

donde $d_{ik}$ es una medida de distancia euclidiana (en un espacio *m*-dimensional de características,$R^m$) entre la *k*-ésima muestra de datos $x_k$ y el *i*-ésimo centro de grupo $v_i$, dado por:

$$d_{ik} = d(x_k - v_i) = \|x_k - v_i\| = \left[\sum_{j=1}^{m} \left(x_{kj} - v_{ij}\right)^2\right]^{\frac{1}{2}} \qquad (1.11)$$

Debido a que cada muestra de datos requiere *m* coordenadas para localizarse en el espacio $R^m$, cada centro de grupo también requiere m coordenadas para describir su localización en el mismo espacio. Por lo tanto, el *i*-ésimo centro de grupo es un vector de longitud *m*:

$$v_i = \{v_{i1}, v_{i2},\ v_{i3}, \dots, v_{im}\}$$

donde la *j*-ésima coordenada es calculada por:

$$v_{ij} = \frac{\sum_{k=1}^{n} X_{ik}.\, x_{kj}}{\sum_{k=1}^{n} X_{ik}} \qquad (1.12)$$

El siguiente paso es buscar la partición óptima, $U^*$, que produzca el mínimo valor de la función $J(U, v)$. Esto es, $J(U^*, v^*) = \min J(U, v)$.

A continuación se describe paso a paso un algoritmo desarrollado por Bezdek [36] para optimizar la búsqueda de la mejor partición:

1. Fijar el valor de $c$ (2≤$c$≤n) e inicializar la matriz $U$:

$$U^{(0)} \in M_C$$

2. Calcular el valor de los centros de grupo:

$$\{V_i{}^{(r)}\ con\ U^{(r)}\}$$

3. Actualizar la matriz $U$:

$$X_{ik}{}^{(r+1)} = \begin{cases} 1, & d_{ik}{}^{(r)} = min\{d_{jk}{}^{(r)}\}\ para\ todo\ j \in C \\ 0, & de\ otra\ forma \end{cases} \qquad (1.13)$$

4. Si

$$\left\|U^{(r+1)} - U^{(r)}\right\| \leq \varepsilon\ (nivel\ de\ tolerancia) \qquad (1.14)$$

se detiene el algoritmo; de otro modo se hace $r = r + 1$ y se regresa al paso 2.

En la Figura 2, se muestra la idea de “hard clustering” en un espacio tridimensional (m=3) de características, en esta figura se observa cada grupo de datos en una forma hiperesférica con un hipotético centro de grupo [34].

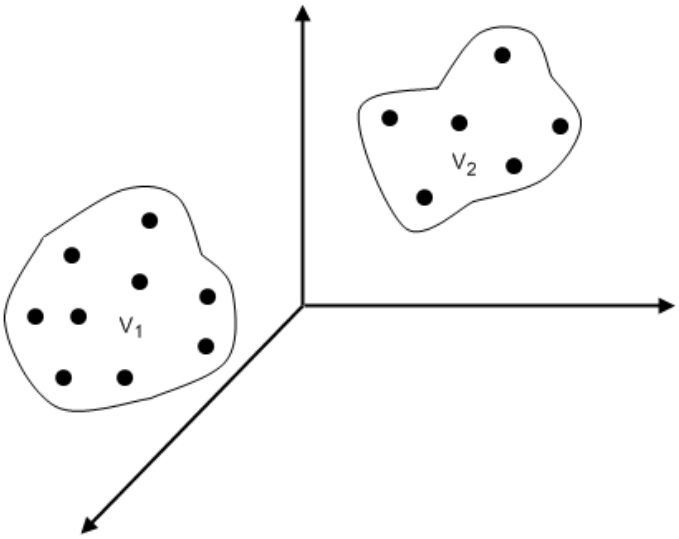



Figura 2. “Hard Clustering” en un espacio tridimensional. Adaptada de [34].

#### 1.3.1.2 Fuzzy c-means (FCM)

Al igual que el método hard c-means, este método busca clasificar $n$ datos en $c$ particiones. Para introducir este método, se define una familia de conjuntos fuzzy $\{A_i,\ i = 1,2,\dots,c\}$ como una partición fuzzy sobre un universo de datos, $X$. Debido a que los conjuntos fuzzy manejan grados de membresía, se puede extender la idea de clasificación “hard” a “fuzzy”, es decir, se pueden asignar valores de membresía a los datos para cada conjunto. De ahí que, cada dato puede tener una membresía parcial en más de una clase [37].

El valor de membresía que el *k*-ésimo dato tiene en el *i*-ésimo grupo se expresa con la siguiente notación [34]:

$$\mu_{ik} = \mu_{A_i}(\boldsymbol{x}_k) \ \in \ [0,1]$$

Con la restricción de que la suma de todos los valores de membresía, para cada dato, en todas las clases debe ser la unidad, es decir [34]:

$$\sum_{i=1}^{C} \mu_{ik} = 1 \quad para\ k = 1,2,\dots,n \qquad (1.15)$$

Al igual que en clasificación “hard” no pueden haber grupos vacios ni grupos que contengan todos los puntos de datos, esto es:

$$0 < \sum_{k=1}^{n} \mu_{ik} < n \qquad (1.16)$$

Debido a que los datos pueden tener parcial membresía en más de una clase, la regla descrita en la ecuación 1.5 no se cumple en el caso de clasificación fuzzy:

$$X_{ik} \wedge X_{jk} \neq 0 \qquad (1.17)$$

Ahora se define una familia de matrices de particiones fuzzy, $M_{fc}$, para la clasificación involucrando $c$ clases y $n$ puntos de datos.

$$M_{fc} = \left\{ U \middle| \mu_{ik} \in [0,1], \sum_{i=1}^{c} \mu_{ik} = 1\,, 0 < \sum_{k=1}^{n} \mu_{ik} < n \right\} \qquad (1.18)$$

Cualquier matriz $U \in M_{fc}$ es una *fuzzy c-partition*. Debido al solapamiento entre clases y al infinito número de valores para describir su membresía, la cardinalidad de $M_{fc}$ es infinita [34].

La función objetivo se define para fuzzy c-means, como:

$$J_m(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{C} (\mu_{ik})^{m'} (d_{ik})^2 \qquad (1.19)$$

donde

$$d_{ik} = d(x_k - v_i) = \left[ \sum_{j=1}^{m} \left( x_{kj} - v_{ij} \right)^2 \right]^{\frac{1}{2}} \qquad (1.20)$$

siendo $\mu_{ik}$ la membresía del *k*-ésimo punto de dato en el *i*-ésimo grupo y $m'$ el parámetro de ponderación que controla la cantidad de incertidumbre en el

proceso de clasificación [36].

Cada coordenada del centro para cada grupo puede ser calculada de la siguiente forma:

$$v_{ij} = \frac{\sum_{k=1}^{n} {\mu_{ik}}^{m'} . x_{kj}}{\sum_{k=1}^{n} {\mu_{ik}}^{m'}} \qquad (1.21)$$

donde *j* es una variable sobre el espacio de características, con $j = 1,2, \dots, m.$

La mejor partición fuzzy será la que genere el menor valor de la función objetivo, es decir:

$${J_m}^{*}(U^{*}, v^{*}) = min\, J_m(U, v)$$

Los pasos para encontrar la mejor partición en un algoritmo de calsificación fuzzy son los siguientes [34]:

1. Fijar el valor de $c$, seleccionar un valor para el parámetro $m' \in [1, \infty)$ e inicializar la matriz de partición $U^{(0)}$.
2. Calcular los $c$ centros de grupo $\{V_i^{(r)}\}$.
3. Actualizar la matriz de partición:

$$\mu_{ik}^{(r+1)} = \left[ \sum_{j=1}^{c} \left( \frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{\frac{2}{(m'-1)}} \right]^{-1} \quad para\ I_k = 0 \qquad (1.22)$$

o

$$\mu_{ik}^{(r+1)} = 0 \quad para\ todas\ las\ clases\ i\ donde\ i \in \tilde{I}_k \qquad (1.23)$$

donde

$$I_k = \left\{ i \middle| 2 \le c \le n; d_{ik}^{(r)} = 0 \right\} \qquad (1.24)$$

y

$$\tilde{I}_k = \{1,2, \dots, C\} - I_k \qquad (1.25)$$

y

$$\sum_{i \in I_k} \mu_{ik}^{(r+1)} = 1 \qquad (1.26)$$

4. Si $\left\|U^{(r+1)} - U^{(r)}\right\| \leq \varepsilon_L$, se detiene el proceso iterativo, de otra forma se hace $r = r + 1$ y se regresa al paso 2.

### 1.3.2 Reconocimiento de patrones

El reconocimiento de patrones es un proceso en el cual se busca identificar características de un conjunto de datos sobre aquellas previamente conocidas. Las características conocidas se obtienen a través de procesos de clasificación, tal como el clustering. El propósito del reconocimiento de patrones es asignar cada nueva entrada a un grupo o patrón de datos previamente clasificado [34][38].

Los datos usados para diseñar el sistema de reconocimiento de patrones son divididos en dos grupos: datos para el diseño y datos para la prueba, el primer grupo es usado para determinar los parámetros algorítmicos del sistema y el segundo para probar el rendimiento del sistema [39].

La clasificación y el reconocimiento de patrones son procesos complementarios, por una parte la clasificación se encarga de agrupar los datos encontrando las características de cada uno de los grupos, y el reconocimiento asigna un conjunto de nuevos datos a cada grupo con base en la información obtenida de las características. En pocas palabras, la clasificación define los patrones y el reconocimiento asigna datos a estos [34].

En este trabajo se resumen dos métodos para realizar el reconocimiento de patrones con más de una característica.

#### 1.3.2.1 Clasificador del vecino más cercano (NNC[9])

[9] En inglés Nearest Neighbor Classifier

Para ilustrar este método se consideran *m* características de cada dato, de modo que cada dato $(x_i)$ es un vector de características, $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$.

Ahora, se supone que se tienen *n* muestras de datos conocidas en un universo $X = \{x_1, x_2, \dots, x_n\}$. Utilizando un método de agrupamiento se encuentran los $c$ grupos adecuados para los datos en estudio. Si se usa clasificación fuzzy, se debe hacer el proceso de convertir la partición resultante en una partición "hard" donde se cumplan las siguientes condiciones [34]:

$$\bigcup_{i=1}^{C} A_i = X$$

$$A_i \cap A_j = \emptyset \quad \text{para todo } i \neq j$$

Posteriormente, si se tiene un nuevo dato $\boldsymbol{x}$, entonces el clasificador del vecino más cercano está dado por la siguiente medida de distancia:

$$d(x, x_{k^*}) = \min_{1 \leq k \leq n} \left( d(x, x_k) \right) \qquad (1.27)$$

Para cada uno de los *n* datos donde $x_{k^*}$ pertenece al grupo $A_i$. Esto significa que $x$ y $x_{k^*}$ pertenecen al mismo grupo.

#### 1.3.2.2 Clasificador del centro más cercano (NCC[10])

De nuevo se inicia con *n* datos conocidos, $X = \{x_1, x_2, \dots, x_n\}$, y cada dato es descrito por *m* características. Por medio de un método de agrupamiento se encuentran los $c$ grupos en los que se clasifican los datos. Cada grupo tiene un centro, de modo que: $v = \{v_1, v_2, \dots, v_c\}$, donde $v$ es un vector de $c$ centros de clases. Si se tiene un nuevo dato $x$, el clasificador del centro más cercano esta dado por:

$$d(x, v_{k^*}) = \min_{1 \leq k \leq c} \left( d(x, v_k) \right) \qquad (1.28)$$

[10] En inglés Nearest Center Classifier

Por tanto, $x$ pertenece al grupo $A_i$ que tiene como centro $v_{k^*}$

### 1.3.3 Sistema de Inferencia Fuzzy

Un sistema de inferencia fuzzy es un proceso que formula la asignación de una determinada entrada a una salida usando lógica fuzzy. Esta determinación, proporciona un punto de partida desde el cual se pueden tomar decisiones o distinguir patrones [34][39].

Pueden implementarse varios tipos de sistemas de inferencia fuzzy, los más usados son: Mamdani y Sugeno. Estos sistemas se diferencian en los pasos que siguen para determinar la salida. En el sistema de inferencia tipo Mamdani, se espera que la salida este representada por funciones de membresía. De esta manera, después del proceso de agregación existe un conjunto fuzzy para cada variable de salida que seguidamente se defuzzifica. Los sistemas de inferencia tipo Sugeno, en lugar de integrar a través de la función de salida para encontrar el centro, utilizan el promedio ponderado de los puntos de algunos datos. En general, los sistemas de tipo Sugeno se pueden utilizar para modelar cualquier sistema de inferencia en el que las funciones de membresía de salida son lineales o constantes [34].

Una regla típica en un modelo fuzzy tipo Sugeno tiene la forma:

si la entrada 1 es $x$ y la entrada 2 es $y$, la salida es $z = f(x, y)$

donde $z = f(x, y)$ es una función crips a la salida de cada regla. Generalmente, $f(x, y)$ es una combinación lineal de $x$ y $y$. El nivel de salida de cada regla $z_i$ es ponderado por el peso $w_i$ de la regla. Por ejemplo, para una regla AND con una entrada 1 igual a $x$, y una entrada 2 igual a $y$, el peso es:

$$w_i = min(F_1(x), F_2(y)) \qquad (1.29)$$

donde, $F_1(x)$, $F_2(y)$, son funciones de membresía para las entradas 1 y 2. La

salida final del sistema es la media ponderada de todas las reglas de salida, y se obtiene a partir de la siguiente ecuación:

$$Salida\ final = \frac{\sum_{i=1}^{N} w_i\, z_i}{\sum_{i=1}^{N} w_i} \qquad (1.30)$$

Una regla Sugeno opera como se muestra en el siguiente diagrama:

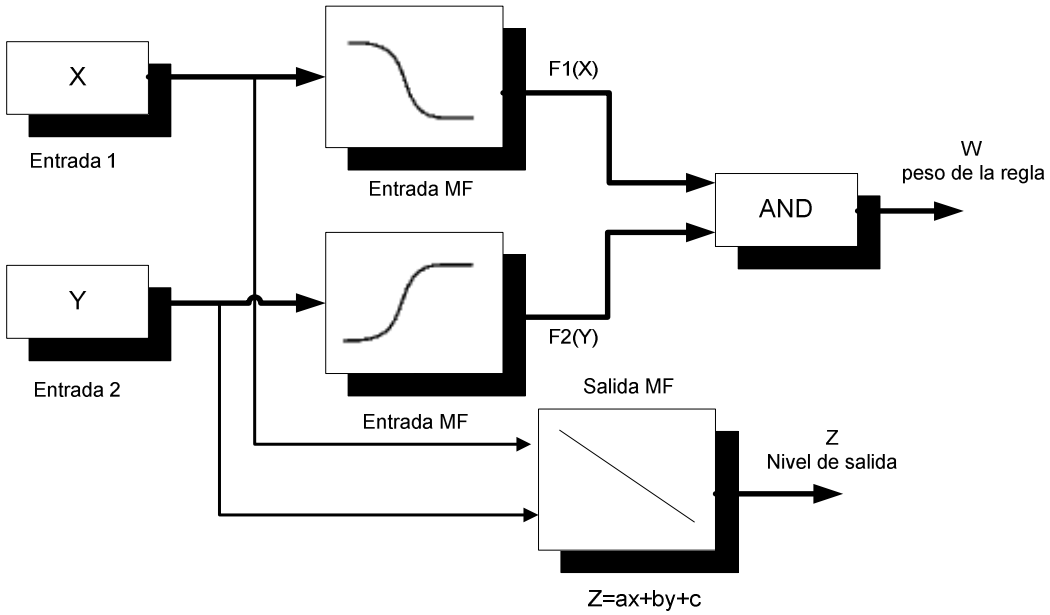



Figura 3. Diagrama de representación de una regla Sugeno. Adaptada de [41].

Finalmente, al comparar los tipos de sistema de inferencia fuzzy, se encuentra que el tipo Sugeno es computacionalmente eficiente, funciona bien con técnicas lineales, de optimización y adaptación, garantiza la continuidad de la salida y es muy adecuado para el análisis matemático; mientras que el tipo Mamdani, presenta ventajas respecto a que es intuitivo, cuenta con una amplia aceptación y se adapta muy bien a la intervención humana.

Posteriormente, se detalla el proceso que debe seguirse para construir un sistema de inferencia fuzzy, usado en el desarrollo de este proyecto. Una estructura básica de un sistema de dos entradas una salida, es como se muestra a continuación:

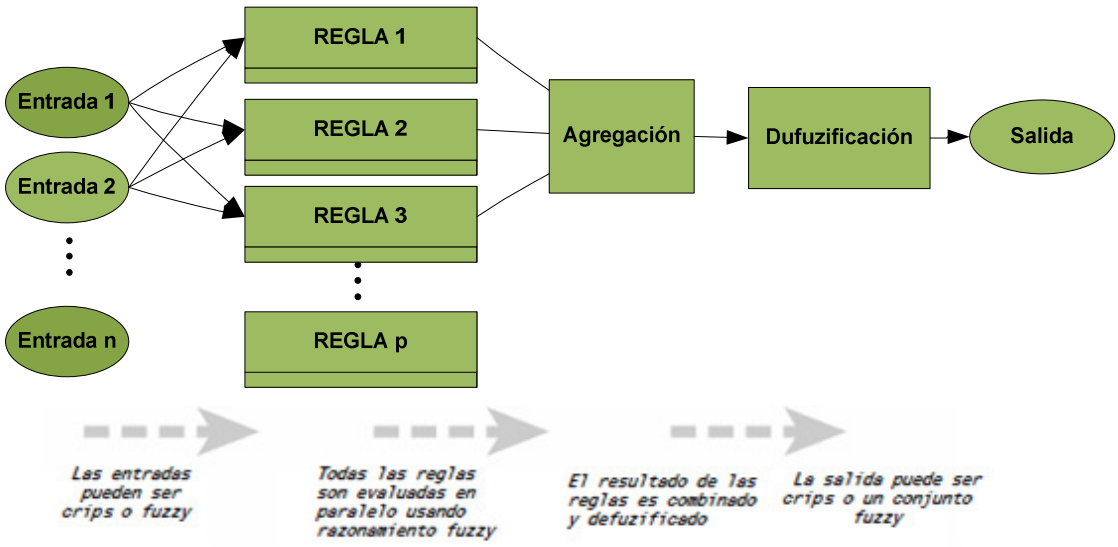



Figura 4. Diagrama de bloques FIS de dos entradas-una salida. Adaptada de [41].

Como se observa en la Figura 4, la información fluye de izquierda a derecha, de dos entradas a una sola salida. La naturaleza paralela de las reglas es uno de los aspectos más importantes en los sistemas de lógica fuzzy. En vez de cambiar bruscamente entre los modos de las entradas con base en puntos de ruptura, la lógica fluye desde las regiones donde el comportamiento del sistema es dominante por una regla u otra.

El proceso de crear un sistema de inferencia fuzzy viene dado por los siguientes pasos [42]:

1. *Fuzificar las variables de entrada*. Consiste en tomar cada entrada y determinar los grados apropiados de ella, a través de funciones de membresía.
2. *Aplicar el operador fuzzy (OR o AND) a los grados de las entradas*. Después de que las entradas son fuzificadas, es conocido el grado en que cada parte de la entrada se cumple para cada regla. Si una regla dada tiene más de una variable, el operador fuzzy es aplicado para obtener un número

que representa el grado de dichas variables en esa regla. Este número se aplica luego a la función de salida obteniendo un único valor. Las operaciones lógicas AND y OR son sustituidas por operaciones min (mínimo) o producto y max (máximo) o probor (OR probabilístico) respectivamente. El método OR probabilístico, también conocido como suma algebraica se calcula según la ecuación:

$$probor\ (a, b) = a + b - ab \qquad (1.31)$$

3. *Aplicar el método de Implicación*. Inicialmente, se debe determinar el peso de la regla que corresponde a un valor entre 0 y 1, dependiendo de la lógica de las entradas. Después de establecida una ponderación adecuada a cada regla, el método de implicación puede aplicarse. La entrada es un número único y la salida es un conjunto fuzzy. La implicación es aplicada para cada regla y consiste en dos pasos: min (mínimo), que trunca la salida del conjunto fuzzy, y producto que escala la salida del conjunto fuzzy.
4. *Agregar todas las salidas.* Las reglas de un FIS[11], deben ser combinadas de alguna manera con el fin de tomar una decisión. El proceso de agregación recibe la lista de funciones de salida truncadas obtenidas por la fase de implicación para cada regla, entregando un conjunto fuzzy para cada variable de salida. Son usados tres métodos: max (máximo), probor (OR probabilístico) y sum (la suma de los conjuntos de salida de cada regla).
5. *Defuzificación.* La entrada para el proceso de defuzificación es un conjunto fuzzy y la salida es un único número. El método de defuzificación más común es el cálculo del centroide, que devuelve el centro del área bajo la curva (función de membresía). Existen otros métodos de defuzificación como: media del máximo, el máximo más grande, y el máximo más pequeño.

[11] Sistema de Inferencia Fuzzy

## 1.4 PARÁMETROS DE VALIDACIÓN: ESPECIFICIDAD, SENSIBILIDAD Y ÁREA BAJO LA CURVA ROC

La curva ROC fue desarrollada inicialmente durante la segunda guerra mundial para fines militares. El análisis ROC ha sido usado desde entonces en medicina, radiología y otras áreas para juzgar la habilidad de discriminación de varios métodos estadísticos que combinan varias características, resultados de pruebas, etc, para propósitos predictivos [43].

En medicina, el análisis de la curva ROC ayuda a evaluar el desempeño diagnóstico de pruebas médicas para discriminar casos no saludables de casos saludables, es decir, proporciona una representación global de la exactitud diagnóstica.

Gráficamente la curva ROC es representada por la fracción de verdaderos positivos (FVP) en ordenadas contra la fracción de falsos positivos (FFP) en abscisas, o lo que es equivalente la sensibilidad contra 1-especificidad. La sensibilidad es un parámetro que describe la probabilidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un paciente enfermo se obtenga en la prueba un resultado positivo. De ahí que, sea llamada también fracción de verdaderos positivos, tal como se muestra en la tabla 2. La especificidad describe la probabilidad de clasificar correctamente un individuo sano, es decir, la probabilidad de que para un paciente sano se obtenga un resultado negativo. La especificidad es también conocida como tasa de verdaderos negativos, definida en la Tabla 2 [44].

La curva ROC puede ser dibujada punto a punto a partir de los valores de sensibilidad y especificidad obtenidos a medida que se varía un umbral, que

corresponde al traslape de dos distribuciones normales, una para las pacientes enfermas y otra para las pacientes sanas, razón por la cual, existe un compromiso entre sensibilidad y especificidad. Es decir, si se modifica el umbral para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad. Esta curva toma valores desde 0 hasta 1 en sus dos ejes, como se representa en la Figura 5. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo [45]. Esto sugiere que el área bajo la curva ROC[12] puede emplearse como un índice conveniente de la exactitud global de una prueba, donde la exactitud máxima corresponde a un valor de AUC de 1, y la mínima a una de 0.5, que corresponde a la diagonal. Si el área es menor de 0.5 debe invertirse el criterio de positividad de la prueba [44].

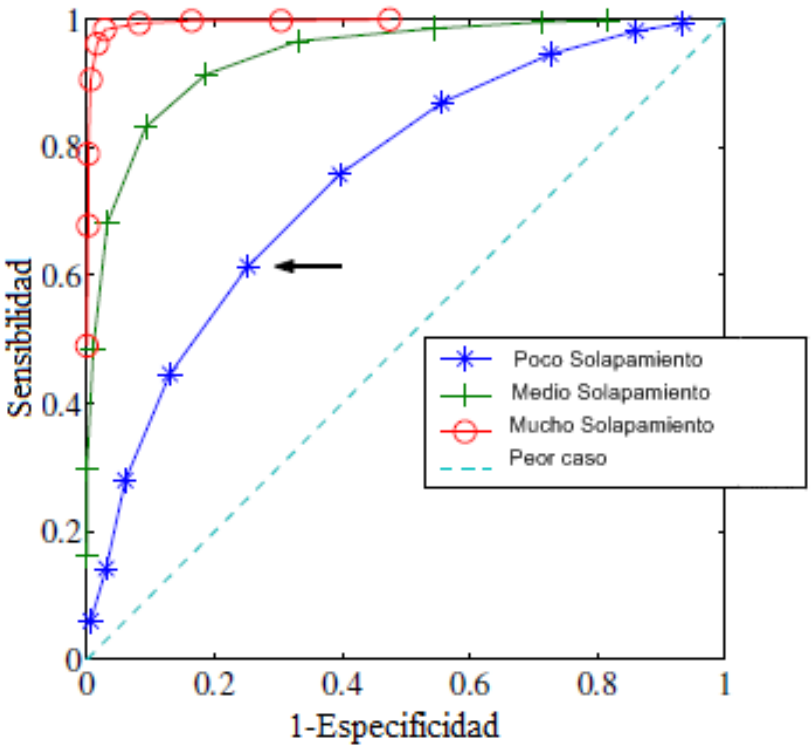



Figura 5. Curva ROC para diferentes valores de umbral. Adaptada de [45]

[12] En siglas AUC

Al considerar un problema de clasificación binaria, en el cual las salidas pueden estar asociadas a positivo (p) ó negativo (n), existen cuatro posibles resultados: si el valor real es positivo y el valor predicho es positivo, entonces se obtiene un verdadero positivo; sin embargo, si el valor predicho es negativo se dice que es un falso negativo. De la misma manera, si el valor real es negativo y el valor predicho es positivo, se obtiene un falso positivo, pero si el valor predicho es negativo el resultado es un verdadero negativo [45]. A manera de ejemplo, en la Tabla 2 se clasifican los datos de una muestra de pacientes de acuerdo al resultado de una prueba y su estado respecto a la enfermedad:

| | | **Verdadero Diagnóstico** | |
|---|---|---|---|
| | | Enfermo (p) | Sano (n) |
| **Resultado de la Prueba** | Prueba Positiva (p') | Verdadero Positivo() | Falso Positivo |
| | Prueba Negativa (n') | Falso Negativo | Verdadero Negativo |
| | | VP+FN | FP+VN |

| | |
|---|---|
| **Sensibilidad** | =VP/(VP+FN)=FVP(fracción de verdaderos positivos) |
| **Especificidad** | =VN/(VN+FP)=FVN(fracción de verdaderos negativos)=1 – FFP |

Tabla 2. Resultado de una prueba y su estado respecto a una enfermedad. Adaptada de [45].

# 2 METODOLOGÍA PARA EL TAMIZAJE DE PACIENTES CON SOSPECHA DE CANCER DE MAMA MEDIANTE UN SISTEMA FUZZY USANDO VARIABLES CLÍNICAS.

Este capítulo describe la metodología que se propone a lo largo del trabajo de investigación desarrollado. El objetivo principal de este estudio es tamizar pacientes con sospecha de cáncer de mama mediante un sistema fuzzy usado variables clínicas.

La adquisición de los datos, es la fase más complicada y delicada del proyecto, debido a que se necesita una gran gestión médica para la participación y continua presencia de pacientes en el estudio [27]. Para esto, se realiza la preparación de la paciente, que incluye un consentimiento informado en el cual se comunica a la paciente sobre el estudio que se está realizando y la importancia de su colaboración. Asimismo, se recolecta la información socio-demográfica, la historia familiar de cáncer, los antecedentes hormonales y los factores clínicos, a través de dos encuestas que la paciente llena personalmente.

Una vez se tiene la información de las pacientes, se registra en una base de datos. Posteriormente se exportan los datos al formato de hoja de cálculo .xls de Excel (Microsoft, derechos reservados®) y a un documento de texto. En el formato de hoja de cálculo, se almacenan los datos tabulados, lo que permite hacer una preselección de las variables que serán usadas en Matlab®.

Con los datos escogidos hasta el momento, se procede a realizar la selección de variables que posteriormente se utilizarán como entradas al sistema de clasificación, según la relación existente con el resultado histopatológico. En el desarrollo de este proyecto, se proponen tres técnicas. La primera de ellas es la *correlación lineal*, que se realiza mediante los coeficientes de Pearson y Spearman. Este método supone una independencia entre cada variable, que no necesariamente ocurre, pero es muy simple y eficaz. La segunda técnica es *la selección secuencial de variables*, que en este caso adhiere parámetros al modelo, dependiendo de un criterio determinado. Y por último, *el ranking de las variables*, que se efectúa por medio de dos técnicas diferentes: la prueba T y la prueba de Wilcoxon. De esta forma, se emplea un método que mide la relación entre variables con el diagnóstico de la biopsia. En la Figura 6 se ilustran las tres técnicas mencionadas anteriormente.

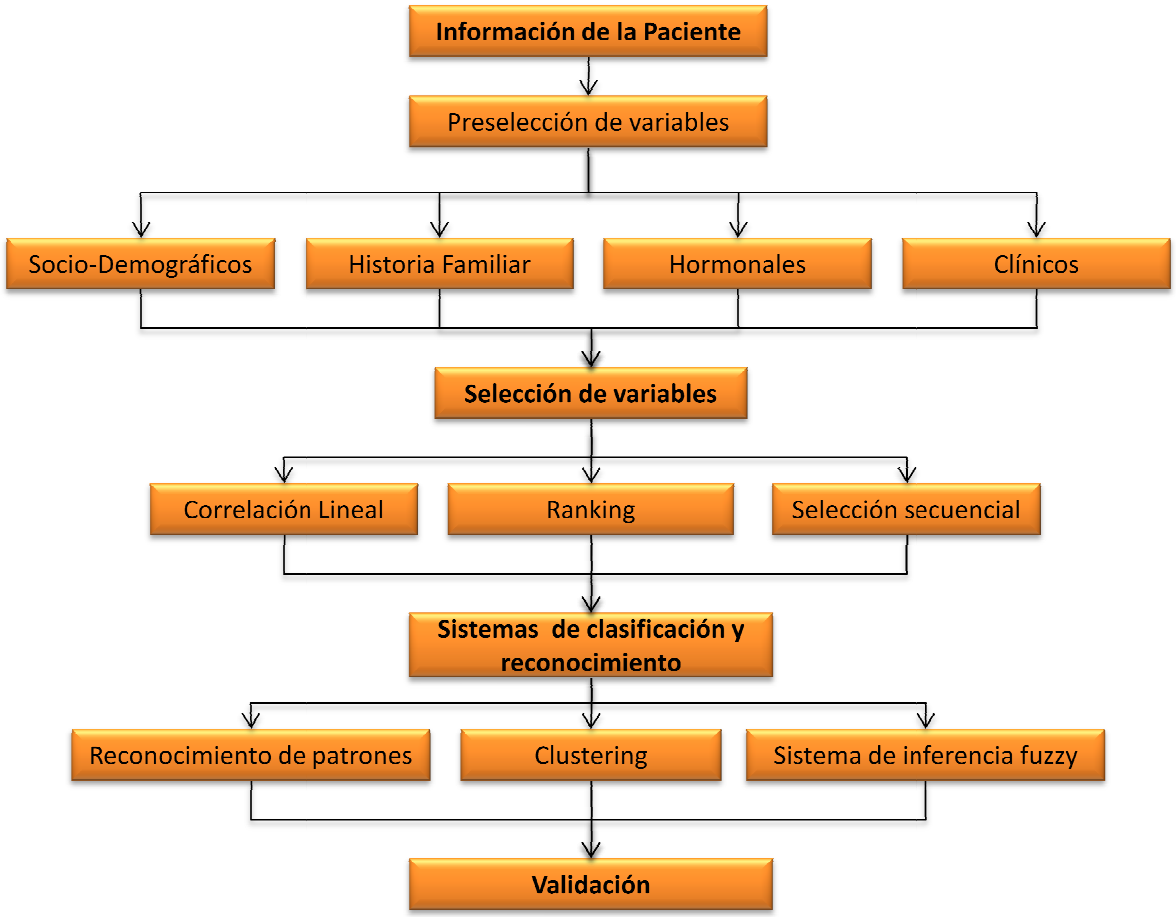



Figura 6. Metodología para el tamizaje de pacientes con sospecha de cáncer de mama mediante un sistema fuzzy usando variables clínicas. Adaptada de [27].

La selección definitiva se realiza para el conjunto de datos pre-seleccionados. A partir de aquí, se aplican cinco diferentes algoritmos de clasificación: las técnicas de agrupamiento “hard c-means” y “fuzzy c-means”, el reconocimiento de patrones usando dos conceptos diferentes: clasificador del vecino más cercano y clasificador del centro más cercano, y un sistema de inferencia fuzzy en el cual se ha utilizado la experiencia de un especialista y cirujano de mama experto. Finalmente, se determinan los parámetros de sensibilidad, especificidad y área bajo la curva ROC para validar los sistemas y realizar una comparación entre ellos.

# 3 ADQUISICIÓN DE LA INFORMACIÓN

Una de las fases más extensa y complicada es la adquisición de la información, ya que no existen antecedentes de estudios en la región sobre diagnóstico médico del cáncer de mama usando variables clínicas. En este trabajo, los datos provienen de la información suministrada por pacientes reales con sospecha de carcinoma de glándula mamaria. Esta información es obtenida por medio de una serie de preguntas contenidas en dos encuestas sobre: datos socio-demográficos, historia familiar de cáncer, datos hormonales y clínicos. El proceso de adquisición de la información comienza desde la asignación de la cita a la paciente hasta el registro de la información. Inicialmente, se cita la paciente y se le informa sobre la consistencia y desarrollo del proyecto mediante un consentimiento para posteriormente, acceder a llenar las encuestas antes mencionadas. Una vez recopilados los datos, estos deben ser almacenados en una base de datos, construida con el fin de organizar la información.

La muestra con que se trabajó hasta el momento de la culminación de este proyecto es de 50 pacientes, la mayor parte, provenientes de consulta médica con el especialista y cirujano de mama Dr. Álvaro Niño. Del total de esta muestra, 27 han sido diagnosticadas con cáncer de mama.

## 3.1 CONSENTIMIENTO INFORMADO

Para tomar parte en el proyecto, los pacientes deben participar de forma voluntaria, conociendo la información respecto al estudio. Siempre debe respetarse el derecho de las participantes en la investigación a proteger su integridad y deben tomarse toda clase de precauciones para resguardar la intimidad de los individuos, la confidencialidad de la información de la paciente y para reducir al mínimo las consecuencias de la investigación sobre su integridad

física, mental y su personalidad [27].

Esta investigación, por formar parte del proyecto macro, desarrollado junto a un equipo médico y epidemiológico, adscritos a la investigación de COLCIENCIAS, se fundamenta en la declaración de Helsinki, que la Asociación Médica Mundial ha promulgado. Dicha Declaración es una propuesta de principios éticos que sirven para orientar a los médicos y a otras personas que realizan investigación médica en seres humanos, incluyendo la investigación del material humano o de información identificables. Además, se enfatiza en que la investigación debe ejecutarse dentro del marco de la normatividad vigente en el país donde se realiza y que cada individuo participante debe recibir información adecuada acerca de los objetivos, métodos, fuentes de financiamiento, posible conflictos de intereses, afiliaciones institucionales del investigador, beneficios calculados, riesgos previsibles e incomodidades derivadas. La persona debe ser informada del derecho de participar o no en la investigación y de retirar su consentimiento en cualquier momento, sin exponerse a represalias. Después de asegurarse que el individuo ha comprendido la información, el médico debe obtener entonces, preferiblemente por escrito, el consentimiento informado y voluntario de la persona. Es por esto que, mediante el consentimiento informado como procedimiento médico formal se aplica el principio de autonomía de la paciente [46].

En el ANEXO A se muestra el consentimiento informado elaborado en el proyecto macro.

## 3.2 RECOPILACIÓN DE LA INFORMACIÓN SOCIO-DEMOGRÁFICA, LA HISTORIA FAMILIAR, LOS DATOS HORMONALES Y LA INFORMACIÓN CLÍNICA.

En el capítulo 1, se mencionaron varios factores de riesgo, que pueden incidir en el desarrollo del cáncer de mama. Estas variables, son fundamentales para revelar

un posible diagnóstico de la enfermedad. En este trabajo se han elaborado dos encuestas: la primera involucra la recopilación de algunos datos socio-demográficos, historia familiar respecto al cáncer, antecedentes hormonales y acceso a la atención médica. La segunda, reúne información semejante a una historia clínica: presencia de masas, dolor, anomalías en la piel, etc. Ambas fueron elaboradas con la asesoría epidemiológica y clínica involucrada en el proyecto de COLCIENCIAS, teniendo en cuenta diferentes estudios sobre los factores de riesgo del carcinoma de glándula mamaría, como señala la Tabla 1. La Tabla 3 muestra los esquemas resumidos de las dos encuestas, mientras que en el ANEXO B se encuentra una copia completa de cada una.

## 3.3 CREACIÓN DE BASE DE DATOS

La información obtenida a través de los formularios, fue digitalizada mediante el software gratuito *EpiData 3.1*[13]. Este programa está diseñado para la entrada y documentación de datos.

*EpiData 3.1*, fue seleccionado, debido a su facilidad de acceso, y a la sencillez para convertir líneas de texto simple  a una forma explícita de entrada de datos. En este software, pueden definirse las escalas de los datos de entrada, así como las restricciones que pueda tener un dato. Las fechas se guardan fácilmente, por ejemplo, 2301 será el formato de 23/01/2010 si se ha introducido en el año 2010 en un campo "dd/mm/aaaa".

Una de las ventajas que se aprovechó de *Epidata* al registrar los datos, es que se puede verificar la información digitada por dos personas diferentes haciendo una comparación.

[13] Software libre, disponible en www.epidata.dk

<table>
<tr><th colspan="4">Primera Encuesta</th><th colspan="2">Segunda Encuesta</th></tr>
<tr><td rowspan="5">Datos Personales</td><td colspan="5">Cédula</td></tr>
<tr><td colspan="5">Nombre</td></tr>
<tr><td colspan="5">Dirección</td></tr>
<tr><td colspan="5">Teléfono</td></tr>
<tr><td colspan="5">Talla del brasier</td></tr>
<tr><td rowspan="2">Registro de Termografía</td><td colspan="5">Fecha del registro</td></tr>
<tr><td colspan="5">Código de las imágenes</td></tr>
<tr><td rowspan="8">Datos socio-demográficos</td><td colspan="3">Edad</td><td rowspan="8">Índice de masa corporal</td><td rowspan="4">Estatura</td></tr>
<tr><td colspan="3">Lugar de residencia</td></tr>
<tr><td colspan="3">Estrato</td></tr>
<tr><td colspan="3">Raza</td></tr>
<tr><td colspan="3">Presencia de pareja estable</td><td rowspan="4">Peso</td></tr>
<tr><td colspan="3">Nivel de estudios</td></tr>
<tr><td colspan="3">Profesión</td></tr>
<tr><td colspan="3">Ingresos económicos</td></tr>
<tr><td rowspan="2">Acceso a la atención médica</td><td colspan="3">Datos del SISBEN</td><td rowspan="2">Dolor en los senos</td><td>Mama(s)</td></tr>
<tr><td colspan="3">Información sobre EPS</td><td>Región(es)</td></tr>
<tr><td rowspan="5">Historia familiar del cáncer</td><td rowspan="3">De primer grado</td><td>Mamá</td><td colspan="3" rowspan="3">Asimetría en los senos</td></tr>
<tr><td>Hermanas</td></tr>
<tr><td>Hijas</td></tr>
<tr><td rowspan="2">De segundo grado</td><td>Tías</td><td rowspan="2">Telorrea</td><td>Espontaneidad</td></tr>
<tr><td>Primas</td><td>Tipo de secreción</td></tr>
<tr><td rowspan="6">Antecedentes hormonales</td><td colspan="3">Edad de la menarquia</td><td rowspan="6">Masas en los senos</td><td rowspan="3">Mama(s)</td></tr>
<tr><td colspan="3">Número de embarazos</td></tr>
<tr><td colspan="3">Edad del primer embarazo</td></tr>
<tr><td colspan="3">Tiempo de lactancia</td><td rowspan="3">Región(es)</td></tr>
<tr><td colspan="3">Edad de la menopausia</td></tr>
<tr><td colspan="3">Tratamientos hormonales</td></tr>
<tr><td rowspan="4">Antecedentes clínicos</td><td colspan="3">Mamografías previas</td><td rowspan="4">Anormalidad de piel en las mamas</td><td>Tipo de anormalidad</td></tr>
<tr><td colspan="3">Biopsias anteriores</td><td rowspan="3">Región(es)</td></tr>
<tr><td colspan="3">Antecedentes propios de cáncer</td></tr>
<tr><td colspan="3">Radioterapias anteriores</td></tr>
<tr><td colspan="4" rowspan="3"></td><td rowspan="2">Anomalías de nódulos</td><td>Nódulos linfoides axilares</td></tr>
<tr><td>Nódulos supraclaviculares</td></tr>
<tr><td colspan="2">Resultados de la mamografía</td></tr>
</table>

Tabla 3. Esquemas de los formularios que recopilan la información socio-demográfica, hereditaria, hormonal (Primera encuesta) y clínica (Segunda encuesta). Adaptada de [27].

Los pasos seguidos para la creación de la base de datos fueron:

1. Entrada de los datos: en este paso se escribe el cuestionario asignándole a cada pregunta una variable, y definiendo el tipo para cada una de ellas, por ejemplo, numérico, texto, fecha, etc.
2. Añadir controles: Una herramienta importante de *EpiData* es la posibilidad de especificar reglas y cálculos durante la entrada de datos. En este punto se le da un valor numérico a las opciones de respuesta y además se pueden realizar cálculos entre las entradas. Por ejemplo, el IMC[14] puede determinarse a partir del peso y la estatura como (peso/estatura$^2$). Igualmente es posible especificar una secuencia de la entrada de los datos, a través de saltos condicionales.
3. Al culminar el paso anterior, la encuesta ha sido creada en el programa, así, esta debe guardarse para luego entrar los datos.
4. La entrada de los datos se realiza con la información recolectada en las encuestas, llenando los campos en cada pregunta de acuerdo al valor asignado a su respectiva respuesta. Esta entrada se realizó dos veces por diferentes personas a fin de corroborar que los datos hayan sido correctamente digitados.
5. Finalmente, se comparan los dos archivos guardados que contienen los 50 registros de las pacientes y se realizan las correcciones que sean necesarias hasta coincidir la información.

En la Figura 7 se muestra una parte del primer cuestionario en *EpiData* que contiene información de una de las pacientes. Desde *EpiData* se exportaron los datos a formato de texto y luego fueron retomados en MATLAB®. La información recopilada es bastante amplia y variada. Todos los datos no son usados en el desarrollo de éste proyecto, pero si almacenados para posteriores estudios que puedan involucrar un mayor número de variables.

[14] Índice de masa corporal

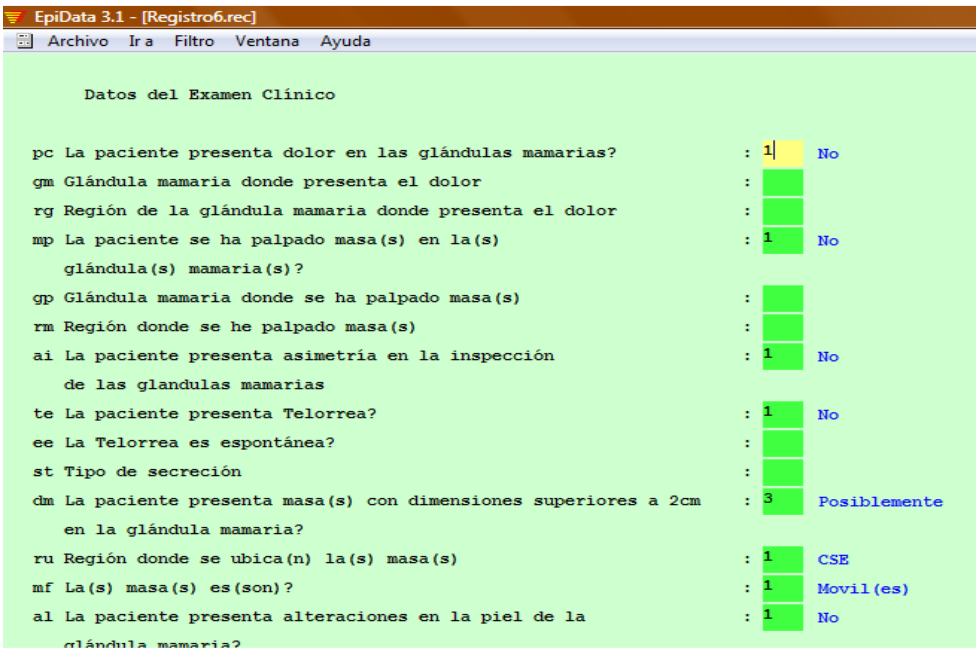


Figura 7. Registro de una de las pacientes en Epidata 3.1 Fuente: el autor

# 4 SELECCIÓN DE LAS VARIABLES

Antes de realizar un análisis estadístico de las variables, se hace necesario seleccionar algunas de estas para reducir su número y hacer más eficiente el rendimiento de los sistemas a implementar. A este proceso se le denomina en este libro preselección. Seguidamente se estudian las variables seleccionadas a través de tres métodos de correlación estadística con el fin de descartar las variables que no sirvan como criterio para determinar la presencia o ausencia de carcinoma de glándula mamaria.

## 4.1 PRESELECCIÓN DE LAS VARIABLES

El procedimiento de preselección, consiste en escoger las variables que según el criterio médico y los resultados que existen de investigaciones desarrolladas (ver Tabla 1) incidan en el padecimiento del cáncer de mama. Esta preselección se ve limitada por la falta de información en alguna de las encuestas. Por ejemplo, la

variable IMC[15], no pude ser considerada en el estudio, debido a que no se cuenta con información de la estatura de algunas pacientes.

| | |
|---|---|
| **Variables preseleccionadas** | Edad |
| | Lugar de residencia |
| | Raza |
| | Historia familiar de primer grado |
| | Historia familiar de segundo grado |
| | Menarquia |
| | Embarazos |
| | Edad del primer embarazo |
| | Edad Menopausia |
| | Tratamiento Hormonal |
| | Dolor en los senos |
| | Asimetría en las mamas |
| | Masas en los senos (percepción del cirujano) |
| | Masas en los senos (percepción de la paciente) |
| | Telorrea |
| | Anormalidad en la piel de las mamas |

Tabla 4. Variables preseleccionadas. Fuente: el autor.

La Tabla 4. muestra las variables preseleccionadas de acuerdo a las consideraciones anteriormente nombradas.

## 4.2 SELECCIÓN ESTADÍSTICA DE LAS VARIABLES

Para la selección de variables[16], como uno de los objetivos de este trabajo se plantean tres técnicas: *correlación lineal* mediante los coeficientes de Pearson y Spearman, *ranking de variables* mediante las pruebas T y de Wilcoxon, y *la selección secuencial* de variables.

### 4.2.1 Correlación lineal mediante los coeficientes de Pearson y Spearman

En el análisis de estudios médicos, específicamente epidemiológicos, es necesario

[15] Índice de masa corporal
[16] Conocida en inglés como Feature Selection

encontrar el nivel en el que dos variables se relacionan para un conjunto de pacientes. Para esto se presenta la correlación lineal, la cual busca determinar el grado en que dos variables están correlacionadas. Una aplicación de la correlación lineal es encontrar variables que permitan predecir el valor de otra variable [44]. Se debe aclarar que, el hecho de obtener un coeficiente bajo no significa que las variables no estén correlacionadas, ya que se puede presentar el caso en que exista un comportamiento no lineal.

El coeficiente de correlación de Pearson es un índice de fácil interpretación. Los valores entre los cuales oscila este índice están dentro del rango -1 y +1. La magnitud del coeficiente es la que define si la correlación es fuerte o no, mientras el signo indica la dirección de dicha correlación. Por lo tanto, un valor de -1 es tan importante como un valor de +1. Un signo positivo indica correlación positiva, esto es, que los valores de las dos variables aumentan o disminuyen en el mismo sentido y un signo negativo indica correlación negativa, es decir que el valor de una variable se incrementa cuando el de la otra disminuye. Un valor de 0 en el coeficiente de correlación de Pearson significa que no hay correlación entre las variables [47].

El coeficiente de correlación de Pearson se calcula por medio de la siguiente expresión:

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{XY}}{S_x S_y} \qquad (1.35)$$

donde $X$ y $Y$ son las variables, $N$ es el número de datos para las variables y $S$ es la desviación de cada una de ellas.

El coeficiente de correlación de Spearman es una medida de asociación lineal que utiliza los rangos y los números de orden de cada grupo de datos, y compara dichos niveles [44].

El cálculo del coeficiente de correlación de Spearman viene dado por la siguiente expresión:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \qquad (1.32)$$

donde $d_i = r_{xi} - r_{yi}$ es la diferencia entre los rangos de $X$ y $Y$, y $n$ es el número de datos.

Para determinar la relación que existe de forma independiente, entre el resultado de la biopsia y cada una de las variables preseleccionadas, se determinaron los coeficientes de Pearson y Spearman. Los resultados obtenidos se muestran en la Tabla 5 y 6, donde aparecen las 16 variables ordenadas de acuerdo a ambos criterios. Se puede concluir entonces que existe una correlación lineal mayor con la edad, presencia de masas, menopausia y tratamiento hormonal que con el resto de variables.

| Variable | Spearman |
|---|---|
| Edad | 0,579 |
| Masas | -0,400 |
| Menopausia | 0,246 |
| Trat_horm | 0,246 |
| N°_emb | 0,224 |
| Asimetría | 0,218 |
| Alt_piel | 0,156 |
| Historia_fam1 | -0,142 |
| Telorrea | -0,142 |
| Raza | -0,139 |
| Historia_fam2 | 0,091 |
| Urb_Rur | 0,084 |
| Dolor | -0,080 |
| Pri_emb | -0,066 |
| Masas2cm | -0,052 |
| Menarquia | 0,040 |

Tabla 5. Correlación lineal mediante coeficiente de Spearman entre las variables y el resultado histopatológico. Fuente: el autor.

| Variable | Pearson |
|---|---|
| Edad | 0,545 |
| Masas | -0,400 |
| Menopausia | 0,246 |
| Trat_horm | 0,246 |
| N°_emb | 0,226 |
| Asimetría | 0,218 |
| Alt_piel | 0,156 |
| Telorrea | -0,142 |
| Historia_fam1 | -0,142 |
| Raza | -0,111 |
| Urb_Rur | 0,084 |
| Dolor | -0,080 |
| Pri_emb | -0,066 |
| Masas2cm | -0,052 |
| Menarquia | 0,040 |
| Historia_fam2 | 0,026 |

Tabla 6. Correlación lineal mediante coeficiente de Pearson entre las variables y el resultado histopatológico. Fuente: el autor.

### 4.2.2 Ranking de las variables a través de la prueba T y de Wilcoxon

Otro método empleado para la selección de variables es el ranking, que consiste en posicionar variables usando un criterio de evaluación independiente para clasificación binaria. El criterio varía de acuerdo a la técnica que se seleccione. Entre estas técnicas, se encuentran la prueba T, entropía, ROC y Wilcoxon.

La prueba T asume la hipótesis nula: las variables y el resultado histopatológico son muestras aleatorias independientes con distribuciones normales, medias y varianzas iguales pero desconocidas, contra la alternativa de que las medias no sean iguales. Si el resultado de la prueba T es 1, indica un rechazo de la hipótesis nula a un nivel de significancia determinado, pero si es 0 indica que no se rechaza la hipótesis nula [41].

| Variable | Resultado del ranking |
|---|---|
| Edad | 4,513 |
| Masas | 3,023 |
| Trat_horm | 1,759 |
| Menopausia | 1,759 |
| Asimetria | 1,549 |
| N°_emb | 1,612 |
| Telorrea | 1,000 |
| Historia_fam1 | 1,000 |
| Alt_piel | 1,095 |
| Raza | 0,778 |
| Urb_Rur | 0,585 |
| Dolor | 0,556 |
| Pri_emb | 0,462 |
| Masas2cm | 0,361 |
| Menarquia | 0,280 |
| Historia_fam2 | 0,183 |

Tabla 7. Selección de las variables mediante prueba T. Fuente: el autor.

En la Tabla 7 se muestra un ejemplo de utilización del ranking mediante la prueba T. Se aplicó este método paramétrico sobre todas las variables preseleccionadas, observando nuevamente que las primeras variables, en su mayoría, corresponden a las descritas por el coeficiente de Spearman y Pearson.

La técnica de Wilcoxon para dos muestras no apareadas, es una prueba no paramétrica, que se utiliza para comparar la media de cada una de las variables con la media del resultado histopatológico. Por lo general este método se utiliza cuando los datos son ordinales. En este método no se asume alguna distribución de la muestra. Sin embargo, se supone que es aleatoria, que las variables son independientes y que su escala de medida es ordinal.

| Variable | Resultado de la prueba Wilcoxon |
|---|---|
| Edad | 0,314 |
| Masas | 0,180 |
| N°_emb | 0,108 |
| Asimetría | 0,080 |
| Raza | 0,089 |
| Trat_horm | 0,060 |
| Menopausia | 0,060 |
| Dolor | 0,060 |

| | |
|---|---|
| Telorrea | 0,040 |
| Historia_fam1 | 0,040 |
| Pri_emb | 0,040 |
| Alt_piel | 0,040 |
| Masas2cm | 0,040 |
| Historia_fam2 | 0,013 |

Tabla 8. Selección de las variables mediante prueba Wilcoxon. Fuente: el autor.

En la Tabla 8 se muestran las variables seleccionadas usando la prueba de Wilcoxon. Se ve claramente que la edad y la presencia de masas son nuevamente las variables con mayor relación al resultado histopatológico, pero además aparecen el número de embarazos, la asimetría y la raza como factores incidentes.

### 4.2.3 Selección secuencial de variables

La selección secuencial de variables es un método que reduce la dimensionalidad de los datos, es decir, que simplifica el número de características que describen un dato, seleccionando un subconjunto de factores predictores para crear un modelo. El criterio de selección, también llamado función objetivo, usualmente abarca la minimización de una medida específica de error predictivo para modelos [41]. Este método puede ser aplicado de dos maneras: añadiendo variables al modelo *selección secuencial hacia delante* (SFS[17]), o quitando variables *selección secuencial hacia atrás* (SBS[18]).

Para la selección de variables se usó selección secuencial hacia adelante, la cual agrega variables por cada iteración, tomando como criterio la desviación estándar de la regresión logística, tal como se muestra en la Figura 8.

[17] En inglés, Sequential Forward Selection
[18] En inglés, Sequential Backward Selection

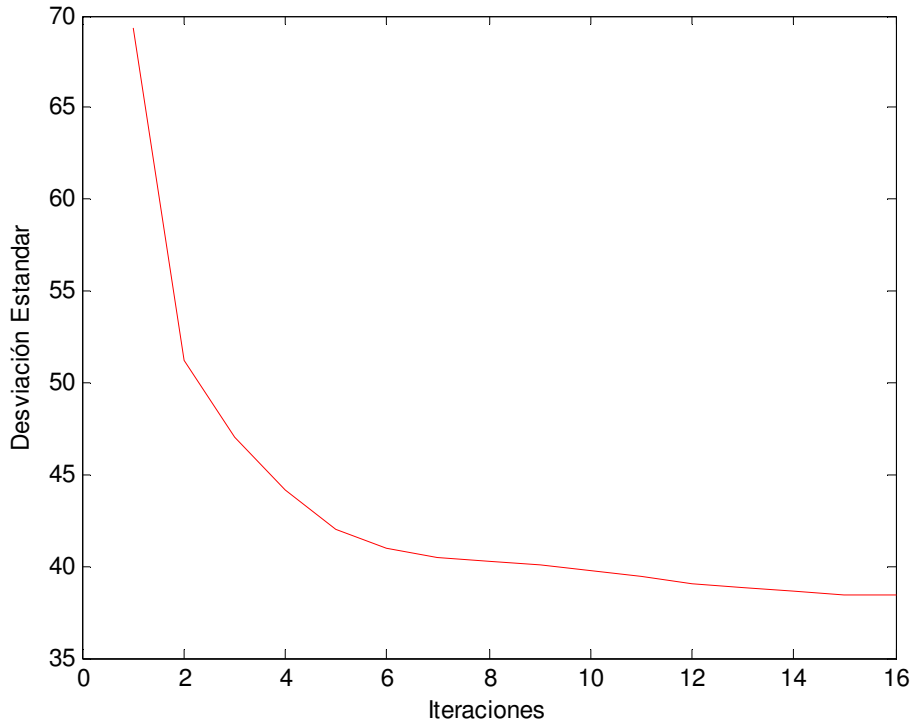



Figura 8. Resultado de la función criterio para cada iteración usando selección secuencial. Fuente: el autor.

En la Figura 8 se observa el comportamiento del criterio para cada iteración, en el caso en que se adhieren las 16 variables, resultando un cambio insignificante en la desviación estándar a partir de la sexta iteración. De esta manera, seis variables son suficientes para describir los datos que se van a ingresar a los sistemas de inteligencia artificial a implementar.

| Variable | Desviación Estándar |
|---|---|
| Edad | 51,247 |
| Urb_Rur | 47,043 |
| Raza | 44,174 |
| Historia_fam1 | 41,998 |
| Historia_fam2 | 41,002 |
| Menarquia | 39,687 |
| N°_emb | 39,214 |
| Pri_emb | 38,863 |
| Menopausia | 38,620 |

| | |
|---|---|
| Trat_horm | 38,324 |
| Dolor | 38,038 |
| Masas | 37,605 |
| Asimetria | 37,297 |
| Telorrea | 37,095 |
| Masas2cm | 36,973 |
| Alt_piel | 36,936 |

Tabla 9. Selección de variables usando SFS. Fuente: el autor.

La Tabla 9 muestra las variables adheridas en cada iteración y el respectivo valor de la función objetivo, obteniéndose nuevamente la edad como la variable más correlacionada con la biopsia.

Los métodos anteriormente propuestos, pueden ser utilizados para cualquier estudio que requiera reducir el número de variables o características de una muestra para optimizar un modelo, resultando eficientes cualquiera de las tres formas señaladas en este capítulo. Esta eficiencia depende además del tamaño de la muestra y de si las variables son paramétricas o no. Por ejemplo en la prueba T, se asume distribución normal debido al tamaño de la muestra, y la prueba Wilcoxon asume que los datos son ordinales, por tanto no paramétricos.

Para la implementación de los algoritmos analizados en el capitulo siguiente, no es relevante escoger uno de los métodos de selección debido a que seis de las variables se repiten en las primeras posiciones de cada uno. Estas variables son: la edad, presencia de masas, edad de la menarquía, número de embarazos, edad del primer embarazo y edad de la menopausia.

# 5 ANÁLISIS DE RESULTADOS

Hasta este punto de la investigación, se ha logrado recopilar la información socio-demográfica, la historia familiar, los datos hormonales y clínicos, a partir de las encuestas. Posteriormente, se fijaron las variables a utilizar, para luego seleccionar estadísticamente las más importantes referentes al resultado

histopatológico de las pacientes en estudio. Ahora, se debe diseñar el sistema de clasificación, para lo cual, se implementaron varios tipos de algoritmos. A continuación se muestran los resultados y la validación para cada uno.

## 5.1 VALIDACIÓN DEL ALGORITMO HARD C - MEANS

El algoritmo para realizar la clasificación por medio de este método se presentó en el capítulo 1 de este libro. Allí se mencionó la función objetivo y su uso como criterio para determinar cual partición es la mejor, la forma como se calculan los centros de los grupos y los valores de membresía asociados a cada paciente que indican la pertenencia a un grupo.

En este trabajo, el uso de hard c-means busca comparar sus resultados con otros métodos, especialmente fuzzy c-means. La muestra, objeto de este estudio, son pacientes con sospecha de cáncer de mama. Debido a que este método no es supervisado, no es necesario dividir la muestra, y se usa todo el conjunto para realizar la clasificación. Cada paciente está caracterizada por *m* variables.

Los parámetros seleccionados para el algoritmo hard c-means, son los siguientes:

- El número de grupos en el que se divide la muestra, $c$. En este caso $c = 2$, ya que corresponde a los grupos que se necesitan; uno para clasificar las pacientes enfermas y otro para clasificar las pacientes sanas.
- El valor del nivel de tolerancia, $\varepsilon = 10^{-5}$, que establece cuando el cambio en la función objetivo es lo suficientemente pequeño para detener el proceso iterativo del algoritmo, determinando así, que se ha encontrado la mejor partición posible.

Las variables que se usan en el algoritmo para realizar el agrupamiento, fueron

seleccionadas por medio de análisis estadístico. Estas variables no necesitan ser normalizadas ya que los rangos en los que se encuentran no son extremadamente diferentes. Cada paciente, tiene 6 variables que la describen y permiten estimar si tiene o no cáncer.

El algoritmo hard c-means, clasifica la muestra en los grupos que se requiere. Esta división la hace únicamente observando las variables de las pacientes de modo que la pertenencia a un grupo u otro, aún así, no determina si una paciente está enferma o sana. Para conocer a qué grupo le asignamos la condición enferma y a cual la condición sana, se realizó la validación del sistema con cada grupo, es decir, se plantearon dos casos: (1) en el que se asigna al primer grupo la condición enferma y, (2) en el que se da la condición enferma al segundo grupo. Los resultados de la validación en cada caso se muestran en la Figura 9. En esta figura se puede observar que el primer caso tiene mejor resultado ya que presenta la mayor tasa de pacientes clasificadas correctamente, de ahí se deriva que al primer grupo pertenecen las pacientes con cáncer.

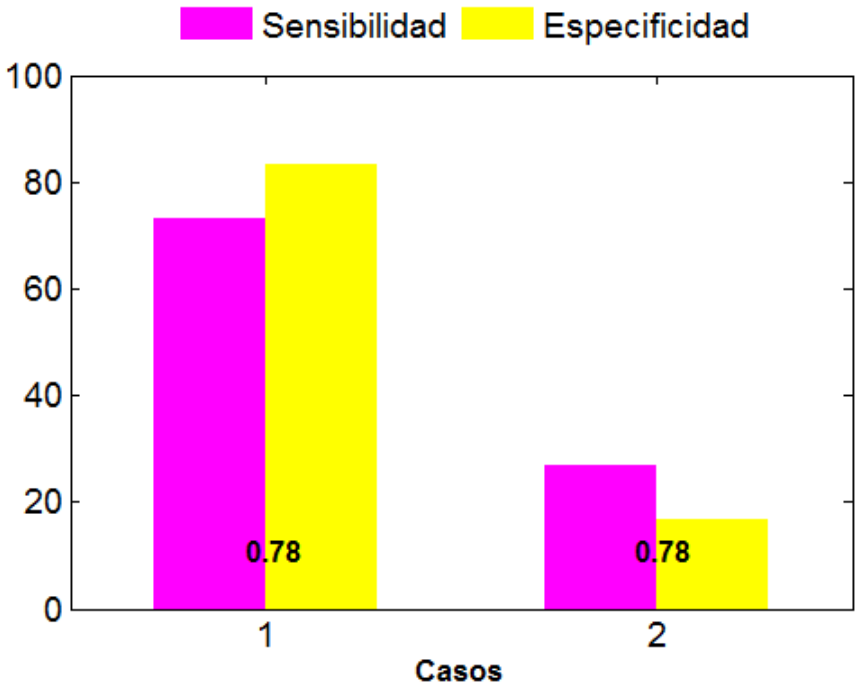



Figura 9. Sensibilidad, especificidad y área bajo la curva ROC del resultado obtenido con el algoritmo hard c-means. Fuente: el autor.

El resultado obtenido con el algoritmo hard c-means es aceptable, puesto que en ambos casos el área bajo la curva ROC es mayor a 0.5. Además se observa que las dos son iguales, esto se debe a que un grupo es el complemento del otro, es decir, que las pacientes supuestas como enfermas en un caso son las pacientes supuestas como sanas en el otro. Este sistema de clasificación en general presenta una sensibilidad del 73%, una especificidad del 83% y un área bajo la curva ROC de 0.78.

## 5.2 VALIDACIÓN DEL ALGORITMO DE CLASIFICACIÓN FUZZY C-MEANS

Fuzzy c-means agrupa datos acorde al grado de similitud que ellos tengan, para lo cual mide la distancia que existe entre los datos y los centros de los grupos de forma que en cada grupo se encuentren los puntos de datos más cercanos. La

información de cada paciente representa un dato de *m* coordenadas. Para este estudio se formaron dos grupos en los cuales se clasifican las pacientes enfermas y las pacientes sanas.

FCM es un algoritmo no supervisado, por tanto, no es necesario dividir la muestra, usándola toda para realizar el proceso de agrupamiento. El algoritmo fuzzy c-means se presentó en el capítulo 1 de este libro. Allí se explicó el procedimiento para crear los grupos. Los parámetros seleccionados para realizar el agrupamiento son los siguientes:

- Se fija $c = 2$, es decir, se hacen dos grupos, uno que contiene las pacientes enfermas y el otro las sanas.
- Se selecciona $m' = 1.5$, este valor es un parámetro que indica la cantidad de incertidumbre en el proceso de clasificación y a medida que se aproxima a uno, se da más peso al dato que esté más cerca del centro del grupo [48].
- Por último, se da un valor al umbral mínimo de cambio en la función objetiva, $\varepsilon = 10^{-5}$, el cual determina cuando debe finalizar el algoritmo.

La normalización de la muestra es un proceso que se usa en los métodos de agrupamiento para que las características no presenten rangos extremadamente diferentes, así se evita la pérdida de interpretación del sistema. En este trabajo, las escalas definidas para cada variable no difieren significativamente y por tanto se plantea un procedimiento sin normalización.

Las variables seleccionadas por medio de los estudios estadísticos presentados anteriormente son ingresadas al sistema de agrupamiento. Luego, por medio del algoritmo presentado se realiza la división de los datos en los dos grupos seleccionados: condición enferma y sana. Como resultado, FCM entrega *m* coordenadas de los centros de cada grupo, el valor de la función objetivo para cada iteración y el valor de membresía asociado a cada paciente indicando la

pertenencia a cada grupo. Para hacer la comparación con la biopsia, primero se debe asignar la total pertenencia a un grupo u a otro, es decir, se debe asignar el dato al grupo que presenta mayor membresía con 1 y al de menor membresía (el resto para casos con más de dos grupos) con 0, a este proceso en fuzzy se le conoce como “Hardering the fuzzy c-partition” [34]. Seguidamente, se procede a identificar entre los grupos el de las pacientes enfermas y el de las pacientes sanas, para esto se realizó la validación de los resultados obtenidos con la biopsia de las pacientes, alcanzándose valores de sensibilidad, especificidad y área bajo la curva ROC en cada caso. Estos resultados se muestran en la Figura 10.

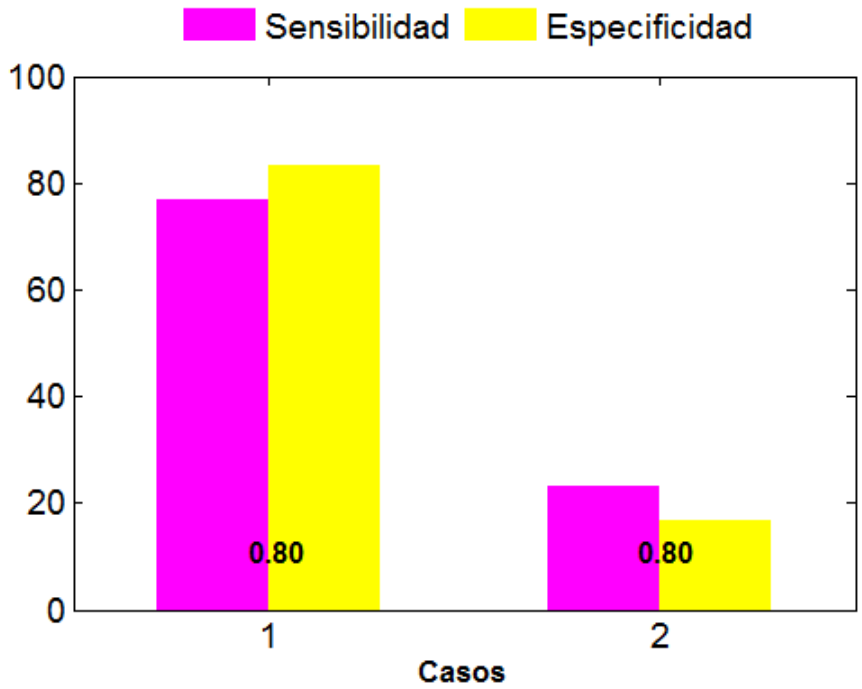



Figura 10. Sensibilidad, especificidad y área bajo la curva ROC del resultado obtenido con el algoritmo Fuzzy c-means. Fuente: el autor.

Los resultados obtenidos por medio del algoritmo corresponden a dos situaciones: (1) cuando se asigna la condición enferma al primer grupo y (2) cuando se establece la condición enferma al segundo grupo. Es evidente que el primer caso corresponde a la mejor situación, es decir, que el grupo 1 es el correspondiente al de las pacientes enfermas y el grupo 2 a las pacientes sanas. El resultado de área

bajo la curva ROC obtenido con Fuzzy c-means es bueno, ya que está por encima de 0.5. En general, este sistema de clasificación presenta una sensibilidad del 77%, una especificidad del 83% y un área bajo la curva ROC de 0.8.

## 5.3 VALIDACIÓN DEL ALGORITMO DE RECONOCIMIENTO DE PATRONES

El reconocimiento de patrones está dividido en dos fases: en la primera se encuentran los patrones por medio de cualquier método de agrupamiento con una parte de la muestra y en la segunda se verifica mediante la asociación de los datos de la muestra restante a cada patrón conocido.

El método propuesto para la división de la muestra se realiza por medio de un análisis discriminante usando validación cruzada. Con este método se crea una división aleatoria para el conjunto de validación sobre todos los datos, es decir, se crean dos conjuntos, uno para entrenar o construir y otro para validar [49]. Usualmente se emplea el 10% de la muestra para validar. La división de la muestra se realizó de modo que el 80% de ella se utilizará en la clasificación y el 20% en el reconocimiento.

La lógica fuzzy  presenta dos alternativas para realizar el reconocimiento: NCC y NNC. A continuación se presenta cada uno de los métodos y se muestran los resultados obtenidos.

### 5.3.1 Clasificador del vecino más cercano (NNC)

En la primera fase de este método, se usó FCM sobre una parte de la muestra para encontrar los patrones de dos grupos, uno de pacientes enfermas y otro de pacientes sanas. Este algoritmo entrega *m* coordenadas de los centros de los dos grupos formados, el valor de la función objetivo, y el valor de membresía asociado

a cada paciente que indica la pertenencia a cada grupo. De manera similar a la clasificación realizada con Fuzzy c-means, se realiza la validación con cada grupo.

Los resultados obtenidos en la clasificación, se muestran en la Figura 11, en donde se puede observar la sensibilidad y la especificidad de cinco casos correspondientes a diferentes divisiones aleatorias de la muestra. En esta se puede observar que la sensibilidad oscila entre el 60% y el 80% y la especificidad es del 80% en la mayoría de los casos, por tanto, los patrones descritos son un buen criterio para hacer el reconocimiento. En esta fase no se calcula el área bajo la curva ROC como criterio de validación porque hasta este punto solo se ha realizado la clasificación y no el reconocimiento, proceso que se desea evaluar.

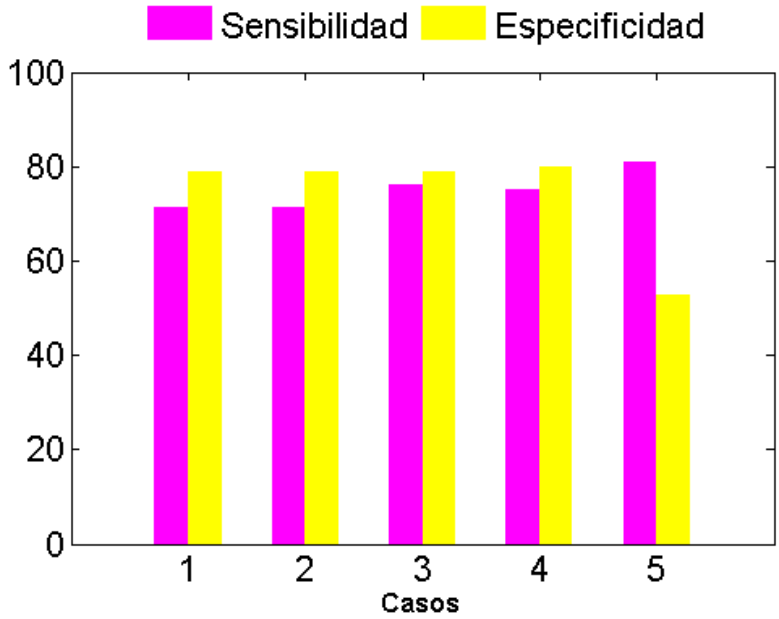



Figura 11. Sensibilidad y especificidad de la fase 1 de reconocimiento de patrones para el caso del clasificador del vecino más cercano. Fuente: el autor.

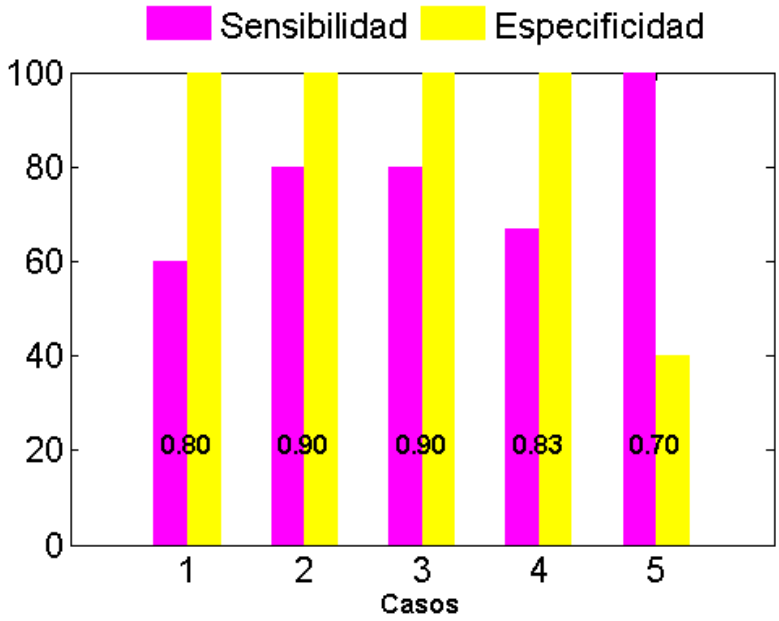


Figura 12. Sensibilidad, especificidad y área bajo la curva ROC del reconocimiento de patrones usando el clasificador del vecino más cercano. Fuente: el autor.

La segunda fase consiste en determinar el grupo al que pertenecen las pacientes de la muestra que no han sido manipuladas. Para esto, el método del clasificador del vecino más cercano usa la distancia euclidiana medida entre los datos de las pacientes que ya fueron clasificadas y los nuevos datos, asociando de acuerdo a la menor distancia, cada nuevo dato con el grupo al que pertenece la paciente clasificada.

Los resultados obtenidos por medio de este método de reconocimiento se muestran en la Figura 12. En esta ilustración, se puede observar que la sensibilidad oscila entre el 60% y el 100%, la especificidad es del 100% en la mayoría de los casos y el área bajo la curva ROC es en promedio 0.82. Esta última medida de desempeño en los sistemas de clasificación indica que el sistema de reconocimiento de patrones por medio del clasificador del vecino más cercano es bueno.

### 5.3.2 Clasificador del centro más cercano (NCC)

La primera fase del reconocimiento se hace de manera similar a la que se mostró en el método del clasificador del vecino más cercano. En ella se determinan los patrones por medio de FCM usando parte de la muestra. En este caso la información más importante son las coordenadas de los centros.

En la Figura 13, se muestra la sensibilidad y especificidad obtenida en el proceso de clasificación para cinco casos en los que se dividió aleatoriamente la muestra. En esta ilustración, se observa que la sensibilidad y la especificidad oscilan entre el 65% y el 90%, esto indica que las características de los datos permiten diferenciar bien las pacientes enfermas de las pacientes sanas y por tanto se pueden establecer los patrones para cada grupo.

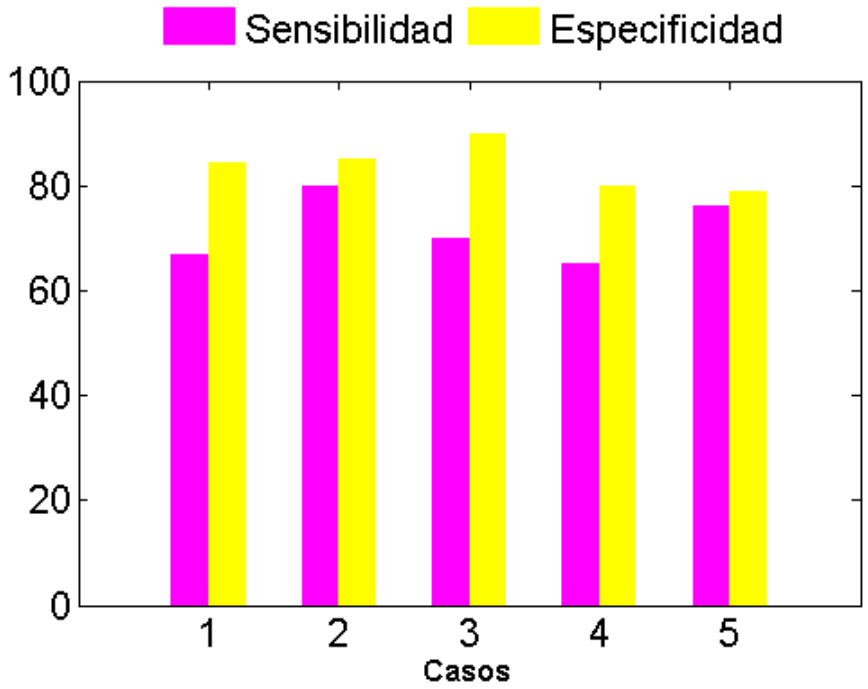



Figura 13. Sensibilidad y especificidad de la fase 1 de reconocimiento de patrones para el caso del clasificador del centro más cercano. Fuente: el autor.

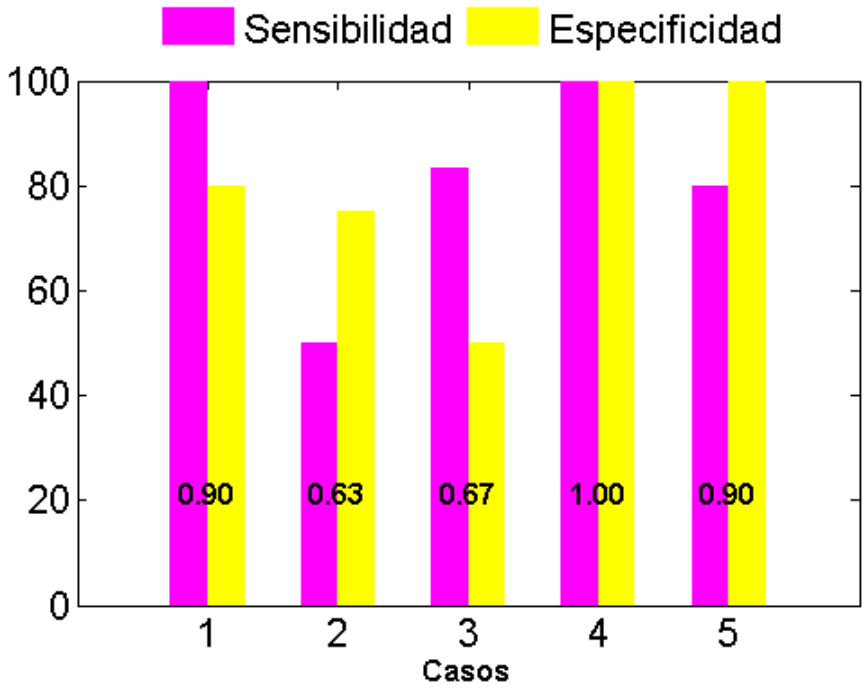



Figura 14. Sensibilidad, especificidad y área bajo la curva ROC del reconocimiento de patrones usando el clasificador del centro más cercano. Fuente: el autor.

Este método de reconocimiento consiste en encontrar la distancia euclidiana entre los datos de la muestra que no se han tocado y los centros de los grupos encontrados en el proceso de clasificación. De modo que, para cada nuevo dato se encuentran dos distancias, donde la menor determina si pertenece al grupo de las pacientes enfermas o al grupo de las sanas.

Los resultados obtenidos en esta técnica son mostrados en la Figura 14. En esta se muestra la especificidad, la sensibilidad y el área bajo la curva ROC para cinco casos en los cuales se dividió la muestra aleatoriamente. En tres casos se obtienen valores de sensibilidad y de especificidad entre el 80% y el 100% y áreas bajo la curva ROC de 0.9 y 1. Esto significa que el sistema clasifica con muy buena eficiencia.

## 5.4 VALIDACIÓN DEL SISTEMA DE INFERENCIA FUZZY

El sistema de inferencia fuzzy implementado busca deducir la presencia o ausencia del carcinoma de glándula mamaria a partir de las variables clínicas seleccionadas. Para este algoritmo se utiliza toda la muestra, aprovechando que es un sistema no supervisado, es decir no necesita entrenarse. Como se mencionó en el capítulo 1 el sistema de inferencia puede ser de tipo Sugeno o Mamdani. Los resultados acá mostrados se limitan al FIS tipo Sugeno, debido al comportamiento lineal en la salida, pero se aclara que el algoritmo desarrollado permite escoger el tipo de sistema que se necesite.

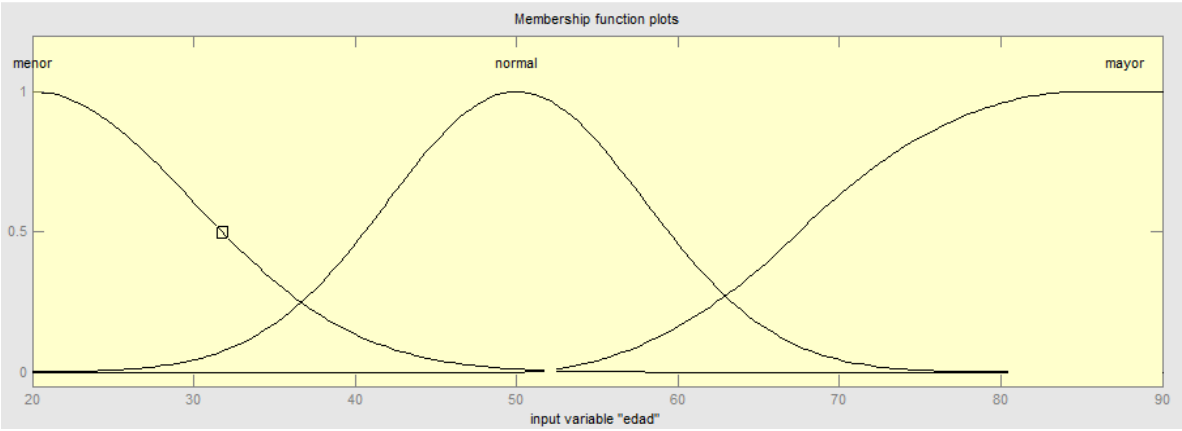
Las variables usadas como entradas al sistema de inferencia fuzzy fueron modificadas con la finalidad de trabajar conjuntos fuzzy. Esto conduce a que exista un rango en cada variable, de modo que puedan formarse grupos que la caractericen, frente al posible diagnostico. Para esto, se calculan dos nuevas variables: una llamada ventana estrogénica, dada por la diferencia entre la edad de la menopausia y la edad de la menarquía. Para pacientes que aún no han llegado al periodo de la menopausia, se toma la diferencia entre la edad actual y la edad de la menarquía. Esta variable, brinda información del tiempo expuesto a hormonas de la paciente. La otra variable, corresponde a la región donde se encuentra la(s) masa(s) superior(es) a 2 cm a criterio del cirujano. La edad, la historia familiar de primer grado, y el número de embarazos son las otras entradas al sistema de inferencia.

Una vez definidas las entradas al sistema se procede a fuzificarlas, es decir, que para cada variable se arman grupos apropiados a través de las funciones de membresía. La forma de las funciones de membresía usadas son campanas gaussianas, descritas por la siguiente ecuación:

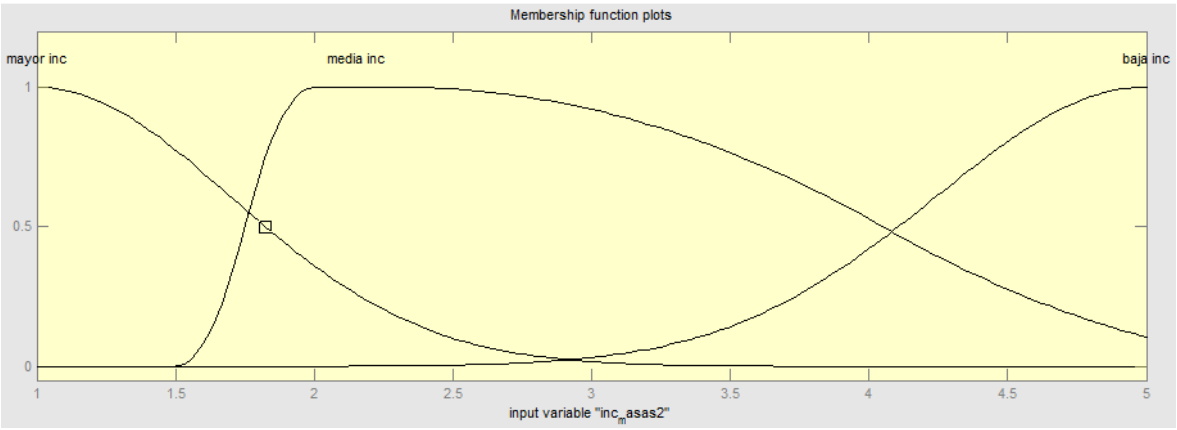
$$f(x, \sigma, c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \qquad (1.32)$$

donde, $c$ representa el centro de la campana y $\sigma$ la apertura [34][41]. En la Figura

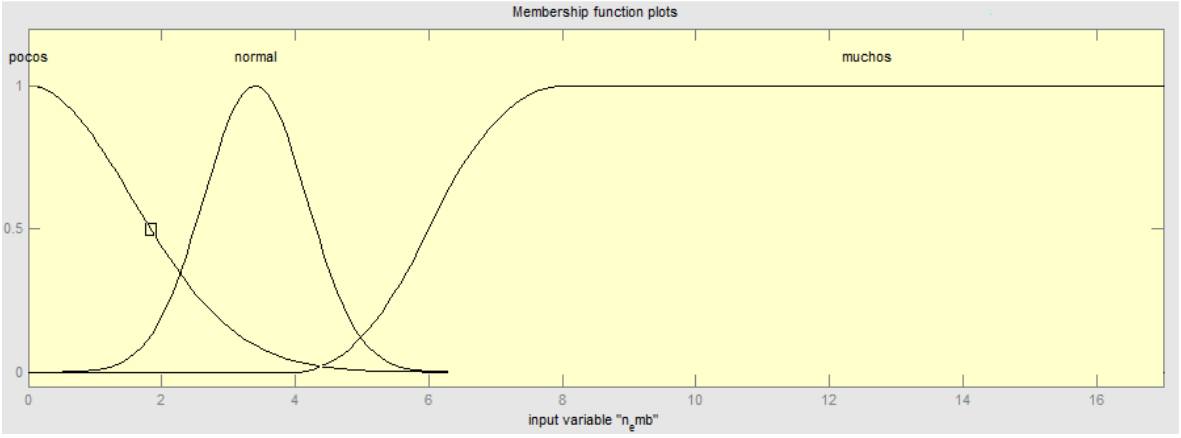
15. se muestra cada variable fuzificada.



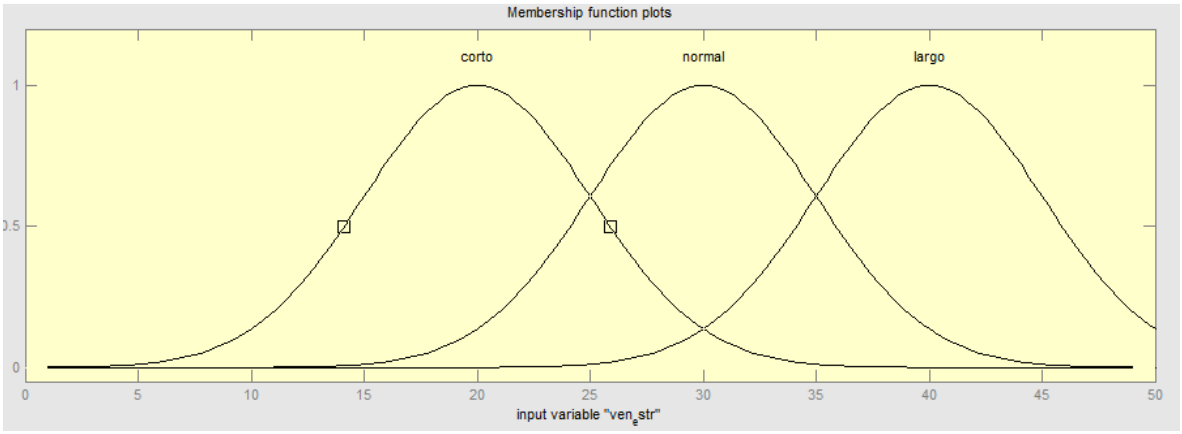

(a) Funciones de membresía de la variable “edad”



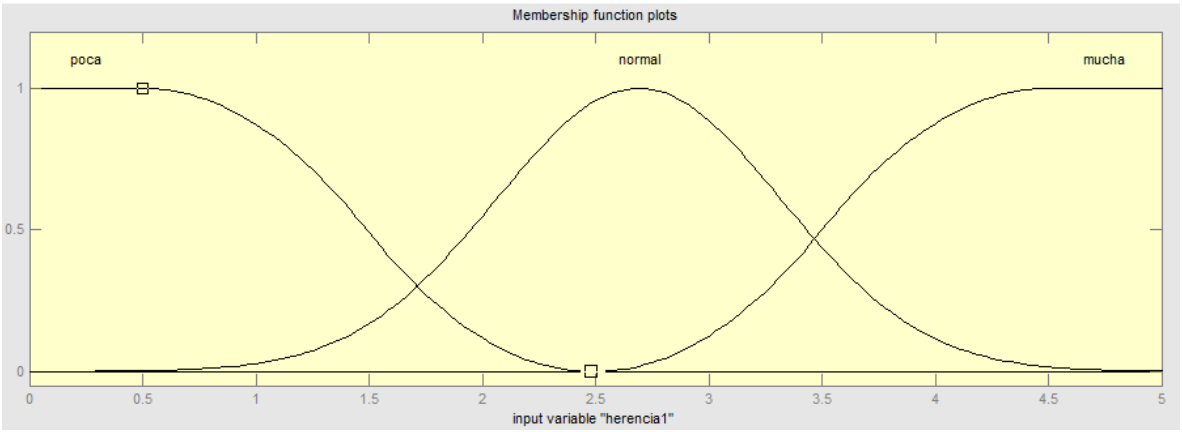

(b) Funciones de membresía de la variable “inc masas2”




(c) Funciones de membresía de la variable “n_emb”




(d) Funciones de membresía de la variable “ven_estr”




(e) Funciones de membresía de la variable “herencia1”

Figura 15. Funciones de membresía de las variables de entrada al sistema FIS. Fuente: el autor.

La Tabla 10. especifica los grupos formados para cada variable describiendo los criterios usados para cada uno.

| **Nombre de la variable independiente** | **Descripción** | **Fuzificación** |
|---|---|---|
| Edad | Edad de la paciente en años | Menor: edad inferior a 35 años |
| | | Normal: edad entre 35 y 65 años |
| | | Mayor: edad superior a 65 años |
| Herencia 1 | Suma de familiares en primer grado (mamá, hijas y hermanas) con cáncer de mama | Poca: Si el rango está entre 0 y 1 familiar |
| | | Normal: Si el rango está entre 1 y 4 familiares |
| | | Mucha: Si son más de 4 familiares |

| | | |
|---|---|---|
| N_emb | Cantidad de embarazos | Pocos: inferior a 2 embarazos |
| | | Normal: Si el rango está entre 2 y 5 embarazos |
| N_emb | Cantidad de embarazos | Muchos: si son más de 5 embarazos |
| Ven_estr | Ventana estrogénica: tiempo correspondiente a la diferencia entre la edad de la menopausia y la edad de la menarquía. | Corto: entre 10 y 25 años |
| | | Normal: entre 25 y 35 años |
| | | Largo: entre 35 y 50 años |
| Inc_masa2 | Incidencia del cáncer de acuerdo a la región donde se encuentran la(s) masa(s) superior(es) a 2cm a percepción del cirujano. | Mucha: Si la(s) masa(s) se ubica(n) en la región CSE ó Centro |
| | | Normal: Si la(s) masa(s) se ubica(n) en la región Centro, CSI ó CIE |
| | | Poca: Si la(s) masa(s) se ubica(n) en la región CIE ó CII |

Tabla 10 . Descripción de las variables de entrada y sus rangos para la fuzificación. Fuente: el autor.

A partir del conocimiento basado en investigaciones [ver Tabla 1] y en la experiencia médica, y conociendo la incidencia que tiene cada grupo de cada variable en el posible diagnóstico de cáncer de mama, se crean las reglas lógicas que modelan el comportamiento del sistema. La cantidad de reglas posibles esta dado por la productoria de los rangos o categorías establecidos para las variables de entrada [50]. En la Figura 16 se muestran las reglas del sistema de inferencia fuzzy, las cuales se crearon por medio de operaciones lógicas AND (mínimo). A dos de estas se les da una ponderación de 0.8, mayor al de las demás, por ser los casos en los que se cumplen todas las características que sugieren uno de los dos diagnósticos. Dado que, el cáncer de mama es considerado una enfermedad heterogénea, puede ser que una paciente no cumpla con la lógica que se plantea en cada regla y resulte en una mala clasificación, esto se verá más adelante en la validación. Con este sistema se consigue disminuir la subjetividad que introduce un médico al momento del diagnóstico de la enfermedad.

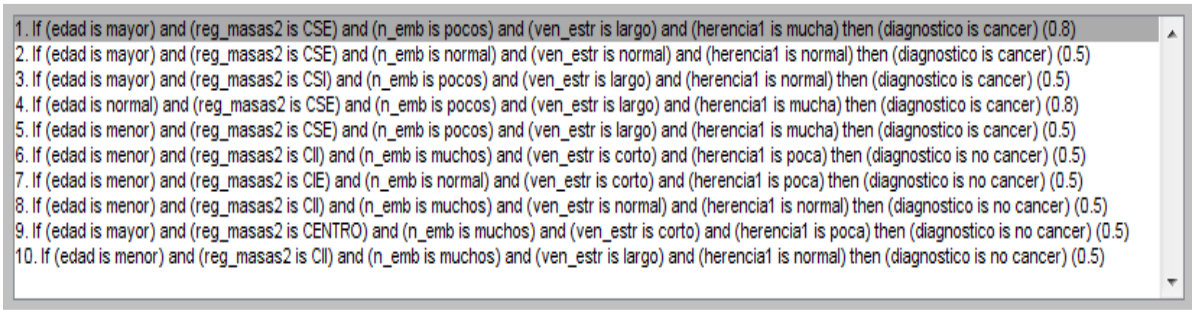



Figura 16. Reglas del sistema FIS. Fuente: el autor.

Al llevar a cabo el proceso de implicación se obtiene un peso correspondiente a cada regla y el resultado del sistema se obtiene calculando el promedio ponderado de todas las salidas [34].

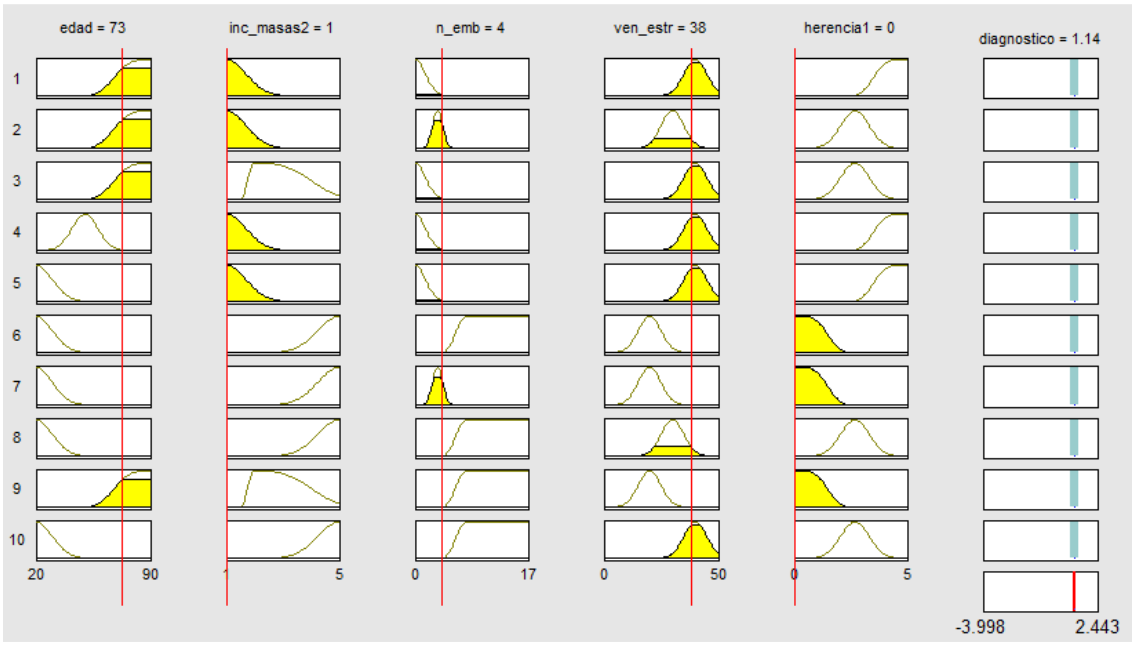



Figura 17. Ejemplo del sistema de inferencia fuzzy para una paciente enferma. Fuente: el autor.

En la Figura 17, se muestra un ejemplo al usar el sistema de inferencia fuzzy implementado, para una paciente enferma de 73 años, que presenta masas en la región más incidente (CSE), 4 embarazos, 38 años expuesta a hormonas y sin historia familiar de cáncer de primer grado. El posible diagnóstico dado por el sistema es que la paciente presenta cáncer de mamá, siendo coherente con el resultado histopatológico.

Debido a que el resultado obtenido con una paciente no es suficiente para evaluar el rendimiento del sistema, se utiliza toda la muestra. Los resultados obtenidos para los parámetros de sensibilidad, especificidad y área bajo la curva ROC se aprecian en la Figura 18.

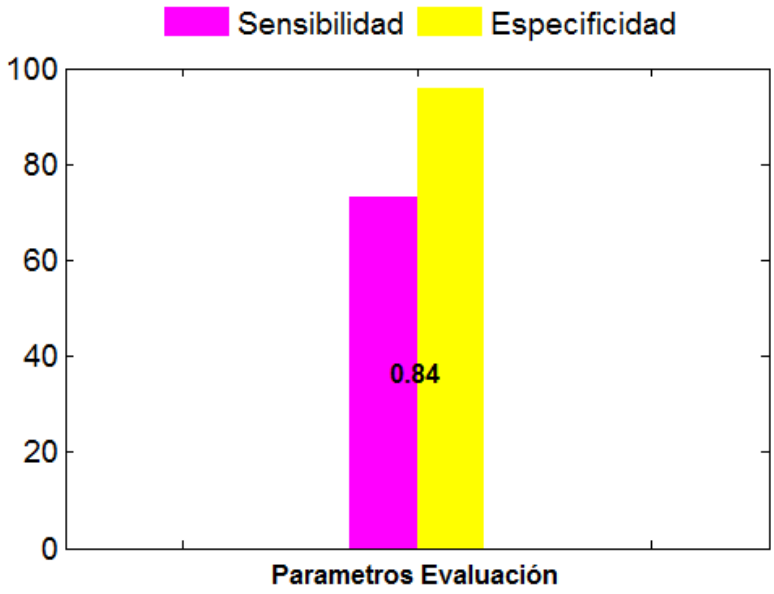



Figura 18. Sensibilidad, especificidad y área bajo la curva ROC para el sistema de inferencia fuzzy. Fuente: el autor.

Los parámetros de evaluación calculados para el sistema de inferencia fuzzy muestran un rendimiento bueno, con una especificidad del 95.8% una sensibilidad del 73.1% y un área bajo la curva ROC de 0.84. Estos resultados indican que el sistema en la mayoría de los casos es capaz de diferenciar las pacientes sanas de las enfermas.

## 5.5 COMPARACIÓN DE LOS ALGORITMOS DE CLASIFICACIÓN

Los resultados de sensibilidad, especificidad y área bajo la curva ROC promediados para los cinco sistemas utilizados en el tamizaje de pacientes con

sospecha de cáncer de mama, se resumen en la Tabla 10. El mejor resultado lo obtiene el sistema de inferencia fuzzy, en el que se resalta el uso de toda la muestra en la validación. El reconocimiento de patrones, produce resultados parecidos para los dos casos: NNC y NCC. Sin embargo, el análisis realizado para varios casos tomando divisiones aleatorias de la muestra, evidencia que el reconocimiento de patrones NNC presenta un mejor desempeño.

Los sistemas fuzzy c-means y hard-c-means, por ser técnicas de inteligencia artificial no supervisadas, usan toda la muestra para el proceso de clasificación, a diferencia del reconocimiento de patrones que es una técnica supervisada. Teniendo en cuenta esto, los resultados obtenidos para estos sistemas son buenos ya que el área bajo la curva ROC se encuentra alrededor de 0.8, presentando el peor comportamiento hard c-means.

| Algoritmo | Sensibilidad | Especificidad | AUC |
|---|---|---|---|
| **HCM** | 73,0 | 83,0 | 0.78 |
| **FCM** | 77,0 | 83,0 | 0,80 |
| **RP NNC**[19] | 78,0 | 88,0 | 0,83 |
| **RP NCC**[20] | 82,0 | 81,0 | 0,82 |
| **FIS** | 73,1 | 95,8 | 0,84 |

Tabla 11. Sensibilidad, especificidad y AUC promediadas a través de todas las pruebas para cada algoritmo de clasificación. Fuente: el autor.

# 6 CONCLUSIONES Y RECOMENDACIONES

El estudio realizado permite determinar ciertos factores, asociados a mujeres, que inciden en la aparición del cáncer de mama en un grupo de pacientes de la región de Santander. Los resultados obtenidos muestran que la edad, el número de embarazos, la edad de la menopausia, la existencia de masas y la presencia de tratamiento hormonal, son las más incidentes y por tanto permiten realizar un

[19] Reconocimiento de patrones usando el concepto del vecino más cercano.
[20] Reconocimiento de patrones usando el concepto del centro más cercano.

proceso de tamizaje de las pacientes.

Las conclusiones obtenidas una vez desarrolladas cada una de las etapas de este trabajo de investigación se presentan a continuación.

Se mostró un procedimiento para la adquisición de los datos. En este se incluyen las encuestas realizadas directamente a las pacientes, el proceso de plasmar en forma digital esta información en una base de datos médica, la corroboración de los datos por comparación y la exportación de los mismos a un documento de texto, con el fin de, ordenar y verificar la información antes de hacer cualquier análisis.

Se analizó y se probó cada uno de los métodos de selección de variables propuestos en este trabajo. Todos los métodos usados posicionan a la edad en primer lugar, es decir, se muestra como el factor de riesgo más incidente en la presencia o ausencia del cáncer de mama para la población en estudio. De esta forma, cualquiera de estos métodos es un buen criterio de selección para la muestra de pacientes con que se realizó la investigación.

Se encontró el valor de especificidad, sensibilidad y área bajo la curva ROC para validar los algoritmos implementados. Se mostró que el sistema de inferencia fuzzy, presenta levemente un mejor rendimiento, entregando una sensibilidad del 73.1% y una especificidad del 95.8% donde se usa toda la muestra para la validación. El reconocimiento de patrones presentó áreas bajo la curva ROC de 0.9 en varios casos y reveló en un caso una clasificación perfecta, al obtenerse un área bajo la curva ROC igual a la unidad usando el 20% de la muestra para la validación. Se resalta que los otros sistemas demuestran una buena eficiencia aún cuando utilizan toda la muestra para la clasificación.

La adquisición de la información fue un proceso arduo que demandó mucho

tiempo, debido a la dificultad para conseguir pacientes dispuestas a participar en este estudio y que presentaran un resultado histopatológico. Además existe muy poca información disponible sobre la detección del carcinoma de glándula mamaria usando variables clínicas. Se sugiere la permanencia en la búsqueda de pacientes, con el fin de ampliar la muestra para estudios posteriores, aprovechando los contactos en la parte médica que hasta el momento se han gestionado.

Se desarrolló una metodología aplicable a diferentes áreas que requieran un proceso de clasificación y reconocimiento. El área bajo la curva ROC se presentó como un parámetro de validación adicional a lo planteado inicialmente, que permite una mejor interpretación de los resultados de los sistemas.

En este trabajo se crearon las reglas del sistema de inferencia a partir del conocimiento del experto y las investigaciones mencionadas. Para lograr que las reglas tengan una base matemática más sólida, se recomienda utilizar un sistema ANFIS, el cual configura tanto las reglas a aplicar como las funciones de membresía por medio de una realimentación entre la salida a optimizar y los datos de entrada.

# 7 BIBLIOGRAFÍA


[1] Organización Mundial de la Salud, Cáncer, Nota descriptiva No. 297. [Citado 23 de abril de 2010]. Disponible en internet: http://www.who.int/mediacentre/factsheets/fs297/es/index.html.

[2] Asociación Ámese, Cáncer de seno en Colombia. [Citado 23 de abril de 2010] Disponible en internet: http://www.amesecolombia.com/home.php?id=22.

[3] Botero J Natalia, Mantilla S. Juan Carlos y Rey S. Juan José. Hallazgos clínicos, mamográficos y ecográficos en un programa comunitario de tamizaje para detección temprana de cáncer de seno en la ciudad de Bucaramanga. Revista Med UNAB vol. 10 No.1 Bucaramanga, abril de 2007. [citado: 23 de Abril de 2010]. Disponible en internet: http://caribdis.unab.edu.co/pls/portal/docs/PAGE/REVISTAMEDUNAB/NUMEROSANTERIORES/VOL%2010%20NO%201%20MAYO%202007/CA%20SENO.PDF

[4] Revista Panamericana de Salud Pública, El cáncer de mama en América Latina y el Caribe, vol.12 no.2 Washington, agosto de 2002, [citado: 23 de Abril de 2010]. Disponible en internet: http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1020-49892002000800016.

[5] Sanabria Álvaro, Romero Javier. La mamografía como método de tamizaje para el cáncer de seno en Colombia. Revista Colombiana de Cirujía vol.20 N°.3 Bogotá, septiembre de 2005, [citado: 25 de Abril de 2010]. Disponible en internet: http://www.encolombia.com/medicina/cirugia/Ciru20305-mamografia.htm

[6] Mushlin AI, Kouides RW Y Shapiro DE. Estimating the Accuracy of Screening Mammography: a meta-analysis. American Journal of Preventive Medicine vol.14 no.2, New York febrero de 1998. [citado 25 de Abril de

2010].

[7] Fletcher Robert Fletcher Suzanne y Wagner Edward. Epidemiología clínica: aspectos fundamentales. Segunda Edición. Google Books, [citado 25 de Abril de 2010]. Disponible en internet: http://books.google.com.co/books?ei=J7vZS66wLoGC8gbV3YR_&ct=result&q=epidemiolog%C3%ADa+cl%C3%ADnica&btnG=Buscar+libros.

[8] Moreno A., Cano V. y García M. Epidemiología clínica. 2ª ed. México: Interamericana. McGraw-Hill; 1994.

[9] Brownson R, Remigton P. y Davis J. Chronic disease epidemiology and control. Baltimore: American Public Health Association, 1993.

[10] Jenicek M. y Cleroux R. Epidemiología. Principios-Técnicas-Aplicaciones. Barcelona: Salvat, 1987.

[11] Merlin J, Barberi-Heyob M y Bachmann N. In vitro comparative evaluation of trastuzumab (Herceptin®) combined with paclitaxel (Taxol®) or docetaxel (Taxotere®) in Her-2-expressing human breast cancer cell lines. Annals of Oncology vol.13 n°.11, France 2002.

[12] Peto J and Mack TM. High constant incidence in twins and other relatives of women with breast cancer. Vol.26 n°.4, Nat Genet 2000.

[13] Pike MC, Spicer DV, Dahmoush L, et al. Estrogens, progestogens, normal breast cell proliferation and breast cancer risk. Rev Epidemiol vol.15, 1993.

[14] Factores de riesgo. [citado 9 de abril de 2010]. Disponible en internet:http://www.elmundo.es/elmundosalud/especiales/cancer/mama6.html

[15] Kelsey J, Fischer D, Holford T, LiVoisi V, Mostow E, Goldenberg IS and White C. Exogenous estrogens and other factors in the epidemiology of breast cancer. Journal of the National Cancer Institute vol 67 n°.2, agosto de 1981.

[16] National Center for Health Statistics. Seer cancer statistics review, 1973-1995. Bethesda, MD: U.S. National Cancer Institute, 1998.

[17] Sturgeon S, Schairer C and Gail M. Geographic variation in mortality from breast cancer among white women in the United States. Journal National Cancer Institute vol.87, 1995.

[18] Rosen P, Groshen S. and Kinne DW. Factors influencing prognosis in node-negative breast carcinoma: analysis of 767 T1N0M0/T2N0M0 patients with long-term follow up. Journal Clinic Oncology. vol.11 1993.

[19] Slamon D, Clark G and Woung S. Human breast cancer: correlation of relapse and survival with amplification of the Her-2/neu oncogene. Science 1987; vol.235 n°4785, enero 1987.

[20] Hsieh C-C, Trichopoulos D, Katsouyanni K, et al. Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study. Institute Journals Cancer vol.46, 1990.

[21] Kelsey J, Gammon M and John E. Reproductive factors and breast cancer. Epidemiol Rev vol.15 1993.

[22] Brinton L, Schairer C, Hoover R. Menstrual factors and risk of breast cancer. Cancer Investigation vol.6, 1988.

[23] Layde P, Webster L and Baughman A. The independent associations of parity, age at first full term pregnancy, and duration of breastfeeding with the risk of breast cancer. Cancer and Steroid Hormone Study Group. Journal Clinic Epidemiology vol.42, 1989.

[24] Zheng T, Holford T and Mayne S. Lactation and breast cancer risk: a case-control study in Connecticut. Br J Cancer vol.84, 2001.

[25] Collaborative Group on Hormonal Factors in Breast Cancer (Writing Committee: Beral V, Bull D, Peto R). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease vol.360 Lancet 2002.

[26] Jernstrom H, Lubinski J and Lynch HT. Breast-feeding and the risk of breast

cancer in BRCA1 and BRCA2 mutation carriers. Journal National Cancer Institute vol.96 2004.

[27] Alférez Edwin Santiago. Detección del carcinoma de glándula mamaria fusionando variables clínicas y termográficas. Trabajo de Investigación de maestría. Universidad Industrial de Santander. Bucaramanga, abril de 2010.

[28] Medex Behavioral Science Page. [citada 22 de abril de 2010]. Disponible en internet: http://faculty.washington.edu/alexbert/MEDEX/.

[29] Ismail Saritas, Novruz Allahverdi and Ibrahim Unal, A Fuzzy Expert System Design for Diagnosis of Prostate Cancer. Ney York 2003.

[30] Seker H, Odetayo M, Petrovic D and Naguib R. A Fuzzy Logic Based Method for Prognostic Decision Making in Breast and Prostate Cancers, IEEE Trans. on Information Technology in Biomedicine, (In Press), 2003.

[31] Abbod M, Von Keyserlingk D, Linkens D and Mahfouf M. Survey of Utilization of Fuzzy Technology in Medicine and Healthcare, Fuzzy Sets and Systems, vol.120, 2001.

[32] Nguyen H, and Kreinovich V, Fuzzy Logic and Its Applications in Medicine, International Journal of Medical Informatics vol.62, p.165–173, 2001.

[33] E. Y. K Ng, and E. C. Kee, Advanced integrated technique in breast cancer thermography, Journal of Medical Engineering & Technology, p.1-12, 2007.

[34] Timothy J. Ross, Fuzzy Logic with Engineering Applications. Second edition. McGraw Hill. New York. 2004.

[35] Höppner Frank, Fuzzy cluster analysis methods for classification, data analysis, and images. Editorial John Wiley & Sons Ltd. New York 2000.

[36] Bezdek, J. Pattern recognition with fuzzy objective function algorithms. Plenum. New York, 1981.

[37] Sato Mika, C. Jain Lakhmi, Innovations in Fuzzy Clustering theory and applications. Editorial Springer. The Netherlands, 2006.

[38] Fukunaga Keinosuke, Introduction to statistical pattern recognition, Second edition, Academic press, USA 1990.

[39] Wang, P. (1983). ‘‘Approaching degree method,’’ in Fuzzy sets theory and its applications, Science and Technology Press, Shanghai, PRC (in Chinese).

[40] Jaimes Rosas O, García Alonso A, Mas Oliva J, Alvarez Icaza L, Diagnóstico de riesgo de aterogénesis asistido por lógica borrosa, Instituto de Ingeniería, UNAM, Instituto de Fisiología Celular, UNAM, Mayo 2006.

[41] Matlab® Versión 7.6.0 R2008a, Fuzzy Logic Toolbox™ User’s Guide, The MathWorks, Inc 1995-2010.

[42] Ferreira A, Fuentes R, Ambiente de desarrollo iteractivo para lógica difusa, UAM-Azcapotzalco, departamento de electrónica, México D.F

[43] Hanley James A, McNeil Barbara J, The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. Departament of Epidemiology and Health. McHill University, Montereal, Canadá (J.A.H.) and the Departament of Radiology, Harvard Medical School and Brigham and Women’s Hospital, Boston, MA (B.J.M.) Revision Receiver Dic. 15 de 1981.

[44] Fernández Pita y Díaz Pértegas, Investigación Pruebas Diagnósticas, Unidad de epidemiología clínica y bioestadística, Complexo Hospitalario-Universitario Juan Canalejo. A Coruña España, 2003.

[45] Westin Lena K. Receiver Operating Characteristic Analysis, Evaluating discriminance effects among decision support systems, Department of Computing Science. Umea University.

[46] Grupo de Conectividad y Procesado de Señal (CPS), Grupo de Investigación en Patología Oncológica (ONCOPAT), Innovación y Desarrollo tecnológico de Unisangil (IDENTUS). Evaluación de la termografía infrarroja en la detección del carcinoma de glándula mamaria**.** Propuesta a COLCIENCIAS. Universidad Industrial de Santander – Fundación Universitaria de San Gil, 2006.

[47] Marques Dos Santos María José, Estadística Básica un enfoque no

paramétrico, Universidad Nacional Autónoma de México, Facultad de estudios superiores Zaragoza.

[48] Mamdani E. and S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, Vol. 7 N°.1, p.1-13, 1975.

[49] Hastie T, Tibshirani R. and Friedman J, The Elements of Statistical Learning, editorial Springer, 2001.

[50] Devore J. Probability and Statistics for Engineering and the Sciences, fourth edition, Duxbury, Press, Belmont, CA, 1995.

# ANEXO A. CONSENTIMIENTO INFORMADO

Estimada Paciente:

Las enfermedades de los senos son cada vez más frecuentes en las mujeres en todo el mundo. Por esto, los investigadores del proyecto de investigación "**EVALUACIÓN DE LA TERMOGRAFÍA INFRARROJA EN LA DETECCIÓN DEL CARCINOMA DE GLÁNDULA MAMARIA**" realizado por miembros del Departamento de Patología de la Universidad Industrial de Santander (UIS) y de la Unidad de Oncología del Hospital Universitario (HUS) de Santander, en unión con el Grupo de Investigación de Conectividad y Procesado de Señal (CPS) y el grupo de investigación en Innovación y Desarrollo Tecnológico de Unisangil (IDENTUS) preocupados ésta situación estamos realizando este estudio que tiene como objetivo conocer la temperatura de las distintas partes del seno y sobre todo la temperatura específica del trastorno que pudiera tener el seno enfermo. Esto permite saber si éste método es útil para detectar el cáncer de seno en mujeres de cualquier edad.

La técnica de termografía infrarroja, es un examen que se puede realizar en cualquier parte del cuerpo, pero en éste estudio lo vamos a realizar a los senos únicamente. Se hace mediante el uso de una cámara diseñada especialmente para esto. La cámara se utiliza como una cámara fotográfica, de manera que registra la temperatura superficial de los senos. Lo que se observa después de la toma es una imagen de los senos con distintos colores en donde las zonas más calientes se ven de un color diferente al de las zonas más frías. Eso hace que si en el seno hay alguna masa, ésta pudiera resultar más fría o más caliente que el resto del seno y haría pensar que existe enfermedad en ese sitio.

La toma de la imagen es únicamente de los senos y no de su rostro ni do otro sitio del cuerpo. La cámara no estará en contacto físico con usted, no genera rayos x y es un procedimiento rápido e indoloro.

Este estudio se realizará en 200 pacientes, que es el número de personas necesarias para poder hacer un análisis valedero de los resultados. Su participación es completamente voluntaria y la realización de la técnica no implica ningún costo para usted como tampoco habrá compensación económica alguna. Los miembros del grupo investigador estarán en disposición de brindarle ahora y en el futuro cualquier información o pregunta que le surja acerca de los resultados o del procedimiento; para eso se suministrará en el momento de su valoración los datos de los responsables de aclarar dudas al respecto.

La información generada por este estudio es estrictamente confidencial y se mantendrá su privacidad. La información del estudio no será utilizada para generar beneficios económicos. Usted es libre de rehusar a participar en este estudio en cualquier momento sin que esto conlleve a cambios en su futuro cuidado.

Usted tiene derecho a conocer los resultados de los estudios realizados cuando lo desee, una vez se haya realizado el análisis de las imágenes así como de solicitar que no sean incluidos en las conclusiones del trabajo. De igual forma el grupo investigador podrá tomar la decisión de retirarla del estudio si lo considera conveniente.

**PARTICIPANTE**

Yo _____ Firma _____

Cédula No. _____ de _____

He leído y recibido copia del presente consentimiento informado. Habiendo comprendido el significado de la investigación declaro estar debidamente informada y consiento en participar en este el estudio.

Ciudad: _____ Fecha: _____Hora:_____

**TESTIGOS**

Nombre _____ Firma _____

Nombre:_____ Firma _____

Ciudad: _____ Fecha: _____ Hora:_____

**INVESTIGADOR QUE BRINDA EL CONSENTIMIENTO**

Nombre:_____ Firma _____

Telefóno: _____

Ciudad: _____ Fecha: _____ Hora:_____

# ANEXO B.FORMATOS DE RECOLECIÓN DE LA INFORMACIÓN

ENCUESTA SOBRE LOS FACTORES ASOCIADOS SOCIODEMOGRÁFICOS, HEREDITARIOS Y HORMONALES ENDÓGENOS EN LAS PACIENTES CON CARCINOMA INFILTRANTE DE LA GLÁNDULA MAMARIA EN EL DEPARTAMENTO DE SANTANDER

Se guardará la confidencialidad de los datos y en ningún momento se revelará la identificación de los pacientes.

NOTA: Es importante que al registrar los datos tenga en cuenta las unidades de medida, marque con una X la respuesta en la casilla que corresponda y evite dejar espacios en blanco.

Nombre (s) y Apellidos:
Dirección: Barrio:
Tél. Fecha hoy: dd/mm/aa
( / / )
Código de las Imágenes: Frontal:_____
Lateral Der _____ Lateral Izq _____
Oblicua Der._____ Oblicua Izq_____
No. de Historia Clínica: Cédula:
Talla del Brasier: □ □ □

DATOS SOCIODEMOGRAFICOS

1. ¿Cuántos años cumplidos tiene?: 2. ¿Cuál es su fecha de nacimiento?
□□ años ( / / ) dd/mm/aa
3. En el último año, ¿dónde ha residido? Nombre del lugar (municipio y/o vereda)_____

Área urbana□ Área rural □

4. De acuerdo a su recibo de luz, ¿A cual estrato corresponde su vivienda?
1□ 2□ 3□ 4□ 5□ 6□
5. En su opinión, ¿a cuál de las siguientes razas pertenece usted?
1. Blanca □ 2. Mestiza □ 3. Negra □ 4. No sabe □ 5. No responde □
6. ¿Cuál es su Estado Civil?:

Soltero □
Casado □
Divorciado □
Unión Libre □
Otro □

7. ¿Convive con pareja estable y permanente?
Si □ No □
8. ¿Cuál fue el último grado de estudios que usted aprobó?

| Tipo de enseñanza | No. de años |
|---|---|
| Ninguna | 0□ |
| Primaria | 1□ 2□ 3□ 4□ 5□ |

| | |
|---|---|
| Secundaria | 6□ 7□ 8□ 9□ 10□ 11□ |
| Técnica o Universitaria, incluyendo Postgrados | 1□. 2□. 3□. 4□. 5□. 6□. 7□. 8□. 9□. 10□ |

9. ¿A qué se dedicó la mayor parte del tiempo en el ultimo año?

| | |
|---|---|
| Trabajó | □ |
| Trabajó y estudió | □ |
| Estudió (a) | □ |
| Actividades del hogar | □ |
| Buscó trabajo | □ |
| Pensionado (a) | □ |
| Retirado sin pensión | □ |
| Otra | □ ¿Cuál? |

10. ¿En cuál de los siguientes rangos está el ingreso mensual de su familia (personas que aportan económicamente para el sostenimiento de su hogar)?

1. $0-$496.800□ 2. $496.800-$993.600□ 3. $993.600-$1987.200□ 4. $1987.200-$3.974.400□ 5. $3.974.400 o más □ 6.No sabe □ 7. Rehúsa contestar □

COBERTURA Y ACCESO A LA ATENCIÓN MÉDICA

11. ¿A su familia alguna vez le aplicaron la encuesta del SISBEN?

Si □ No □ □Si “No” pase a la pregunta 14 No sabe/No recuerda □ □Si “No” pase a la pregunta 14

12. Después del año 2001, ¿Le han aplicado la encuesta del SISBEN a su familia?

Si □ No □ No sabe/No recuerda □

13. ¿En que nivel del SISBEN está clasificado? □

14. ¿En el ultimo año ha estado o estuvo asegurado o afiliado a un plan de salud como cotizante?

Si □ No □ Si, pero no siempre □ No sabe/No recuerda □

15. ¿Actualmente esta asegurado o afiliado a un plan de salud como cotizante?

Si □ No □ No sabe/No recuerda □

16. ¿En el ultimo año ha estado o estuvo asegurado o afiliado a un plan de salud como beneficiario?

Si □ No □ Si, pero no siempre □ No sabe/No recuerda □

17. ¿Actualmente esta asegurado o afiliado a un plan de salud como beneficiario?

Si □ No □ No sabe/No recuerda □

18. ¿Actualmente a que entidad de salud está afiliado o es beneficiario?

| | | | |
|---|---|---|---|
| 1.Nueva EPS (ISS) | □ | 6. Fuerzas Militares, Policía Nacional | □ |
| 2.Administradora de régimen Subsidiado (ARS) | □ | 7. ECOPETROL | □ |
| 3.Empresa promotora de | □ | 8. Magisterio | □ |

Salud (EPS)

4.Empresa de Medicina prepagada □ 9. Ninguna □

5.Empresa Solidaria □ 10. Otra, ¿Cuál?

19. Nombres de la entidad de salud a la cual está afiliado

_____

HISTORIA DE SALUD FAMILIAR

20. ¿Su mamá biológica tiene o tuvo cáncer?

Si □ No □ □Si “No” pase a la pregunta 23 No sabe □ □ Si “No sabe” Pase a la pregunta 23

21. ¿Qué edad tenía su mamá cuando le diagnosticaron cáncer?

□□ años

22. ¿En qué sitio/órgano su mamá tiene o tuvo cáncer?

_____

23. ¿Su mamá biológica está viva?

Si □ No □ No sabe □

24. ¿Su papá biológico tiene o tuvo cáncer?

Si □ No □ □Si “No” pase a la pregunta 27 No sabe □ □ Si “No” pase a la pregunta 27

25. ¿Qué edad tenía su papá cuando le diagnosticaron cáncer?

□□ años

26. ¿En qué sitio/órgano su papá tiene o tuvo cáncer?

_____

27. ¿Cuántas hermanas tiene usted? □ □ Si no tiene hermanas pase a la pregunta 31

28. ¿Alguna de sus hermanas tiene o tuvo cáncer de mama?

Si □ No □ □Si “No” pase a la pregunta 31 No sabe □

29. ¿Cuántas hermanas tuvieron cáncer de mama? □ □

30. ¿Qué edad tenía(n) su(s) hermana(s) cuando le diagnosticaron cáncer de mama?

1. □□ años 2. □□ años 3. □□ años

31. ¿Algunas de sus hijas tiene o tuvo cáncer de mamá?

Si □ No □ □Si “No” pase a la pregunta 34 No sabe □

32. ¿Cuántas hijas tienen o tuvieron cáncer de mama? □□

33. ¿Qué edad tenía(n) su(s) hija(s) cuando le(s) diagnosticaron cáncer de mama?

1. □□ años 2. □□ años 3. □□ años

34. ¿Alguna de sus tías o primas tiene o tuvo cáncer de mama?

Si □ No □ □ Si “No” pase a la pregunta 39 No sabe □

35. ¿Cuántas tías tienen o tuvieron cáncer de mama? □□

36. ¿Qué edad tenía(n) su(s) tía(s) cuando le(s) diagnosticaron cáncer de mama?

1. □□ años 2. □□ años 3. □□ años

37. ¿Cuántas primas tienen o tuvieron cáncer de mama?

□□ Ninguna □Si “Ninguna” pase a la pregunta 39

38. ¿Qué edad tenía(n) su(s) prima(s) cuando le(s) diagnosticaron cáncer de mama?

1. □□ años 2. □□ años 3. □□ años

ANTECEDENTES PERSONALES

39. ¿Qué edad tenía usted cuando tuvo su primera menstruación, regla o periodo?

□□ años

40. ¿Ha tenido embarazos?

Si □ No □ □ Si “No” pase a la pregunta 45 ¿Cuántos? □□

41. ¿Ha tenido recién nacidos muertos?
Si □ No □ Si la respuesta es “si” ¿cuantos?_____

42. ¿Ha tenido abortos?
Si □ No □ Si la respuesta es “si” cuantos abortos?_____

43. ¿A qué edad tuvo su primer embarazo?
□□ años

44. ¿Alimentó con leche materna a su(s) hijo(s)?
Si □ No □
¿Cuánto tiempo alimento con leche materna a su(s) hijo(s)?
1. Hijo_____(meses)
2. Hijo_____(meses)
3. Hijo_____(meses)
4. Hijo_____(meses)
5. Hijo_____(meses)
6. Mas? Si □ No □ Si la respuesta es “si”, sumé el tiempo (aproximadamente) de los restantes en que los amamanto_____

45. ¿Ha sido diagnosticada alguna vez alguno de estos canceres?:
Mama Si □ No □
Ovario Si □ No □
Útero Si □ No □

46. ¿Ha transcurrido más de 12 meses desde su última menstruación?
Si □ No □

47. ¿Cuál fue la fecha de su última regla (el primer día de sangrado o menstruación)?
( / / ) dd/mm/aa

48. ¿Recibió tratamiento hormonal para la menopausia?
Si □ No □
Si su respuesta es “Si”, ¿Durante cuanto tiempo los uso o los ha venido usando?_____
Si recuerda, qué tipo de medicamento ¿usó?_____

49. ¿Recibe tratamiento hormonal para la menopausia?
Si □ No □
Si su respuesta es “Si”, ¿Durante cuanto tiempo los ha venido usando?_____
Si recuerda, qué tipo de medicamento ¿usó?_____

50. ¿Le han tomado mamografías en los últimos 3 años?
Si □ No □
Si su respuesta es “Si”, ¿la última mamografía fue tomada?
En el último año □Entre 1 y 2 años □Entre 2 y 3 años □

51. ¿Alguna vez le han practicado cirugía o biopsia en los senos?
Si □ No □ □ Si “No” pase a la pregunta 54

52. ¿Hace cuanto tiempo le realizaron la última intervención (cirugía o biopsia)?
En los últimos 3 meses □ Entre 3 y 12 meses □Entre 1 y 2 años □
¿Más años?¿Cuantos? _____

53. ¿En que seno le realizaron el procedimiento (cirugía o biopsia)?
Derecho □ Izquierdo □

54. ¿Ha recibido alguna vez tratamiento de radioterapia?

Si □ No □ Si la respuesta es "si", por cuanto tiempo?_____

INFORMACION DEL ESTADO DURANTE LA TOMA DE LA TERMOGRAFIA

55. ¿Ha realizado alguna actividad (caminata, bronceado, etc.) donde se haya expuesto prolongadamente al sol durante los últimos 5 días?
Si □ No □
Si la respuesta es "si" ¿Por cuánto tiempo? _____horas

56. ¿Hoy usó lociones, cremas, polvos o algún tipo de maquillaje en el área?
Si □ No □ ¿Cuál? _____
57. ¿Hoy usó desodorante o antitranspirante?
Si □ No □
58. ¿Realizó alguna terapia física en las últimas 24 horas?
Si □ No □
59. ¿Realizó algún ejercicio físico 4 horas antes del examen?
Si □ No □
60. ¿Ha tomado algún medicamento para el dolor o vaso dilatador el día del examen?
Si □ No □

**REGISTRO DE LA INFORMACIÓN CLÍNICA**

Se guardará la confidencialidad de los datos y en ningún momento se revelará la identificación de los pacientes.

NOTA: Es importante que al registrar los datos tenga en cuenta las unidades de medida, marque con una X la respuesta en la casilla que corresponda y evite dejar espacios en blanco.

Nombre (s) y Apellidos:
Dirección: Barrio:
Tél: Fecha hoy dd/mm/aa
( / / )
Código de las Imágenes:Frontal:_____Lateral
Der_____ Lateral Izq _____ Oblicua
Der._____Oblicua Izq_____
No. de Historia Clínica: Cédula:

Estatura: _____ Peso:_____ IMC:_____

**DATOS DEL EXÁMEN CLÍNICO**

1. ¿La paciente presenta dolor en las glándulas mamarias?
   Si□ No □ Si la respuesta es “si”, ¿En cuál glándula mamaria?
   Derecha □ Izquierda.
   En cuál región de la glándula mamaria?
   CSE □ CSI □ CIE □ CII □ Centro □
2. ¿La paciente se ha palpado masa(s) en la(s) glándula(s) mamaria(s)?
   Si□ No □ Si la respuesta es “si”, ¿En cuál glándula mamaria?
   Derecha □ Izquierda
   En cuál región de la glándula mamaria?
   CSE □ CSI □ CIE □ CII □ Centro □
3. La paciente presenta asimetría en la inspección de las glándulas mamarias?
   Si□ No □
4. ¿La paciente presenta Telorrea?
   Si□ No □ Si la respuesta es “si” ¿La Telorrea es espontánea? Si□ No □,
   Si hay telorrea, tipo de secreción: Sanguinolenta □ Verdosa□ Blanquecina □
   Serosa □
5. ¿La paciente presenta masa(s) con dimensiones superiores a 2 *cm* en la glándula mamaria?
   Si□ Posiblemente □ No □ Si la respuesta es “si” o “Posiblemente”,
   ¿En que región se ubica(n) la(s) masa(s)?
   CSE □ CSI □ CIE □ CII □ Centro □
   ¿Las masa(s) es (son)?
   Móvil(es) □ Fija(s) □
6. ¿La paciente presenta alteraciones en la piel de la glándula mamaria?
   Si□ No □ Si la respuesta es “si”,
   ¿Qué tipo de anormalidad existe?
   Eritema □ Edema(piel de naranja) □ Nodulaciones □ Retracción □
Ulceración□
   ¿En qué región presenta la anormalidad de la piel?
   CSE □ CSI □ CIE □ CII □ Pezón y areola □
7. ¿La paciente presenta anomalías en los nódulos linfoides axilares?
   Si□ No □ Si la respuesta es “si”,
   ¿Qué nivel de anomalía se encontró?
   N1(menores a 2 *cm*) □ N2 (mayores a 2 *cm*) □ N3 □
8. ¿La paciente presenta nódulo (s) supraclavicular(es)?
   Si□ No □ Si la respuesta es “si”,
   ¿Qué nivel de anomalía se encontró?
   N1(menores a 2 *cm*) □ N2 (mayores a 2 *cm*) □ N3 □
9. Tiene resultado de mamografía?
   Si□ No □ Si la respuesta es “si”,
   Cuál es el resultado de la mamografía?
   BIRADS 0 □ BIRADS I □ BIRADS II □ BIRADS III □ BIRADS IV □ BIRADS V□