Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Académico Profesional de Estadística

Clasificación de tumores de mama usando métodos estadísticos

TESINA

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Carla Patricia ZÚÑIGA VILCA

ASESOR

Ana María CÁRDENAS ROJAS

Lima, Perú

2009



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Zúñiga, C. (2009). *Clasificación de tumores de mama usando métodos estadísticos*. Tesina para optar el título profesional de Licenciada en Estadística. Escuela Académico Profesional de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

I N D I C E

Pág.

INTRODUCCION 1

CAPITULO I: Aspectos Generales 3

1.1 Planteamiento del Problema 3
1.2 Importancia del Problema 4
1.3 Justificación del Problema 4
1.4 Ámbito Geográfico 4
1.5 Objetivos 5
1.5.1 Objetivo General 5
1.5.2 Objetivos Específicos 5
1.6 Metodología 5
1.6.1 Tipo de Investigación 5
1.6.2 Población de Estudio 5
1.6.3 Variables de Estudio 6
1.6.4 Metodología a Usar 6

CAPITULO II: Marco Teórico 8

2.1 Análisis de Regresión Logística 10
2.1.1 Modelo de Regresión Logística 11
2.1.2 Estimación de los Parámetros del Modelo 12
2.1.3 Evaluación del Modelo Ajustado 14
2.2 Redes Neuronales Artificiales 16
2.2.1 Historia de las Redes Neuronales 18
2.2.2 Red Neuronal Biológica 21
2.2.3 Red Neuronal Artificial 23
2.2.3.1 Elementos de una Red Neuronal Artificial 23
2.2.3.2 Modelo de una Red Neuronal Artificial 25
2.2.3.3 Arquitectura de una Red Neuronal Artificial 27
2.2.3.4 Aprendizaje de una Red Neuronal Artificial 28
2.2.3.5 Perceptrón 30
2.2.3.5.1 Definición 31
2.2.3.5.2 Aprendizaje del Perceptrón 32
2.2.3.6 Red Neuronal Multicapa 34
2.2.3.6.1 Definición 35

2.2.3.6.2 Estructura de la Red Multicapa 35
2.2.3.6.3 Reglas de Aprendizaje de Redes Multicapa 36
2.2.3.7 Backpropagation 41
2.2.3.8 Sobreentrenamiento 45
2.2.3.9 Evaluación de Modelos 45
2.2.3.9.1 Error de Entrenamiento y Error de Prueba 46
2.3 Comparación Teórica de Regresión Logística y Redes Neuronales Artificiales 48

CAPITULO III: Aplicación: Clasificación de Tumores de Mama de Pacientes de la Universidad de Wisconsin 49

CAPITULO IV: Análisis y Resultados 56

4.1 Análisis Descriptivo 58
4.2 Análisis de regresión Logística 61
4.3 Redes Neuronales Artificiales 65
4.4 Comparación de los Modelos de Clasificación 76

CONCLUSIONES 78

ANEXO A: Análisis Descriptivo 79
ANEXO B: Análisis de Regresión Logística 85
ANEXO C: Manual de Software Tanagra 87

REFERENCIAS BIBLIOGRAFICAS 93

RESUMEN

CLASIFICACION DE TUMORES DE MAMA USANDO METODOS ESTADISTICOS

CARLA PATRICIA ZUÑIGA VILCA

DICIEMBRE – 2009

Orientador: Mg. Ana María Cárdenas Rojas
Título Obtenido: Licenciado en Estadística

En este trabajo el objetivo es clasificar la presencia de tumores malignos y benignos en la detección de cáncer de mama en los pacientes del Hospital de la Universidad de Wisconsin a través de las características de las células, usando métodos estadísticos de clasificación.

Los métodos estadísticos que se utilizarán para la clasificación de los tumores benigno y maligno en cáncer de mama son: análisis de regresión logística, y redes neuronales artificiales. Una vez obtenida la clasificación por estos dos métodos, se realizará una comparación y se verificará qué método nos da una mejor clasificación.

PALABRAS CLAVES: TUMORES DE MAMA
CLASIFICACIÓN ESTADISTICA
REGRESION LOGISTICA
REDES NEURONALES ARTIFICIALES

ABSTRACT

BREAST TUMORS' CLASSIFICATION USING STATISTICAL METHODS

CARLA PATRICIA ZUÑIGA VILCA

DECEMBER – 2009

Tutor: Mg. Ana María Cárdenas Rojas
Título: Degree in Statistics

The objective in this work is to classify the presence of malignant and benign in the breast cancer's detection in patients in the Hospital of the University of Wisconsin through the cell's characteristics, using statistical methods of classification.

The statistical methods are used for classification of benign and malignant tumours in breast cancer are logistic regression analysis and neural networks. Once the classification is obtained by the two methods, there will be a comparison and verify what method gives us a better classification.

KEY WORDS: BREAST TUMORS
STATISTICAL CLASSIFICATION
LOGISTIC REGRESSION
NEURAL NETWORKS

INTRODUCCIÓN

Los problemas de clasificación ocurren en las diferentes áreas de la investigación científica. Por ejemplo en el área de medicina, “diagnosticar” es equivalente a clasificar a un sujeto en una determinada enfermedad.

Desde el punto de vista estadístico clasificar a un sujeto en un grupo determinado tiene un cierto grado de incertidumbre, lo que hace pensar en la utilización de una metodología probabilística, que nos permita cuantificar esa incertidumbre.

Existen muchos métodos de clasificación estadística que han sido utilizados en diversas investigaciones científicas tales como el análisis discriminante, el análisis clúster, análisis de regresión logística, redes neuronales y otros.

En este trabajo se aplicarán dos métodos de clasificación estadística que son: el análisis de regresión logística y redes neuronales artificiales.

Para la aplicación del análisis de regresión logística en la clasificación de tumores de mama se dispone de datos nacionales; no así para la aplicación de las redes neuronales dado que se requiere un número de observaciones grande y en nuestro medio no existen suficientes observaciones para aplicarlas; debido a que la técnica de punción de aspiración de aguja fina para detectar la presencia de tumores de mama en el país, se viene realizando desde el año 2008.

Siendo el interés dar a conocer las redes neuronales con una buena alternativa frente a los métodos de clasificación estadística clásicos; la aplicación se realiza al conjunto de datos "Wisconsin Breast Cancer", éste conjunto de datos fue obtenido por el Dr. William H. Wolberg en la investigación que llevó a cabo en el Hospital de la Universidad de Wisconsin en Noviembre de 1995.

El objetivo de este trabajo es clasificar la presencia de tumores malignos y benignos en los pacientes del Hospital de la Universidad de Wisconsin a través de las características de las células, usando los métodos estadísticos de clasificación, cuyos datos son motivo de análisis

Este trabajo está estructurado en cuatro capítulos. El primero, corresponde a los aspectos generales dentro de este capítulo se muestra el planteamiento, la importancia, la justificación y los objetivos del problema. El segundo, contiene el marco teórico en el cual se difundirá la teoría del análisis de regresión logística y redes neuronales artificiales. El tercero, corresponde a la aplicación en la cual se presenta el desarrollo práctico de las metodologías aplicadas. El cuarto, incluye el análisis y resultados. Finalmente se encuentran las conclusiones y el anexo con las respectivas salidas del SPSS, Tanagra y Microsoft Office Excel 2007 a fin de comprobar los resultados.

CAPITULO I

ASPECTOS GENERALES

1.1 PLANTEAMIENTO DEL PROBLEMA

A pesar de que el cáncer de mama es de fácil diagnóstico y puede ser controlada con un tratamiento adecuado es la causa más frecuente de mortalidad en mujeres en todo el mundo, principalmente por la falta de conciencia de las mujeres sobre su enfermedad y por las limitaciones en los despistajes de mamografías.

De acuerdo U.S Cancer Statistics Working Group [21], a 186 467 mujeres se les diagnóstico cáncer de mama de las cuales 41 116 murieron. En el Perú según la "Revista Diagnóstico [17], el cáncer de mama en mujeres es la primera causa de cáncer y la cuarta causa de mortalidad, teniendo una incidencia es 12.4 casos por cada 100,000 habitantes"**.**

Por esta razón, nos planteamos la interrogante ¿Qué métodos estadísticos de clasificación serán los más adecuados para clasificar los tumores de mama en malignos y benignos en los pacientes del Hospital de la Universidad de Wisconsin?

1.2 IMPORTANCIA DEL PROBLEMA

Es importante prevenir el cáncer de mama para evitar que la tasa de mortalidad siga creciendo, una forma de prevenir es que las mujeres visiten a su médico en forma anual para su correcta clasificación en caso que el tumor existiera usando métodos probabilísticos.

La detección temprana, a través de la mamografía, permitirá extender el período de vida de las mujeres que presenten esta enfermedad.

1.3 JUSTIFICACIÓN DEL PROBLEMA

Es de suma responsabilidad que los oncólogos y patólogos realicen una correcta clasificación de tumores de sus pacientes porque una detección temprana de cáncer permitirá extender el período de vida de las mujeres que presenten esta enfermedad así como mejorar la calidad de vida mientras haya sido detectada la presencia de cáncer, evitando que la tasa de mortalidad siga en aumento.

Así mismo, como estadístico se busca brindar información a los oncólogos y patólogos trabajando en forma conjunta para encontrar respuesta a los diferentes problemas de clasificación y dar una solución utilizando el método científico.

1.4 ÁMBITO GEOGRÁFICO

La aplicación se realizará con la base de datos “Wisconsin Diagnostic Breast Cancer” perteneciente al Hospital de la Universidad de Wisconsin, esta base de datos fue tomada del “Machine Learning Database Repository” [http://www.ics.uci.edu/~mlearn/MLRepository.html] proveniente de la Universidad de California en Irvine.

.

Dichos hospitales cuentan con más de 60 lugares en todo el estado de Wisconsin. Este hospital es reconocido como un líder nacional en el tratamiento del cáncer.

1.5 OBJETIVOS

1.5.1 OBJETIVO GENERAL

El objetivo principal de este estudio es clasificar correctamente la presencia de tumores de mama malignos y benignos en los pacientes del Hospital de la Universidad de Wisconsin a través de las características de las células, usando los métodos estadísticos de clasificación como son redes neuronales de reciente aplicación y el análisis de regresión logística de aplicación clásica.

1.5.2 OBJETIVOS ESPECÍFICOS

Comparar los dos métodos de clasificación y verificar cual nos da una mejor clasificación para el problema propuesto.

1.6 METODOLOGIA

1.6.1 TIPO DE INVESTIGACIÓN

Se utilizarán los datos de un estudio básico y no experimental, de tipo descriptivo transversal.

1.6.2 POBLACIÓN DE ESTUDIO

La población corresponde a todos los pacientes pertenecientes al servicio de cirugía y oncología del Hospital de la Universidad de Wisconsin realizada en Noviembre de 1995.

El tamaño de base de datos a utilizar es de 561 pacientes. Estos son casos atendidos por el Dr. William H. Wolberg en los hospitales de la Universidad de Wisconsin.

1.6.3 VARIABLES DE ESTUDIO

1) Diagnóstico (M = malignos, B = benigno)

A continuación se menciona 9 características citológicas de células de la muestra proporcionada por el especialista, las cuales serán explicadas detalladamente en el capítulo III.

2) Radio
3) Textura
4) Perímetro
5) Área
6) Grado de Suavidad
7) Compactación
8) Concavidad
9) Puntos de concavidad
10) Simetría

1.6.4 METODOLOGIA A USAR

Existen muchos métodos científicos para realizar una clasificación estadística correcta de los tumores de mama, los cuales son análisis discriminante, regresión logística, árboles de clasificación, redes neuronales y otros.

Para la aplicación se usará dos métodos de clasificación estadística, previa verificación de los supuestos que deben de cumplir los datos.

El primer método que se utilizará será el análisis de regresión logística, con este método se identificara, estimará los parámetros y se evaluará el modelo.

El segundo método que se utilizará será las redes neuronales, para ello se tendrá un vector de entrada, se desarrollará el método de aprendizaje supervisado y la estructura de la red será multicapa, el algoritmo que se utilizará para la clasificación de tumores será el backpropagation o retropropagación, mediante el cual se estimará los pesos de salida de la red neuronal artificial, terminando con la evaluación del modelo

Una vez obtenida la clasificación por estos dos métodos se realizará una comparación y se verificará cual nos da una mejor clasificación.

CAPITULO II

MARCO TEORICO

Existen muchos métodos de clasificación estadística que han sido utilizados en diversas investigaciones científicas. En este trabajo se utilizarán el análisis de regresión logística y redes neuronales.

Antes de explicar estos métodos de clasificación estadística se mencionan algunas aplicaciones recientes de estos métodos:

1. "Hernández, Torrealba y Reigosa [9] realizaron una investigación de un sistema automático de reconocimiento basado en contornos de los núcleos de las células, el cual utiliza una red neuronal para la clasificación entre carcinoma intraductal tipo comedo[1] y no comedo[2] de la mama, dentro de una imagen histológica digitalizada. La arquitectura de la red utilizada fue del tipo perceptrón multicapa, entrenada con el algoritmo de retropropagación del error. El sistema se validó mediante el análisis de 11 casos de carcinoma intraductal tratados en el Instituto Oncológico "Dr. Miguel Pérez Carreño", Valencia-Venezuela".

[1] **Comedo:** es una condición donde las células cancerosas crecen para bloquear los conductos de la mama con células muertas (necrosis). Debido a esto la necrosis, el cáncer de tipo comedón es más agresivo y tiene una mayor probabilidad de propagación.

[2] **No Comedo:** no presenta necrosis,no se esparce tan rápidamente y es menos maligno que del tipo comedo.

2. “Ferrero [6], presentó una aplicación experimental para la detección de posibles áreas con lesiones en la mama mediante el uso de redes neuronales en imágenes médicas. En dicha aplicación se observa la generación de diferentes topologías con el fin de mejorar redes en uso y obtener nuevas arquitecturas que permitan mejorar la detección”.

3. “Jan, Norup, Dounias, Bjerregaard [11], desarrollaron un estudio de casos en el cual se proporcionan datos y una línea de base para comparar métodos de clasificación. Los datos consisten de 917 imágenes de células de papanicolau, clasificadas cuidadosamente por técnicos especialistas en citología y por doctores. Cada célula es descrita por 20 características numéricas y las células son dividas en 7 clases. Los datos fueron recolectados en el Hospital de la Universidad de Herley - Dinamarca. Este estudio incluye un análisis básico a los datos incluyendo gráficos de dispersión y el resultado de la clasificación lineal”.

4. “Torres [19], seleccionó 80 pacientes del Servicio de Cirugía del Hospital Dr. Peset – Valencia, para comprobar la correlación entre la anatomía patológica del ganglio centinela y de la linfadenectomía axilar. Para ello realizó un análisis descriptivo de las variables, calculó índices de diagnósticos globales como sensibilidad, especificidad, cociente de probabilidades entre otros; además realizó un análisis multivariante utilizando la regresión logística simple para detectar qué variables a nivel individual pueden alterar los índices de diagnósticos globales”.

5. “Rojas [18], evalúo el efecto de la lactancia materna sobre el riesgo de desarrollar cáncer de mama en una población peruana. Realizó un estudio de tipo caso-control en el Hospital Nacional Arzobispo Loayza, en los servicios de Oncología Médica y Radiología (sección de Mamografía), entrevistando 103 pacientes con cáncer de mama como casos y a 208 pacientes con mamografías Birads 1 y 2 como casos control. En este trabajo utilizó estadísticas descriptivas, análisis bivariado y análisis de regresión logística, obteniendo como conclusión que la lactancia materna, en las

mujeres que tuvieron hijos, muestra protección sobre el cáncer de mama, sobre todo si el período de lactancia materna total fue de tres meses a más".

Después de haber citado éstas aplicaciones, procederé a explicar estos dos métodos de clasificación:

2.1 ANÁLISIS DE REGRESIÓN LOGÍSTICA

La bibliografía revisada para el análisis de regresión logística es: Catena [2] y Uriel [20].

El análisis de regresión logística se originó en la década de los 60 con el trabajo de Cornfield, Gordon y Smith; Walter y Duncan utilizan este análisis en la forma que se conoce actualmente para estimar la probabilidad de ocurrencia de un suceso en función de ciertas variables.

En 1972 Nelder y Wedderburn plantean que el análisis de regresión logística es un caso particular de los modelos lineales generalizados (GLM).

Desde principios de los 80 el uso del análisis de regresión logística ha ido incrementándose como consecuencia del aumento en la velocidad de procesamiento de las computadoras.

La Regresión Logística se usa para modelar la relación entre una variable respuesta binaria (Y) y una ó más variables regresoras (variables explicativas). Además las variables regresoras pueden ser cuantitativas (covariables) o cualitativas (factores) y la variable respuesta tiene dos posibles resultados etiquetados como 0 y 1 con $P(Y = 1|x) = \pi(x)$ y $P(Y = 0|x) = 1 - \pi(x)$.

El modelo de regresión logística permite clasificar a los individuos en uno de los dos grupos establecidos de acuerdo a los valores de la variable dependiente.

2.1.1 Modelo de Regresión Logística

El modelo de regresión logística puede ser simple cuando presenta una variable regresora X. Asimismo este modelo puede extenderse para k variables regresoras $X_1, X_2, \dots, X_k$ que se denomina modelo de regresión múltiple.

Los modelos de regresión logística simple se expresa en (1) y el modelo de regresión logística múltiple se expresa en (2).

$$E(Y/X = x) = \pi(X) = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}} \qquad (1)$$

$$E(Y/X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \pi(X) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} \qquad (2)$$

Se necesita realizar una transformación que permita expresar el modelo como una relación lineal entre el predictor[3] y la respuesta, la transformación que se realiza es tomar logaritmo natural a ambos miembros de la igualdad que se denomina logit (.), como se muestra en la ecuación:

$$logit\big(\pi(X)\big) = ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (3)$$

donde: $\left(\frac{\pi(X)}{1-\pi(X)}\right)$ es una razón de probabilidades (*odds)*.

[3] **Predictor:** Corresponde a la combinación lineal de las variables regresoras $X_1, X_2, \dots, X_k$. El predictor se define como: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

Los supuestos del modelo son:

1. Las variables regresoras son independientes entre sí pues, si este supuesto no se cumple estaría habiendo colinealidad en las variables explicativas.

2. La variable respuesta tiene distribución Bernoulli.

2.1.2 Estimación de los Parámetros del Modelo

En la regresión logística el método usado para la estimación de parámetros es el de máxima verosimilitud, que consiste en maximizar la función de verosimilitud de la muestra. El procedimiento consta de los siguientes pasos:

1. Para una muestra de tamaño n, se tiene el vector $y' = (y_1, \dots, y_n)$ formado por ceros y unos.

2. Como las observaciones son independientes; entonces la función de verosimilitud es obtenida como el producto de los n términos:

$$L(\beta, y, X) = \prod_{i=1}^{n} (\pi_i(x_i))^{y_i} (1 - \pi_i(x_i))^{1-y_i}$$

Al calcula el logaritmo de la función de verosimilitud se obtiene:

$$Ln[L(\beta, y, X)] = \sum_{i=1}^{n} y_i Ln\left(\frac{\pi_i(x_i)}{1 - \pi_i(x_i)}\right) + \sum_{i=1}^{n} Ln(1 - \pi_i(x_i))$$

Usando el enlace canónico, $Ln\left(\frac{\pi_i(x_i)}{1-\pi_i(x_i)}\right) = x_i^T \beta$ se tiene que la función de verosimilitud es:

$$L(\beta) = Ln[L(\beta, y, X)] = \sum_{i=1}^{n} y_i x_i^T \beta + \sum_{i=1}^{n} Ln\left(1 - exp(x_i^T \beta)\right) \qquad (4)$$

3. El vector $L(\beta)$ debe derivarse respecto al vector de parámetros β, de modo que se utilizan métodos numéricos para calcular los estimadores de este vector. El método más utilizado es el de Score Fisher que se basa en la función score obtenida de la función de verosimilitud.

La interpretación de los parámetros se realiza mediante el odds ratio, que es una razón de chances[4], en dos valores diferentes de la variable regresora.

$$Ln\big(Chance(x_i + 1)\big) - Ln\big(Chance(x_i)\big) = Ln\left(\frac{\big(Chance(x_i + 1)\big)}{\big(Chance(x_i)\big)}\right) = \sum_{i=1}^{n} \hat{\beta}_i$$

Tomando el antilogaritmo obtenemos el odds ratio (OR) estimado, como muestra la ecuación:

$$OR_{Estimado} = \left(\frac{\big(Chance(x_i + 1)\big)}{\big(Chance(x_i)\big)}\right) = e^{\sum_{i=1}^{n} \widehat{\beta_i}} \qquad (5)$$

El odds ratio estimado se puede interpretar como el incremento en el "chance" de tener el evento dado que la variable es $x_i + 1$ comparando con el "chance" de tener el evento dado que la variable es x_i. Además el odds ratio estimado es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes.

[4] **Razón de Chances:** es el cociente entre $P(Y = 1|x) = \pi(x)$ y $P(Y = 0|x) = 1 - \pi(x)$

2.1.3 Evaluación del Modelo Ajustado

Es uno de los aspectos más importantes del análisis de regresión, comprende:

1. **Bondad de Ajuste:** El ajuste del modelo es bueno si, las distancias entre los valores de la variable respuesta observada con respecto a la ajustada son pequeñas. Las siguientes medidas, permiten medir la bondad del ajuste del modelo:

   Estadística de Wald: Permite contrastar las hipótesis de que los parámetros individuales del modelo son iguales a cero. Las hipótesis a evaluar son:

   $$H_0: \beta_j = 0$$
   $$H_1: \beta \neq 0$$

   La estadística de prueba se muestra en:

   $$w = \left(\frac{\hat{\beta}_j}{Se(\hat{\beta}_j)}\right)^2 \sim \chi^2_{\alpha,1} \qquad (6)$$

   Donde $Se(\hat{\beta}_j)$ es el error estándar de $\hat{\beta}_j$ y α es el nivel de significación.

   Decisión: Si $w \geq \chi^2_{\alpha,1}$ rechazamos H_0 y concluimos que los parámetros individuales del modelo son diferentes de cero.

   Desvianza: mide la discrepancia entre el modelo saturado[5] y un modelo de p parámetros:

   $$D(\hat{\mu}, y) = 2\{Ln[L(modelo\ saturado)] - Ln[L(\hat{\beta})]\} \qquad (7)$$

[5] **Modelo Saturado:** Este modelo tiene n observaciones entonces tiene n parámetros lineales. Por lo tanto el modelo reproduce exactamente los datos.

Si el valor de la desvianza es igual a cero, indicaría que el modelo con p parámetros es igual al modelo saturado; por ello se busca que el valor de la desvianza sea un valor pequeño, que indicaría que es el mejor modelo ajustado.

Estadística Ganancia: Permite comparar la desvianza para el modelo que solo tiene la constante y la desvianza para el modelo con p parámetros. La expresión de la ganancia está dada en:

$$G(\hat{\mu}, y) = D(\hat{\mu}, y, \beta_0) - D(\hat{\mu}, y, \beta_p) \qquad (8)$$

Valores grandes de *G* indican que las variables adicionales incluidas en el modelo contribuyen en la explicación de la variable respuesta. Por tanto, se busca que el valor de la ganancia sea un valor grande, que indicaría que es el mejor modelo ajustado.

2. **Análisis de Residuos:** Ayuda a evaluar si el modelo es adecuado y a detectar las deficiencias y observaciones de casos atípicos en el modelo de regresión logística, mediante la obtención de los valores de los residuos y sus gráficos respectivos. Los residuos más utilizados para este análisis son: residuos estandarizados, residuos estudentizados, entre otros.

3. **Análisis de Influencia:** Ayuda a detectar observaciones influyentes sobre el ajuste del modelo de regresión logística y evaluar qué efectos produce sobre el modelo ajustado. Para el análisis de influencia se usan los leverage y la distancia de Cook.

4. **Puntos de Corte:** Un criterio para determinar un punto de corte para clasificar individuos es: Si la probabilidad estimada es menor o igual a 0.5 asignar Y=0 y si la probabilidad estimada es mayor que 0.5 asignar a Y=1. Este criterio permite evaluar el modelo mediante el Cuadro 2.1.

Cuadro 2.1: Tabla de Clasificación

Observado	Pronóstico		
	Y=1	Y=0	Porcentaje Correcto
Y=1	a	b	a/(a+b) = Especificidad
Y=0	c	d	d/(c+d) = Sensibilidad
Porcentaje Global			(a+d)/(a+b+c+d) Capacidad Predictiva

Especificidad: Es la probabilidad de clasificar correctamente a un individuo sano; es decir, la probabilidad de que para un individuo sano se obtenga un resultado médico negativo. En otras palabras, se puede definir la especificidad como la capacidad para detectar a los individuos sanos.

Sensibilidad: Es la probabilidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un individuo enfermo se obtenga en la prueba un resultado médico positivo. La sensibilidad es, por lo tanto, la capacidad del test para detectar la enfermedad.

2.2 REDES NEURONALES ARTIFICIALES

Para las redes neuronales artificiales se consultó la bibliografía: Bishop [1], De la Torre [4], Duda [6], Hastie [8] y Hernández [10].

El desarrollo de redes neuronales artificiales se inició aproximadamente hace 50 años, motivada por el deseo de comprender el funcionamiento del cerebro y de emular algunas características como la capacidad de memorizar y asociar hechos.

Las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano, que es el ejemplo más perfecto siendo capaz de adquirir conocimiento a través de la experiencia.

Las Redes Neuronales Artificiales (RNA) fueron originalmente una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "neuronas" o "nodos" conectadas unas con otras. Estas conexiones tienen una gran semejanza con las dendritas y los axones en los sistemas nerviosos biológicos.

Recientemente las redes neuronales han cobrado interés y se puede atribuir a varios factores: como técnicas de entrenamiento, estas están siendo desarrolladas para arquitecturas de redes más sofisticadas que son capaces de superar las deficiencias de las primeras redes neuronales y con el aumento en la velocidad de procesamiento de las computadoras hacen posible la simulación de los procesos neuronales más factible reduciendo el tiempo.

Las RNA surgen como una alternativa de estudio en ‘’problemas de clasificación’’, ante las limitaciones de los modelos clásicos, rígidos y con características de tipo lineal.

Como modelos estadísticos las redes neuronales destacan ante problemas no lineales. Por ello las redes neuronales han sido aplicadas a problemas de estadística como predicción y clasificación.

Las aplicaciones más exitosas de las RNA son:

1. Procesamiento de imágenes y de voz
2. Reconocimiento de patrones
3. Planeamiento
4. Interfaces adaptivas para sistemas Hombre/máquina
5. Predicción
6. Control y optimización
7. Filtrado de señales

2.2.1 Historia de las Redes Neuronales

Los primeros trabajos que se desarrollaron sobre las redes neuronales artificiales estuvieron relacionadas con el campo de la inteligencia artificial, es así que en 1963 Alan Turing fue el primero en estudiar el cerebro como una forma de ver el mundo de la computación, pero quienes primero concibieron algunos fundamentos de la computación neuronal fueron Warren McCulloch y Walter Pitts (1943).

Donald Hebb en 1949, fue el primero en explicar los procesos del aprendizaje desde un punto de vista psicológico, desarrollando una regla de como el aprendizaje ocurría intentando encontrar semejanzas entre el aprendizaje y la actividad nerviosa.

En 1957 Frank Rosenblatt comenzó el desarrollo del modelo perceptrón que consistía en reconocer patrones similares a los que ya conocía. Dos años después Bernard Widrow y Marcial Hoff desarrollaron el modelo ADALINE (ADAptative LINear Elements), éste modelo fue aplicado a un problema real denominado “filtros adaptativos para eliminar ecos en las líneas telefónicas”.

Paul Werbos en 1974 desarrolló la idea básica del algoritmo de aprendizaje backpropagation (propagación hacia atrás).

En Japón Kunihiko Fukushima en el año 1980 desarrolló un modelo para el reconocimiento de patrones visuales.

Pero no fue hasta 1985, con el trabajo de Hinton, Rumelhart y Williams quienes presentaron el algoritmo de ‘’Backpropagation’’ para entrenar perceptrones multicapa. Desde entonces comenzaron a consolidarse los congresos más importantes como Neuronal Networks for Computing, la Neural Information Processing Systems, entre otros.

Desde un **enfoque estadístico** destacan los siguientes trabajos:

- Ripley [16], realizó la investigación "Statistical aspects of neural networks".

- "Chen & Titterington [3], realizaron el trabajo titulado "Neural Networks: A Review from a Statistical Perspective". Este trabajo da una visión general sobre las redes neuronales y el objetivo de los autores es alentar la investigación multidisciplinaria entre los investigadores de redes neuronales y estadísticos; ya que los estadísticos deben mejorar el conocimiento de ésta interesante e importante área.

  Este trabajo responde a dos preguntas: ¿Qué pueden aprender los estadísticos acerca de los investigadores de redes neuronales? y ¿Qué pueden aprender los investigadores de redes neuronales acerca de los estadísticos?

  Algunas respuestas a la primera pregunta fueron: Los estadísticos deben abordar problemas difíciles de datos reales, como algunos de los abordados por los investigadores de redes neuronales por ejemplo el reconocimiento de voz. Además, los modelos con un gran número de parámetros pueden ser útiles para la predicción sobre todo para grandes conjuntos de datos y para problemas que presenten un alto porcentaje de error.

  Algunas respuestas a la segunda pregunta fueron: Los investigadores de redes neuronales deberían preocuparse más sobre obtener estadísticas óptimas para comparar sus métodos con los métodos estadísticos tradicionales.

  En conclusión, en los últimos años las redes neuronales han mejorado sus resultados mediante la aplicación de técnicas estadísticas clásicas,

por lo que los autores sugieren que ambos campos deben escuchar y aprender unos de otros".

- Kuan & White [12], realizaron el trabajo titulado "Artificial Neural Networks: an Econometric Perspective".

- "Warner & Misra [22], realizaron la investigación titulada "Understanding Neural Networks as Statistical Tools", en esta investigación encontraron muchas aplicaciones en los diferentes campos de estudio, centrándose en la utilización de redes neuronales para la predicción y la clasificación de problemas.

  Los investigadores usaron paralelamente la terminología de las redes neuronales y la terminología estadística aplicada a los modelos de regresión. La notación de las redes neuronales es diferente al análisis de regresión estadística; pero la mayoría de las ideas son muy parecidas. Por ejemplo, en lugar de los coeficientes y observaciones en el análisis de regresión, la red neuronal utiliza el término pesos y patrones respectivamente.

  A la conclusión que llegaron los investigadores es que las redes neuronales se pueden ver como un método no paramétrico de regresión".

- Zapranis & Refenes [23], realizaron el trabajo titulado "Principles of neural model identification, selection and adequacy with applications to financial econometrics".

Actualmente, son numerosos los trabajos que se realizan y publican. Revistas como Neural Networks, Transactions on Neural Networks, entre otros, son las encargadas de la publicación de los últimos avances.

El Departamento de Defensa de los Estados Unidos, la Sociedad Europea de Redes Neuronales son algunos de los ejemplos de investigación sobre redes neuronales.

2.2.2 Red Neuronal Biológica

Se estima que el cerebro humano contiene más de cien mil millones de neuronas y 100 billones de sinapsis en el sistema nervioso.

La Neurona es la célula fundamental y básica del sistema nervioso, las neuronas son “unidades celulares” (entidades completas y autónomas) capaces de conectarse con otras neuronas, ya sea para inhibirlas, excitarlas o simplemente para re-transmitirles el impulso nervioso.

Las partes fundamentales de una neurona (Ver Figura 2.1) son:

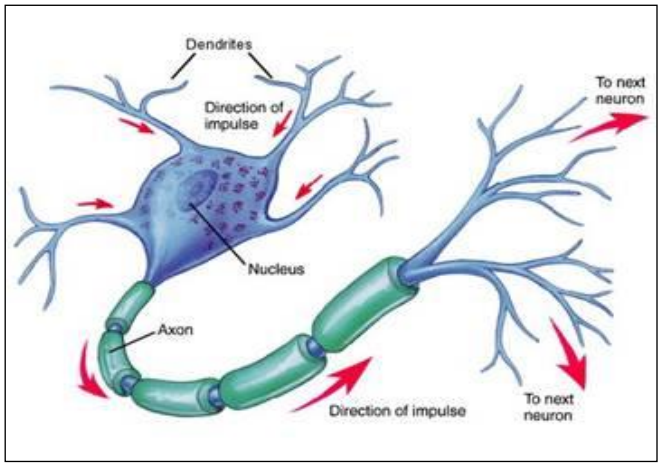
- **El Citón, Soma o Cuerpo Celular:** Se refiere al cuerpo de la célula.

- **El Núcleo:** Situado en el cuerpo celular y suele ocupar una posición central.

- **Las Dendritas:** Son prolongaciones cortas que se originan en el soma o cuerpo celular, cuya función es recibir los impulsos de otras neuronas y enviarlas al soma de la neurona.

- **Axón:** Es una prolongación única y larga que puede medir hasta un metro de longitud y cuya función es sacar el impulso desde el soma neuronal y conducirlo hasta otro lugar del sistema o hasta un órgano receptor, por ejemplo un músculo.

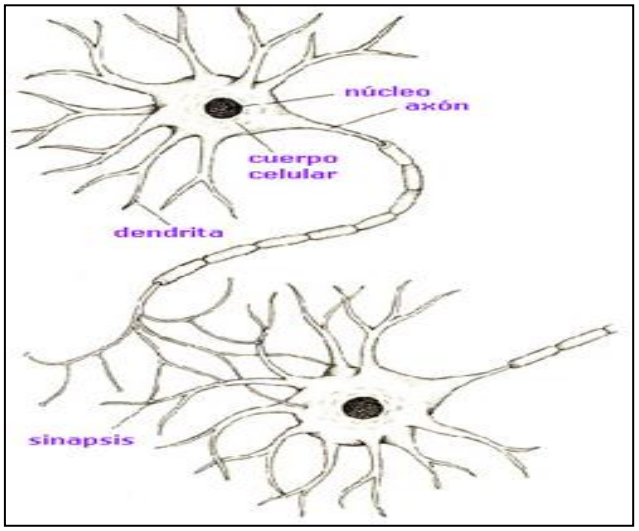
Figura 2.1: Partes de una Neurona



El objetivo principal de las redes neuronales de tipo biológico es desarrollar operaciones de síntesis y procesamiento de información, relacionadas con los sistemas biológicos. Para ello, la neurona recibe información a través de la sinapsis de sus dendritas, cada sinapsis representa la unión de un axón con las dendritas de otra neurona.

La información es transmitida a lo largo de las dendritas hasta que alcanza el cuerpo de la célula, aquí se junta toda la información a través de impulsos y aplican algún tipo de función de activación que enviará una señal a lo largo del axón con la finalidad de comunicarse con otras neuronas (Ver Figura 2.2).

Figura 2.2: Sinapsis de una Neurona



2.2.3 Red Neuronal Artificial (RNA)

Pitarque, Roy y Ruiz [14], han conceptualizado a las redes neuronales como modelos estadísticos no paramétricas al estar libres del cumplimento de los supuestos teóricos de la estadística paramétrica.

El enfoque estadístico ayuda al entendimiento de las capacidades y las limitaciones que presentan las RNA.

Existen términos en redes neuronales que se relacionan con términos estadísticos como se observa en el Cuadro 2.2:

Cuadro 2.2: Relación de Términos Neuronales con Términos Estadísticos

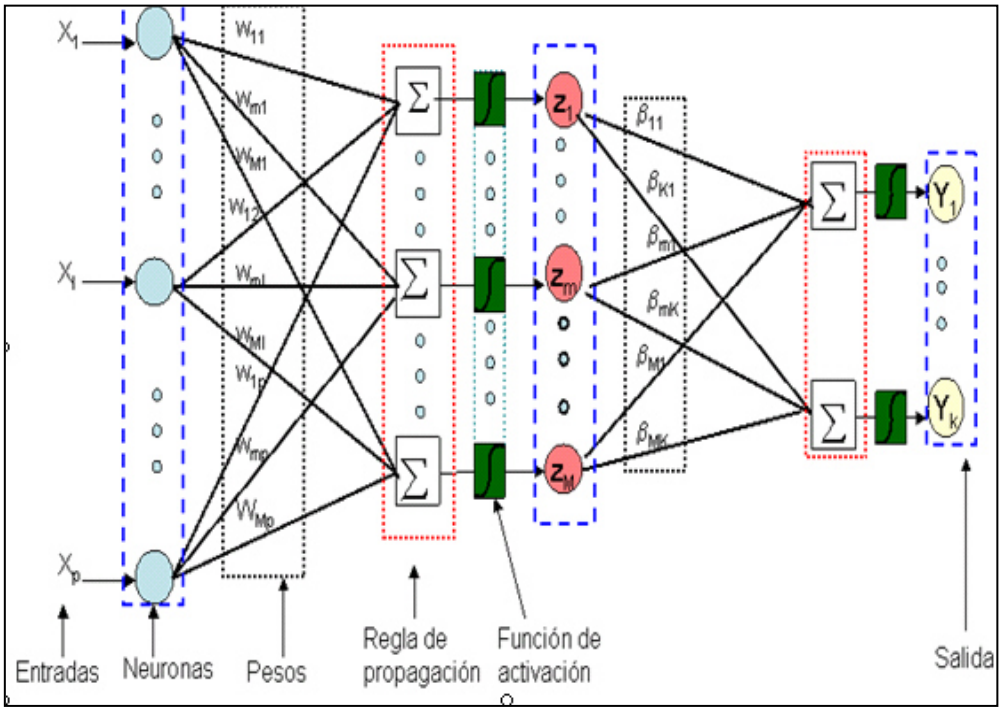
REDES NEURONALES	ESTADISTICA
Entrada	Variable Independiente
Salida	Variable Dependiente
Pesos	Parámetros
Aprendizaje	Estimación
Entrenamiento	Ajuste
Función de Pérdida, Función de Error	Error

En una red neuronal artificial al conocimiento adquirido por la red se le denomina APRENDIZAJE y en términos estadísticos corresponde a la estimación de los pesos de las conexiones de la red neuronal.

2.2.3.1 ELEMENTOS DE UNA RED NEURONAL ARTIFICIAL

Los elementos de una red neuronal artificial (Ver Figura 2.3) son:

Figura 2.3: Elementos de una Red Neuronal Artificial



1. **Neurona:** son los elementos de un proceso. Cuando estos elementos se agrupan forman capas. Se tiene tres tipos que son: capas de entrada, capas ocultas y capas de salida.

2. **Entrada:** la información entrante se realiza a través del vector aleatorio $X^T \in \Re^P$.

3. **Pesos:** A cada neurona se le asigna un factor de importancia. El peso se va modificando durante el proceso de aprendizaje de la RNA y luego almacenará la información para determinados propósitos.

4. **Regla de propagación:** una vez que se dio la entrada y se asignó los pesos se realizan diferentes operaciones para reunir la información, la más usada es la suma. **A la combinación de entradas se le denomina Net.**

5. **Función de Activación o Función de Transferencia:** una vez obtenido el valor usando la regla de propagación se filtra a través de una función de activación, la cual proporcionará la salida de la red. Las principales funciones son: función escalón, función identidad y función exponencial.

6. **Salida:** Es el resultado de la función de activación, ver la siguiente ecuación:

$$f(X) = g(Net) \tag{9}$$

2.2.3.2 MODELO DE UNA RED NEURONAL ARTIFICIAL

Se tiene el vector aleatorio $X \in \Re^p$ y la variable dependiente $Y \in \Re$, una red neuronal es un modelo estadístico no lineal, definido por:

$$Y = f(X) + \varepsilon = g^s[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)] + \varepsilon \tag{10}$$

donde:

g^s: son funciones de transferencia de la capa de salida

σ : son funciones de activación de la capa oculta

w : son vectores de pesos de la capa oculta

β : son vectores de pesos de la capa salida

s : indica el número de capas ocultas

X : vector de entrada

Y : salida de la red

ε : es el error.

Para el desarrollo de este trabajo se utilizará las siguientes notaciones:

- X representará al vector aleatorio y una observación de X puede escribirse como x_i.

- La salida puede denotarse por Y si es una variable cuantitativa y G si es una variable categórica.

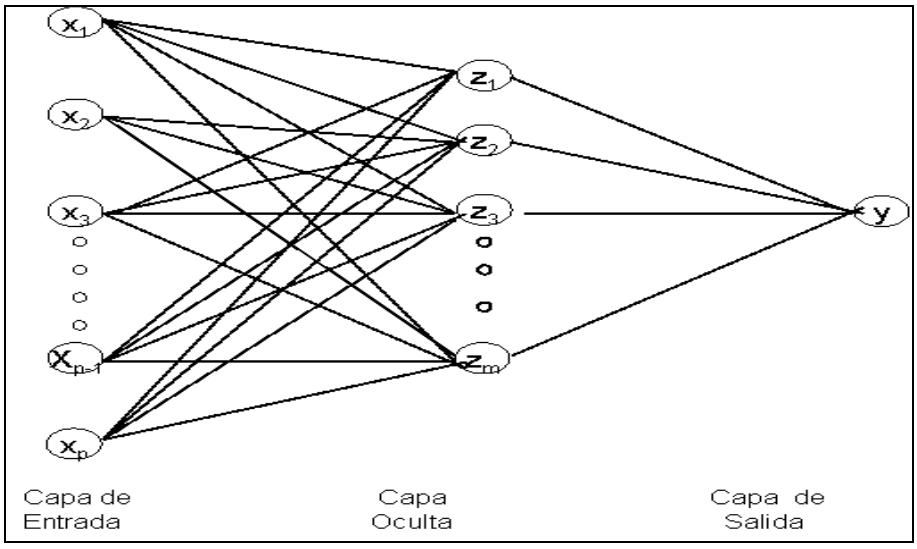
- La net para la capa oculta es $Net_m = w_0 + w_m^T X$, para $w^T = (w_{1m}, w_{2m}, \dots, w_{pm})$

- La net para la capa de salida es $Net_k = \beta_0 + \beta_k^T Z$, para $\beta^T = (\beta_{1k}, \beta_{2k}, \dots, \beta_{Mk})$

En muchos problemas, la red de una sola capa oculta es suficiente para describir el comportamiento de los datos. El modelo estadístico no lineal para este caso, está definido por:

$$Y = f(X) + \varepsilon = g\left[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)\right] + \varepsilon \text{ , cuando } s = 1$$

Figura 2.4: Representación de la Red Neuronal con una Capa Oculta



En la figura 2.4, cada salida Z_m de cada unidad de la capa oculta se puede escribir como:

$$Z_m = \sigma(w_{0m} + w_m^T X) = \sigma(Net_m) \qquad m = 1, \dots, M$$

Y la salida de la red está dada por la ecuación:

$$Y = f(X) = g\left[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)\right] = g\left(\beta_{0k} + \beta_k^T Z_m\right) = g(Net_k) \;\; k = 1 \qquad (11)$$

2.2.3.3 ARQUITECTURA DE LA RED NEURONAL ARTIFICIAL

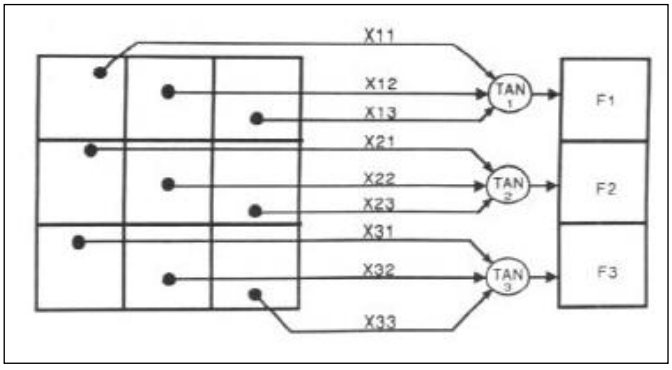
La arquitectura de una red es una estructura o patrón de conexión. En ésta arquitectura las neuronas se organizan en una secuencia de capas formando un patrón de interconexiones entre los diferentes elementos dependiendo de la posición de las capas.

Existen tres tipos de capas: de entrada, salida y ocultas. La capa de entrada está compuesta por neuronas que reciben información proveniente del entorno. La capa de salida es la que proporciona la respuesta de la red neuronal. La capa oculta es la que no tiene relación alguna con el exterior. Este tipo de capa agrega a la red, grados de libertad adicionales, con los cuales es posible mejorar la capacidad del modelo.

La manera como se interconectan las neuronas puede ser:

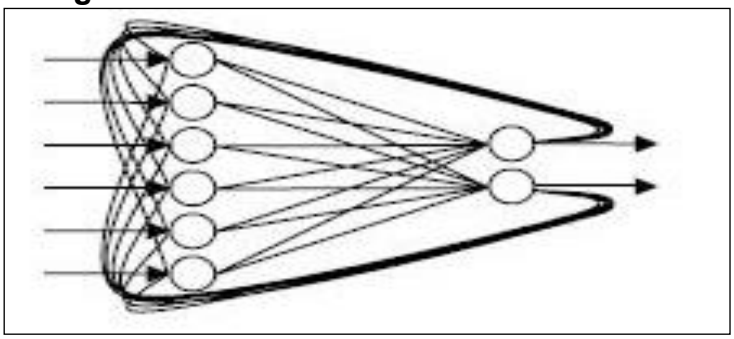
- Conexiones hacia delante (*feedforward* o *redes unidireccionales*), la información fluye en un único sentido, de las neuronas de entrada a las de salida. Este tipo de conexiones se usan en aplicaciones de clasificación, regresión y aproximación de funciones.

Figura 2.5: Redes Neuronales Feedforward



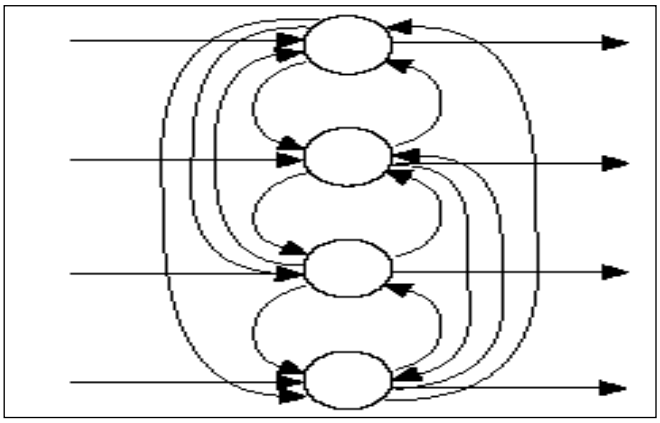
- Conexiones hacia atrás (*feedback o redes recurrentes*).

Figura 2.6: Redes Neuronales Feedback



- Conexiones laterales dentro de la misma capa. Este tipo de conexiones se usan para formar conglomerados y para la extracción de componentes principales.

Figura 2.7: Redes Neuronales con Conexiones Laterales



- Conexiones con retardo temporal. Estas conexiones presentan aplicaciones en predicción y clasificación.

2.2.3.4 APRENDIZAJE DE UNA RED NEURONAL ARTIFICIAL

Las redes neuronales artificiales tienen la capacidad de ‘‘aprender de su ambiente’’ a través de un proceso iterativo de ajustes de sus pesos. Este proceso recibe el nombre de *regla de aprendizaje* o simplemente aprendizaje.

Para realizar el aprendizaje de una RNA, primero se alimenta a la red con los datos de entrenamiento. Cada observación de entrenamiento se propaga a través de la red para producir una respuesta en la capa de salida, esta respuesta se compara con la salida real.

Luego se obtiene una diferencia que se usa para calcular una función de pérdida del error de la respuesta de la red con un algoritmo apropiado permitiendo modificar los valores de los pesos con el fin de reducir el error. Estas correcciones deben realizarse varias veces, para todo el conjunto de observaciones de entrenamiento.

Existen dos tipos de aprendizaje en las RNA:

1. **Aprendizaje no Supervisado**: Es cuando no existe un agente externo que indique si la respuesta es la adecuada para una observación de entrenamiento. En este caso la red debe de auto enseñarse.

2. **Aprendizaje Supervisado:** Este tipo de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor comprueba la salida de la red y en el caso de que ésta no coincida con la salida correcta se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la correcta.

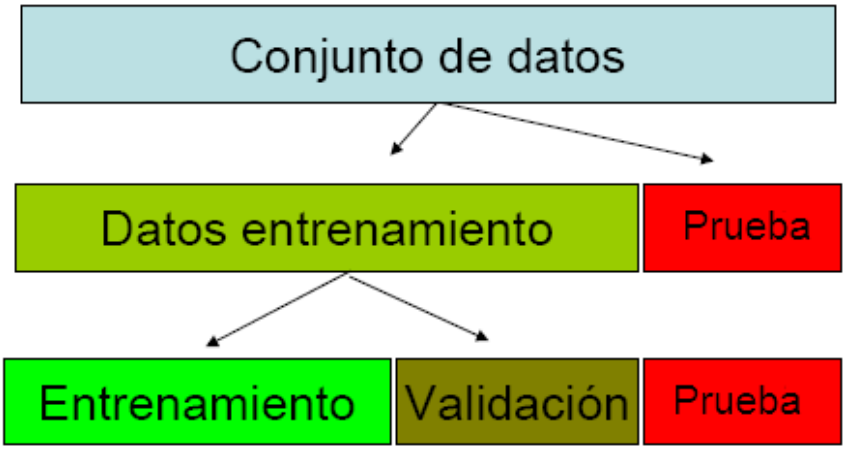
En el aprendizaje supervisado la base de datos se divide en conjunto de entrenamiento y conjunto de prueba. (Ver Figura 2.8)

El conjunto de entrenamiento a su vez suele dividirse en dos subconjuntos que son: entrenamiento y validación. En el subconjunto de

entrenamiento se suele utilizar el 50% de los datos que servirán para ajustar el modelo y en el subconjunto de validación se suele utilizar el 25% de los datos que servirán para estimar el error del modelo ajustado.

El conjunto de prueba utiliza el 25%de los datos restantes para estimar la tasa de clasificación incorrecta.

Figura 2.8: Conjunto de Datos de una Red Neuronal



2.2.3.5 PERCEPTRON SIMPLE

El perceptrón fue el primer modelo de red neuronal desarrollado por Rosenblatt en 1962. Este modelo en los años 60 despertó gran interés, debido a su capacidad para aprender a reconocer comportamientos sencillos.

Al mismo tiempo que Rosenblatt fue desarrollando el perceptrón, Widrow y sus compañeros de trabajo realizaban una investigación similar conocida como ADALINE (ADAptative LINear Element) que se refiere a una sola unidad de procesamiento al igual que el perceptrón.

Se ha visto que una red con una sola capa de pesos tiene capacidades limitadas; por ello para mejorar el desempeño del perceptrón, Rosenblatt usó una capa de elementos fijos de procesamiento para transformar los datos de entrada. Este procesamiento de elementos se puede

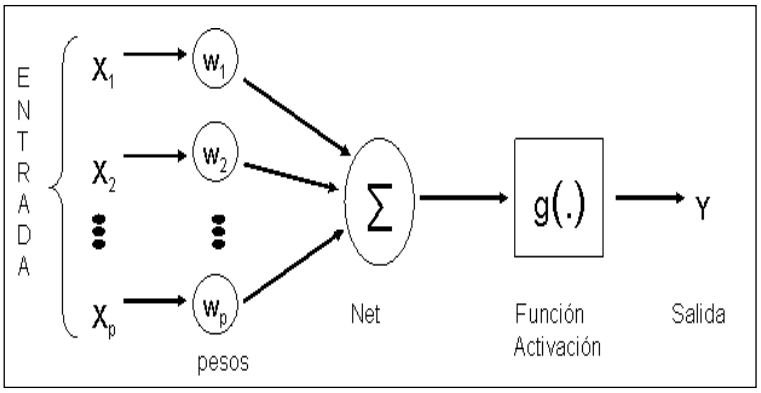
considerar como las funciones base de una discriminación lineal generalizada.

2.2.3.5.1 Definición

Dado $X \in \Re^P$ e $Y \in \Re$, el perceptrón simple es la red neuronal que no presenta neuronas en la capa oculta (Ver Figura 2.9) y está definida por:

$$Y = f(X) + \varepsilon = g(w_0 + w^T X) + \varepsilon \qquad Net = w_0 + w^T X \quad (12)$$

Figura 2.9: Modelo Perceptrón



El perceptrón es el modelo más simple de red neuronal que consiste de una combinación lineal, formada por varias neuronas para recibir entradas a la red y una neurona de salida.

El perceptrón actúa como un clasificador lineal dado por:

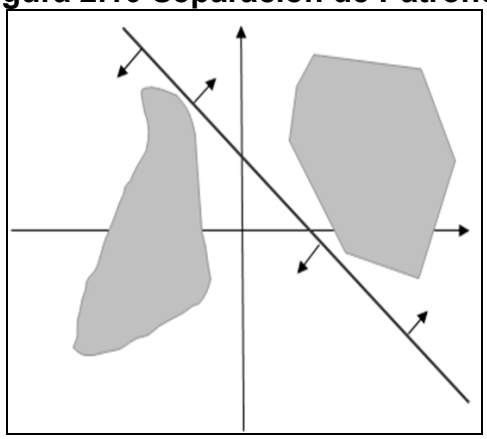
$$g(Net) = \begin{cases} 0 & cuando \quad Net < 0 \\ 1 & cuando \quad Net \geq 0 \end{cases}$$

La finalidad del perceptrón es clasificar uno de los dos grupos G_1 o G_2, la regla de decisión que se utiliza para la clasificación es asignar X al grupo G_1 si la salida del perceptrón es 1 y al grupo G_2 si es 0.

2.2.3.5.2 Aprendizaje del Perceptrón

El perceptrón realiza su aprendizaje tratando de encontrar una separación del espacio de patrones de entrada mediante un hiperplano en dos categorías, dependiendo dicha separación de los valores de los pesos, tal como se observa en la Figura 2.10

Figura 2.10 Separación de Patrones



El algoritmo de aprendizaje del perceptrón simple busca minimizar la función de pérdida del error, que está definida por la ecuación:

$$R = \frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2 \qquad (13)$$

donde:

R: función de pérdida del error para el conjunto de entrenamiento

y_i: variable salida

$f(x_i) = g(Net) = w^T x$: es el resultado de la función de transferencia.

El método gradiente descendiente nos permite encontrar los valores de todos los pesos que minimizan la función de pérdida del error.

La idea es hacer una modificación de los pesos que sea proporcional a la derivada de la función de pérdida del error con respecto a cada uno de los pesos:

$$\Delta w_i = -\gamma \frac{\partial R}{\partial w_l} = -\gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial w_l} \qquad l = 1, \dots, p$$

donde:

$\Delta w_l = w_l^{(r+1)} - w_l^{(r)}$: modificación de los pesos

γ: Factor de aprendizaje, este valor varía entre 0 y 1.

Luego, se aplica la regla de la cadena para el cálculo de $\frac{\partial R_i}{\partial w_l}$ y se obtiene la siguiente expresión:

$$\frac{\partial R_i}{\partial w_l} = \frac{\partial R_i}{\partial Net} \frac{\partial Net}{\partial w_l} \qquad (14)$$

A continuación se calcula la expresión (14):

$$\frac{\partial R_i}{\partial Net} = \frac{\partial}{\partial Net}\left[\frac{1}{2}\sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2\right]$$

$$\frac{\partial R_i}{\partial Net} = \sum_{i=1}^{N}\left[\frac{\partial}{\partial Net}\left(\frac{1}{2}\left(y_i - f(x_i)\right)^2\right)\right]$$

$$\frac{\partial R_i}{\partial Net} = -\sum_{i=1}^{N}\left(y_i - f(x_i)\right) = -\sum_{i=1}^{N}\delta_i$$

$$\frac{\partial Net}{\partial w_l} = \frac{\partial}{\partial w_l}\left[\sum_{l=1}^{p} w_l x_{il}\right] = x_{il}$$

Reemplazando estos resultados en (14) se tiene:

$$\frac{\partial R_i}{\partial w_l} = -\sum_{i=1}^{N} \delta_i x_{il}$$

Finalmente se tiene que la actualización de los pesos en la r+1-ésima iteración para el conjunto de entrenamiento son mostrados en (15):

$$\Delta w_l = w_l^{(r+1)} - w_l^{(r)}$$
$$w_l^{(r+1)} = w_l^{(r)} + \Delta w_l$$
$$w_l^{(r+1)} = w_l^{(r)} + \gamma \sum_{i=1}^{N} \delta_i x_{il} \qquad (15)$$

2.2.3.6 RED NEURONAL MULTICAPA O PERCEPTRÓN MULTICAPA

Minsky y Papert en 1969 mostraron que se podía aumentar la funcionalidad del perceptrón simple usando más capas, formando de esta manera el perceptrón multicapa.

En 1974, Paul Werbos desarrolla el primer algoritmo de entrenamiento para redes multicapa y en 1985 -1986 Hinton propuso una regla de aprendizaje que permite que el perceptrón multicapa aprenda para cualquier tipo de red con diferentes estructuras de conexión, o arquitecturas.

La regla de propagación hacia atrás o Backpropgation fue propuesta por Rumelhart y en forma independiente por Parker y Le Cun, ésta regla se utiliza para una red de tipo Feedback (Hacia atrás) con una o más capas ocultas de forma que las neuronas de una misma capa no estén conectadas entre sí.

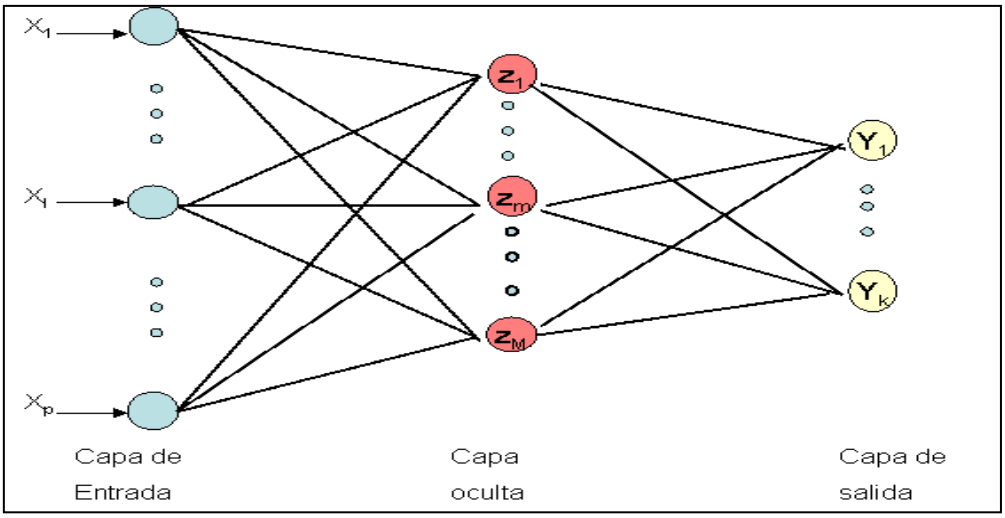
2.2.3.6.1 Definición

El perceptrón multicapa está formado por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón simple.

2.2.3.6.2 Estructura de la Red Multicapa

La estructura típica de una red multicapa se observa en la figura 2.11:

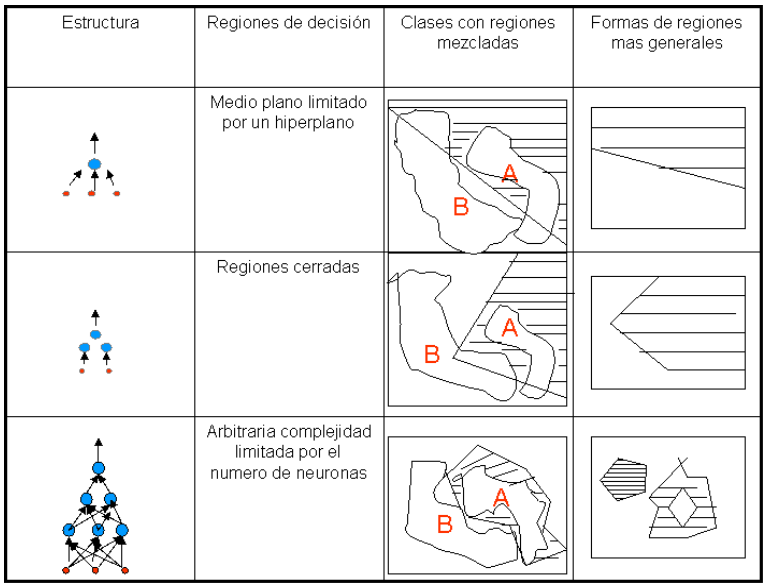
Figura 2.11: ***Representación de la Red Multicapa***



La primera capa es llamada capa de entrada que toma datos del conjunto de entrenamiento y la última capa es denominada de salida, cuyos pesos adquieren, mediante un proceso de aprendizaje, distintos valores para observaciones distintas.

De esta manera situando los perceptrones en cascada mediante una estructura de múltiples capas se puede implementar fronteras de decisión mucho más complejas como se ilustra en la Figura 2.12

Figura 2.12: ***Fronteras de Decisión***



2.2.3.6.3 Reglas de Aprendizaje de Redes Multicapa

El entrenamiento de una red neuronal multicapa se realiza mediante un proceso de aprendizaje.

Para realizar el proceso de aprendizaje se debe tener inicialmente definida la topología de la red, es decir, se debe conocer el número de neuronas en la capa de entrada la cual dependerá del número de componentes del vector de entrada, también la cantidad de capas ocultas, como el número de neuronas en la capa de la salida la cual dependerá del número de componentes del vector de salida y finalmente las funciones de transferencia requeridas en cada capa. Una vez escogida la topología de la red se asignan valores iniciales a cada uno de los pesos (parámetros) que conforma la red.

Es importante indicar que no existe una técnica para determinar el número de capas ocultas, ni el número de neuronas que debe contener cada una de ellas para un problema específico, esta elección es determinado por la experiencia del investigador.

Por ejemplo desarrollaremos un procedimiento para obtener una red con una capa de entrada, una oculta y una de salida. La salida de la red está dada por:

$$f_k(X) = g^s[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)] = g(Net_k) \qquad (16)$$

donde:

Net_k: net para la capa de salida.

Para iniciar el entrenamiento de la red se ingresa una observación x_i que pertenece al conjunto de entrenamiento $E = \{x_i, y_i / \ i = 1,2,\dots,N\}$ donde $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ al ingresar la observación se propagara a través de las conexiones existentes en la red, produciendo una net en cada una de las M neuronas de la capa oculta, la cual está dada por:

$$Net_{mi} = \sum_{l=1}^{p} w_{0m} + w_{ml} x_{il}$$

$$Net_{mi} = w_{0m} + w_m^T x_i \qquad (17)$$

$$m = 1,2,\dots,M \qquad i = 1,2,\dots,N \quad w_m^T = (w_{m1}, w_{m2}, \dots, w_{mp})$$

donde:

Net_{mi}: net para la capa oculta

w_{ml}: pesos de la capa oculta

w_{0m}: ganancia (umbral) de la neurona *m* de la capa oculta.

Utilizando la regla de propagación obtenemos el valor de la net para la capa oculta, luego utilizando la función de activación para la capa oculta obtenemos una salida para cada una de las neuronas de la capa oculta, que se denotará por:

$$Z_{mi} = \sigma(Net_{mi})$$
$$Z_{mi} = \sigma\left(\sum_{l=1}^{p} w_{0m} + w_{ml}x_{il}\right)$$
$$Z_{mi} = \sigma(w_{0m} + w_m^T x) \quad (18)$$

Las salidas de la capa oculta Z_{mi} (M componentes), ahora pasaran a ser las entradas de las conexiones de la capa de salida utilizando nuevamente la regla de propagación obtenemos la net para la capa de salida que está denotada por:

$$Net_{ki} = \sum_{m=1}^{M} \beta_{0k} + \beta_{km} Z_{mi}$$
$$Net_{ki} = \beta_{0k} + \beta_k^T Z_i$$
$$\beta_k^T = (\beta_{1m}, \beta_{2m}, \dots, \beta_{km}) \quad (19)$$

donde:

β_{km}: pesos de la capa de salida

β_{0k}: Ganancia de la neurona k de la capa de salida

Z_{mi}: Salida de la neurona

Net_{ki}: Net de la capa de salida.

Para producir la salida final para cada una de las observaciones debemos aplicar la función de transferencia a la net correspondiente a la capa de salida, la cual se denota por:

$$f_k(x_i) = g(Net_{ki}) \qquad k = 1,2,\dots,K \quad (20)$$

.

Se sabe que $Net_{ki} = \beta_{0k} + \beta_k^T Z_i$ y $Z_{mi} = \sigma(w_{0m} + w_m^T x)$. Finalmente, las expresiones anteriores reemplazamos en (20) para obtener la siguiente ecuación:

$$f_k(x_i) = g(Net_{ki})$$
$$f_k(x_i) = g(\beta_{0k} + \beta_k^T Z_i)$$
$$f_k(x_i) = g\left(\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T x)\right) \qquad (21)$$

La salida de la red $f_k(x_i)$ se compara con el valor real y_{ik} para calcular una medida de error llamado función de pérdida de error $L(y_{ik}, f_k(x_i))$. Cuando la salida y_{ik} es una variable aleatoria continua, la función de pérdida comúnmente está dada por el error cuadrático, tal como se muestra en la siguiente expresión:

$$R = L(y_{ik}, f_k(x_i)) = \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2 \qquad (22)$$

Cuando la salida G es una variable categórica, se requiere de otra función de pérdida del error denominada cross-entropía definida por:

$$R = L(G_{ik}, f_k(x_i)) = -\sum_{k=1}^{K} G_{ik} log f_k(x_i) \qquad (23)$$

donde $f_k(x_i) = p_x(x_i)$ son probabilidades que representan la salida de la red.

Los valores de los pesos que minimizan la función de pérdida del error son obtenidos por el método de gradiente descendiente, para ello debemos actualizar los pesos para la capa oculta y para la capa de salida en la r+1 - ésima iteración, tal como se muestra en las ecuaciones:

$$w_{ml}^{(r+1)} = w_{ml}^{(r)} - \gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial w_{ml}}$$

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}} \qquad (24)$$

donde:

γ: factor de aprendizaje, que varía entre 0 y 1.

El desarrollo del método gradiente descendiente se verá de forma más amplia cuando se desarrolle el tema de backpropagation.

También se sede tener en cuenta las siguientes observaciones:

- Se debe elegir un incremento adecuado que influya en la velocidad de convergencia del algoritmo, esta velocidad se controlará a través del factor de aprendizaje γ. Por lo general se escoge un número pequeño, para asegurar que la red encuentre una solución adecuada teniendo que hacer un gran número de iteraciones.

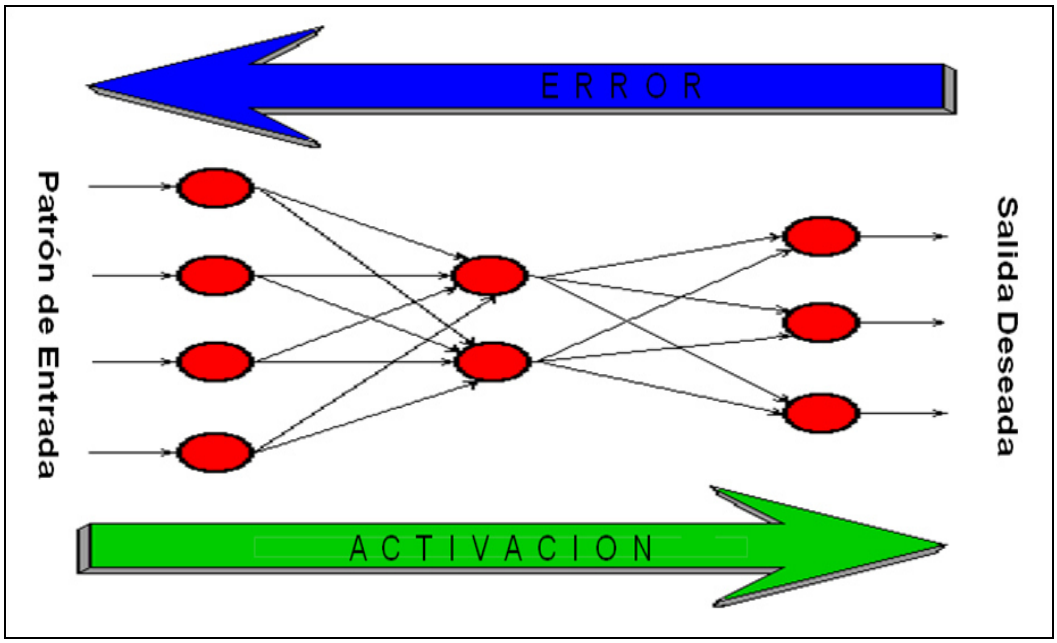
- Es recomendable aumentar el valor de γ a medida que disminuye el error de la red durante la fase de entrenamiento, para garantizar así una rápida convergencia, teniendo la precaución de no tomar valores demasiado grandes que hagan que la red oscile alejándose demasiado del valor mínimo sin poder alcanzarlo.

2.2.3.7 BACKPROPAGATION

El backpropagation es un algoritmo de aprendizaje supervisado, que emplea un ciclo de propagación, propagándose desde la primera capa a través de las capas superiores de la red, hasta generar una salida. Esta salida se compara con la salida real y se calcula la señal de error para cada una de las observaciones. Los errores se propagan hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida.

Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido una señal de error que describa su contribución relativa al error total. Basándose en la señal de error percibida, se actualizan los pesos de conexión de cada neurona, para hacer que la red converja hacia un estado que permita clasificar correctamente todas las observaciones de entrenamiento.

Figura 2.13: Propagación de las señales en un Perceptrón Multicapa



La importancia de este proceso consiste en que a medida que se entrena la red las neuronas de las capas intermedias se organizan a sí mismas de tal modo que las distintas neuronas aprenden a reconocer distintas características del espacio total de entrada.

A continuación se explicará el algoritmo de backpropagation orientado a la clasificación de problemas y cuya finalidad será modelar las probabilidades de salida.

La salida de la red que está definida por:

$$f_k(X) = g\left(\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)\right) = g(Net_k) \tag{25}$$

En este caso para la capa de entrada se usará como función de activación a la función exponencial, que está definida por:

$$\sigma(w_{0m} + w_m^T X) = \sigma(Net_m) = \frac{1}{1 + exp(-Net_m)} \tag{26}$$

y para la capa oculta se usará la función exponencial normalizada o función de activación softmax (Bridle 1990) que está dada por:

$$g(Net_{ki}) = \frac{e^{Net_{ki}}}{\sum_{k=1}^{K} e^{Net_{ki}}}$$

$$Net_k = \beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X) \tag{27}$$

donde k es el número de clases.

Bajo estas condiciones las salidas de la red $f_k(X)$ con $k = 1,2,\dots,K$ representa una distribución de probabilidad, porque:

$$0 \leq f_k(X) \leq 1 \quad \text{y} \sum_{k=1}^{K} f_k(X) = 1$$

Suponiendo que la salida esperada, está dada por:

$$E[Y_k/X] = f_k(X) = g\left(\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)\right) \qquad (28)$$

para ajustar este modelo, se compara la salida real G con la salida de $f_k(X) = g(Net_k)$ mediante la función de pérdida del error:

$$L\left(G, f_k(X)\right) = -logP(G)$$

$$L\left(G, f_k(X)\right) = -log \prod_{k=1}^{K} \left(f_k(X)\right)^{G_k}$$

$$L\left(G, f_k(X)\right) = -\sum_{k=1}^{K} G_k . log\left(f_k(X)\right)$$

$$L\left(G, f_k(X)\right) = -\sum_{k=1}^{K} I(G = k) . log\left(f_k(X)\right)$$

Ahora el objetivo será minimizar la función de error dada por:

$$R = -\sum_{i=1}^{N} \sum_{k=1}^{K} G_{ki} . log\left(f_k(X)\right) \qquad (29)$$

<u>Para la capa de salida</u>

Utilizando el método de gradiente descendiente:

$$\Delta\beta_{km} = -\gamma \frac{\partial R}{\partial \beta_{km}} = -\gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}}$$

$$\frac{\partial R_i}{\partial \beta_{km}} = -\frac{\partial R_i}{\partial Net_{ki}}\frac{\partial Net_{ki}}{\partial \beta_{km}}$$

$$\frac{\partial R_i}{\partial \beta_{km}} = -(G_{ik} - f_k(x_i))Z_{mi} = \delta_{ki}Z_{mi}$$

Actualizando los pesos de la capa de salida, luego de presentar todas las observaciones de entrenamiento se tiene la expresión:

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}} \tag{30}$$

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} + \gamma \sum_{i=1}^{N} (G_{ik} - f_k(x_i))Z_{mi}$$

<u>Si la neurona pertenece a la capa Oculta.</u>

Se tiene:

$$\Delta w_{ml} = -\gamma \frac{\partial R}{\partial w_{ml}} = -\gamma$$

$$\frac{\partial R_i}{\partial w_{ml}} = -\frac{\partial R_i}{\partial Net_{ki}}\frac{\partial Net_{ki}}{\partial Z_{mi}}\frac{\partial Z_{mi}}{\partial Net_{mi}}\frac{\partial Net_{mi}}{\partial w_{ml}}$$

$$\frac{\partial R_i}{\partial w_{ml}} = -x_{il}Z_{mi}(1 - Z_{mi})\sum_{k=1}^{K}(G_{ik} - f_k(x_i))\beta_{km}$$

Finalmente los pesos de la capa oculta son actualizados iterativamente mediante la siguiente relación:

$$w_{ml}^{(r+1)} = w_{ml}^{(r)} - \gamma \sum_{i=1}^{N} \frac{\partial R_i}{\partial w_{ml}}$$

$$w_{km}^{(r+1)} = w_{km}^{(r)} + \gamma \sum_{i=1}^{N} \sum_{k=1}^{K} (G_{ik} - f_k(x_i)) \beta_{km}^{(r+1)} Z_{mi}(1 - Z_{mi}) x_i \qquad (31)$$

2.2.3.8 SOBREENTRENAMIENTO *(OVERFITTING)*

Cuando la red neuronal presenta un número grande de capas y neuronas y el número de parámetros a ajustar durante el entrenamiento es muy elevado, puede que la red sea capaz de aprender de los ejemplos de entrenamiento pero, no será capaz de generalizar, es decir que estaría existiendo un sobreentrenamiento en la red.

2.2.3.9 EVALUACIÓN DEL MODELO

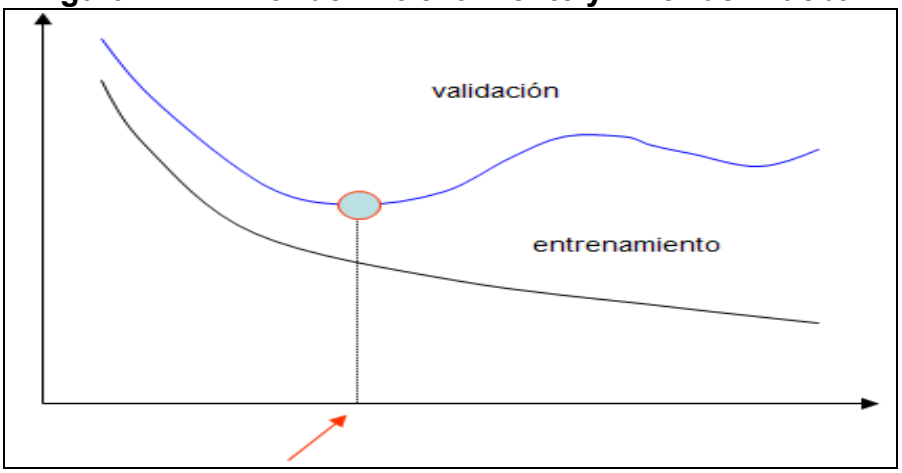
Para la evaluación de modelos se utilizará el método de validación cruzada (cross-validation) que fue introducido por T. M. Cover. Dicho método consiste en dividir los datos en dos partes; una parte se utiliza como conjunto de entrenamiento para ajustar el modelo y la otra parte, llamada conjunto de prueba o test, se utiliza para estimar la tasa de clasificación incorrecta del clasificador con datos diferentes a los utilizados en el proceso de entrenamiento.

Para validar un modelo de redes neuronales se debe conseguir un error de prueba que sea pequeño.

2.2.3.9.1 Error de Entrenamiento y Error de Prueba

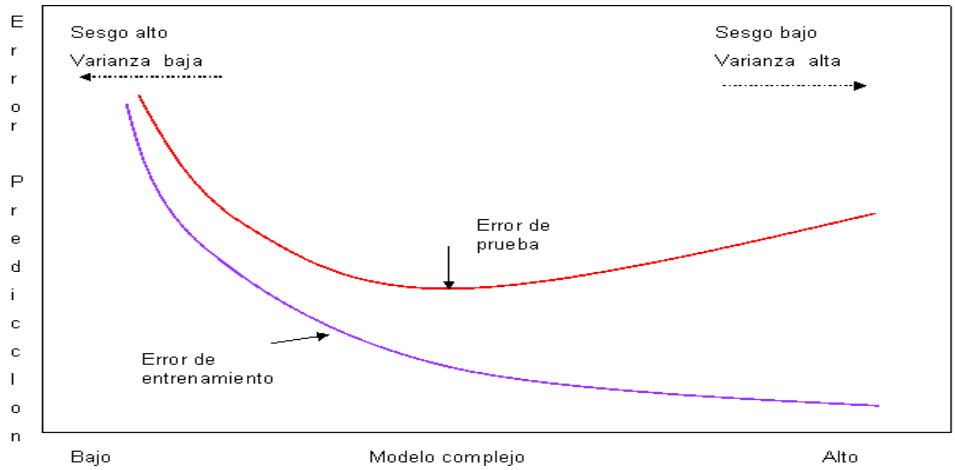
El comportamiento típico del error de entrenamiento de un clasificador decrece monótonamente durante la fase de entrenamiento, mientras que el error sobre el conjunto de validación decrece hasta un punto a partir del cual crece, lo que indica que a partir del mismo el clasificador realiza un *superajuste* sobre los datos de entrenamiento. Por ello, el proceso de entrenamiento debe finalizar cuando se alcance el primer mínimo de la función del error de validación.

Figura 2.14: Error de Entrenamiento y Error de Prueba



Además, el error de entrenamiento decrece a medida que el modelo tiene mayor grado de complejidad, mayor número de neuronas en la capa oculta, en este caso la red presenta mayor número de parámetros y grados de libertad y esto evidentemente se relaciona con un incremento en la varianza y una reducción del sesgo, debido a que el modelo se ajustará más al conjunto de observaciones de entrenamiento.

Figura 2.15: Comportamiento del Error de Prueba y de Entrenamiento



El error de entrenamiento frecuentemente puede ser menor que el error de prueba (Ver Figura 2.15), puesto que los datos de entrenamiento son inicialmente usados para el ajuste y evaluación del error.

Como la respuesta es categórica, entonces $G \in C = \{1,2,\dots,K\}$ donde los números son etiquetas que representan clases $\{G_1,\dots,G_k\}$, y la función de pérdida de error comúnmente usada para la clasificación es definida por:

$$L\left(G,\hat{G}(X)\right) = I\left(G \neq \hat{G}(X)\right) = \begin{cases} 1 & G \neq \hat{G}(X) \\ 0 & G = \hat{G}(X) \end{cases}$$

donde: $\hat{G}(X) = arg\ max_k\ \hat{p}_k(X)$

El error de prueba se define mediante la siguiente relación:

$$Err = E\left[L\left(G,\hat{G}(X)\right)\right] \qquad (32)$$

El error de entrenamiento se define por:

$$\overline{err} = \frac{\sum_{i=1}^{N} L\left(G,\hat{G}(X)\right)}{N} \qquad (33)$$

2.3 COMPARACIÓN TEÓRICA DE REGRESIÓN LOGÍSTICA Y REDES NEURONALES ARTIFICIALES EN PROBLEMAS DE CLASIFICACIÓN

Regresión Logística	Redes Neuronales Artificiales
Modelo: $E(Y/X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \pi(X) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$	Modelo: $Y = f(X) + \varepsilon = g[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)] + \varepsilon$
Supuestos: 1. Existe una regla que aconseja disponer de al menos $10 * (k+1)$ casos es decir, al menos 10 casos por cada variable que interviene en el modelo, considerando también la variable dependiente. 2. Las variables regresoras pueden ser variables cualitativas o cuantitativas. 3. Las variables regresoras son independientes entre sí. 4. La variable respuesta tiene distribución Bernoulli.	Supuestos: 1. El uso de redes neuronales es recomendable en problemas con bases de datos de tamaños grandes, donde la distribución de los datos no es conocida. 2. Las variables de entrada pueden ser variables cualitativas o cuantitativas. 3. Las variables de entrada no presentan el supuesto de independencia entre sí. 4. Las variables de salida son variables cualitativas.
Estimación de Parámetros: Se utiliza el método de máxima verosimilitud.	Aprendizaje de la Red Neuronal: Utiliza el backpropagation que es un algoritmo de aprendizaje supervisado.
Presenta una solución única ante unos mismos datos.	No garantiza que la solución sea única, porque está sujeta a las diferentes topologías como número de capas ocultas, funciones de activación entre otros hasta encontrar una solución adecuada.
Tiene menos posibilidades de una sobreestimación de parámetros.	Tienen posibilidades de una sobreestimación de parámetros debido a que no existe una técnica para determinar el número de capas ocultas y la elección dependerá de la experiencia del investigador.

CAPITULO III

CLASIFICACIÓN DE TUMORES DE MAMA DE PACIENTES DE LA UNIVERSIDAD DE WISCONSIN

La bibliografía revisada para las redes neuronales artificiales son: González [7], Martínez [154] y Rivera [17].

El cáncer de mama es la neoplasia maligna más frecuente en la mujer. La incidencia de este cáncer está en aumento en la mayoría de países, a pesar de la disminución de la tasa de mortalidad en los últimos años.

Cada año en Estados Unidos más de 180 000 mujeres se enteran de que padecen esta enfermedad. La detección temprana aumenta las probabilidades de supervivencia. Las mujeres con cáncer mamario pequeño y localizado (sin diseminación más allá del seno) tienen una excelente probabilidad de vivir muchos años luego del tratamiento del cáncer. En el Perú según la Revista Diagnóstico [15], es la primera causa de cáncer en mujeres y la cuarta causa de mortalidad. Su incidencia es 12.4 casos por cada 100,000 habitantes.

Existen muchos factores asociados al incremento de riesgo para el desarrollo del cáncer de mama, tales como el incremento de edad, parientes cercanos de primer

grado con cáncer de mama, primer embarazo en mujeres mayores de 35 años, menopausia tardía, menarquia temprana, obesidad en mujeres post-menopausicas, etc.

La detección del cáncer por medio de mamografías es el mejor método de diagnóstico por imágenes que se dispone en la actualidad sobre todo para la detección de lesiones mamarias no palpables durante el examen médico.

Tradicionalmente, el diagnóstico de cáncer de mama se realiza mediante biopsias, que constituyen un método quirúrgico invasivo. Una variante es la punción aspiración de aguja fina (PAAF). Esta técnica, proporciona una vía para examinar las características individuales de las células y utilizan técnicas de procesamiento de imágenes para la extracción de las características de las células.

En el Perú, la técnica punción aspiración de aguja fina se viene realizando desde el año 2008 por lo que no existe suficientes datos para poder desarrollar la clasificación de tumores usando redes neuronales; por ello la investigación se realizará con la base de datos "Wisconsin Diagnostic Breast Cancer" perteneciente al Hospital de la Universidad de Wisconsin.

Esta importante técnica constituye el primer paso de detección del cáncer mamario; sin embargo, para que sea totalmente efectiva, requiere ser correlacionada con otros métodos, como el examen físico, radiodiagnóstico y el estudio histopatológico, siendo este último más confiable para el diagnóstico porque nos permite determinar si hay benignidad o malignidad.

Los materiales necesarios para realizar la PAAF son:

- **Agujas:** Las agujas pueden ser de longitud variable aunque siempre es preferible utilizar las de menor longitud con objeto de reducir la desviación durante el abordaje de la lesión. El calibre oscila entre 21 y 25 G. Las más usuales son las de 21 G ya que las de mayor calibre provocan una muestra

excesivamente traumática hemorrágica. Los conos son transparentes para poder controlar la llegada del material aspirado y saber que se tiene suficiente y así poder interrumpir el proceso.

Figura 3.1: Agujas Para Realizar la PAAF



- **Jeringas:** Pueden utilizarse jeringas de 5, 10 o 20 ml. montadas en la pistola que va a permitir realizar la aspiración manteniendo la presión negativa de forma continua, con una sola mano y sin gran esfuerzo. Si se carece de pistola se puede mantener la presión negativa interponiendo el protector plástico de la aguja en el embolo de la jeringa, pero es un sistema más difícil de manejar.

Figura 3.2: Jeringa y Pistola Para Realizar PAAF

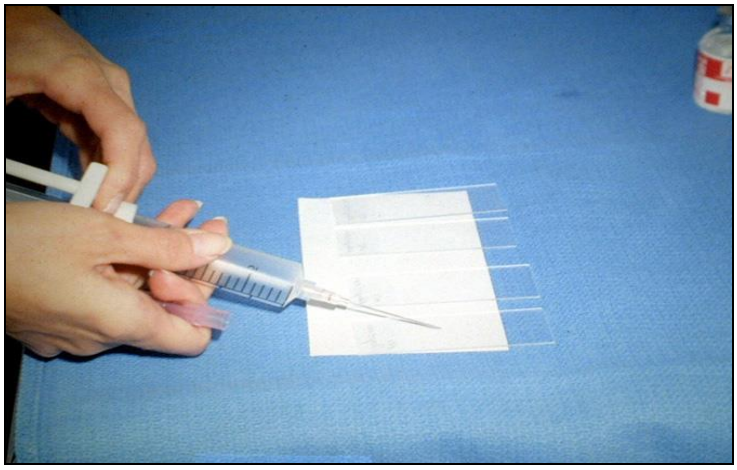


- **Solución desinfectante:** Servirá para limpiar la piel. Es aconsejable aplicarlo en dos tiempos. Realizando una primera limpieza y pasados unos minutos hacer una segunda aplicación. Si el método empleado es la guía ecográfica, la solución desinfectante servirá también como gel transmisor

evitando las interferencias de la interfase aire-piel. En caso de alergia al yodo se puede utilizar otro tipo de solución alcohólica.

- **Portas y tubos de ensayo:** Normalmente se utilizan 4 portas sobre los que se hacen las extensiones, dos se introducen en una solución alcohólica y los otros dos se deja que sequen al aire. Esta metodología puede variar dependiendo del servicio de Anatomía Patológica.

Figura 3.3: Procedimiento Para Colocar la Muestra en los Porta Objetos



El procedimiento para realizar la PAAF son:

1. Antes de empezar, se debe hacer un examen clínico cuidadoso para confirmar la presencia de lesiones o masas mamarias.

2. Se coloca a la paciente en una posición confortable donde se resalta la prominencia de la lesión de donde se extraerá la muestra.

3. Se explica el procedimiento a la paciente.

4. Se prepara el sitio de la punción y se limpia la piel con un líquido desinfectante.

5. Se aspira 5 ml de aire en la jeringa antes de conectar la aguja a la jeringa. Este procedimiento permite que una vez llevada a cabo la aspiración queden 5 ml de aire en la jeringa que pueden ser utilizados para expulsar el contenido de la aguja en el portaobjetos sin necesidad de quitar la aguja de la jeringa, este procedimiento se realiza para obtener una distribución más uniforme del aspirado en los portaobjetos.

6. Es importante que al estar extrayendo la muestra no se mueva con la aguja para poder obtener resultados óptimos.

7. La inserción de la aguja debe estar en un ángulo de 90 a 45 grados con la piel, se introduce la aguja y una vez que la aguja se encuentra en el seno de la masa mamaria, se debe realizar el movimiento de la aguja en un solo trayecto para obtener una muestra citológica satisfactoria.

8. Una vez que se observa la presencia de material necesario en la jeringa, se interrumpe la aspiración y se retira la aguja que todavía se encuentra conectada a la jeringa y se expulsa el material obtenido en los portaobjetos.

9. Luego el material se extiende en los portaobjetos y a continuación se procede al estudio de la muestra.

Figura 3.4: Técnica de Punción de Aspiración de Aguja Fina



Preparación de la Imagen

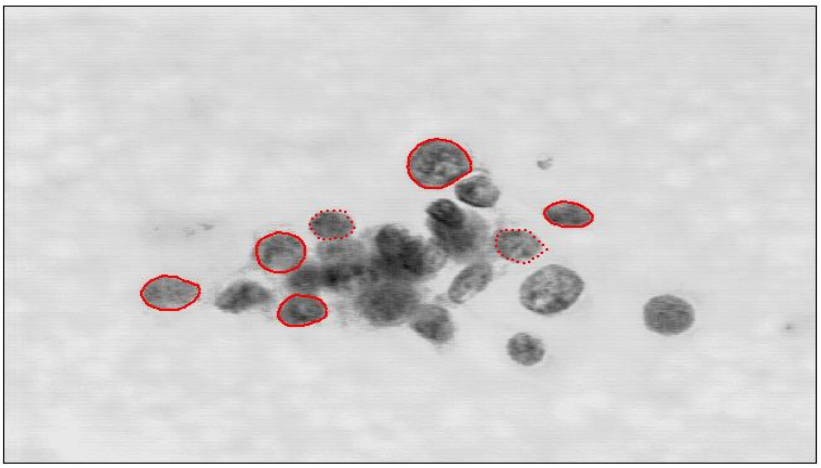
Para preparar la muestra el médico selecciona una región de la muestra, la cual es digitalizada utilizando una cámara de vídeo montada en un microscopio.

El área observada de las muestras fue visualmente seleccionada para un mínimo solapamiento nuclear de las células. El análisis de la imagen digital ha sido generado por una cámara de color de vídeo JVC TK-1070 montada sobre un microscopio Olympus y la imagen se proyectó en la cámara con 63X de objetivo[6] y 2.5 X de ocular[7]. La imagen fue capturada por una ComputerEyes/ RT color con capturador de imágenes con 640*400.

Se utilizó un programa de computación gráfica llamado Xcyt, que permite al usuario introducir la ubicación aproximada de un número suficiente de núcleos (10 a 20) para proporcionar una muestra representativa.

El mouse se utiliza para trazar un esbozo de núcleos de células en el monitor del ordenador. A partir de este esbozo, el límite actual del núcleo de la célula fue localizada por un modelo de contorno conocido como una "curva".

Figura 3.5: Imagen Digitalizada de la Muestra



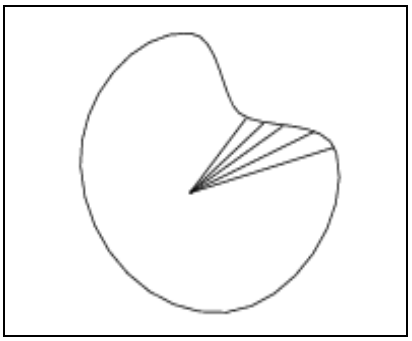
[6] **Objetivo:** Es una lente que amplía la imagen y es un elemento vital que permite ver a través de los oculares.

[7] **Ocular: L**ente situada cerca del ojo del observador. Capta y amplia la imagen formada en los objetivos.

Las variables que se identificaron en este estudio son:

1. **Radio:** el radio del núcleo de la célula individual es medido por conteo de los pixeles desde el centro a los puntos de la curva y luego obtener el promedio. Se sabe que un pixel es igual 0.353 mm (Ver figura 3.6)

Figura 3.6: Radio de la Célula



2. **Textura:** la textura del núcleo de la célula es medido por la desviación estándar de la escala de intensidades de grises en los componentes de los pixeles.

3. **Perímetro:** el número de pixeles que forma el límite de la curva es usado para medir el perímetro.

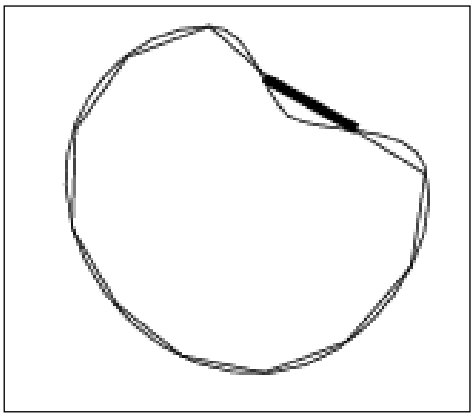
4. **Área:** el área es medido simplemente contando el número de pixeles en el interior de la curva y añadiendo un pixel y medio al valor del perímetro.

5. **Grado de Suavidad:** es la diferencia promedio en las variaciones de las longitudes del radio.

6. **Compactación:** el perímetro y el área están combinados para dar una medida de compactación del núcleo celular usando la fórmula perimetro^2/área. Estos resultados obtenidos son adimensionales.

7. **Concavidad:** mide la severidad de la concavidad, para ello se dibuja un segmento entre los puntos de la curvatura no adyacente tal como se muestra en la figura 3.7

Figura 3.7: Concavidad de la Célula



8. **Puntos de Concavidad:** es el número de porciones cóncavas del contorno.

9. **Simetría:** La diferencia relativa de la longitud entre los segmentos de recta perpendicular a uno y otro lado del eje mayor.

CAPITULO IV

ANÁLISIS Y RESULTADOS

Cabe indicar que esta aplicación se realiza con la base de datos "Wisconsin Diagnostic Breast Cancer" debido a que en Perú la técnica punción de aspiración de aguja fina se viene realizando desde el año 2008 y aun no se cuenta con datos suficientes para realizar una correcta clasificación de tumores de mama utilizando las redes neuronales artificiales.

Está base de datos consta de 561 datos y fue obtenida por el Dr. William H. Wolberg en los hospitales de la Universidad de Wisconsin para el diagnóstico de cáncer de mama en Noviembre de 1995.

Cada caso corresponde a la descripción de características citológicas de células constituyentes de un tumor, obtenidas mediante la observación microscópica.

Para el análisis de los resultados se utilizó los softwares estadísticos SPSS v 13, Tanagra v 1.4 y Microsoft Office Excel 2007.

4.1 Análisis Descriptivo

La base de datos consta de 561 casos y 10 variables las cuales son: diagnóstico del tumor, radio, textura, perímetro, área, grado de suavización, compactación, concavidad, puntos de concavidad y simetría. Los resultados del análisis descriptivo se encuentran en el Anexo A.

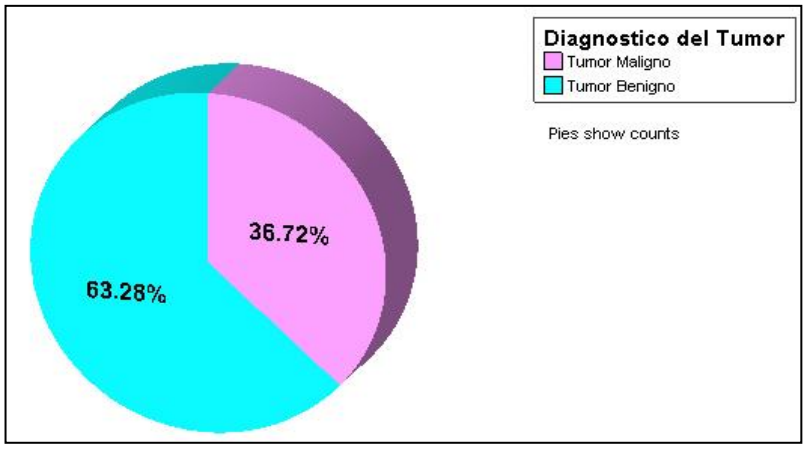
En las variables radio, textura, perímetro, área, grado de suavización, compactación, concavidad, puntos de concavidad y simetría se observaron la presencia de datos atípicos; una vez verificado los supuestos del modelo se analizará si son influyentes o no.

Las variables anteriormente mencionadas presentan una distribución asimétrica con cola a la derecha esto nos indica que la mayoría de los datos se encuentran alejados de la media hacia la derecha y que existe diferencias entre los tumores de mama en cuanto al radio, textura, perímetro, área, grado de suavización, compactación, concavidad, puntos de concavidad y simetría.

A continuación se explica algunas características de algunas variables debido a que la interpretación es similar.

Diagnóstico de las Células del Tumor: Se tiene 355 diagnósticos de tumores de mama benignos que representa el 63.28% y 206 tumores de mama malignos que representa el 36.72%. (Ver Figura 4.1.1)

Figura 4.1.1: Gráfico de Diagnostico del Tumor



Radio de las Células del Tumor: El promedio del radio de las células del tumor es de 14.13 pixeles (4.99 mm) con una variación de 3.52 pixeles (1.25 mm) por encima y por debajo del promedio.

El radio mínimo de las células del tumor es 6.98 pixeles (2.46 mm) y el máximo 28.11 pixeles (9.92 mm).

Se sabe que el 50% de los pacientes presentan tumores de mama de radio igual a 13.30 pixeles (4.69mm) y el 75% de los pacientes presentan tumores de mama de radio igual o inferior a 15.82 pixeles (5.66 mm).

En el grupo de pacientes con tumor de mama benigno el promedio del radio de las células del tumor es de 12.13 pixeles (4.28 mm), el tamaño mínimo de radio es 6.98 pixeles (2.46 mm) y el máximo 17.85 pixeles (6.30 mm). El 50% de los pacientes presentan tumores de mama de radio igual a 12.19 pixeles (4.30 mm) y el 75% de los pacientes presentan tumores de mama de radio igual o inferior a 13.37 pixeles (4.71 mm).

Mientras que en el grupo de pacientes con tumores de mama maligno el promedio medio del radio de las células del tumor es 17.58 pixeles (6.20 mm), el tamaño mínimo de radio es 10.95 pixeles (3.86 mm) y el máximo 28.11 pixeles (9.92 mm). El 50% de los pacientes presentan tumores de mama de radio igual a 17.47 pixeles (6.17 mm) y el 75% de los pacientes presentan tumores de mama de radio igual o inferior a 19.68 pixeles (6.95 mm).

Área del Tumor: El promedio del área de las células del tumor de mama es de 655.83 pixeles^2 con una variación de 354.15 pixeles^2 por encima y por debajo del promedio.

El área mínima del tumor es 143.50 pixeles^2 y el máximo 2501 pixeles^2.

El 50% de los pacientes presentan tumores de mama de área igual a 546.30 pixeles^2 y el 75% de los pacientes presentan tumores de mama de área igual o inferior a 793.20 pixeles^2.

En el grupo de pacientes con tumor de mama benigno el promedio del área de las células del tumor es de 461.72 pixeles^2, el tamaño mínimo del área es 143.5 pixeles^2 y el máximo 992.10 pixeles^2. El 50% de los pacientes presentan tumores de mama de área igual a 457.90 pixeles^2 y el 75% de los pacientes presentan tumores de mama de área igual o inferior a 546.40 pixeles^2.

Mientras que en el grupo de pacientes con tumores de mama maligno el promedio medio del área de las células es 990.33 pixeles^2, el tamaño mínimo del área es 361.60 pixeles^2 y el máximo 2501 pixeles^2. El 50% de los pacientes presentan tumores de mama de área igual a 949.80 pixeles^2 y el 75% de los pacientes presentan tumores de mama de área igual o inferior a 1207 pixeles^2.

Una vez terminado el análisis univariado se procedió a realizar un análisis bivariado observando que algunas variables tienden a presentar relación lineal las cuales son: área y puntos de concavidad, grado de suavización y puntos de concavidad, textura y área, finalmente área y grado de suavización. La relación que presentan están variables no es muy fuerte.

Una vez finalizado el análisis exploratorio, las variables que se utilizarán para realizar la clasificación de tumores según los supuestos requeridos por cada uno de los métodos de clasificación descritos anteriormente son: diagnóstico, textura, área, grado de suavización y puntos de concavidad de las células de los tumores de mama.

A continuación se realizará los métodos de clasificación utilizando Análisis de Regresión Logística y Redes Neuronales

4.2 Análisis de Regresión Logística

Al realizar el análisis exploratorio se comprobó los supuestos del análisis de regresión logística. Los resultados fueron obtenidos con el software SPSS y se utilizó el método Enter para ajustar el modelo de regresión. Los resultados de éste análisis se encuentran en el Anexo B.

Para el análisis de regresión logística, según el cumplimiento de los supuestos las variables independientes que fueron seleccionadas son textura, área, grado de suavización y puntos de concavidad de las células de los tumores de mama y como variable dependiente la variable diagnóstico del tumor de mama con las categorías de tumor benigno y tumor maligno.

Se incluyeron 561 casos en el análisis, no hubo ningún caso perdido y ningún caso no seleccionado.

El modelo que solo incluye el término constante realizó tres iteraciones, obteniendo un valor de desvianza de 737.658, la estimación del término constante es $\hat{\beta}_0 = 0.544$ y el error estándar es 0.088. Este modelo es:

$$\pi(X) = \frac{exp(\beta_0)}{1 + exp(\beta_0)}$$

$$\hat{\pi}(X) = \frac{exp(0.544)}{1 + exp(0.544)}$$

Luego en el siguiente paso ingresan las variables textura, área, grado de suavización y puntos de concavidad, realizando 10 iteraciones para ajustar el modelo, el cual es:

$$\pi(X) = \frac{exp(\beta_0 + \beta_1 Text + \beta_2 Area + \beta_3 Gsuav + \beta_4 PConcav)}{1 + exp(\beta_0 + \beta_1 Text + \beta_2 Area + \beta_3 Gsuav + \beta_4 PConcav)}$$

$$\hat{\pi}(X) = \frac{exp\left(31.3 - 0.48\widehat{Text} - 0.01\widehat{Area} - 75.05\widehat{GSuav} - 116.22\widehat{PConcav}\right)}{1 + exp\left(31.3 - 0.48\widehat{Text} - 0.01\widehat{Area} - 75.05\widehat{GSuav} - 116.22\widehat{PConcav}\right)}$$

Este modelo puede linealizarse, obteniendo:

$$Ln\left(\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)}\right) = 31.3 - 0.48\widehat{Text} - 0.01\widehat{Area} - 75.05\widehat{GSuav} - 116.22\widehat{PConcav}$$

La incorporación de las variables textura, área, grado de suavización y puntos de concavidad en el modelo mejora la explicación del diagnóstico de las células del tumor de los pacientes, obteniendo una ganancia de 630.656.

Todas las variables bajo estudio son significativas a un nivel de significación del 5%.

El modelo ajustado nos sirve también para predecir la probabilidad de tener pacientes con diagnóstico de células de tumores en función de las variables independientes.

Así por ejemplo la variable textura tiene un OR = 0.620, lo cual nos indica que la presencia de la variable textura ofrece un riesgo, 0.620 veces mayor en pacientes que presentan tumores de mama malignos con respecto a los pacientes que presentan tumores de mama benignos.

La variable área tiene un OR = 0.987, lo cual nos indica que la presencia de la variable área ofrece un riesgo, 0.987 veces mayor en pacientes que presentan tumores de mama malignos con respecto a los pacientes que presentan tumores de mama benignos.

Para la evaluación del modelo se utilizó la prueba de Hosmer y Lemeshow, que presenta las siguientes hipótesis:

$$H_0: El\ modelo\ se\ ajusta\ correctamente$$
$$H_1: El\ modelo\ no\ se\ ajusta\ correctamente$$

El p-valor asociado es 0.885 y como es mayor a $\alpha = 0.05$, entonces no se rechaza la hipótesis nula, por consiguiente el modelo se ajusta correctamente a los datos.

Mientras que para la evaluación de los coeficientes individuales del modelo ajustado se utilizó la estadística de prueba de Wald, que presenta las siguientes hipótesis:

$$H_0: \beta_j = 0 \quad j = 1,2,3\ y\ 4$$
$$H_1: \beta_j \neq 0$$

Los p-valores asociados son menores al valor de $\alpha = 0.05$, entonces se rechaza la hipótesis nula, por consiguiente los coeficientes individuales son diferentes a cero.

Cuadro 4.2.1: Tabla de Clasificación[a]

Observado			Pronosticado		
			Diagnostico del Tumor		
			Tumor Maligno	Tumor Benigno	Porcentaje correcto
Paso 1	Diagnostico del Tumor	Tumor Maligno	192	14	93.2
		Tumor Benigno	12	343	96.6
	Porcentaje global				95.4

a. El valor de corte es .500

En el cuadro 4.2.1 se observa una especificidad del 96.6%, que es la probabilidad que el modelo clasifique correctamente a los pacientes que tienen tumores de mama benignos. También se observa una sensibilidad del 93.2% que es la probabilidad que el modelo clasifique correctamente a los pacientes que tienen tumores de mama malignos.

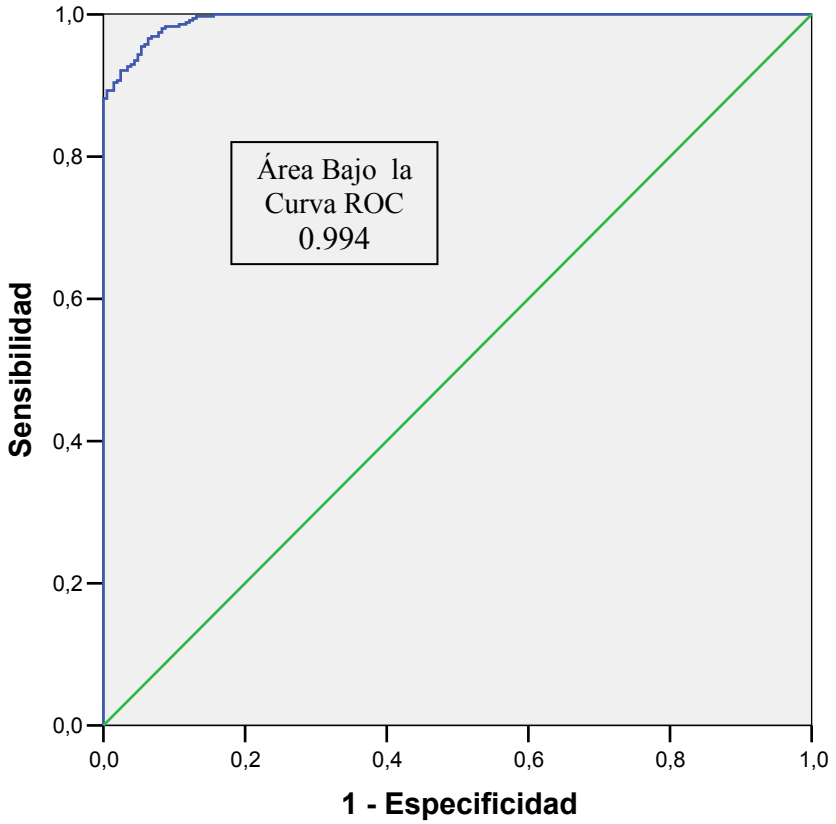
El porcentaje del diagnóstico de los tumores de mama correctamente clasificados es de 95.4%.

Al evaluar los residuos no se observó la presencia de datos atípicos y se confirmó que el modelo es adecuado. (Ver Anexo B – Cuadro 4.2.11)

Luego se procedió a realizar el análisis de influencia, utilizando la distancia de Cook y el valor leverage. Al realizar este análisis no se detectaron observaciones que influyeran en la estimación de los parámetros y en la calidad del ajuste del modelo. (Ver Anexo B – Cuadro 4.2.12)

La Figura 4.2.1 nos muestra la curva ROC cuya área es 0.994, como el valor está cercano a uno nos indica que el modelo utilizado tiene gran capacidad para clasificar las células de los tumores benignos y malignos.

Figura 4.2.1: Curva ROC



4.3 Redes Neuronales Artificiales

Como este método no requiere la comprobación de supuestos se procede a realizar la clasificación, para ello se procede a formar el conjunto de datos de entrenamiento el cual nos sirve para ajustar el modelo utilizando el 75% de los datos, esto quiere decir que el tamaño de muestra para el entrenamiento será igual a 420 (Ver Cuadro 4.3.1). Los resultados fueron obtenidos con el software Tanagra.

Cuadro 4.3.1: Proporción de Datos para el Ajuste del Modelo

Muestra Para el Ajuste
Parámetros
Proporción de Muestreo : 75.00%

El vector de entrada está formado por las variables: radio, textura, perímetro, área, grado de suavización, compactación, puntos de concavidad y concavidad.

La arquitectura que se utilizó fue el perceptrón multicapa con una sola capa escondida, el tipo de aprendizaje fue supervisado; además, se utilizó el algoritmo de backpropagation con una proporción de aprendizaje de 0.15 (Ver Cuadro 4.3.2)

Cuadro 4.3.2: Resumen de los Parámetros de la RNA

Aprendizaje Supervisado (Perceptrón Multicapa)	
Parámetros	
Arquitectura del MLP	
Uso de Capa Escondida	Si
Neuronas en la capa escondida	1
Parámetros de Aprendizaje	
Proporción de Validación	0.2
Tasa de Aprendizaje	0.15
Atributo de Transformación	Ninguna

El cuadro 4.3.3 presenta una especificidad del 97.78% que es la probabilidad que el modelo clasifique en el grupo de tumores benignos a los pacientes que tienen tumores de mama benignos y una sensibilidad de 96% que es la probabilidad que el modelo clasifique en el grupo de tumores malignos a los pacientes que tienen células de tumores malignos.

El porcentaje de los diagnósticos de tumores correctamente clasificados es de 97.14%.

Cuadro 4.3.3: Tabla de Clasificación para el Ajuste del Modelo

Tasa de Error: 0.0286			
Matriz de Confusión			
	M	**B**	**Sum**
M	144	6	150
B	6	264	270
Sum	150	270	420

En el cuadro 4.3.4 se muestra la tasa de error del perceptrón cuando se extrae un atributo, comparando la tasa de error de cada atributo respecto a la tasa de error de todo el modelo, con el fin de evaluar la importancia de cada atributo en la predicción.

Por ejemplo, la tasa de error de todo el modelo es 0.0286, al eliminar el atributo textura se obtendría una tasa de error de 0.0524 de todo el modelo lo que implica que estaría aumentando la tasa de error; es decir que éste atributo contribuye al modelo y no debe ser eliminado. Las variables que más contribuyen en la clasificación son textura, grado de suavización y concavidad debido a que si se eliminan estas variables la tasa de error aumentaría y lo que se busca es obtener la menor tasa de error

Cuadro 4.3.4: Contribución de los Atributos

Atributo Excluido	Tasa de Error	Diferencia
Ninguno	0.0286	
Radio	0.0381	0.0095
Textura	0.0524	0.0238
Perímetro	0.0357	0.0071
Area	0.0381	0.0095
GSuavización	0.0500	0.0214
Compactación	0.0405	0.0119
Concavidad	0.0595	0.031
PConcavidad	0.0333	0.0048
Simetria	0.0333	0.0048

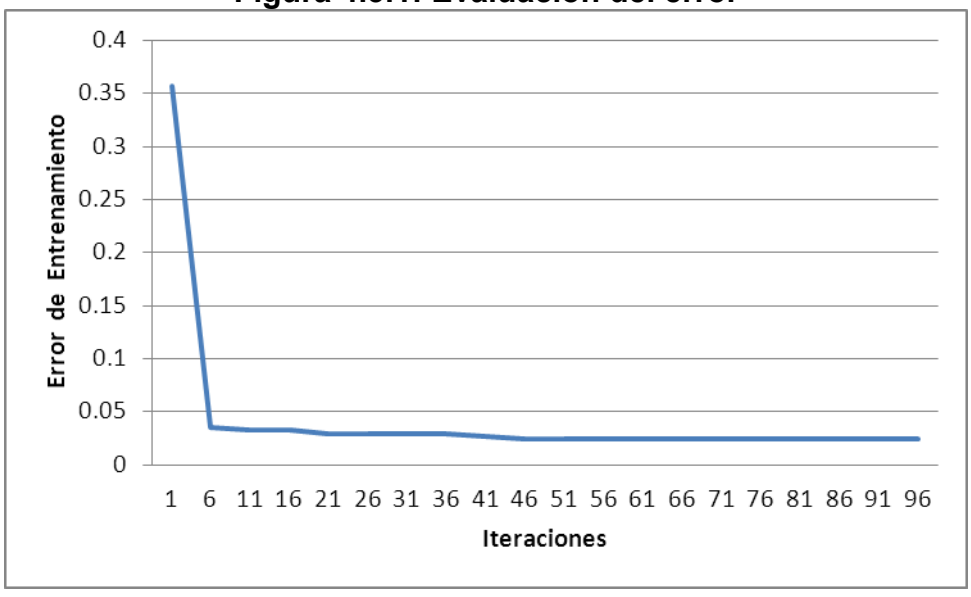
El error de entrenamiento decrece hasta la iteración 46 (repetición de entrenamiento de la red), después de esta iteración tiende a estabilizarse, por tanto, con 46 iteraciones ya se puede estimar el modelo de la red. (Ver Cuadro 4.3.5 y Figura 4.3.1)

El error de entrenamiento decrece debido a que los pesos que se da para cada observación se modifican en función al resultado que se obtuvo en el en la iteración anterior, de esta forma la red tiende a reconocer más el comportamiento de los datos para poder realizar una mejor clasificación.

Cuadro 4.3.5: Evaluación del error

Iteración	Error de Entrenamiento	Error de Prueba
1	0.3571	0.3571
6	0.0357	0.0238
11	0.0327	0.0357
16	0.0327	0.0357
21	0.0298	0.0357
26	0.0298	0.0357
31	0.0298	0.0476
36	0.0298	0.0476
41	0.0268	0.0476
46	0.0238	0.0476
51	0.0238	0.0476
56	0.0238	0.0476
61	0.0238	0.0476
66	0.0238	0.0476
71	0.0238	0.0476
76	0.0238	0.0476
81	0.0238	0.0476
86	0.0238	0.0476
91	0.0238	0.0476
96	0.0238	0.0476

Figura 4.3.1: Evaluación del error



Cuadro 4.3.6: Pesos de la Capa Oculta

-	Neurona "1"
Radio	-2.33675
Textura	-1.63861
Perímetro	-2.38919
Area	-2.61925
GSuavización	-2.40752
Compactación	1.24646
Concavidad	-2.77077
PConcavidad	-0.63651
Simetria	-0.92971
bias	0.27774

Cuadro 4.3.7: Pesos de la Capa de Salida

-	M	B
Neurona "1"	-7.20138	7.20138545
bias	3.231312	-3.2313138

El cuadro 4.3.6 y 4.3.7 nos indica los valores de los pesos de la capa oculta y la capa de salida, obteniendo el siguiente modelo:

$$Y = g[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)]$$

$$Y = g[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T [Text + Area + GSuav + PConcav])]$$

$$\hat{Y} = g\left[-3.23 + 7.20\sigma\left(0.28 - 2.34\widehat{Radio} - 1.64\widehat{Text} - 2.39\widehat{Perim} - 2.62\widehat{Area} - 2.41\widehat{GSuav} + 1.25\widehat{Compac} - 2.77\widehat{Concav} - 0.64\widehat{PConcav} - 0.93\widehat{Sim}\right)\right]$$

Cuadro 4.3.8: Tabla de Clasificación para la Prueba de Datos

Tasa de Error: 0.0993			
Matriz de Confusión			
	M	**B**	**Sum**
M	47	9	56
B	5	80	85
Sum	52	89	141

El cuadro 4.3.8 se evalúa el modelo de redes neuronales utilizando la muestra de prueba que corresponde a 141 diagnósticos y la tasa de error en la prueba es de 0.0993.

Tiene una especificidad del 94.12% que es la probabilidad que el modelo clasifique en el grupo de tumores benignos a los pacientes que tienen tumores de mama benignos y una sensibilidad de 83.93% que es la probabilidad que el modelo clasifique en el grupo de tumores malignos a los pacientes que tienen tumores de mama malignos.

El porcentaje de los diagnósticos de tumores correctamente clasificados es de 90.07%.

Como mi objetivo es realizar una comparación entre el análisis de regresión logística y las redes neuronales, utilizaré las mismas variables que se obtuvieron en el análisis de regresión logística y comprobar cuál nos da un mejor resultado para la clasificación de tumores de mama.

El procedimiento a seguir es el siguiente:

Se procede a formar el conjunto de datos de entrenamiento el cual nos sirve para ajustar el modelo utilizando el 75% de los datos, esto quiere decir que el tamaño de muestra para el entrenamiento será igual a 420 (Ver Cuadro 4.3.9). Los resultados fueron obtenidos con el software Tanagra.

Cuadro 4.3.9: Proporción de Datos para el Ajuste del Modelo

Muestra Para el Ajuste
Parámetros
Proporción de Muestreo : 75.00%

El vector de entrada está formado por las variables: textura, área, grado de suavización y puntos de concavidad.

La arquitectura que se utilizó fue el perceptrón multicapa con una sola capa escondida, el tipo de aprendizaje fue supervisado; además, se utilizó el algoritmo de backpropagation con una proporción de aprendizaje de 0.15 (Ver Cuadro 4.3.10)

Cuadro 4.3.10: Resumen de los Parámetros de la RNA

Aprendizaje Supervisado (Perceptrón Multicapa)	
Parámetros	
Arquitectura del MLP	
Uso de Capa Escondida	Si
Neuronas en la capa escondida	1
Parámetros de Aprendizaje	
Proporción de Validación	0.2
Tasa de Aprendizaje	0.15
Atributo de Transformación	Ninguna

El cuadro 4.3.11 presenta una especificidad del 97.33% que es la probabilidad que el modelo clasifique en el grupo de tumores benignos a los pacientes que tienen tumores de mama benignos y una sensibilidad de 91.77% que es la probabilidad que el modelo clasifique en el grupo de tumores malignos a los pacientes que tienen células de tumores malignos.

El porcentaje de los diagnósticos de tumores correctamente clasificados es de 95.24%.

Cuadro 4.3.11: Tabla de Clasificación para el Ajuste del Modelo

Tasa de Error: 0.0476			
Matriz de Confusión			
	M	**B**	**Sum**
M	145	13	158
B	7	255	262
Sum	152	268	420

En el cuadro 4.3.12 se muestra la tasa de error del perceptrón cuando se extrae un atributo, comparando la tasa de error de cada atributo respecto a la tasa de error de todo el modelo, con el fin de evaluar la importancia de cada atributo en la predicción.

Por ejemplo, la tasa de error de todo el modelo es 0.0476, al eliminar el atributo textura se obtendría una tasa de error de 0.0881 de todo el modelo lo que implica que estaría aumentando la tasa de error; es decir que éste atributo contribuye al modelo y no debe ser eliminado. Las variables que más contribuyen en la clasificación son el área, textura y puntos de concavidad debido a que si se eliminan estas variables la tasa de error aumentaría y lo que se busca es obtener la menor tasa de error

Cuadro 4.3.12: Contribución de los Atributos

Atributo Excluido	**Tasa de Error**	**Diferencia**
Ninguno	0.0476	-
Textura	0.0881	0.0405
Area	0.0976	0.05
GSuavización	0.0548	0.0071
PConcavidad	0.0881	0.0405

El error de entrenamiento decrece hasta la iteración 16 (repetición de entrenamiento de la red), después de esta iteración tiende a estabilizarse, por tanto, con 16 iteraciones ya se puede estimar el modelo de la red. (Ver Cuadro 4.3.13 y Figura 4.3.2)

El error de entrenamiento decrece debido a que los pesos que se da para cada observación se modifican en función al resultado que se obtuvo en el en la iteración anterior, de esta forma la red tiende a reconocer más el comportamiento de los datos para poder realizar una mejor clasificación.

Cuadro 4.3.13: Evaluación del error

Iteración	Error de Entrenamiento	Error de Prueba
1	0.378	0.369
6	0.0446	0.0714
11	0.0476	0.0595
16	0.0446	0.0595
21	0.0446	0.0595
26	0.0446	0.0595
31	0.0446	0.0595
36	0.0446	0.0595
41	0.0446	0.0595
46	0.0446	0.0595
51	0.0446	0.0595
56	0.0446	0.0595
61	0.0446	0.0595
66	0.0446	0.0595
71	0.0446	0.0595
76	0.0446	0.0595
81	0.0446	0.0595
86	0.0446	0.0595
91	0.0446	0.0595
96	0.0446	0.0595

Figura 4.3.2: Evaluación del error



Cuadro 4.3.14: Pesos de la Capa Oculta

-	Neurona "1"
Textura	-1.98974
Area	-3.06495
GSuavización	-1.13367
PConcavidad	-3.47049
Bias	0.794366

Cuadro 4.3.15: Pesos de la Capa de Salida

-	M	B
Neurona "1"	-6.89699	6.896993
bias	3.2741	-3.2741

El cuadro 4.3.14 y 4.3.15 nos indica los valores de los pesos de la capa oculta y la capa de salida, obteniendo el siguiente modelo:

$$Y = g[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T X)]$$

$$Y = g[\beta_{0k} + \beta_k^T \sigma(w_{0m} + w_m^T [Text + Area + GSuav + PConcav])]$$

$$\hat{Y} = g\left[-3.27 + 6.89\sigma\left(0.79 - 1.98\widehat{Text} - 3.06\widehat{Area} - 1.13\widehat{GSuav} - 3.47\widehat{PConcav}\right)\right]$$

Cuadro 4.3.16: Tabla de Clasificación para la Prueba de Datos

Tasa de Error: 0.0284			
Matriz de Confusión			
	M	**B**	**Sum**
M	46	2	48
B	2	91	93
Sum	48	93	141

El cuadro 4.3.16 se evalúa el modelo de redes neuronales utilizando la muestra de prueba que corresponde a 141 diagnósticos y la tasa de error en la prueba es de 0.0284.

Tiene una especificidad del 97.85% que es la probabilidad que el modelo clasifique en el grupo de tumores benignos a los pacientes que tienen tumores de mama benignos y una sensibilidad de 95.83% que es la probabilidad que el modelo clasifique en el grupo de tumores malignos a los pacientes que tienen tumores de mama malignos.

El porcentaje de los diagnósticos de tumores correctamente clasificados es de 97.16%.

4.4 *Comparación de Modelos de Clasificación*

Método	*Variables Utilizadas*	*Correcta Clasificación*	*Especificidad*	*Sensibilidad*
Regresión Logística	***Variable Dependiente*** • *Diagnóstico: presenta dos categorías tumores benignos y malignos.* ***Variable Dependiente*** • *Textura* • *Área* • *Grado de Suavización* • *Puntos de Concavidad*	*95.40%*	*96.60%*	*93.20%*
Redes Neuronales	***Variable de Salida*** • *Diagnóstico: presenta dos categorías tumores benignos y malignos.* ***Variable de Entrada*** • *Radio* • *Textura* • *Perímetro* • *Área* • *Grado de Suavización* • *Compactación* • *Concavidad* • *Puntos de Concavidad* • *Simetría*	*90.07%*	*94.12%*	*83.93%*
	Variable de Salida • *Diagnóstico: presenta dos categorías tumores benignos y malignos.* ***Variable de Entrada*** • *Textura* • *Área* • *Grado de Suavización* • *Puntos de*	*97.16%*	*97.85%*	*95.83%*

	Concavidad			

Al comparar los métodos de clasificación utilizando el análisis de regresión logística y redes neuronales para las cuatro variables ya antes mencionadas, resulta muy adecuada la clasificación de tumores de mama porque presentan porcentajes similares en la clasificación de tumores benignos y malignos. Esto nos indica que para este problema ambos métodos son adecuados.

Sin embargo, cuando se utiliza el modelo de redes neuronales con las 10 variables el porcentaje obtenido es mayor del 90%, lo que nos indica que para este caso la clasificación de tumores de mama es adecuada y además incluye variables que aportan más información en clasificación de tumores debido a la no exigencia de supuestos teóricos en las redes neuronales.

Por ello, para este caso se sugiere que ambos métodos pueden complementarse convenientemente para obtener resultados óptimos.

CONCLUSIONES

1. El porcentaje de correcta clasificación del diagnóstico de tumores en cáncer de mama obtenido por el análisis de regresión logística es de 95.4%, considerando las variables textura, área, grado de suavización y puntos de concavidad.

2. El porcentaje de correcta clasificación del diagnóstico de tumores en cáncer de mama obtenido por las redes neuronales es de 97.16%, considerando las variables textura, área, grado de suavización y puntos de concavidad.

3. Las variables que más influyen en la clasificación de tumores de mama son el área y la textura utilizando los métodos de clasificación de regresión logística y redes neuronales.

4. Los métodos de clasificación utilizando el análisis de regresión logística y redes neuronales resultan muy adecuados para clasificar los tumores benignos y malignos en cáncer de mama, presentando porcentajes similares.

ANEXO A

Análisis Descriptivo

Cuadro 4.1: Conteo del Diagnóstico de las Células del Tumor

		Recuento
Diagnostico del Tumor	Tumor Maligno	206
	Tumor Benigno	355

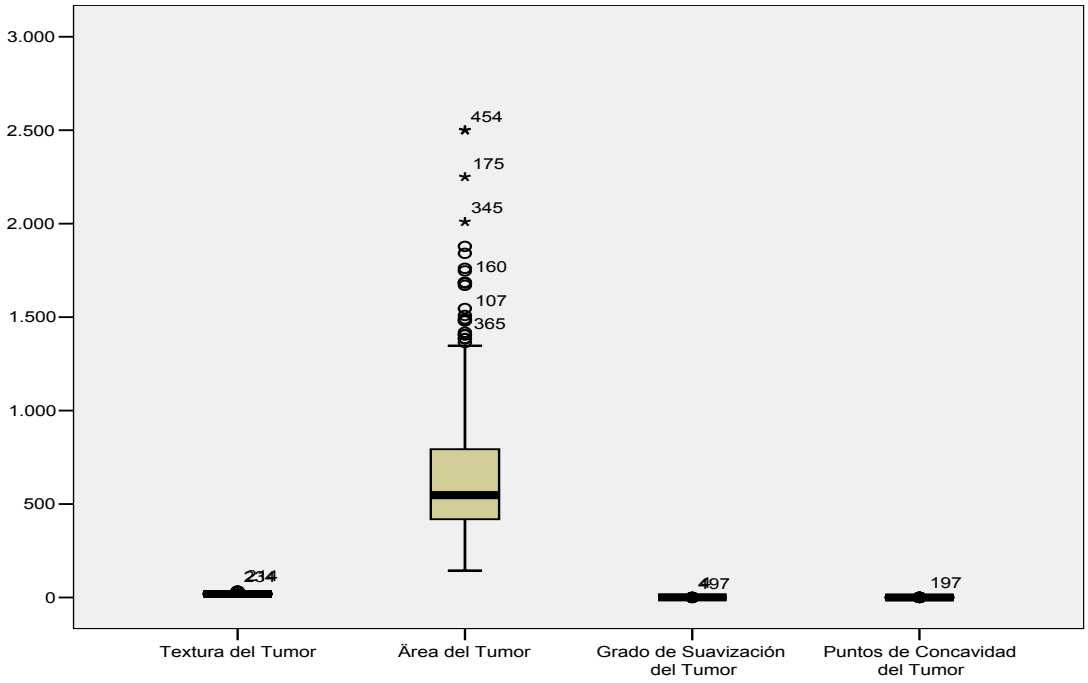
Cuadro 4.2: Estadísticas Descriptivas

	Mínimo	Máximo	Media	Mediana	Percentil 75	Desviación típica
Radio del Tumor	6.98	28.11	14.13	13.30	16.02	3.55
Textura del Tumor	9.71	39.28	19.29	18.83	21.81	4.32
Perímetro del Tumor	43.79	188.50	92.00	86.24	104.70	24.45
Ärea del Tumor	143.50	2501.00	655.83	546.30	793.20	354.15
Grado de Suav ización del Tumor	.05	.16	.10	.10	.11	.01
Compactación del Tumor	.02	.35	.10	.09	.13	.05
Concav idad del Tumor	.00	.43	.09	.06	.13	.08
Puntos de Concav idad del Tumor	.00	.20	.05	.03	.07	.04
Simetría del Tumor	.11	.30	.18	.18	.20	.03

Cuadro 4.3: Estadísticas Descriptivas Según Diagnóstico de las Células del Tumor

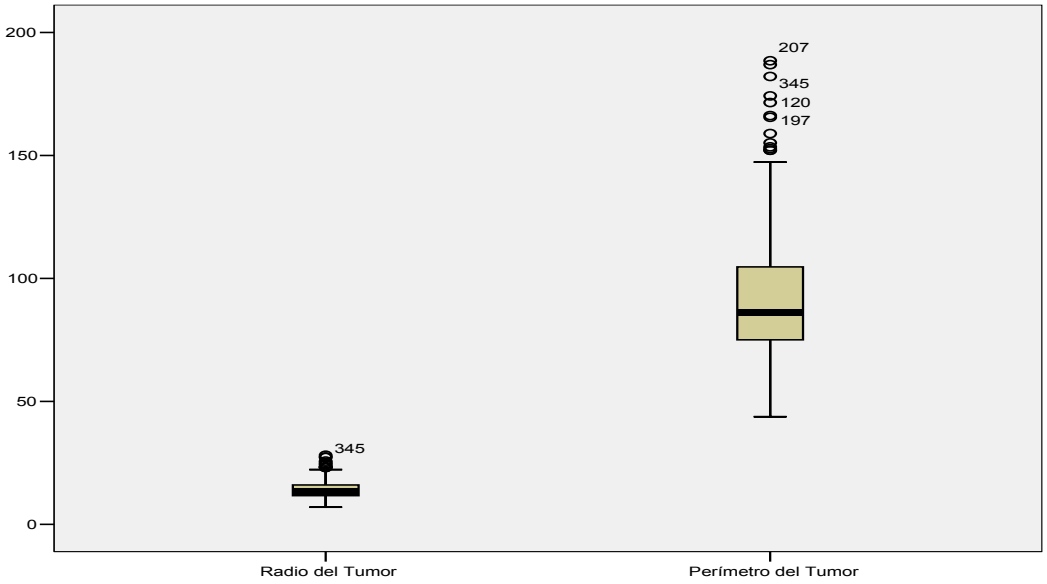
			Mínimo	Máximo	Media	Mediana	Percentil 75	Desviación típica
Diagnostico del Tumor	Tumor Maligno	Radio del Tumor	10.95	28.11	17.58	17.47	19.68	3.17
		Textura del Tumor	10.38	39.28	21.67	21.49	23.84	3.80
		Perímetro del Tumor	71.90	188.50	116.21	114.75	130.00	21.56
		Ärea del Tumor	361.60	2501.00	990.33	949.80	1207.00	366.03
		Grado de Suavización del Tumor	.07	.14	.10	.10	.11	.01
		Compactación del Tumor	.05	.35	.15	.13	.18	.05
		Concavidad del Tumor	.02	.43	.16	.15	.20	.07
		Puntos de Concavidad del Tumor	.02	.20	.09	.09	.10	.03
		Simetría del Tumor	.13	.30	.19	.19	.21	.03
	Tumor Benigno	Radio del Tumor	6.98	17.85	12.13	12.19	13.34	1.78
		Textura del Tumor	9.71	33.81	17.91	17.39	19.83	4.00
		Perímetro del Tumor	43.79	114.60	77.96	78.11	85.98	11.74
		Ärea del Tumor	143.50	992.10	461.72	457.90	546.40	133.90
		Grado de Suavización del Tumor	.05	.16	.09	.09	.10	.01
		Compactación del Tumor	.02	.22	.08	.08	.10	.03
		Concavidad del Tumor	.00	.41	.05	.04	.06	.04
		Puntos de Concavidad del Tumor	.00	.08	.03	.02	.03	.02
		Simetría del Tumor	.11	.27	.17	.17	.19	.02

Figura 4.1: Diagrama de Cajas Agrupadas de las Variables Textura, Área, Grado de Suavización, y Puntos de Concavidad de las Células del Tumor



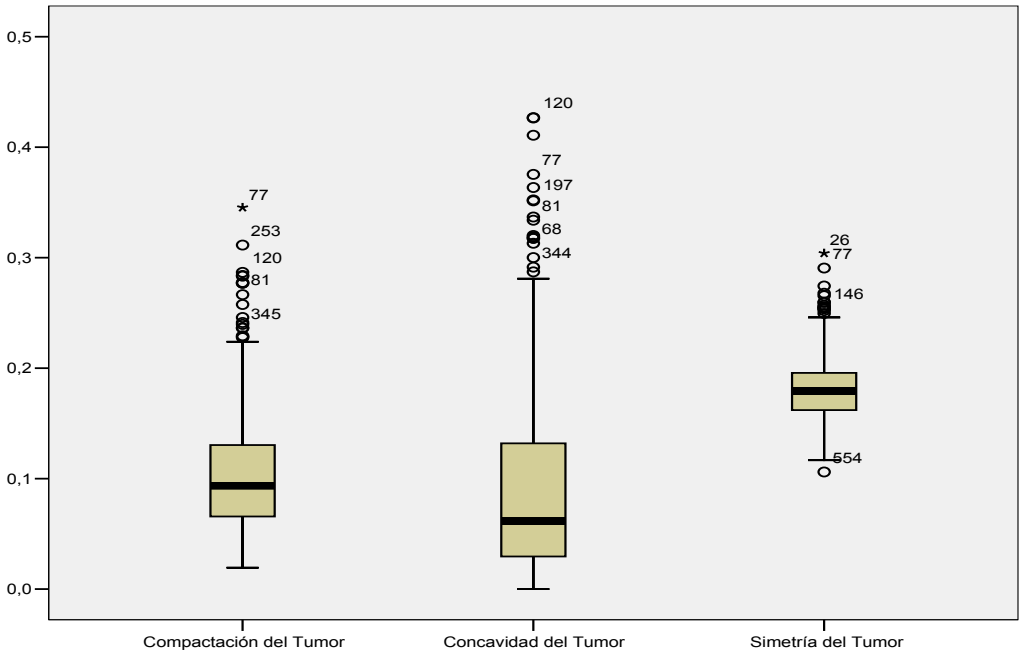
La figura 4.1 nos indica que la variable área del tumor presenta más variabilidad respecto a las variables textura, grado de suavización y puntos de concavidad.

Figura 4.2: Diagrama de Cajas Agrupadas de las Variables Radió y Perímetro de las Células del Tumor



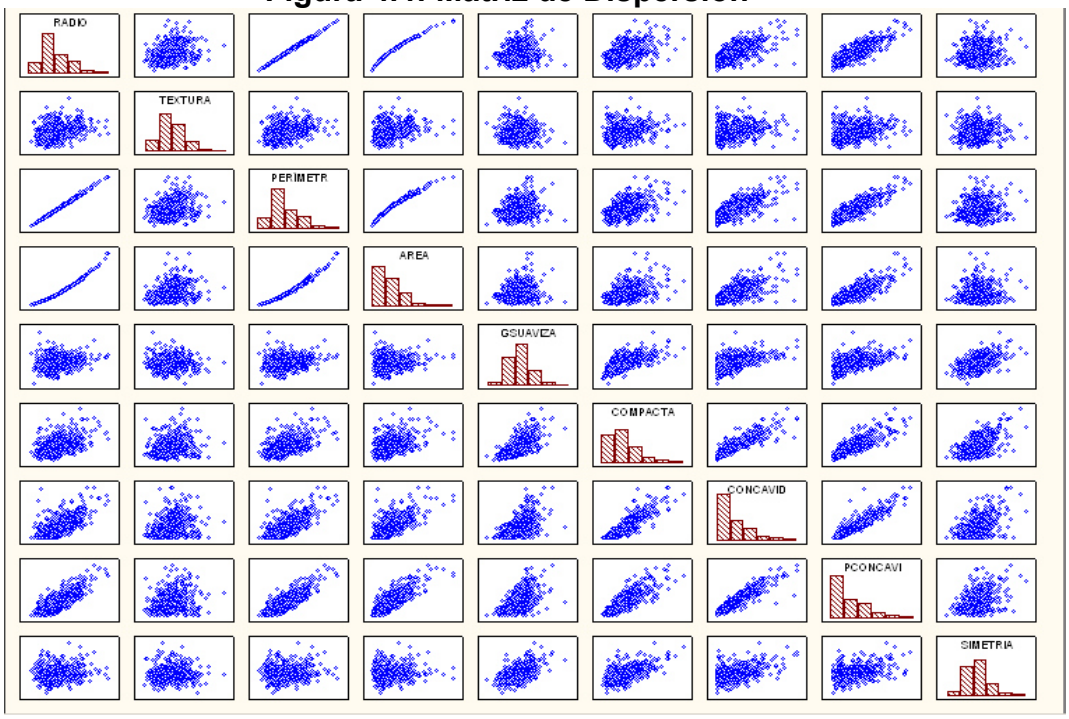
La figura 4.2 nos indica que la variable perímetro del tumor presenta más variabilidad respecto a la variable radio del tumor.

Figura 4.3: Diagrama de Cajas Agrupadas de las Variables Radio y Textura de las Células del Tumor



La figura 4.3 nos indica que las variables compactación y concavidad del tumor presenta más variabilidad respecto a la variable simetría del tumor.

Figura 4.4: Matriz de Dispersión



En la figura 4.4 se observa que algunas variables presentan relación lineal y la mayoría tiene una distribución asimétrica con cola a la derecha, de la

gráfica se observa que la mayoría de las distribuciones de las variables no son normales.

Cuadro 4.4: Matriz de Correlación

		Textura del Tumor	Ärea del Tumor	Grado de Suav ización del Tumor	Puntos de Concav idad del Tumor
Textura del Tumor	Correlación de Pearson	1	.322**	-.020	.297**
	Sig. (bilateral)		.000	.643	.000
	N	561	561	561	561
Ärea del Tumor	Correlación de Pearson	.322**	1	.178**	.825**
	Sig. (bilateral)	.000		.000	.000
	N	561	561	561	561
Grado de Suav ización del Tumor	Correlación de Pearson	-.020	.178**	1	.555**
	Sig. (bilateral)	.643	.000		.000
	N	561	561	561	561
Puntos de Concav idad del Tumor	Correlación de Pearson	.297**	.825**	.555**	1
	Sig. (bilateral)	.000	.000	.000	
	N	561	561	561	561

**. La correlación es signif icativ a al niv el 0,01 (bilateral).

La matriz de correlaciones nos ayuda a corroborar que variables presentan una relación lineal entre ellas.

ANEXO B

Análisis de Regresión Logística

Cuadro 4.2.1: Tabla de Clasificación[a]

			Pronosticado		
			Diagnostico del Tumor		
	Observado		Tumor Maligno	Tumor Benigno	Porcentaje correcto
Paso 1	Diagnostico del Tumor	Tumor Maligno	192	14	93.2
		Tumor Benigno	12	343	96.6
	Porcentaje global				95.4

a. El valor de corte es .500

Cuadro 4.2.2: Resumen del Procesamiento de los casos

Casos no ponderados [a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	561	100.0
	Casos perdidos	0	.0
	Total	561	100.0
Casos no seleccionados		0	.0
Total		561	100.0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

Cuadro 4.2.3: Historial de iteraciones[a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes Constante
Paso 0	1	737.680	.531
	2	737.658	.544
	3	737.658	.544

a. En el modelo se incluye una constante.

b. -2 log de la verosimilitud inicial: 737.658

c. La estimación ha finalizado en el número de iteración 3 porque las estimaciones de los parámetros han cambiado en menos de .001.

Cuadro 4.2.4: Tabla de clasificación[a,b]

			Pronosticado		
			Diagnostico del Tumor		
	Observado		Tumor Maligno	Tumor Benigno	Porcentaje correcto
Paso 0	Diagnostico del Tumor	Tumor Maligno	0	206	.0
		Tumor Benigno	0	355	100.0
	Porcentaje global				63.3

a. En el modelo se incluye una constante.

b. El valor de corte es .500

Cuadro 4.2.5: Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	.544	.088	38.611	1	.000	1.723

Cuadro 4.2.6: Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Textura	98.334	1	.000
		Area	290.930	1	.000
		GSuavización	77.008	1	.000
		PConcavidad	354.210	1	.000

a. No se calculan los chi-cuadrado residuales a causa de las redundancias.

Bloque 1:

Cuadro 4.2.7: Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes				
			Constante	Textura	Area	GSuavización	PConcavidad
Paso 1	1	318.085	4.449	-.085	-.001	-2.591	-28.961
	2	197.442	8.494	-.148	-.003	-10.960	-46.673
	3	142.767	13.443	-.226	-.005	-23.185	-63.747
	4	118.136	19.402	-.313	-.007	-40.305	-80.912
	5	109.084	25.456	-.398	-.010	-58.156	-98.264
	6	107.134	29.747	-.457	-.012	-70.575	-111.216
	7	107.003	31.204	-.477	-.013	-74.704	-115.815
	8	107.002	31.327	-.479	-.013	-75.048	-116.216
	9	107.002	31.327	-.479	-.013	-75.051	-116.219
	10	107.002	31.327	-.479	-.013	-75.051	-116.219

a. Método: Introducir

b. En el modelo se incluy e una constante.

c. -2 log de la verosimilitud inicial: 737.658

d. La estimación ha finalizado en el número de iteración 10 porque las estimaciones de los parámetros han cambiado en menos de .001.

Cuadro 4.2.8: Pruebas ómnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	630.656	4	.000
	Bloque	630.656	4	.000
	Modelo	630.656	4	.000

Cuadro 4.2.9: Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	3.683	8	.885

Cuadro 4.2.10: Variables en la Ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95.0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Textura	-.479	.084	32.443	1	.000	.620	.525	.730
	Area	-.013	.003	22.269	1	.000	.987	.982	.992
	GSuavización	-75.051	35.209	4.544	1	.033	.000	.000	.002
	PConcavidad	-116.219	24.576	22.363	1	.000	.000	.000	.000
	Constante	31.327	5.732	29.873	1	.000	4E+013		

a. Variable(s) introducida(s) en el paso 1: Textura, Area, GSuavización, PConcavidad.

Cuadro 4.2.11: Valores de Análisis de Residuos

Residuo studentizados			Residuo normalizado		
Recuento	Mínimo	Máximo	Recuento	Mínimo	Máximo
561	-2.35394	2.05924	561	-3.77632	2.58825

Cuadro 4.2.12: Valores de Análisis de Influencia

Análogo de los estadísticos de influencia de Cook			Valor de influencia Leverage		
Recuento	Mínimo	Máximo	Recuento	Mínimo	Máximo
561	.00000	.39794	561	.00000	.16461

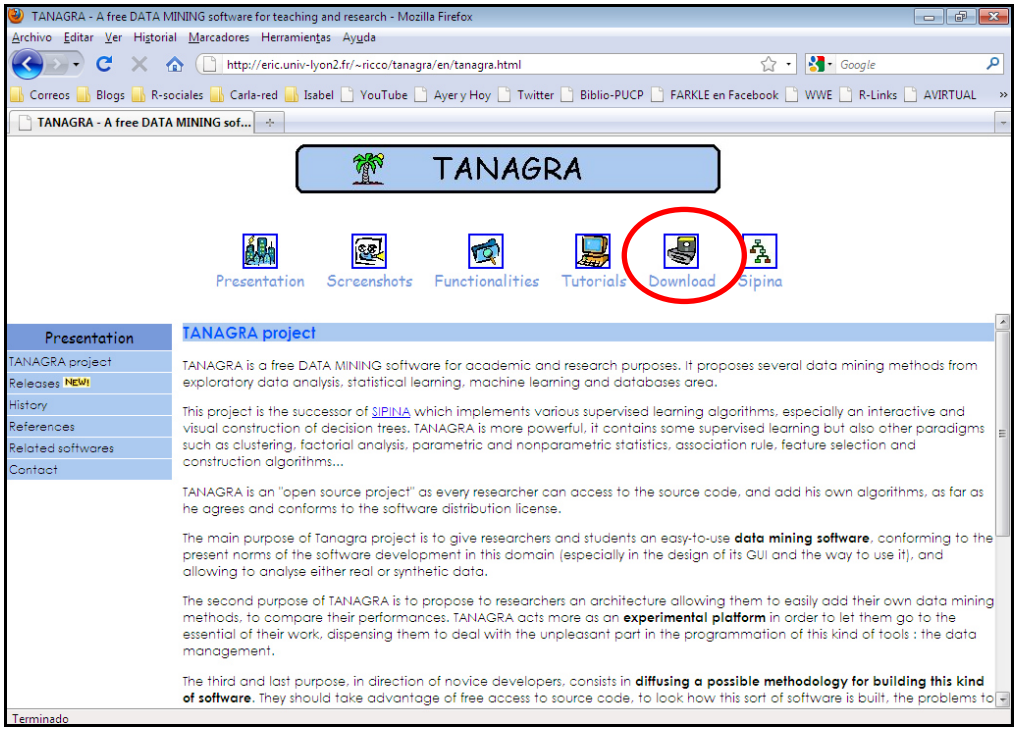
ANEXO C

Manual de Software Tanagra

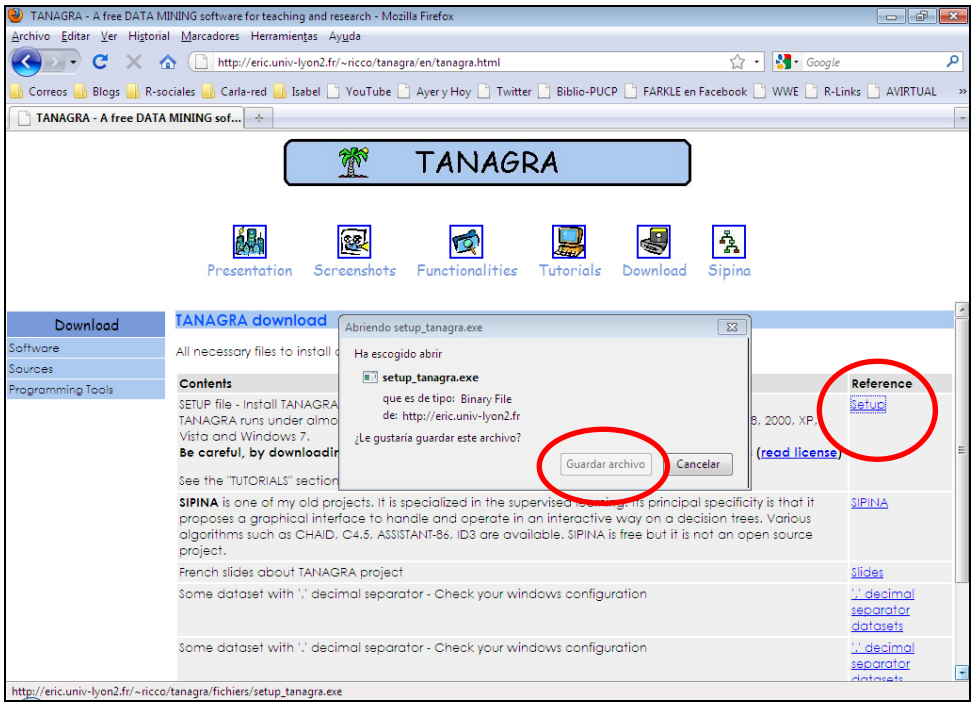
TANAGRA es un software libre de minería de datos para fines académicos y de investigación. Propone varios métodos de extracción de datos a partir del análisis exploratorio de datos y análisis estadísticos; además permite realizar ajustes en modelos de redes neuronales. TANAGRA se distribuye desde diciembre de 2003 y está compilado para la plataforma WIN32.

El software está disponible en: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

Para descargar el software se debe hacer click en la opción Download, tal como se muestra en la siguiente figura:



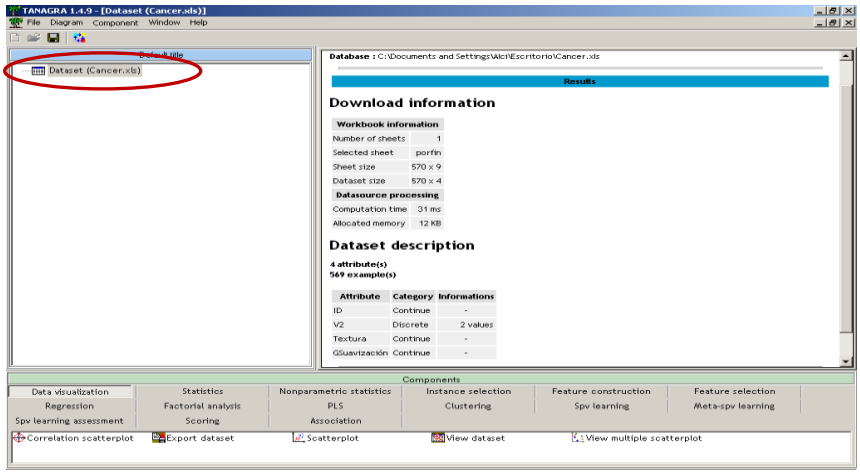
A continuación se debe hacer click en la opción Setup e inmediatamente aparecerá una ventana y se debe hacer click en la opción guardar. Con esta opción el instalador se guardará en el disco duro.



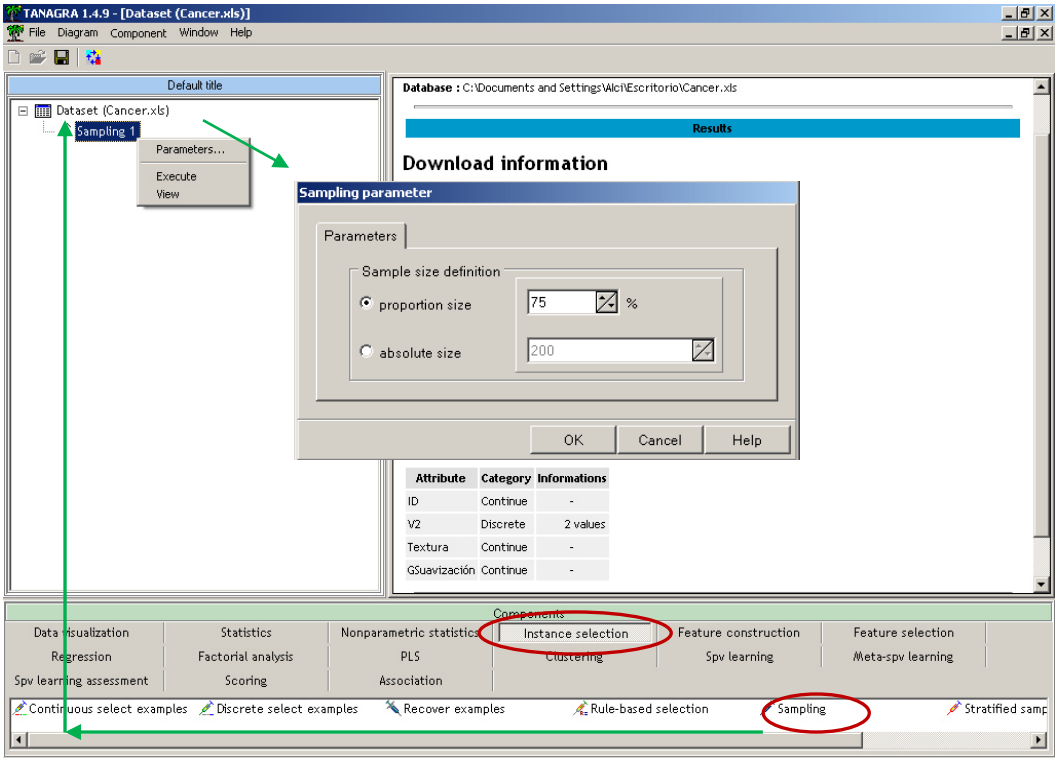
Una vez obtenido el instalador del software se procede a instalarlo como cualquier software.

A continuación procederé a explicar los pasos que se deben seguir para ajustar un modelo de redes neuronales así como la validación de este modelo.

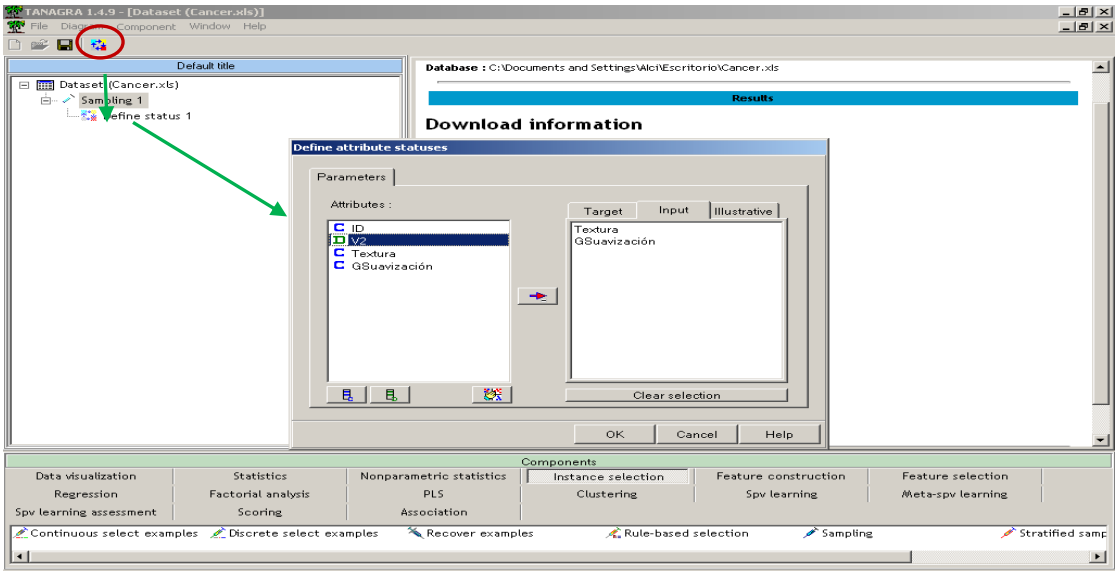
1. Abrir el archivo deseado, al realizar este paso aparecerá la opción Dataset seguido del nombre del archivo.



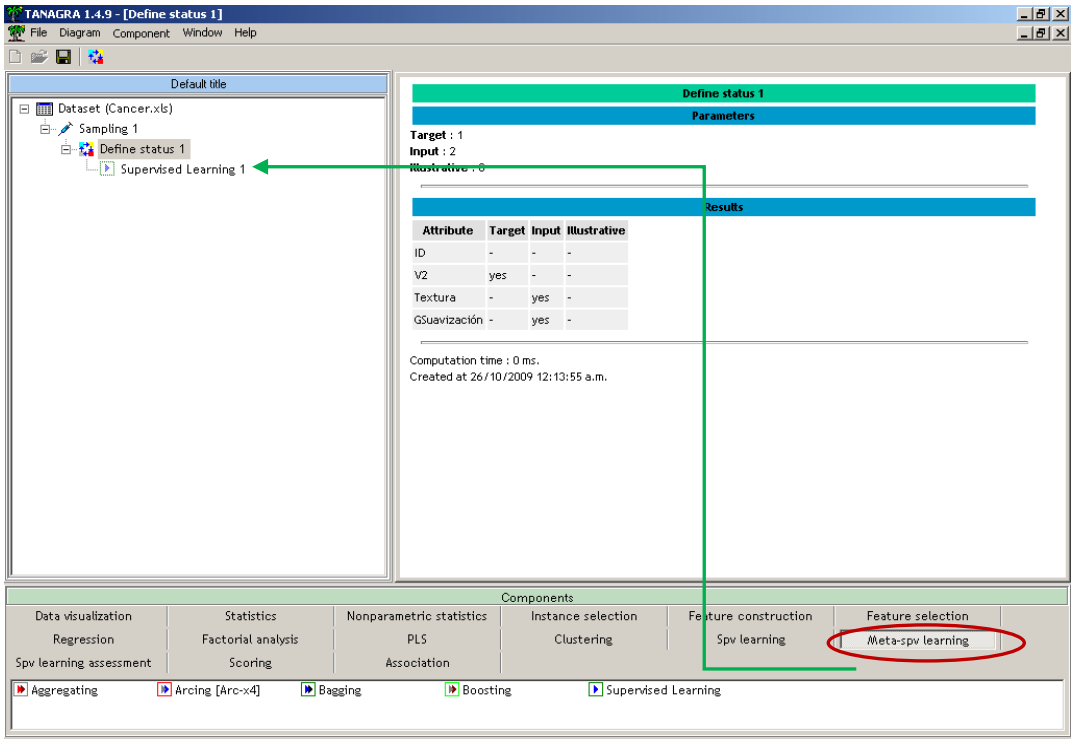
2. Se toma una parte de la muestra como conjunto de entrenamiento y la otra como conjunto de prueba. Para ello se inserta la opción Sampling y se toma el 75% de los datos para el entrenamiento.



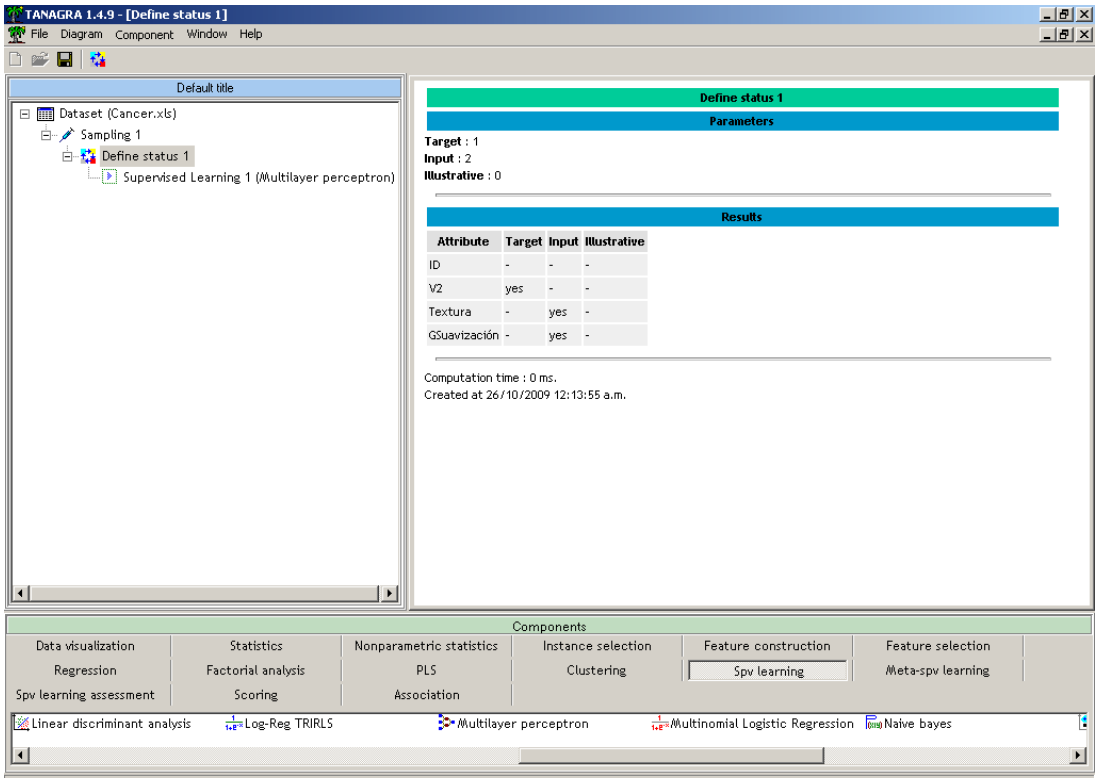
3. Para seleccionar las variables, se inserta la opción Define Status colocándolo debajo del elemento Sampling. En la pestaña Input se incluyen todas las variables independientes y en la pestaña Target se incluye la variable dependiente.



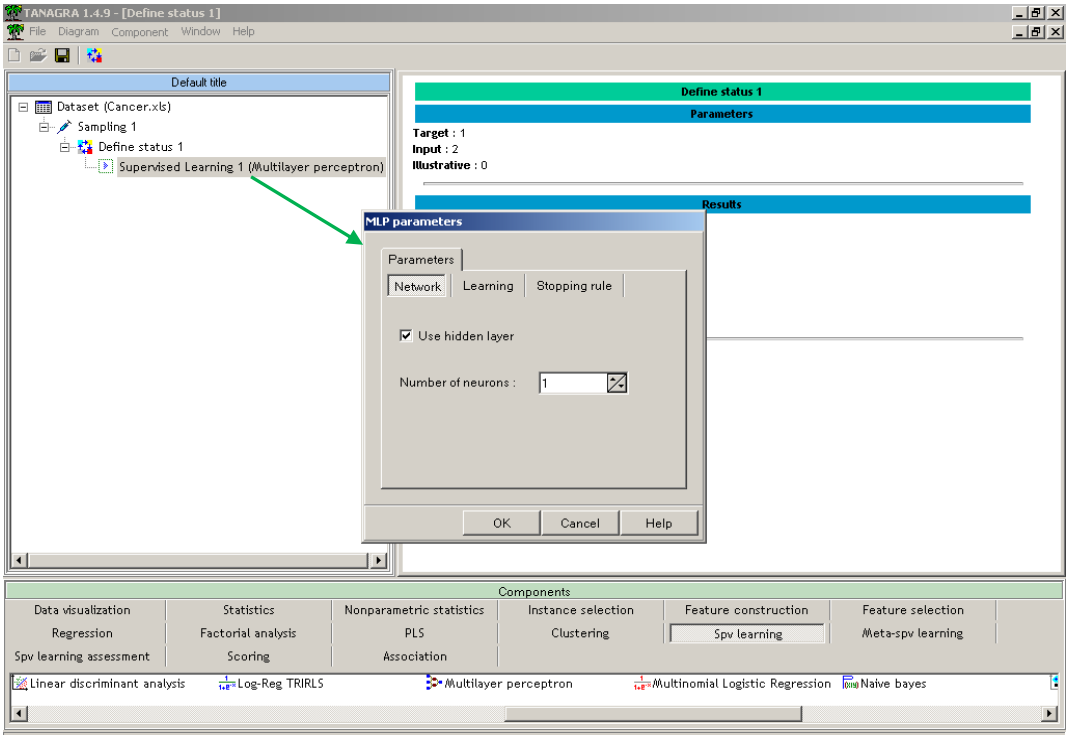
4. A continuación se inserta Supervised Learnig que está ubicado en la opción Meta-spv learning.



5. A lado de la opción Supervised Learnig se coloca Multilayer perceptron que se encuentra ubicado en la opción Spv learning.

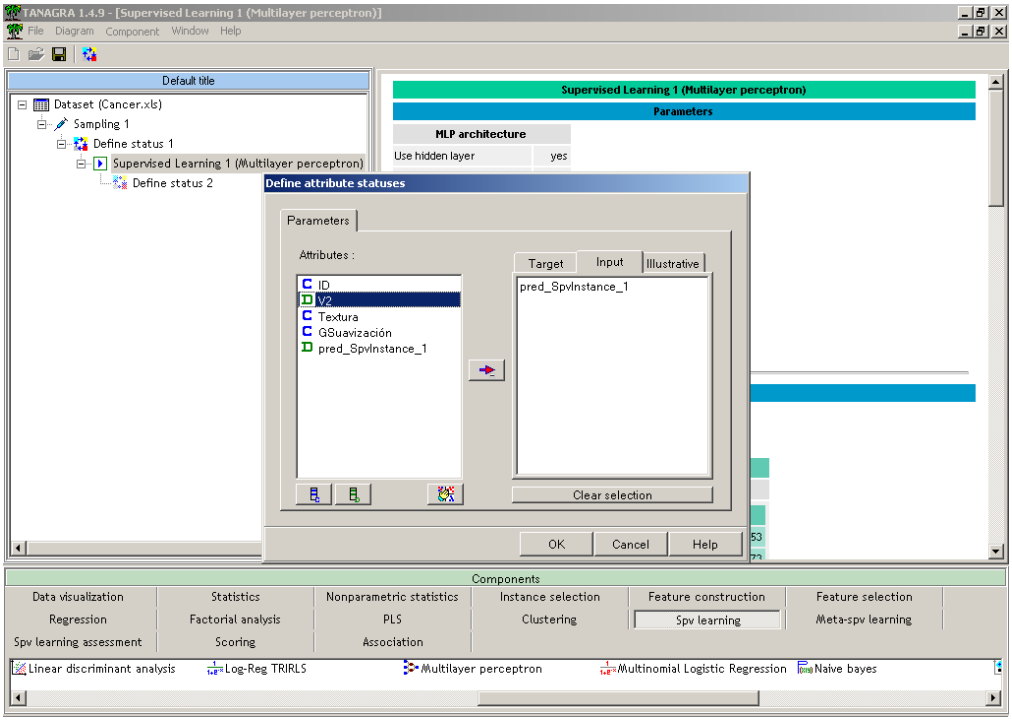


6. Se hace un anticlick en la opción Supervised learning, luego se va a la opción Supervised Parameters. Luego se procede a rellenar la información de acuerdo a la investigación que uno realiza.

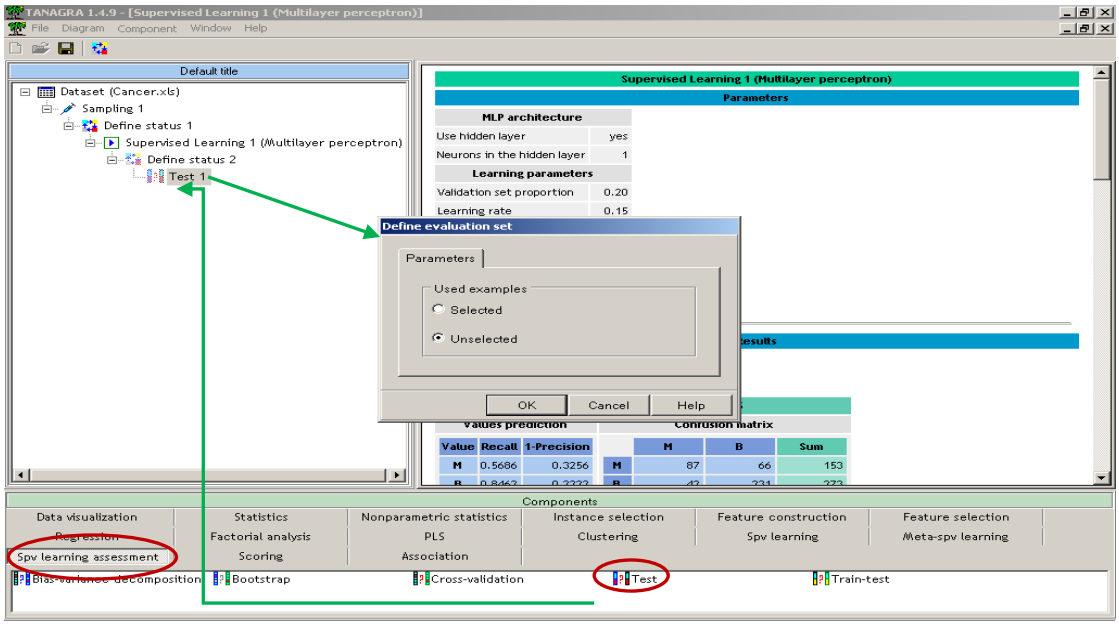


7. Nuevamente se hace un anticlick en la opción Supervised learning, luego se va a las opciones Execute y View respectivamente. De esta forma se realiza el aprendizaje de la red y se pueden observar los resultados obtenidos

8. Luego se calcula la tasa de error en la prueba, para ello añadimos la opción Define Status para analizar los datos de la prueba; a continuación en la pestaña input se coloca la variable pred_SpvInstance y en la pestaña target la variable dependiente.



9. Se añade la componente Test que está ubicado en la opción Spv learning assessment



10. Finalmente se hace un anticlick en la opción Test, luego se va a las opciones Execute y View respectivamente. De esta forma se realiza la validación de la red neuronal.

REFERENCIAS BIBLIOGRAFICAS

1. BISHOP, CHRISTOPHER M.(1995); Neural Network for Pattern Recognition, Clarendon Press – Oxford

2. CATENA ANDRÉS, RAMOS MANUAL M. & TRUJILLO HUMBERTO M. (2003); Análisis Multivariado – Un Manual para Investigadores, Editorial Biblioteca Nueva, S. L., Madrid.

3. CHENG BIN & TITTERINGTON D. M. (1994); Neural networks: A Review from statistical Perspective, Statistical Science Vol. 9 N°1.

4. DE LA TORRE DUEÑAS, CLETO (2007), Combinación De Clasificadores En Redes Neurales, Tesis para Optar el Grado Académico de Magister en Estadística, Universidad Nacional Mayor de San Marcos.

5. DUDA RICHARD O., HART PETER E. & STORK DAVID G. (2001); Pattern Classification, Second Edition, A Wiley – Interscience Publication, Jhon Wiley & Sons INC.

6. FERRERO, GUSTAVO A**.** (2005); Sistema de Ayuda Para la Detección de Cáncer de Mama, Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires.

7. GONZÁLEZ, ALAINA & GLEDHILL TERESA (2008); La Punción Aspiración con Aguja Fina como Herramienta Diagnóstica del Adenocarcinoma de la Glándula Mamaria, Extensos del X Congreso Venezolano de Patología, Caracas, Venezuela.

8. HASTIE TREVOR, TIBSHIRANI ROBERT & FRIEDMAN JEROME (2003); The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer Series in Statistics.

9. HERNÁNDEZ LILIA, TORREALBA VÍCTOR, REIGOSA Y ALDO **(2001).** Clasificación Automática del Carcinoma Intraductal de la Mama, Mediante Un Sistema de Reconocimiento Basado en Redes Neuronales, II Congreso Latinoamericano De Ingeniería Biomédica, La Habana 2001.Disponible en: http://www.hab2001.sld.cu/arrepdf/00250.pdf

10. HERNÁNDEZ ORALLO JOSÉ, RAMIREZ QUINTANA M.J. & FERRI RAMÍREZ CÉSAR (2004); Introducción a la Minería de Datos, Prentice Hall, Pearson educación S.A. Madrid.

11. JAN JANTZEN, JONAS NORUP, GEORGE DOUNIAS & BETH BJERREGAARD (2003); Pap – Smear Benchmark For Pattern Classification, Herlev University Hospital- Denmark.

12. KUANG CHUANG – MING & WHITE HALBERT (1992); Artificial Neural Networks: An Econometrics Perspective, Econometric Review Vol. 13 N°1.

13. MARTÍNEZ ORFILA, JUAN; Punción Aspirativa con Aguja Fina, Hospital de la esperanza. Disponible en: http://www.radiolegsdecatalunya.cat/formacio/resums/GE122EM05_R.pdf

14. PITARQUE ALFONSO, ROY JUAN FRANCISCO Y RUIZ JUAN CARLOS (1998); Redes neurales *vs* modelos estadísticos: Simulaciones sobre tareas de predicción y Clasificación, Psicológica 19, 387-400. Universidad de Valencia. Disponible en http://www.uv.es/revispsi/articulos3.98/pitarque.pdf

15. REVISTA DIAGNÓSTICO, Enero – Marzo 2007. [en línea]. Volumen 46 – N°1. Disponible en: http://www.fihu-diagnostico.org.pe/revista/numeros/2007/ene-mar/34-36.html

16. RIPLEY B. D. (1993); Statistical Aspects of Neural Networks, In networks and Chaos – Statistical and Probabilistic Aspect, Edited by Barnodorff – Nielsen O. E, Jansen J. L. and Kendall W. S., Chapman & Hall.

17. RIVERA NEUTZE, MARIA ISABEL (1999); Comparación de dos Técnicas de Obtención de Material Citológico por Aguja Fina, en el Diagnóstico de Masas Palpables de la Mama, Guatemala. Disponible en: http://www.tesis.ufm.edu.gt/pdf/2853.pdf

18. ROJAS CAMAYO, JOSÉ (2008); Lactancia Materna y Cáncer De Mama: Un Estudio Caso-Control En Pacientes Del Hospital Nacional Arzobispo Loayza, Lima-Perú, Anales de la Facultad de Medicina 69(1):22-8.

19. TORRES SÁNCHEZ, Mª TERESA (2005), Estudio de la Aplicación del Ganglio Centinela en el Diagnóstico y Tratamiento del Cáncer de Mama, Tesis parea Optar el Grado Académico de Doctor, Universidad de Valencia Disponible en: http://www.tesisenxarxa.net/TESIS_UV/AVAILABLE/TDX-1213106-110008//torres.pdf

20. URIEL, EZEQUIEL & ALDÁS, JOAQUÍN (2005); Análisis Multivariado Aplicado: Aplicaciones al Marketing, Investigación de Mercados, Economía, Dirección de Empresas y Turismo, Thomson Editores, España.

21. U.S. CANCER STATISTICS WORKING GROUP. UNITED STATES CANCER STATISTICS (1999–2005); Incidence and Mortality Web-based Report. Atlanta (GA). Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2009. Available at: http://www.cdc.gov/uscs.

22. WARNER BRAD & MISRA MENAVENDRA (1996); Understanding Neural Networks as Statistical Tools, The American Statistician Vol. 50, N° 4.

23. ZAPRANIS ACHILLEAS & REFENES APOSTOLOS P. (1999); Principles of Neural Model identification, Selection and Adequacy with Applications in Financial Econometrics, Springer – Verlag London.