FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA DE SISTEMAS

Modelo predictivo: Aprendizaje Supervisado Aplicado al Diagnostico de Ulceras Estomacales utilizando una red neuronal Perceptrón Multicapa

**Integrantes**

Apestegui Morales Alejandro U201812553

Bryan Jason Bolaños Quispe U201818900

**Curso**

Inteligencia Artificial

**Docente**

Jose Luis Bugarín Peche

**Sección**

T81B

Contenido

[1. Resumen 3](#_Toc13793361)

[2. Introducción 4](#_Toc13793362)

[a. Planteamiento del problema 4](#_Toc13793363)

[3. Estado del arte 5](#_Toc13793364)

[a. Fundamentos Matemáticos 5](#_Toc13793365)

[b. Algoritmo o modelo seleccionado (Perceptrón Multicapa) 7](#_Toc13793366)

[4. Desarrollo de la solución 9](#_Toc13793367)

[a. Explicación detallada del algoritmo o modelo seleccionado 9](#_Toc13793368)

[b. Variantes del algoritmo o modelo 11](#_Toc13793369)

[c. Usos y aplicaciones 11](#_Toc13793370)

[5. Resultados 13](#_Toc13793371)

[a. Línea base de la implementación 13](#_Toc13793372)

[6. Bibliografía 20](#_Toc13793373)

# Resumen

El presente documento tiene como objetivo demostrar como mediante técnicas de inteligencia artificial se puede llegar a predecir comportamientos. El comportamiento que se analizará es la aparición de Ulceras de estómago en pacientes, para este proyecto se toman datos de una encuesta realizada en España el año 2003 por la JUNTA DE CASTILLA Y LEÓN - CONSEJERÍA DE SANIDAD.

# Introducción

## Planteamiento del problema

La aparición de ulceras estomacales no siempre causan síntomas, lo mas común es una punzada o ardor en el abdomen, que se produce entre las comidas y durante las primeras horas en la mañana.

Si a este problema se adiciona la falta de tratamiento adecuado o preventivo, puede incurrir en complicaciones más serias como:

* Hemorragias. A medida que el revestimiento del estómago o de la pared duodenal se erosiona, también se pueden dañar los vasos sanguíneos y causar hemorragia.
* Perforación. Algunas veces, la úlcera erosiona y forma un orificio en la pared del estómago o del duodeno y las bacterias y los alimentos parcialmente digeridos pueden derramarse a través de la abertura en la cavidad abdominal (peritoneo) y causar peritonitis, una inflamación de la pared y de la cavidad abdominal.
* Estrechez y obstrucción. Cuando las úlceras se encuentran al final del estómago, donde se une con el duodeno, pueden causar inflamación y cicatrización que pueden estrechar o cerrar la abertura intestinal. Esta obstrucción puede evitar que los alimentos salgan del estómago y entren en el intestino delgado, lo cual produce el vómito del contenido del estómago.

Los factores que se consideran de relevancia en el desarrollo de ulcera de estomago o duodeno incluyen: estrés, edad, sexo, tabaquismo cafeína.

El presente proyecto pretende crear una red neuronal capaz de predecir la aparición de una ulcera estomacal teniendo como variables factores sociales como el grupo de edad, nivel de estudios, clase social, situación laboral, población de residencia y sexo. Los datos anteriormente mencionados serán obtenidos de una encuesta**1** realizada por JUNTA DE CASTILLA Y LEÓN - CONSEJERÍA DE SANIDAD del año 2003 en España

1. Problemas o enfermedades crónicas o de larga evolución diagnosticadas por el médico (ÚLCERA DE ESTÓMAGO). Números absolutos y distribución porcentual

<https://www.saludcastillayleon.es/sanidad/cm/gallery/ENCUESTA%20REGIONAL%20DE%20SALUD%202003/Castilla%20y%20Leon/Tabla%20I.3.6.HTM>

# Estado del arte

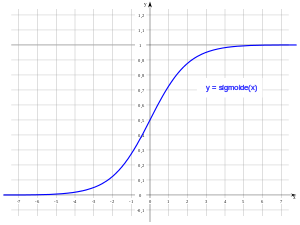
## Fundamentos Matemáticos

**Función sigmoide**

Muchos procesos naturales y curvas de aprendizaje de sistemas complejos muestran una progresión temporal desde unos niveles bajos al inicio, hasta acercarse a un clímax transcurrido un cierto tiempo; la transición se produce en una región caracterizada por una fuerte aceleración intermedia. La función sigmoide permite describir esta evolución. Su gráfica tiene una típica forma de "S". A menudo la función sigmoide se refiere al caso particular de la función logística, cuya gráfica se muestra a la derecha y que viene definida por la siguiente fórmula:

​

Gráfica de una función sigmoide.

[](https://commons.wikimedia.org/wiki/File:Funci%C3%B3n_sigmoide_01.svg)

**Métricas de evaluación**

Se evaluarán los clasificadores obtenidos de Weka para valorar el modelo generado.

Verdaderos positivos TP – True Positives

Instancias positivas (de la clase con ulcera) que fueron correctamente clasificadas como con ulcera.

Falsos positivos FP – False Positives

Instancias negativas (de la clase sin ulcera) que fueron incorrectamente clasificadas como con ulcera. Pacientes para los cuales se predice que sí tienen la enfermedad y no la tienen.

Tasa de verdaderos positivos - TP Rate

Es la proporción de instancias positivas correctamente clasificadas. Si TP es el número de instancias positivas correctamente clasificadas y P el total de instancias positivas.

Tasa de verdaderos negativos -TN Rate

Instancias negativas correctamente clasificadas como negativas. Si TN es el número de instancias negativas correctamente clasificadas y N el total de instancias negativas.

Precisión

Mide el porcentaje de aquellas instancias que son positivas respecto al total predichas como positivas.

Kappa statistic

Mide el nivel de acuerde entre la clasificación predicha por el modelo clasificador y la clasificación obtenida en los datos de prueba, corrigiendo el acuerdo que sucede por azar, es decir mide la probabilidad de que el clasificador clasifique una instancia al azar. Si P observado es el acuerdo observado entre el modelo clasificador y la clasificación real, y P esperado es la probabilidad de acuerdo por casualidad.

## Algoritmo o modelo seleccionado (Perceptrón Multicapa)

El perceptrón multicapa evoluciona el perceptrón simple y para ello incorpora capas de neuronas ocultas, con esto consigue representar funciones no lineales.

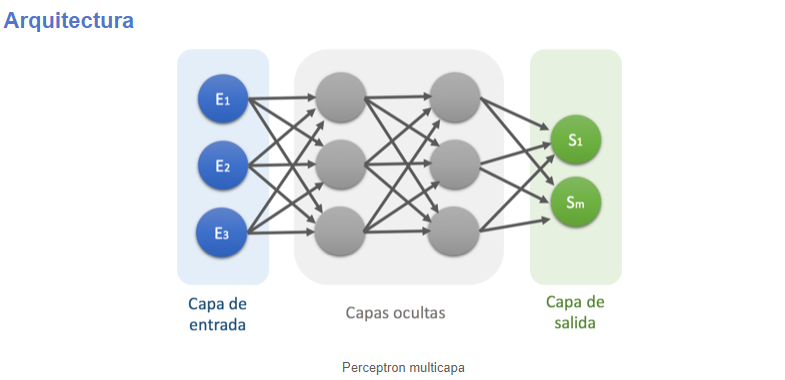
El perceptrón multicapa este compuesto por una capa de entrada, una capa de salida y n capas ocultas entremedias.

Se caracteriza por tener salidas disjuntas pero relacionadas entre sí, de tal manera que la salida de una neurona es la entrada de la siguiente.

En el perceptrón multicapa se pueden diferenciar en 2 fases:

Propagación en la que se calcula el resultado de salida de la red desde los valores de entrada hacia delante.

Aprendizaje en la que los errores obtenidos a la salida del perceptrón se van propagando hacia atrás (backpropagation) con el objetivo de modificar los pesos de las conexiones para que el valor estimado de la red se asemeje cada vez más al real, esta aproximación se realiza mediante la función gradiente del error.



* Capa de entrada: conecta la red con el exterior, cada neurona se corresponde con cada una de las variables de entrada de la red.
* Capas ocultas: son una aglomeración de capas en as que cada activación de una salida procede de la suma ponderada de las activaciones de la capa anterior conectadas, más sus correspondientes umbrales.
* Capa de salida: conecta las capas ocultas con la salida de la red que proporciona los resultados.

# Desarrollo de la solución

## Explicación detallada del algoritmo o modelo seleccionado

El Perceptrón Multicapa (MLP) es una extensión y generalización del perceptrón simple en el que:

* Se añaden una o más capas ocultas. Generalmente una capa oculta es suficiente.
* Se permiten entradas continuas.
* Las funciones de activación son de tipo **sigmoide** (puede ser lineal en la capa de salida).

Su principal ventaja e innovación es su algoritmo de aprendizaje, el algoritmo de retro propagación (en inglés backpropagation, BP), debido a Werbos, Rumelhart y Parker. El incremento de cada peso wij se expresa como:



aj(t) es la salida del elemento de proceso j

 es la derivada de la señal de error en la capa de salida con relación a la función de activación en el elemento de proceso i ( que es, como en el perceptrón simple, el producto escalar de vector de entrada por vector de pesos)

Si el elemento i al que se refiere  es un elemento de salida:



con f’i la derivada de la función de transferencia (sigmoide), con lo que nos queda:



El problema es propagar estas expresiones desde la capa de salida a las capas anteriores. Si el elemento i es de una capa interior (entrada u oculta):



con h=1...n, siendo n=número de elementos de proceso en la capa destino.

Si la función de transferencia es sigmoide, tenemos:



Estas ecuaciones forman la llamada regla delta generalizada.

La regla delta generalizada es en el fondo un algoritmo del gradiente, que trata de minimizar el error cuadrático medio (MSE – Medium Square Error):

1. Se aplica una pareja de ejemplo (x,y)

2. Se propaga hacia adelante y calculamos la salida real y’

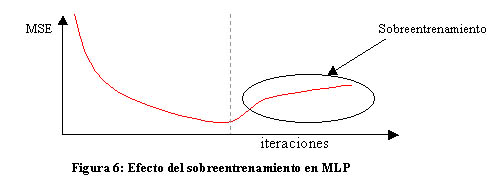
3. Se calcula el error según la diferencia entre y e y’

4. Se van modificando los pesos hacia atrás

5. Volvemos al paso 1 hasta que se alcance un cierto valor mínimo de error.

El proceso de entrenamiento y minimización del error puede sufrir distintos contratiempos. Uno es el sobre entrenamiento, que hace que, cuando la red entrena demasiado, pierda capacidad de generalización.

El otro fenómeno es el sobreajuste, que ocurre cuando hay demasiadas neuronas en la capa oculta, y que implica también pérdida en la capacidad de generalización.

Aquí podemos ver una gráfica con la evolución del error la lo largo del tiempo, y el efecto del sobre entrenamiento:

## Variantes del algoritmo o modelo

Se tiene la variante simplificada que es solo Perceptrón

El Perceptrón Simple (Rosenblatt, 1959), consta de una red con una capa de salida de n neuronas y otra de salida con m neuronas. Utiliza señales binarias, tanto de entrada como de salida de las neuronas y su función de activación es de tipo signo.

Como regla de propagación se usa la suma ponderada del producto escalar por el vector de entrada, debiendo superarse un cierto umbral siendo el factor de aprendizaje (una constante empírica que determina la facilidad con la que la red aprende) y la salida deseada. El perceptrón simple, como otras redes neuronales contemporáneas, se revelaron inútiles debido a sus limitaciones:

El perceptrón simple sólo sirve para clasificar problemas linealmente separables, cosa que ya se podía hacer mediante métodos estadísticos, y de una forma mucho más eficiente.

## Usos y aplicaciones

Se tiene claro que existen dos objetivos básicos para la investigación de redes neuronales.

El modelamiento del cerebro, la meta científica que construye modelos de como realmente funciona el cerebro. Esto significa una ayuda para entender la naturaleza de la inteligencia humana, plantear mejores estrategias de enseñanza, o mejores tratamientos para pacientes con daño cerebral.

Construcción de sistemas artificiales, meta de la ingeniería para crear sistemas eficientes para aplicaciones reales. Con eso se tiene procesamiento aun mas poderoso, disminución de la mano de obra de tareas tediosas, incluso mejorar el rendimiento humano.

Aplicaciones al mundo real se tiene a técnicas de márquetin, teoría de juegos, conducción autónoma, finanzas, reconocimiento visual, detección de patrones, etc. Se detallan en los siguientes libros

Airline Marketing Tactician (Beale & Jackson, Sect. 4.13.2)

Backgammon (Hertz et al., Sect. 6.3)

Data Compression – PCA (Hertz et al., Sect. 6.3; Bishop, Sect. 8.6)

Driving – ALVINN (Hertz et al., Sect. 6.3)

ECG Noise Filtering (Beale & Jackson, Sect. 4.13.3)

Financial Prediction (Beale & Jackson, Sect. 4.13.3; Gurney, Sect. 6.11.2)

Hand-written Character Recognition (Hertz et al., Sect. 6.3; Fausett, Sect. 7.4)

Pattern Recognition/Computer Vision (Beale & Jackson, Sect. 4.13.5)

Protein Secondary Structure (Hertz et al., Sect. 6.3)

Psychiatric Patient Length of Stay (Gurney, Sect. 6.11.1)

Sonar Target Recognition (Hertz et al., Sect. 6.3)

Speech Recognition (Hertz et al., Sect. 6.3)

Text to Phoneme Mapping (Beale & Jackson, Sect. 4.13.1; Bullinaria, 2011)

# Resultados

## Línea base de la implementación

El código fuente es entregado al docente, en adición se ha subido a un repositorio

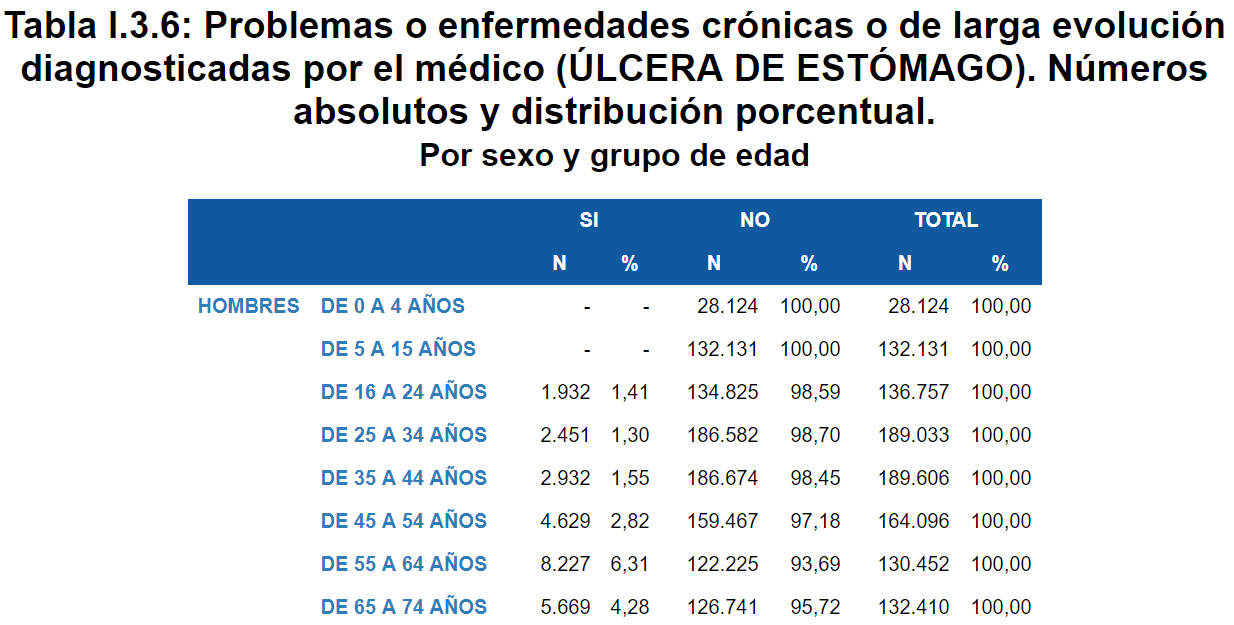
<https://github.com/AlejandroApestegui/MagIA>

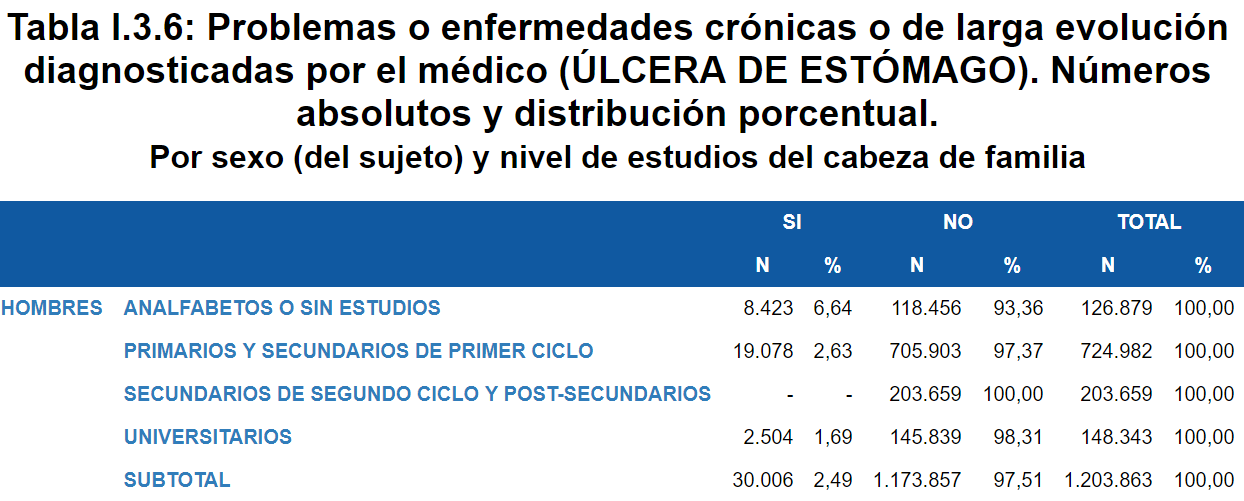
* + 1. Obtención de datos

Para la obtención de datos se usaron los datos obtenidos de 2.432.508 encuestas realizadas por SALUD CASTILLA Y LEON que contiene la siguiente información.

Las imágenes no contienen toda la información, para consulta total puede ingresar al siguiente enlace:

<https://www.saludcastillayleon.es/sanidad/cm/gallery/ENCUESTA%20REGIONAL%20DE%20SALUD%202003/Castilla%20y%20Leon/Tabla%20I.3.6.HTM>.)





En total se tienen 5 tablas

1) Ulcera por sexo y grupo de edad

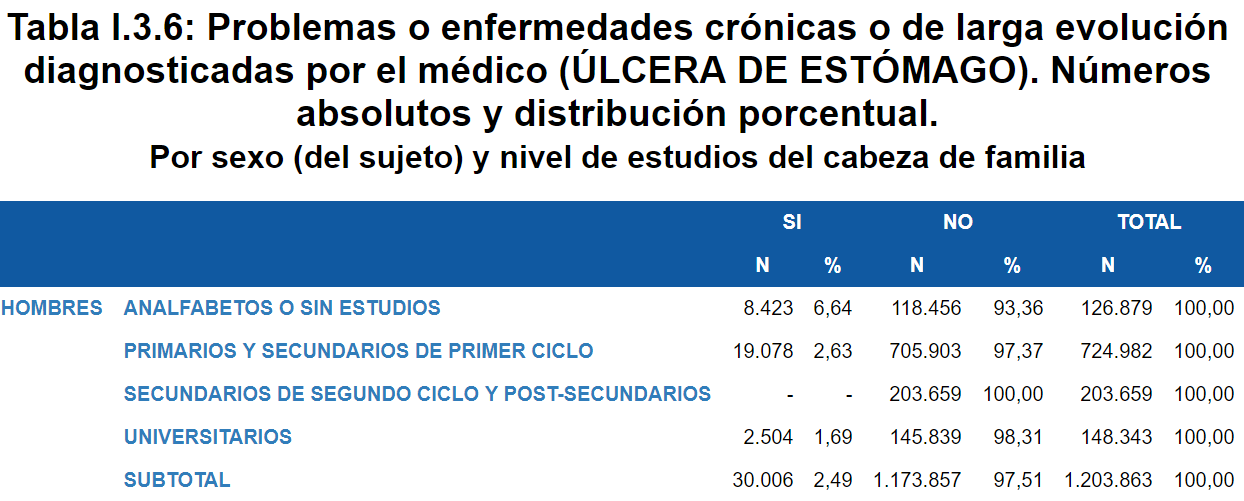
2) Ulcera por sexo y nivel de estudios del cabeza de familia

3) Ulcera por sexo y clase social del cabeza de familia

4) Ulcera por sexo y situación laboral del cabeza de familia

5) Ulcera por sexo y tamaño del municipio de residencia

De las presentes se extrajeron los datos de cantidad Si y No por cada categoría y sexo



* + 1. Transformación de datos

A continuación, se cargaron los datos al proyecto “LecturaDatos” (código fuente en el interior de la clase Main.java)

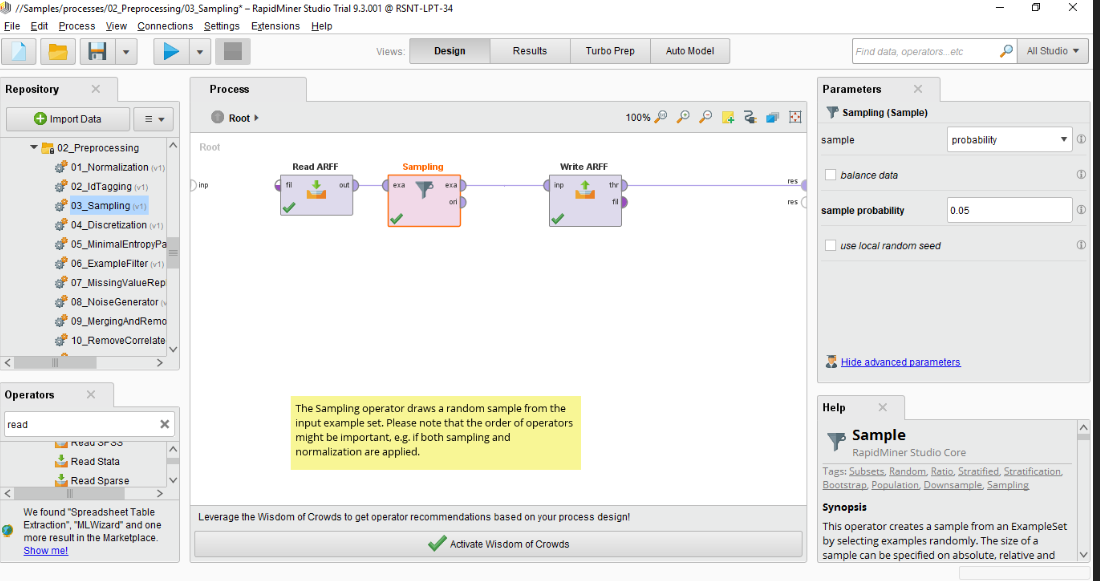
Se crearon las 2.432.508 encuestas para luego generar el archivo arff que sirve como input en Weka y Rapidminer Studio (software para crear redes y minería de datos)

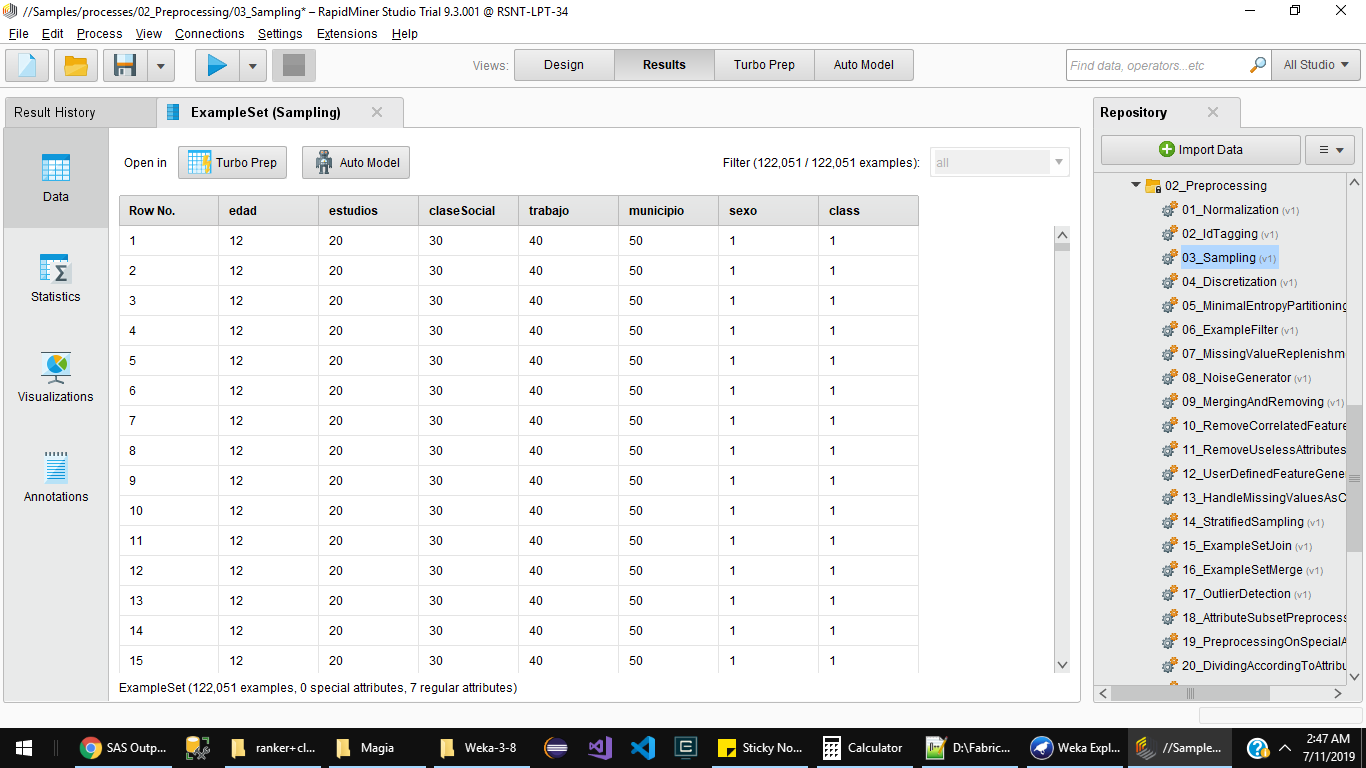
Del procesamiento anterior se logra generar el archivo data\_ulcera.arff con las 2.432.508 encuestas .

* + 1. Reducción de datos

Debido a la gran cantidad de datos obtenidos, el procesamiento sobrepasa los límites de un equipo con 8gb de RAM Core i7 8va gen. Por tales motivos se procede a realizar una reducción de datos.

Se utilizó como herramienta RapidMiner Studio que incluye una función llamada Sampling que obtiene de manera aleatoria muestras de una población, para este caso se aplicó el Sampling Probability para asignar una proporción del 5% del total de registros. Obteniendo así el archivo reduced.arff con ciento veinte mil registros aproximadamente.

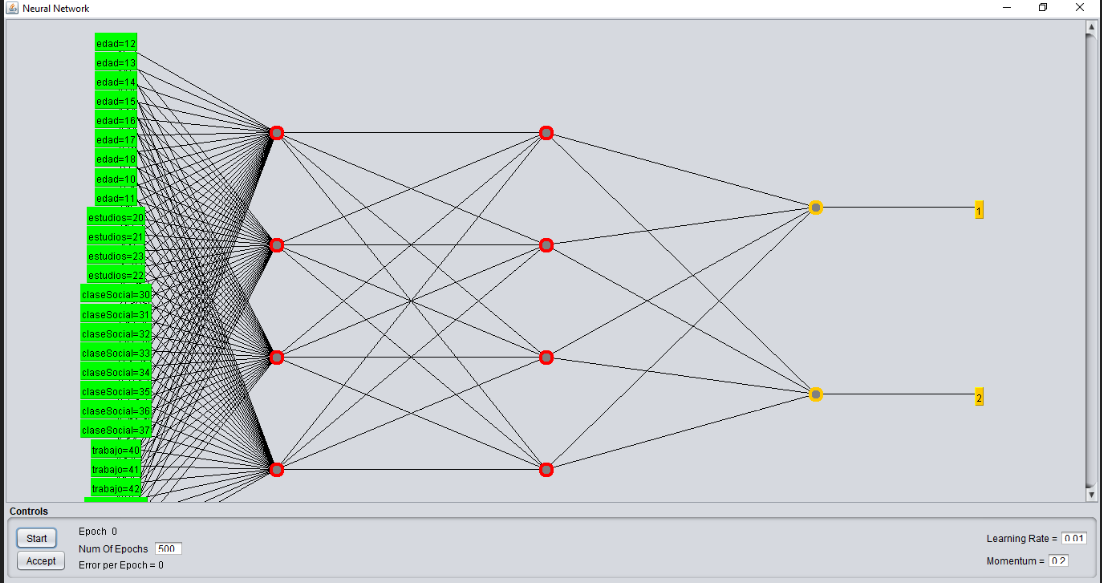




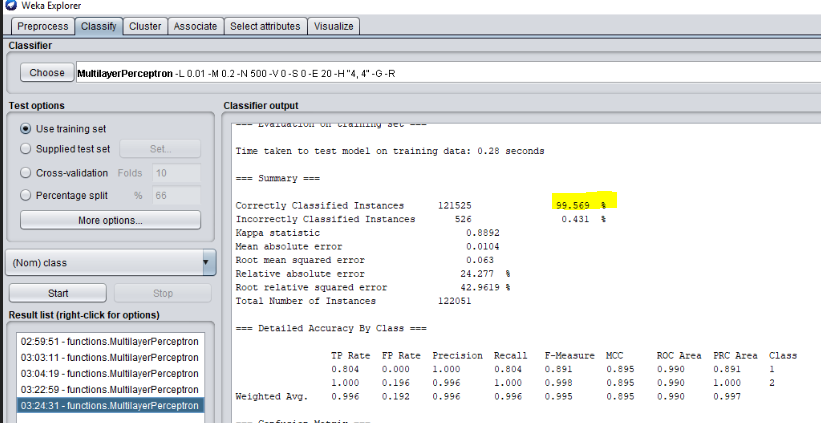
* + 1. Generación del modelo de red neurona multicapa perceptrón con pesos

Con los datos reducidos, se procede a cargar los datos en el Explorer de Weka.

Se selecciona en el menú Classify y elige la opción MultilayerPerceptron. Luego de múltiples pruebas se obtuvo la siguiente configuración de 2 capas ocultas con 4 neuronas cada una.



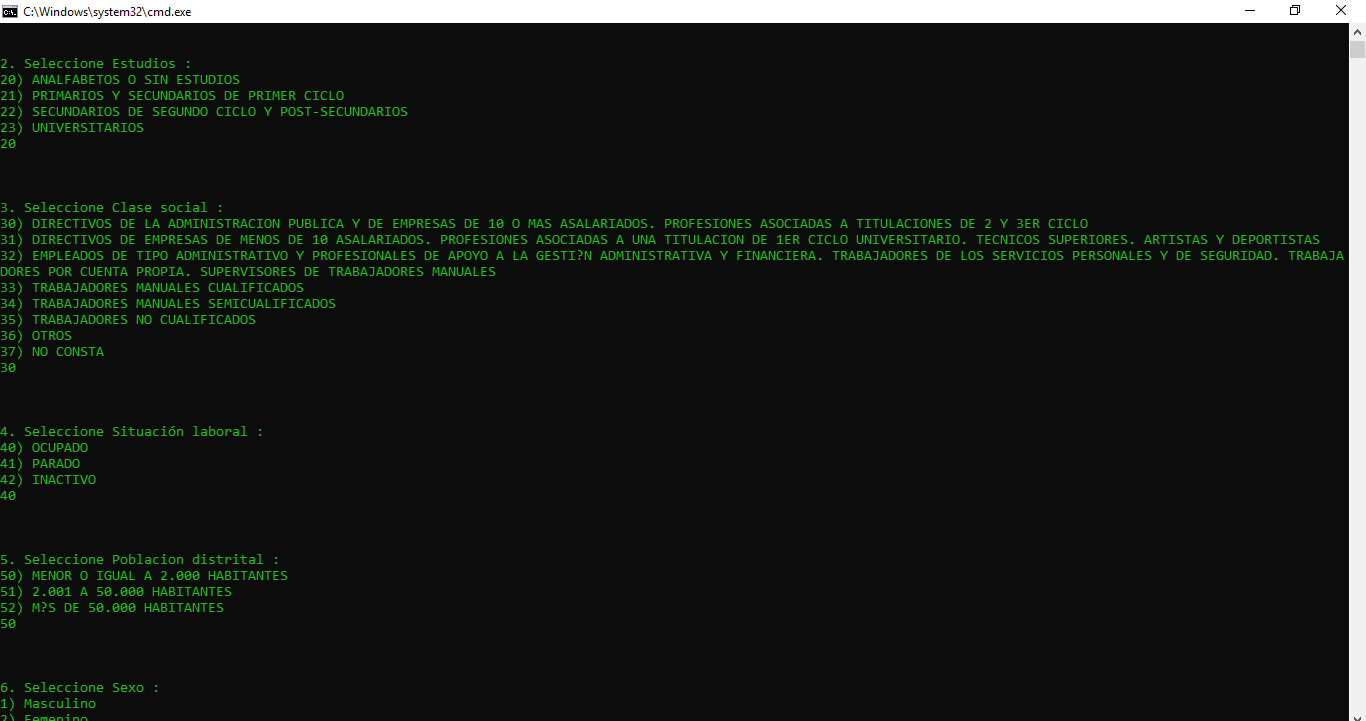
Del resultado obtuvimos una clasificación correcta del 99.56%, con lo cual se escogió este modelo.



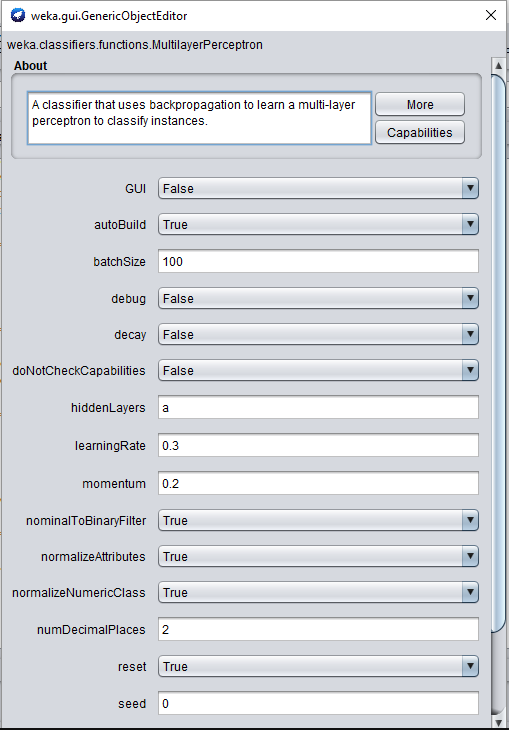
Se guarda el modelo con extensión model que almacenará los pesos calculados de cada variable.

* + 1. Predicción con muestra

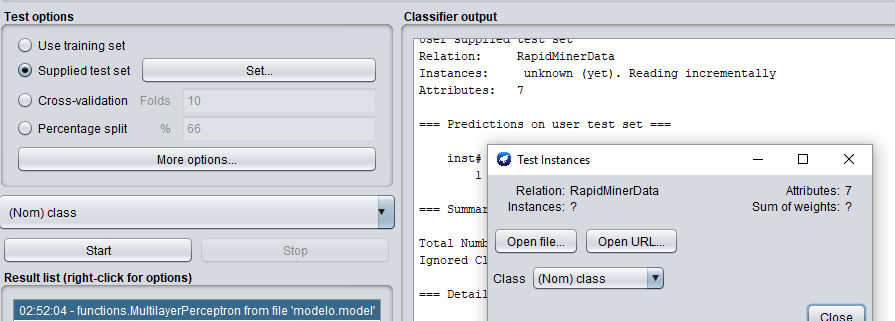
Se inicia con la captura de datos de la muestra a predecir, esto se realiza con el archivo start\_lectura\_datos.bat, lo cual generará un archivo muestra\_ulcera.arff que únicamente almacenará datos de un registro.



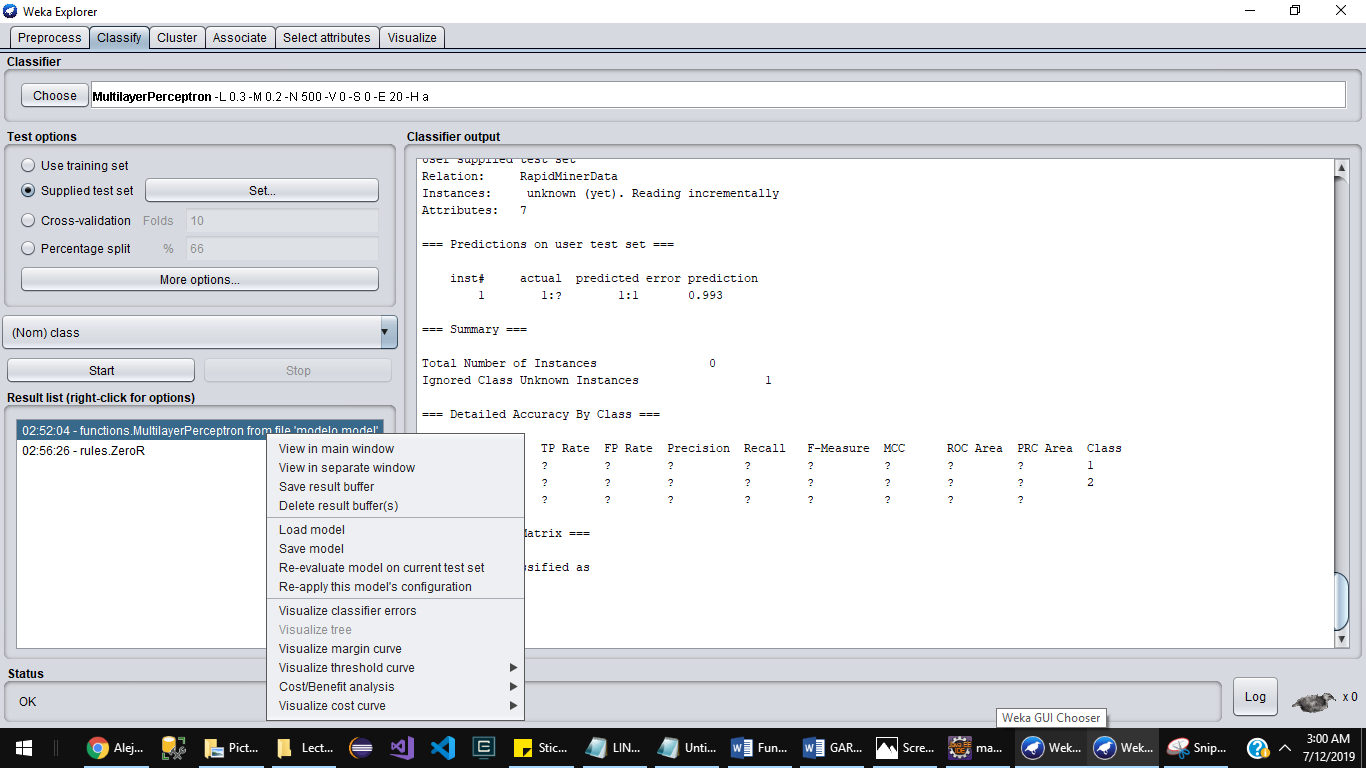
En el menú Classify se carga el modelo que contiene los pesos calculados, se configuran los datos de MultilayerPerceptron.



Luego se carga la muestra al seleccionar Supplied Test Set, activar la opción Plain Text en output Predictions en el apartado de More Options.



Para iniciar la predicción hacer click en el modelo y seleccionar la opción Re evaluate model on current test set.



La predicción aparecerá en el apartado predicted:



Donde

1 aparición de ulcera estomacal

2 ausencia de ulcera estomacal

# Bibliografía

**MACHINE LEARNING – DIEGO CALVO**

Recuperado

<http://www.diegocalvo.es/perceptron-multicapa/>

(Consulta 07/07/19)

**SALUD CASTILLA Y LEON**

Recuperado

<https://www.saludcastillayleon.es/sanidad/cm/gallery/ENCUESTA%20REGIONAL%20DE%20SALUD%202003/Castilla%20y%20Leon/Tabla%20I.3.6.HTM>

(Consulta 07/07/19)

**INTELIGENCIA RNA 50**

Recuperado

<http://rna.50webs.com/tutorial/RNA_intro.html>

(Consulta 07/07/19)

**FUNCIÓN SIGMOIDE**

Recuperado

<https://cesartroyasherdek.wordpress.com/2016/02/28/regresion-logistica-y-funcion-sigmoide/>

(Consulta 07/07/19)