CLASIFICACIÓN DE CLIENTES DE LA INDUSTRIA BANCARIA POR MÉTODOS ESTADÍSTICOS Y REDES NEURONALES ARTIFICIALES USANDO R

CLASSIFICATION OF CUSTOMERS BELONGING TO BANKING INDUSTRY BY STATISTICAL METHODS AND ARTIFICIAL NEURAL NETWORKS USING R

Carlos Eduardo Belman López
Tecnológico Nacional de México en Celaya, México
carlosbelman@gmail.com

José Alfredo Jiménez García
Tecnológico Nacional de México en Celaya, México
alfredo.jimenez@itcelaya.edu.mx

José Antonio Vázquez López
Tecnológico Nacional de México en Celaya, México
antonio.vazquez@itcelaya.edu.mx

Resumen

Esta investigación aborda el problema de la clasificación de clientes pertenecientes al sector bancario utilizando tres métodos diferentes de clasificación supervisada. Los métodos estadísticos utilizados fueron la regresión logística binaria y el análisis discriminante lineal. Adicionalmente, se utilizó un método de Inteligencia Artificial, como son, las redes neuronales artificiales. Se utilizó lenguaje R como herramienta para la construcción y validación de los métodos estadísticos y de Inteligencia Artificial. Como estrategia de validación se dividió el total de conjunto de observaciones en varios subconjuntos para entrenamiento y validación de los modelos ajustados, realizando en cada corrida una prueba de hipótesis que permite comparar las proporciones de clasificaciones correctas y determinar si existe evidencia estadística de que algún método es mejor. Finalmente se presentaron resultados y conclusiones enfocándose en la exactitud de la predicción de la clasificación, las pruebas de hipótesis y los tamaños de muestra utilizado durante el entrenamiento.

Palabras Clave: análisis discriminante, clasificación, lenguaje R, redes neuronales artificiales, regresión logística.

Abstract

This research addressed the problem of the accuracy of the customer classification of customer belonging to the banking sector using three different methods of supervised classification. The statistical methods used were binary logistic regression and linear discriminant analysis. Additionally, an Artificial Intelligence method was used, such as artificial neural networks. R language was used as a tool for the construction and testing of both statistical and Artificial Intelligence methods. As a validation strategy, the total set of observations was divided into several subsets for training and validation of the adjusted models (cross validation), performing in each run a hypothesis test that allows to compare the proportions of correct classifications and determine if there is statistical evidence that some method was better. Finally, results and conclusions were presented focusing on the prediction accuracy in the classification, the hypothesis tests and the sample size used during the training.

Keywords**:** *artificial neural networks, classification, discriminant analysis, logistic regression, R language.*

1. Introducción

El análisis de datos, en su mayoría de casos multivariante, es la parte de la estadística que estudia, analiza, e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás [Cuadras, 2014]. Uno de los problemas principales que se analiza en la estadística multivariante es el de la clasificación de elementos de una población.

El problema de clasificación puede plantearse de varias maneras y aparece en muchas áreas de la actividad humana, desde la diagnosis médica, en los sistemas de concesión de créditos, reconocimiento de obras de arte, etc. El planteamiento del problema es el siguiente: se dispone de un conjunto de elementos que pueden

venir de dos o más poblaciones distintas. En cada elemento se han observado un conjunto de variables aleatorias, cuyos valores son conocidos en las poblaciones de interés. Por lo tanto, el problema consiste en clasificar un nuevo elemento en una de las poblaciones, en base a sus valores de las variables aleatorias. Este enfoque de clasificación a menudo recibe el nombre de clasificación supervisada, para indicar que conocemos una muestra de elementos bien clasificados que sirve de pauta o modelo para predecir la clasificación de las siguientes observaciones o elementos [Peña, 2002].

En este documento se aborda el problema de clasificación de datos multivariante aplicado a la clasificación de clientes para de la industria bancaria utilizando un conjunto de datos multivariante de dominio público [Moro et al., 2014]. Este conjunto de datos es conocido como “Bank Marketing Dataset”, los datos se obtuvieron al realizar llamadas telefónicas de la institución bancaria al cliente. El objetivo de la clasificación era predecir si el cliente suscribirá (“si” o “no”) un depósito a plazo bancario (variable de respuesta). El conjunto de datos está conformado por 45211 observaciones y 16 variables de entrada y 1 de salida. Se utilizaron tres tipos de técnicas diferentes de clasificación supervisada, como son las redes neuronales artificiales (RNA), la regresión logística binaria (RL) y el análisis discriminante lineal (AD).

Cabe mencionar que el problema de la clasificación es un caso especial del desarrollo de análisis predictivos. Los análisis predictivos utilizan patrones que sucedieron en el pasado para predecir lo que sucederá en el futuro basándose en el supuesto de que lo que sucedió en el pasado sucederá de igual o s similar manera en el futuro [Xu & Duan, 2018]. Hoy en día, es de gran interés el desarrollo de análisis predictivos, es decir también, un procedimiento para estimar los valores de una variable de respuesta dado un conjunto de variables independientes. Existen ejemplos de la aplicación de análisis predictivos, en áreas tal como, química, manufactura, ventas [Ruelas Santoyo & Laguna González, 2014], economía [Brummelhuis & Luo, 2017]. A su vez, es un tema de mucho interés en áreas de reciente creación como el aprendizaje estadístico [Du & Swamy, 2014], aprendizaje automático [Lantz, 2013] y minería de datos [Williams, 2011]. Aunque anteriormente

ya han surgido estudios comparativos del rendimiento predictivo entre métodos estadísticos y de redes neuronales artificiales en aplicaciones orientadas a áreas como: el mercado de valores [Adebiyi et al., 2014], de la salud [Shi et al., 2012], educación [Gorr et al., 1994], en la industria avícola [Mehri, 2013], son pocos los estudios que consideran el problema de clasificación y además proporcionen una justificación estadística a sus resultados o reporten aplicaciones en la industria bancaria. Por tanto, la contribución de este documento se centra en la aplicación de modelos estadístico y de redes neuronales artificiales para resolver el problema de la clasificación de clientes de la industria bancaria usando lenguaje R, se utilizó, además, validaciones cruzadas y pruebas de hipótesis sobre las proporciones de clasificaciones correctas para tratar de determinar si algún método proporciona una mejor exactitud. En concreto, se utilizaron tres métodos diferentes de clasificación supervisada. Los métodos estadísticos utilizados fueron la regresión logística binaria y el análisis discriminante lineal. Adicionalmente, se utilizó un método de Inteligencia Artificial, como son, las redes neuronales artificiales, en específico, un Perceptrón Multicapa (MLP) con algoritmo de retropropagación (BP) como estrategia de aprendizaje. El lenguaje R, una herramienta flexible, de código abierto y cada vez más utilizada en el área de ciencia de datos, fue utilizado para entrenar y validar tanto modelos estadísticos como de Inteligencia Artificial.

Como estrategia de validación se dividió el total de conjunto de datos en varios subconjuntos para entrenamiento y validación de los modelos ajustados (validación cruzada), realizando en cada caso una prueba de hipótesis que compara las proporciones de clasificaciones correctas entre los métodos utilizados y permita determinar si existe evidencia estadística para concluir que algún método es mejor. Finalmente se presentaron resultados y conclusiones enfocándose en la exactitud de la predicción de la clasificación, las pruebas de hipótesis y los tamaños de muestra utilizado durante el entrenamiento.

Análisis predictivos

Los análisis predictivos utilizan patrones que sucedieron en el pasado para predecir lo que sucederá en el futuro, basándose en el supuesto de que lo que

sucedió en el pasado sucederá de igual o similar manera en el futuro [Xu & Duan, 2018].

En términos matemáticos, se puede predecir "y", usando $\hat{y} = f(X)$, donde f representa la estimación para la verdadera función F, y $\hat{y}$ representa el resultado estimado o predicho para "y". f es a menudo tratada como una caja negra, en el sentido que uno no suele preocuparse por la forma exacta de f, siempre y cuando arroje predicciones precisas para "y". La precisión de $\hat{y}$ como predicción de "y" depende de dos cantidades que se suelen llamar error reducible y error irreducible. En general, f no será una perfecta estimación para la verdadera F, por lo que esta inexactitud introducirá algún error. Este error es reducible porque se puede potencialmente mejorar la exactitud de f, seleccionando los métodos de aprendizaje más apropiados para estimar la real F, ya sea un método estadístico o uno de Inteligencia Artificial como las RNA [James et al., 2013].

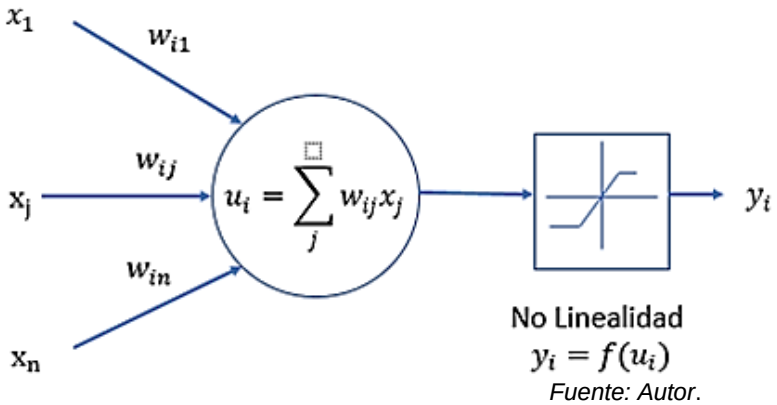
Sin embargo, aunque fuera posible obtener una perfecta estimación para F, tal que la respuesta estimada tomara la forma $\hat{y} = F(X)$, esta, aún tendrá algún error. Esto es porque "y", se encuentra también en función del error aleatorio (ϵ), que no puede ser predicho utilizando el vector de variables explicativas X. Por lo tanto, la variabilidad asociada con ϵ también afecta la precisión de la predicción. El error aleatorio también es conocido como error irreducible, porque no importa que tan bien se estime F, siempre existirá el error introducido por ϵ. Esto se debe a que ϵ puede contener variables que no fueron medidas y que son significantes para predecir "y", que, al no haber sido consideradas, entonces, F no las puede utilizar para predecir [James et al., 2013].

Redes Neuronales Artificiales (RNA)

Las RNA son en esencia estructuras formales de carácter matemático y estadístico con la propiedad del aprendizaje, es decir, la adquisición del conocimiento que en la mayoría de los casos es a partir de ejemplos. Este aprendizaje se produce mediante un estilo de computación denominado en paralelo que intenta simular algunas de las capacidades de nuestro cerebro, por esta razón se les define como redes neuronales artificiales para distinguirlas de los modelos

biológicos. Los 3 elementos clave de los sistemas biológicos que pretenden emular los artificiales son el procesamiento en paralelo, la memoria distribuida y la adaptabilidad [Torras P. & Monte, 2013]. Las RNA han mostrado aplicación en áreas como memoria asociativa, optimización, reconocimiento de patrones, predicción [Belman-López et al., 2018], clasificación [Belman-López et al., 2017] y toma de decisiones [Torras P. & Monte, 2013].

Los elementos que constituyen la estructura genérica de una RNA son: nodos, conjunto de entrada, pesos sinápticos, regla de propagación o función base, función de activación y función de salida. La figura 1 nos muestra el esquema básico de un RNA con sus elementos. El nodo suele definirse como el elemento básico de la red, el cuál recibe un conjunto de entradas (xj) del exterior o desde la salida de otros nodos. Cada entrada posee un peso específico (wij) asociado, que se aumentará o disminuirá en el proceso de aprendizaje. Cada nodo aplica una función base (ui) como por ejemplo la suma de las entradas ponderadas mediante los pesos. El valor de salida de la función base se transforma mediante una función de activación no lineal f(ui). Las funciones de activación más comunes son la función sigmoidal, gaussiana, escalón y tangente hiperbólica. El conjunto de variables de entrada (xj) y salida (yi) pueden ser tanto binarias como continuas, dependiendo del modelo [Martin del Brío & Sanz Molina, 2002].



Fuente: Autor.

Figura 1 La neurona y sus elementos.

Otros aspectos de relevancia para crear un modelo neuronal son: la arquitectura, la tipología de la red, el tipo de conexiones entre las neuronas, y algoritmo de

aprendizaje. La arquitectura se refiere a la forma de las conexiones entre las neuronas. La tipología se refiere al tipo de las unidades de procesamiento, existiendo neuronas visibles y neuronas ocultas. Por neurona visibles se entienden las de entrada y las de salida, en cambio las neuronas ocultas, poseen la función de capturar la representación interna de los datos. El tipo de conexiones que existen entre las neuronas pueden ser, feed-forward, donde las salidas de las neuronas se propagan en una sola dirección hacia adelante, recurrentes, cuyas conexiones se establecen en todas direcciones incluso con procesos de retroalimentación, las redes feedback, que permiten conexiones laterales y hacia atrás. Los algoritmos de aprendizaje son utilizados para estimar los pesos o parámetros de la red siendo el supervisado, no supervisado, reforzado e hibrido los más utilizados [Torras P. & Monte, 2013].

El MLP es una RNA formada por varias capas y que tiene como una de sus características principales que puede actuar como un aproximador universal de funciones mediante el algoritmo de aprendizaje BP, obligando a la red neuronal a contener al menos una capa oculta con suficientes unidades no lineales. El MLP es capaz de aproximar cualquier tipo de función o relación continua entre un grupo de variables de entrada y salida. Por eso considera al modelo MLP como una herramienta de propósito general, no lineal y flexible. El algoritmo de aprendizaje BP tiene la capacidad de dotar a la red de generalización de tal modo que el sistema obtiene una salida correcta para un conjunto de datos de entrada que no se habían usado antes [San Miguel Salas, 2016].

En muchas ocasiones el modelo MLP + aprendizaje BP suele denominarse red de retropropagación (BNN por sus siglas en ingles Backpropagation Neural Network), además varios grupos han postulado teoremas que demuestran matemáticamente que un MLP de una única capa oculta (pueden emplearse más capas ocultas, obteniéndose en ocasiones resultados más eficientes o una mejor generalización) con funciones de activación de tipo sigmoideo, puede aproximar hasta el nivel deseado cualquier función continua en un intervalo [Martin del Brío & Sanz Molina, 2002].

Análisis discriminante (AD)

Las funciones discriminantes son combinaciones lineales de variables que mejor separan en grupos. Se utiliza el termino grupo para representar ya sea a una población o a una muestra de la población. Existen dos objetivos principales en la separación por grupos: El primero es la descripción de la separación del grupo, es decir, creación de funciones lineales de n variables (funciones discriminantes) que son utilizadas para describir las diferencias entre 2 o más grupos. Las funciones discriminantes deben identificar la contribución relativa de las n variables a la separación en grupos y encontrar el plano optimo en que los puntos pueden ser proyectados para ilustrar de la mejor manera la configuración de los grupos.

El segundo es la predicción o asignación de nuevas observaciones en los grupos, en los cuales funciones de n variables lineales o cuadráticas (funciones de clasificación) son empleadas para asignar una unidad de muestra individual a uno de los grupos [Rencher, 2002]. El propósito del análisis discriminante lineal (AD) es encontrar la combinación lineal de las variables originales que de la mejor separación posible entre los grupos en nuestro conjunto de datos. El AD es también conocido como análisis discriminante canónico o simplemente "análisis discriminante". El máximo número de funciones discriminantes útiles que pueden separar un conjunto de datos es el mínimo entre el número de grupos menos uno y el número de variables usadas para la clasificación [Coghlan, 2017].

Regresión logística (RL)

En la vida real hay muchas situaciones en las cuales es evidente que la respuesta no sigue una distribución normal. Por ejemplo, existe una gran cantidad de aplicaciones en las que la respuesta es binaria (0 o 1), por lo que su naturaleza es de Bernoulli. En las ciencias sociales un problema podría ser el de desarrollar un modelo que prediga si un individuo representa riesgos para un crédito (0 o 1), también abundan las aplicaciones en las áreas de manufactura en que ciertos factores controlables influyen en el hecho de que un artículo fabricado este o no defectuoso. El enfoque más popular para modelar respuestas binarias es la técnica llamada regresión logística [Walpole et al., 2012]. Este mismo enfoque se puede

aplicar al problema de clasificación o discriminación entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasificación “y”, que tome el valor “0” cuando el elemento pertenece a la primera población, y el valor de “1” cuando pertenece a la segunda población. Entonces, la muestra consistirá en n elementos del tipo (yi, xi), donde yi es el valor en ese elemento de la variable binaria de clasificación y xi un vector de variables explicativas [Peña, 2002]. El modelo de RL puede generalizarse al caso donde se intenta explicar más de dos opciones discretas. Este tipo de RL, donde la variable dependiente tiene más de dos categorías es conocido como regresión logística multinomial.

2. Métodos

En este caso, se abordó el problema de clasificación de datos multivariable utilizando tres métodos diferentes de clasificación supervisada. La muestra de datos utilizada en esta investigación se encuentra conformada por dos clases, es decir, la variable de respuesta utilizada para predecir la clasificación puede tomar solo 2 valores posibles. Los métodos estadísticos utilizados fueron la regresión logística binaria y el análisis discriminante lineal (ADL). Adicionalmente, se utilizó un método de Inteligencia Artificial, como son, las redes neuronales artificiales (RNA). El método propuesto, fue un método por etapas, que se enumeran a continuación:

- Etapa 1. Selección de la muestra y tamaño de muestra para entrenamiento
- Etapa 2. Construcción de los modelos estadísticos multivariable.
- Etapa 3. Clasificación utilizando los métodos estadísticos.
- Etapa 4. Normalizar los datos y entrenamiento de la RNA.
- Etapa 5. Clasificación utilizando RNA.
- Etapa 6. Resumen de resultados y pruebas de hipótesis.

3. Resultados

En esta investigación se utilizó R [R Core Team, 2017] como herramienta para la construcción y aplicación de los métodos estadísticos regresión logística binaria y análisis discriminante lineal, así como para el entrenamiento y ejecución de las RNA tipo MLP + BP.

Etapa 1 Selección de la muestra y tamaños de muestra para entrenamiento

Para este caso, se utilizó un conjunto de datos multivariante de dominio público [Moro et al., 2014]. Este conjunto de datos es conocido como “Bank Marketing Dataset”, los datos están relacionados con las campañas de marketing directo (llamadas telefónicas) de una institución bancaria portuguesa. El objetivo de la clasificación era predecir si el cliente suscribirá (“si” o “no”) un depósito a plazo bancario (variable y). El conjunto de datos está conformado por 45211 observaciones y 16 variables de entrada y 1 de salida. Debajo se listan las variables que conforman la fuente de datos (datos del cliente del banco):

- Edad (numérico).
- Tipo de trabajo ("admin = 1", "desconocido = 2", "desempleado = 12", "gestión = 6", "empleada doméstica = 5", "emprendedor = 4", "estudiante = 10", "de cuello azul = 3", "negocio propio = 8", "jubilado = 7", "técnico = 11", "servicios = 9").
- Estado civil ("divorciado = 1", “desconocido = 2”, "casado = 3", "soltero = 4"; nota: "divorciado" significa divorciado o viudo).
- Educación ("secundaria = 1", "desconocida = 2", "primaria = 3", "terciaria = 4").
- ¿Tiene crédito? ("sí = 1", "no = 0").
- Saldo medio anual, en euros (numérico).
- Vivienda: ¿tiene préstamo de vivienda? Préstamo: ¿tiene préstamo personal? (binario: "sí = 1", "no = 0")
- Préstamo: ¿tiene préstamo personal? ("sí = 1", "no = 0").
- Tipo de comunicación de contacto ("desconocido"=2, "teléfono"=1, "celular"=3).
- Día de contacto del mes (numérico).
- Mes del año de contacto ("jan =1", "feb = 2", "mar = 3", ..., "nov = 11", "dec = 12").
- Duración del último contacto, en segundos (numérico).
- Número de contactos realizados durante esta campaña y para este cliente (numérico, incluye el último contacto).

- Número de días que pasaron después de que se contactó por última vez con el cliente de una campaña anterior.
- Número de contactos realizados antes de esta campaña y para este cliente (numérico).
- Resultado de la campaña de marketing anterior ("otro = 1", "desconocido = 2", "fracaso = 3", "éxito = 4").
- Variable de respuesta (y), (no utilizada durante el aprendizaje solo para validación) ¿El cliente ha suscrito un depósito a plazo? ("sí = 1", "no = 0").

Se utilizaron los siguientes tamaños de muestra para el entrenamiento: 8, 15, 25, 35. 50, 70, 100, 250, 500, 1000, 2500, 5000, 10000 y 30140 (2/3 del conjunto total). El código en R que se utilizó para la lectura de los datos, fue:

```
# Lectura de las 45211 observaciones de la fuente de datos
> datos<-read.csv("bank.csv")
#Se eligió 30140 observaciones para entrenamiento.
> train<-sample(1:45211, 30140)
# Lectura de las observaciones para entrenamiento.
> training<-datos[train,]
# Lectura de las observaciones restantes para validación.
> test<-datos[-train,]
```

Ejemplo de cómo fueron generadas las muestras para entrenamiento y validación: En este caso particular, se seleccionaron 2/3 equivalente a 30140 observaciones para entrenar los modelos y 1/3 restante para validar y probar los resultados. Se hizo lo mismo para el resto de los tamaños de muestra seleccionados.

Etapa 2 Construcción de los modelos estadísticos multivariable

- **Análisis Discriminante.** El primer propósito del análisis discriminante lineal (ADL), es encontrar la combinación lineal de las variables originales que de la mejor separación posible entre los grupos dentro del conjunto de datos.

  El ADL, es también conocido como análisis discriminante canónico o simplemente "análisis discriminante". El segundo propósito del análisis

discriminante es la predicción o asignación de nuevas observaciones en alguno de los grupos. Se llevó a cabo el ADL utilizando la función 'lda' del paquete "MASS" [Venables & Ripley, 2002] en R, para los diferentes tamaños de muestra para entrenamiento seleccionados, como se muestra:

```
> library(MASS)
> z <- lda(y ~ ., datos, subset = train)
```

- **Regresión logística binaria.** El segundo método estadístico con que se abordó el problema de la clasificación supervisada fue la regresión logística binomial. Se utiliza una muestra de elementos bien clasificados, para generar un modelo que prediga la clasificación de nuevas observaciones en uno de los subgrupos de interés. Se generó el modelo de regresión logística, para los diferentes tamaños de muestra de entrenamiento seleccionados utilizando la función 'glm' como se muestra más adelante. Esta función se encuentra dentro de los paquetes básicos ya incluidos al instalar R por lo que no es necesario importar ningún paquete adicional:

```
> mod<-glm(y~., datos[train, ], family = "binomial")
```

Etapa 3 Clasificación utilizando los métodos estadísticos

Se utilizaron los métodos estadísticos multivariantes construidos en la etapa 2, para la clasificación de las nuevas observaciones en algunos de los subgrupos:

- **Análisis discriminante**. Se utilizó la función "predict" para clasificar las nuevas observaciones de correspondientes a las diferentes muestras para validación como se muestra y se tabula el número de clasificaciones de acuerdo con el ADL contra la clasificación real de la muestra:

```
> preds <- predict(z, datos[-train, ])$class
> actual <-datos[-train,]$y
> xtabs(~actual + preds)
```

- **Regresión logística multinomial**. Se utilizó la función "predict", para clasificar las observaciones de las diferentes muestras para validación utilizando también el modelo por regresión logística binomial de la siguiente manera:

```
> output<-predict(mod,datos[-train, ], type="response")
> xtabs(~actual + round(output))
```

Etapa 4 Normalizar los datos y entrenamiento de la RNA

Las RNA no son fáciles de entrenar, por lo cual, antes de iniciar el entrenamiento de la RNA es necesario realizar algún tipo de preparación de los datos. Es recomendado normalizar los datos antes de entrenar la RNA.

Evitar la normalización puede conducir a resultados inútiles o a un proceso de entrenamiento muy difícil con problemas, por ejemplo, que la mayoría de las veces el algoritmo no convergerá antes del número de iteraciones máximas permitidas. Se pueden elegir diferentes métodos para escalar los datos, por ejemplo, z-normalización o escala min-máx. Se eligió utilizar el método min-máx y escalar los datos en el intervalo [0,1]. Por lo general, la escala en los intervalos [0,1] o [-1,1] tiende a dar mejores resultados [Alice, 2015]. Por lo tanto, el primer paso antes de entrenar la RNA fue normalizar los datos por el método min-máx, como se muestra a continuación:

```
> maxs <- apply(datos, 2, max)
> mins <- apply(datos, 2, min)
> scaled <- as.data.frame(scale(datos, center = mins, scale = maxs - mins))
> train_ <- scaled[train,]
> test_ <- scaled[-train,]
```

Selección de parámetros para la RNA

Para utilizar una RNA de tipo MLP, se tienen que elegir diversos parámetros como son: el número de neuronas en las capas de entrada y de salida, el número de capas ocultas y numero de neuronas por capa oculta, el algoritmo de aprendizaje, la función de activación, función para cálculo del error y función en la capa de salida:

- **Neuronas en la capa de entrada**. Este parámetro se determinó con el número de variables explicativas o independientes que forman los datos de entrenamiento.

- **Neuronas en la capa de salida**. El número de neuronas en la capa de salida fue igual al número de variables de respuesta, es decir, igual al número de variables que se desea predecir.
- **Numero de capas ocultas.** Ya estudios han demostrado que una capa oculta es suficiente para la gran mayoría de problemas [Martin del Brío & Sanz Molina, 2002].
- **Numero de neuronas en la capa oculta.** No hay una regla fija en cuanto al número neuronas a utilizar en la capa oculta, aunque existen varias reglas empíricas para determinar este número. Una de estas reglas con buena aceptación dice que el número de neuronas en la capa oculta es igual al promedio de neuronas en la capa de entrada y de salida [Naved, 2016]. En esta investigación se determinó utilizar esta regla y probar con este número, además de 1 o 2 neuronas adicionales, para finalmente utilizar como numero de neuronas en la capa oculta el número que proporcionó mejores resultados.

Arquitectura seleccionada para la RNA

Finalmente, para esta RNA de tipo Perceptrón multicapa se utilizó 1 capa oculta, con configuración 16:9:1, que indica que la capa de entrada tiene 16 neuronas, 1 capa oculta con 9 neuronas y una neurona en la capa de salida. El resto de los parámetros utilizados fueron: algoritmo de retropropagación como estrategia de aprendizaje, función logística como función de activación, la suma de cuadrados del error (SCE), como función para cálculo del error, y salida lineal en falso, es decir, si se aplicó la función de activación a la neurona de salida.

Se llevó a cabo el entrenamiento de la RNA utilizando la función ‘neuralnet’ del paquete “neuralnet” [Fritsch & Frauke, 2016] de R, para los diferentes tamaños de muestra de entrenamiento seleccionados como se muestra a continuación:

```
> library(neuralnet)
> ann<-neuralnet(y~age+job+marital+education+default+balance+housing+
loan+ contact+day+month+duration+campaign+pdays+previous+poutcome,
train_, stepmax = 1e+08, hidden = 9, act.fct="logistic", threshold =
```

```
0.01,linear.output = FALSE, lifesign="full")
```

Etapa 5 Clasificación utilizando RNA

Se realizó la clasificación de las observaciones para las diferentes muestras para validación, utilizando la RNA entrenada en la etapa 4. Se utilizó la función 'compute' del paquete "neuralnet" [Fritsch & Frauke, 2016] en cada caso, como se muestra a continuación:

```
> pr.nn <- compute(ann, test_[ , c("age", "job", "X3ital", "education",
"default", "balance",
"housing","loan","contact","day","month","duration","campaign","pdays","prev
ious","poutcome")])
```

La red emitió una salida en forma normalizada, por lo que tenemos que escalar de nuevo para poder hacer una comparación significativa:

```
> classRNA <- pr.nn$net.result*(max(datos$y)-min(datos$y)) + min(datos$y)
> real <- (test_$y) * (max(datos$y)-min(datos$y)) + min(datos$y)
> classRNA<-round(classRNA)
```

Finalmente se tabuló el número de clasificaciones de las nuevas observaciones, por la predicción de la RNA, contra la clasificación real de la muestra:

```
> xtabs(~real + classRNA)
```

Etapa 6 Resumen de resultados y pruebas de hipótesis

Como paso final, se presentó el porcentaje de clasificaciones correctas, es decir si el cliente suscribirá ("si" o "no") un depósito a plazo bancario, utilizando los modelos entrenados y las observaciones para validación. La tabla 1, muestra el resumen de corridas para los métodos utilizados, utilizando los siguientes tamaños de muestra para el entrenamiento: 8, 15, 25, 35. 50, 70, 100, 250, 500, 1000, 2500, 5000, 10000 y 30140, en verde se resalta el mejor resultado de cada caso.

En cada corrida se realizó, además, una prueba de hipótesis de comparación de las proporciones de clasificaciones correctas en R.

Tabla 1 Resumen de resultados mostrando en verde el mejor resultado.

Muestra Entrenamiento	Muestra Validación	AD (%)	RL (%)	RNA (%)	p-value
8	45203	0	54.97	82.14	2.20E-16
15	45196	0	49.1	83.75	2.20E-16
25	45186	0	73.06	83.87	2.20E-16
35	45176	0	73.82	85.93	2.20E-16
50	45161	0	85.95	85.5	0.05231
70	45141	85.5	80.74	85.91	2.20E-16
100	45111	88.52	83.46	87.244	2.20E-16
250	44961	88.78	87.633	86.11	2.20E-16
500	44711	89.33	89.25	86.04	2.20E-16
1000	44211	89.46	89.21	87.89	1.58E-12
2500	42711	87.89	89.5	87.48	2.20E-16
5000	40211	89.68	89.66	88.17	2.21E-14
10000	35211	89.77	89.57	88.57	0.6197
30140	15071	89.6	89.5	90.71	0.0002819

Esta prueba permite o no rechazar la hipótesis que las proporciones sean iguales y determinar que existe evidencia estadística que alguna proporción sea mejor. Debajo se muestra un ejemplo, para un cierto tamaño de muestra:

```
> x<-c(13614,13614,13798)
> n<-c(15211,15211,15211)
> prop.test(x,n)
        3-sample test for equality of proportions without continuity correction
data:  x out of n
X-squared = 16.348, df = 2, p-value = 0.0002819
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.8950102 0.8950102 0.9071067
```

Un análisis del tamaño de muestra

La figura 2, muestra de forma gráfica la exactitud de la clasificación de los 3 métodos propuestos a diferentes tamaños de muestra para entrenamiento. Se observó que, de los 3 métodos propuestos, las RNA necesitan menos datos de

entrenamiento y más rápido convergen hacia una muy buena exactitud. Además, se pudo observar que, una vez que cualquiera de los 3 métodos propuestos alcanzó una buena exactitud en un determinado tamaño de muestra, la exactitud de la clasificación se incrementó muy poco al ir aumentando el tamaño de muestra para entrenamiento. Por ejemplo, la eficiencia cuando se utilizó 250 datos para entrenamientos solo varia un 2% o 3%, respecto a cuando se utilizó alrededor de 30000 datos para entrenamiento.

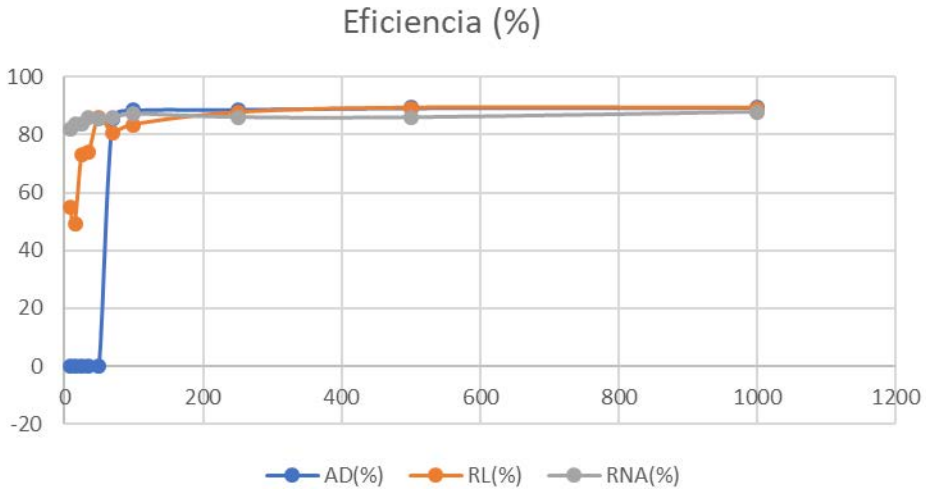


Figura 2 Exactitud de la clasificación de los 3 métodos propuestos para diferentes tamaños de muestra.

4. Discusión

Las RNA son estructuras que tienen un comportamiento de carácter matemático y estadístico con la propiedad del aprendizaje, además de ser herramientas flexibles que ya han mostrado su aplicación como aproximadores de funciones. Hoy en día, las RNA han despertado el interés como herramientas de predicción y clasificación en áreas novedosas, como la minería de datos y el aprendizaje automático. Para estos fines, el lenguaje R es una herramienta flexible, de código abierto y cada vez más utilizada en la ciencia de datos, dotada con la capacidad de entrenar y ejecutar modelos estadísticos y de Inteligencia Artificial. R es un lenguaje sencillo de utilizar, que proporciona muy buen rendimiento sin consumir excesivos recursos computacionales.

En base a los resultados experimentales se concluyó que las tres técnicas de clasificación supervisada utilizadas en esta investigación proporcionaron muy buenos resultados al predecir la clasificación de nuevas observaciones dado los modelos entrenados con los elementos bien clasificados. Es decir, para lograr clasificar si los clientes pertenecientes al sector bancario subscribirían o no un depósito a plazo bancario. Es importante destacar la precisión de la clasificación de nuevas observaciones no utilizadas durante el entrenamiento de los modelos. Las pruebas de hipótesis realizadas nos permitieron comparar las proporciones de clasificaciones correctas y rechazar la hipótesis que las proporciones sean iguales y demostrar que existía evidencia estadística para determinar que una clasificación fue mejor. De manera exploratoria, los resultados parecen indicar que para tamaños de muestra de entrenamiento pequeños y grandes las Redes Neuronales Artificiales son la mejor opción, y para tamaños de muestra de entrenamiento de tamaño mediano el análisis discriminante es mejor opción. Además, el análisis de los diferentes tamaños de muestra de entrenamiento sobre los métodos propuestos mostró que las RNA necesitan menos datos de entrenamiento y convergen más rápido hacia una buena exactitud. También, se pudo observar que una vez que cualquiera de los métodos propuestos alcanzó una buena exactitud en un determinado tamaño de muestra de entrenamiento, la exactitud en la clasificación se incrementó poco al aumentar el tamaño de muestra para entrenamiento. Por ejemplo, la eficiencia cuando se utilizó 250 datos para entrenamientos solo aumento alrededor del 3%, respecto a cuándo se utilizó más de 30 000 datos para entrenamiento.

Los investigadores interesados en continuar esta investigación podrían concentrarse en aplicar los modelos analizados a grandes volúmenes de datos (Big Data) y observar la exactitud en los resultados y el tiempo de entrenamiento para los modelos. Además, existen otras áreas donde se puede resolver el problema de clasificación como en modelos de manufactura, mantenimientos predictivos, etc. También, para el caso de RNA, se puede considerar utilizar algún otro modelo de RNA, como función de base radial (RBF), máquinas de vector de soporte (SVM) o redes neuronales de aprendizaje profundo.

5. Revisores, recepción y aceptación de artículo

Recepción artículo: 30/abril/2019 **Aceptación artículo:** 21/mayo/2019

Revisor 1:

Nombre:	José Emmanuel Franco Barrón
Institución:	GKN Driveline
Cédula Profesional:	10100414
Área de conocimiento:	Ingeniería Industrial y Mecatrónica
Correo electrónico:	harryfranco111@hotmail.com

Revisor 2:

Nombre:	Alejandro Estrada
Institución:	Bodega Aurrera
Cédula Profesional:	11493651
Área de conocimiento:	Redes Complejas, Manufactura, Ingeniería Industrial
Correo electrónico:	al.estrada.o@outlook.com

6. Bibliografía y Referencias

[1] Adebiyi, A., Adewumi, A., & Ayo, C. (2014). Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. Journal of Applied Mathematics, 1-6. doi:http://dx.doi.org/10.1155/2014/614342

[2] Alice, M. (23 de Septiembre de 2015). R-Bloggers: https://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/.

[3] Brummelhuis, R., & Luo, Z. (2017). Cds rate construction methods by Machine Learning Techniques. Data Science Central, 1-51: https://www.datasciencecentral.com/profiles/blogs/choice-of-k-in-k-fold-cross-validation-for-classification-in.

[4] Coghlan, A. (2017). A Little Book of R for Multivariate Analysis. Cambridge: Creative Commons.

[5] Cuadras, C. (2014). Nuevos métodos de análisis multivariante. Barcelona: CMC Editions.

[6] Du, K.-L., & Swamy, M. (2014). Neural Networks and Statistical Learning. London: Springer.

[7] Fritsch, S., & Frauke, G. (2016). neuralnet: Training of Neural Networks: https://CRAN.R-project.org/package=neuralnet.

[8] Gorr, W., Nagin, D., & Szczypula, J. (1994). Comparative study of artificial neural network and statistical models for predicting student grade point averages. International Journal of Forecasting, 17-34. doi:https://doi.org/10.1016/0169-2070(94)90046-9

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer.

[10] Lantz, B. (2013). Machine Learning with R. Birmingham: Packt Publishing.

[11] Martin del Brío, B., & Sanz Molina, A. (2002). Redes neuronales y sistemas difusos. (2a. ed.). Madrid, España: Alfaomega & RA-MA.

[12] Mehri, M. (2013). A comparison of neural network models, fuzzy logic, and multiple linear regression for prediction of hatchability. Poultry Science Association Inc., 1138 - 1142.

[13] Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. Elsevier (62), 22-31: https://archive.ics.uci.edu/ml/datasets/bank+marketing#.

[14] Naved, I. (26/diciembre/2016): https://iqbalnaved.wordpress.com/2016/12/26/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feed forward-neural-network/.

[15] Peña, D. (2002). Análisis de datos multivariantes. Madrid: McGraw-Hill / Interamericana de España, SA.

[16] R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Obtenido de The R Project for Statistical Computing: https://www.R-project.org/.

[17] Rencher, A. (2002). Methods of multivariate analysis (2da ed.). Canadá: John Wiley & Sons, Inc.

[18] Ruelas Santoyo, E., & Laguna González, J. (2014). Predictive comparison based in neural network versus statistical methods to forecast sales. Ingeniería Industrial. Actualidad y Nuevas Tendencias, 91-105.

[19] San Miguel Salas, J. (2016). Desarrollo con MATLAB de una red neuronal para estimar la demanda de energía eléctrica (Tesis de Maestría). Valladolid, España: Universidad de Valladolid.

[20] Shi, H.-Y., Lee, K.-T., Lee, H.-H., Ho, W.-H., Sun, D.-P., Wang, J.-J., & Chiu, C.-C. (2012). Comparison of Artificial Neural Network and Logistic Regression Models for Predicting In-Hospital Mortality after Primary Liver Cancer Surgery. PLoS ONE, 1-6. doi:10.1371/journal.pone.0035781

[21] Torras P., S., & Monte, E. (2013). Modelos neuronales aplicados en economía. Barcelona, España: Addlink.

[22] Venables, W., & Ripley, B. (2002). Modern Applied Statistics with S. (Cuarta ed.). New York: Springer: http://www.stats.ox.ac.uk/pub/MASS4.

[23] Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probabilidad y estadística para ingeniería y ciencias. Ciudad de México: PEARSON.

[24] Williams, G. (2011). Data Mining with Rattle and R. New York: Springer.

[25] Xu, L., & Duan, L. (2018). Big data for cyber physical systems in industry 4.0: a survey. Enterprise Information Systems, 1-23. doi:10.1080/17517575.2018.1442934.