SIMULACIÓN EN LA IDENTIFICACIÓN DE MIRTACEAS BASADO EN REDES NEURONALES ARTIFICIALES SUPERVISADAS

SONIA I. MARIÑO
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste.
simarinio@yahoo.com

Fecha recepción: Enero 2019 - Fecha aprobación: Abril 2019

RESUMEN

Modelar y simular el conocimiento de los especialistas es un área de constante interés científico-tecnológico. En dominios botánicos se aplican tecnologías de la Inteligencia Artificial para apoyar la identificación de especies vegetales, como una estrategia para afrontar complejos procesos decisorios. La Minería de Datos abarca una diversidad de técnicas entre ellas las basadas en tecnologías de la Inteligencia Artificial, como son las Redes Neuronales Artificiales. En el trabajo se proponen y evalúan algunas soluciones inferenciales sustentadas en modelos conexionistas supervisados, como una alternativa de apoyo a la toma de decisiones en la identificación taxonómica. Finalmente, se justifican los resultados obtenidos en las simulaciones y se proponen futuras líneas de trabajo.

PALABRAS CLAVE: Inteligencia Artificial - Minería de Datos - Modelos conexionistas - Modelos supervisados – Simulación - Botánica.

ABSTRACT

Modeling and simulating the knowledge of specialists, is a constant area of scientific and technological interest. In botanical domains technologies are applied in order to support the identification of plant species, as a strategy to deal with complex decision-making processes. The Data Mining covers a range of techniques including those based on the Artificial Intelligence technologies, such as Artificial Neural Networks. In the paper some connectionism supervised models are proposed and evaluated, as an alternative to decision-making support in the taxonomic identification. Finally, the results of simulations are justified and some future lines are proposed.

KEYWORDS: Artificial Intelligence - Data Mining - connectionist models -supervised models – Simulation - Botany.

1. INTRODUCCIÓN

La Inteligencia Artificial (IA), disciplina que surgió en el Simposio de Dartmouth en 1956, se originó en un intento de representar y simular los procesos cognitivos de los sujetos. Es así como mediante sus paradigmas, se vale de modelos, métodos y herramientas que contextualizados a una abstracción de un problema complejo, proponen un planteamiento y su correspondiente solución.

En este trabajo, se presentan algunos antecedentes, el método y los resultados referentes al diseño y evaluación de modelos de Redes Neuronales Artificiales supervisadas y su simulación para apoyar la toma de decisiones en un dominio de la Botánica.

1.1 Redes Neuronales Artificiales

Entre las áreas de la Inteligencia Artificial una de ellas aborda la representación y manipulación del conocimiento de los especialistas (Castillo, Cobo Ortega, Gutierrez Llorente y Pruneda Gonzalez, 1998; Rusell y Norving, 2004). Uno de los métodos de la IA se denomina Descubrimiento de Conocimiento, *Knowledge Discovery in Database* o KDD.

El Descubrimiento de Conocimiento, que data del año 1996, como método define las etapas principales de un proyecto de explotación de información (Frawley, Shapiro y Matheus 1992; Matheus, Chan y Piatetsky-Shapiro, 1993). Establece a la Minería de Datos (MD) como la etapa del proceso en la cual se realiza la extracción de patrones a partir de los datos brindados por el especialista (Moine, Haedo y Gordillo, 2011a, 2011b).

En Moine Haedo y Gordillo, (2011a, 2011b) se especifica que actualmente “el término KDD y Minería de Datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento”. Por otra parte, Martins, Pesado y García-Martínez, (2014) entienden a la Minería de Datos como un subproceso de un proceso general destinado a obtener patrones de conocimiento

La Minería de Datos y el Descubrimiento de Conocimiento contribuyen a la toma de decisiones tácticas y estratégicas, automatizando la generación de conocimiento (Valcárcel Asencios, 2004). La literatura muestra su vasta aplicación en actividades de I+D+i.

Entre los métodos comprendidos en los procesos de KDD y pertenecientes al paradigma conexionista se mencionan las Redes Neuronales Artificiales (RNA). Inspiradas en la Biología, estos modelos conexionistas se propusieron como un sistema de procesamiento de las representaciones distribuidas y no localizadas.

Los inicios de las RNA se remontan a la década de 1940, cuando se difundió un primer modelo computacional: el perceptrón propuesto por Warren McCulloch y Walter Pitts en 1943, constaba de

una capa de neuronas binarias de entrada y una de salida, sólo utilizaba la función de activación de la neurona umbral y no incorporó la ponderación de los pesos de las diferentes entradas. En la década de 1980, resurge el conexionismo con los primeros modelos de RNA multicapas.

La arquitectura de un modelo conexionista o RNA describe el número de capas que la componen, el modo de conexiones entre las neuronas intracapa e intercapa, el método o algoritmo de aprendizaje para obtener los pesos de sus conexiones, y la función de activación que transformará el estado actual de cada neurona en una señal de salida (Hilera y Martinez, 1995; Russell y Norvin, 2004). El problema abordado por el modelo de RNA determina la arquitectura para obtener una solución.

El conocimiento del especialista en torno al dominio se almacena en las instancias de evidencias que conforman los registros del archivo de datos a procesar. Para cada instancia del archivo se puede o no explicitar el valor de la variable objetivo según se trate de aprendizaje supervisado o no supervisado respectivamente.

Una RNA presenta dos modos de operación: entrenamiento y mapeo. En el modo de operación entrenamiento, los modelos de RNA aprenden del conjunto de datos en un proceso que se repite n veces hasta minimizar el error global del modelo y lograr un esquema representativo del conjunto de datos que describen el conocimiento del objeto de estudio. En el modo de operación mapeo, se espera que la RNA brinde una solución correcta ante un nuevo caso en estudio a partir del conocimiento aprendido en el modo de operación entrenamiento.

1.2 Algunos antecedentes del uso de tecnologías informáticas en Botánica

La Biología requiere la identificación de taxones para desarrollar otros estudios. En Mariño y Dematteis (2014) se expone un relevamiento aplicado al mencionado dominio de conocimiento, que permitió determinar los diversos modelos computacionales diseñados para simular el proceso de identificación de entidades.

Esta indagación se actualizó considerando artículos de revistas científicas disponibles en la Web y publicadas en el periodo 1975–2017. Se eligieron sesenta y dos con la finalidad de valorar la aplicación de tecnologías inteligentes artificiales en el área de las ciencias biológicas. Se examinaron los títulos y resúmenes y se determinó que solo un 3,23% incluyó el uso de RNA como se expone en Mariño y Tressens (2001).

2. MÉTODO

En esta sección se describe el método seguido para el logro de los resultados. Cabe aclarar que dado que el proyecto se aborda desde la Minería de Datos, a continuación se explicitan las fases consideradas siguiendo el método CRISP-DM (Chapman *et al.*, 2000) y algunas especificaciones computacionales mencionadas para el dominio como las expuestas en Mariño y Alfonzo (2016) y Mariño y Alfonzo (2017).

2.1 Análisis y comprensión del negocio

En esta fase se entendió el dominio de aplicación, específicamente se decidió simular con métodos inteligentes conexionistas supervisados la identificación de especies vegetales pertenecientes a la familia *Myrtaceae* del NE Argentino. En Botánica el proceso de identificación de taxones se asocia al proceso de clasificación computacional, por ello se optó por un algoritmo de comprendido en la Minería de Datos

2.2 Comprensión de los datos

La fase Comprensión de los datos implicó entenderlos contemplando los objetivos del negocio. El conjunto de datos utilizado para el entrenamiento y comprobación de los modelos se obtuvo de un herbario que posee una nutrida colección de Mirtáceas y un número importante de esos ejemplares han sido identificados por especialistas en la familia, lo que hace que las soluciones computacionales puedan ser corroboradas por comparación con testigos fidedignos o con el experto.

Se incorporaron la mayor cantidad de caracteres posibles para facilitar la identificación de ejemplares en los que, por no poseer frutos, se desconoce el tipo de embrión, carácter de gran importancia para la identificación en esta familia.

La variable objetivo asume distintos valores representativos de las 31 especies de Mirtáceas elegidas en el estudio (Mariño, 2001). El especialista seleccionó *a priori* las variables evidenciales relevantes a partir de su experticia.

2.3 Selección y preparación de los datos

En esta fase se desarrollaron las siguientes actividades:

- Preparación de datos. Esta fase involucró aquellas actividades para construir el conjunto de datos viable de procesar con Weka (manual ver García Morate, 2008), la herramienta de MD seleccionada. Las tareas se pueden aplicar múltiples veces y sin un orden pre-establecido. Incluyen extracción, transformación y carga, proceso conocido como ETL. En Mariño (2001) se describió la conformación del archivo de evidencias denominado Matriz Taxonómica Ampliada o MTA. Este archivo contiene los 424 casos representativos de estas especies del NE Argentino representando el conocimiento del especialista. Lo expuesto es similar al planteamiento de Moret

Bonillo (2014:45), que referencia el problema de la interpretación diferencial como uno de los métodos disponibles para formalizar el razonamiento categórico. Es decir, el proceso consiste en construir todas las posibles combinaciones a partir de las evidencias posibles y todas las hipótesis o interpretaciones disponibles del dominio, que incorpora las heurísticas del experto dotando de mayor valor al conjunto de datos.

- Reducción del número de evidencias. *A priori*, se consideran como variables evidenciales los caracteres seleccionados por el especialista de conocimiento (Mariño, 2001). Para disminuir la dimensionalidad de las variables se aplicó como evaluador de atributos CfsSubsetEval y como método de búsqueda BestFirst – opciones disponibles en la herramienta de MD- . Este proceso redujo el número de variables de entrada o evidencias (Mariño y Tressens, 2001).Las 31 variables evidenciales consideradas en este trabajo se indican en la TABLA 1 del Anexo.
- Estimación de estadísticos sobre los atributos. Registrados los datos, desde la herramienta de MD se procedió a reconocer los atributos y computar algunas estadísticas básicas sobre los mismos. Dado que el conjunto de datos seleccionados son atributos continuos/numéricos, se visualizaron valores mínimo, máximo, media, desviación estándar, entre otros.

2.4 Modelado

El modelado implicó la selección y aplicación de distintas técnicas de modelado y análisis de los datos para el cumplimiento del objetivo del negocio.

- Se optó por modelos de RNA supervisadas.
- Se seleccionó como algoritmo de aprendizaje la Regla Delta Generalizada o Backpropagation. Este algoritmo generaliza la Regla Delta sobre Redes Neuronales de múltiples capas (MLP) y funciones de transferencia no lineales y diferenciables (Freeman y Skapura, 1993; Hilera y Martínez, 1995; Rusell y Norvin, 2004).

  La Regla Delta generalizada (propuesta por Widrow en 1960) se basa en la búsqueda del mínimo de la función de error mediante el descenso por el gradiente de la misma. Un problema es la posibilidad de detectarse un mínimo local de la función error, donde el gradiente vale cero, y por lo tanto los pesos no se modificarán sin embargo el error cometido por la red es significativo.

  El parámetro época. Una época es una iteración realizada sobre el conjunto total de datos de entrenamiento para mejorar los pesos de la RNA.

  El parámetro razón o tasa de aprendizaje, η. Este valor incide en el comportamiento de los algoritmos de aprendizaje. Un valor mínimo incidirá en un mínimo cambio de la magnitud de

los pesos sinápticos y requerirá mayor tiempo de convergencia. Así, un valor grande afectará el comportamiento del algoritmo y dificultaráen la localización del mínimo de la función de error.

El parámetro α o constante de momento. Este valor es responsable del incremento en valor de los pesos w_{ji}en relación con el incremento (positivo o negativo) previo del mismo peso. Se aplica para controlar la velocidad de acercamiento al mínimo error, acelerándola o disminuyéndola en un intento de localizar un mínimo global representativo de todos los datos.

- Se diseñaron diversos modelos, se modificaron en cada uno de ellos ciertos parámetros relacionados con el algoritmo de aprendizaje. Se utilizó como función de activación la sigmoidea. Se establecieron los pesos iniciales con valores aleatorios entre -0.5 y 0.5 (Fausett, 1994).
- Se determinaron las métricas o medidas de calidad. Para evaluar la efectividad en la identificación de especies, se optó por las siguientes métricas de precisión: Kappa Statistic, el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE).

Estadístico Kappa. Es una medidade concordancia entre las categorías pronosticadas por el clasificador (Pr(a)) y las categoríasobservadas. Considera las posibles concordancias debidas al azar (Pr(e)) (Ecuación 1). La valoración del índice Kappa está dada según:

- Si el valor es 1: Concordancia perfecta.
- Si el valor es 0: Concordancia debida al azar.

-Si el valor es negativo: Concordancia menor que la que cabría esperar por azar.

$$K = \frac{Pr(a) - Pr(\mathrm{e})}{1 - \mathrm{Pr(e)}}$$

Ecuación 1. Índice Kappa

RMSEo Root Mean Squared Error. Mide las diferencias entre los valores calculados por un modelo o un estimador y los valores observados. Es una medida de precisión, permite comparar diferentes errores de predicción de un mismo conjunto de datos, dado que depende de la escala de la muestra (Ecuación 2). También se puede denominar RMSD o Root Mean Squared Desviation.

En la Ecuación 2, y_i es el valor observado. Representa la salida de la red para el vector de entrada a_t es el valor de salida deseado para el vector de entrada ap y n es el número de residuales.

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left\| \boldsymbol{y_i} - \widehat{\boldsymbol{y_i}} \right\|^2}{\boldsymbol{n}}}$$

Ecuación 2. Medida de calidad Error Cuadrático Medio

MAE o Mean Absolute Error: Se define como el Error Absoluto Medio a la diferencia entre el valor medio obtenido y el calculado en esa media. El promedio de error absoluto, es la suma de los errores absolutos de clasificación en cada uno de los elementos llevados a promedio. El clasificador que proporcione una mayor cifra (superior a 0.1) define un error de clasificación alto, por lo cual se debe considerar sobre aquellos que brinden una cifra menor (Ecuación 3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\|f_i - y_i\| = \frac{1}{n}\sum_{i=1}^{n}\|e_i\|$$

Ecuación 3. Medida de calidad Error Absoluto Medio

2.5. Evaluación

Los modelos construidos se evaluaron para determinar su utilidad en relación a los objetivos previstos. Se procedió a:

- Elegir los parámetros para la selección de los modelos. Se establecieron los siguientes valores para elegir los modelos: Clasificación correcta > 90 % instancias; Clasificación incorrecta < 10 % instancias; MAE <0.1 ideal; Kappa statistic>0.79, >0.9 ideal; RMSE < 0.3, <1 ideal. El modelo que cumpla con estas especificaciones, se considera representativo del dominio para simular al especialista ante nuevos casos de identificación.
- Seleccionar el modo mapeo o comprobación de los modelos. Modo mapeo en las RNA. Entrenado el conjunto de datos, se procedió a su verificación utilizando la técnica denominada Validación Cruzada disponible en la herramienta de Minería de Datos elegida. Esta técnica calcula el porcentaje de aciertos esperados haciendo n evaluaciones, establecidas en el parámetro hojas, pliegues o*folds*. Se dividen las instancias del archivo de evidencias en tantos pliegues como indica el parámetro, y en cada evaluación se toman las instancias de cada una de ellas como datos de testeo, y el resto se consideran datos de entrenamiento para construir el modelo. Los errores calculados son el promedio de todas las ejecuciones. El número de evaluaciones en cada modelo se estableció en 10. Si el número de instancias es elevado, esta opción es suficiente para estimar con precisión las prestaciones del algoritmo clasificador en el dominio.
- Simular los modelos construidos y sistematizar los resultados proporcionados por la herramienta de MD.
- Interpretar y valorar los resultados, confrontando las distintas medidas de calidad obtenidas para cada uno de los modelos en los modos de operación: entrenamiento y mapeo. Esta actividad permitió corroborar el comportamiento de los modelos en relación

con la actuación del especialista del dominio. Para seleccionar el modelo más representativo, la primera elección se basó en el índice Kappa; ante valores similares la evaluación se centró en la métrica MAE y si esta media brinda valores iguales se elegirá el modelo con menor valor de RMSE.

2.6. Despliegue o Implementación

Esta fase implica el despliegue de la solución tecnológica. Los modelos entrenados y validados se pueden utilizar para inferir nuevos conocimientos desde la herramienta de Minería de Datos seleccionada o desde una interfaz de usuario que incorpore este modelo de aprendizaje.

3. RESULTADOS

En esta sección se exponen los resultados distinguiendo la aplicación de un proceso KDD utilizando una herramienta de Minería de Datos para construir y simular modelos de RNA de apoyo a la toma de decisiones en un dominio de la Botánica.

Se optó por un método de clasificación, dado que el objetivo es determinar la especie botánica a partir de un conjunto de variables evidenciales seleccionadas por el especialista en el dominio de conocimiento. Además, se eligieron técnicas de aprendizaje supervisado para contrastar el aprendizaje automático, es decir, comparar el valor computado que puede asumir la variable objetivo con los valores establecidos por el especialista del dominio.

Como se indicó en trabajos previos (Mariño, 2001; Mariño y Alfonzo, 2016; Mariño y Alfonzo, 2017), la propuesta se validó en la identificación de especies de Mirtáceas del Nordeste Argentino, taxones con presencia tanto en el Microsistema Iberá (Corrientes) como en el Predio Guaraní (Misiones). Numerosas de estas especies se caracterizan por su valor económico, como frutales u ornamentales, y en la medicina popular.

Se definió la arquitectura de los modelos de RNA para apoyar la identificación de taxones (FIGURA 1)**.** Específicamente, la red consta de 3 capas (entrada-oculta-salida), con 31 neuronas en la capa de entrada, cada una correspondea los atributos o variables evidenciales relevantes seleccionadas aplicando métodos automáticos: 10 neuronas en la capa oculta y 31 neuronas en la capa de salida representativas de las posibles especies botánicas a identificar (TABLA 2 del Anexo).

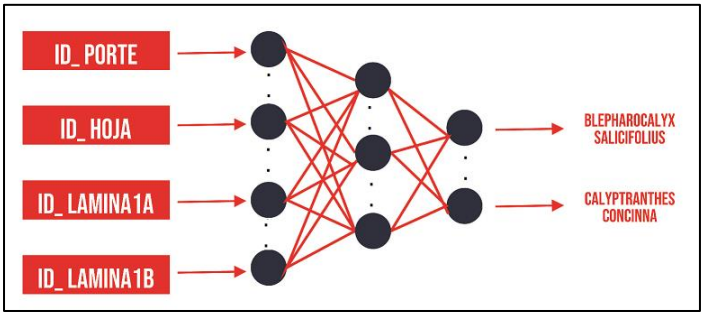

FIGURA 1. Modelo simplificado de Red Neuronal Artificial
(Fuente: elaboración propia)

Definida la arquitectura se construyeron y procesaron los modelos. En los experimentos realizados:

- Se normalizaron los atributos o valores de los nodos de entrada de la RNA,
- Se indició como modo de operación mapeo la evaluación sobre los datos de entrenamiento.
- Se desactivó el parámetro degradación de la velocidad de aprendizaje indicado como dL en las TABLA 1 y TABLA 2.
- Se establecieron -con la finalidad de realizar estudios comparativos- como valores constantes en las configuraciones de los modelos los siguientes parámetros:
  Cantidad de neuronas: 10
  Número de observaciones representativas del dominio: 424
- Se definieron los valores de los parámetros: época, momento y tasa de aprendizaje en los modelos diseñados. Estas cuantías se indican en las columnas **E**, **Lr** y **Mo**de las TABLA 1 y TABLA 2 respectivamente.

En el modo entrenamiento, se procedió experimentalmente, para establecer la mejor configuración de los modelos de RNA, dado que se carece de un procedimiento que la determine *a priori*. Lo expuesto permite afirmar que la simulación se constituyó en una herramienta facilitadora para descubrir conocimiento a través del modelo más representativo del dominio considerando un conjunto de datos o instancias elegidas. En este modo, los modelos de RNA aprenden siendo este conocimiento registrado en el peso de las conexiones entre las neuronas.

La TABLA 1 sintetiza los resultados derivados de la aplicación del método de aprendizaje de retropropagación en el modo de operación entrenamiento. A efectos de comprobar la fiabilidad de los modelos logrados con el aprendizaje automático (modo mapeo), se

utilizó la técnica de validación cruzada (TABLA 2) estableciendo el número de pliegues, opción recomendada en operaciones de entrenamiento y mapeo si se dispone de un mínimo conjunto de datos.

Cabe aclarar que en el entrenamiento y mapeo de los modelos**,** al analizar los resultados se establece que aquel nodo de salida con un mayor valor estimado se correspondería con el nombre de la especie botánica estimado por el modelo de RNA, considerando los datos evidenciales característicos del nuevo caso en proceso de identificación.

Para la elección de los modelos más representativos se evaluaron las metricas establecidas en las columnas: el porcentaje de predicciones correctas (Corr), el índice Kappa (Ka), y los errores MAE y RMSE, valores expuestos en las TABLA 1 y TABLA 2.

Para argumentar la elección del mejor modelo, se optó por aquél con mayor valor de índice Kappa, e inversamente proporcional al valor de la raíz del error cuadrático medio (RMSE) o Error Absoluto Medio (MAE). Siendo el valor de este indice similar en los distintos modelos, la decisión final se sustentó en el valor proporcionado por el Error Absoluto Medio.

TABLA 1 Modo entrenamiento. Modelos diseñados.

Modelos	**Parámetros de la RNA**				**Medidas de evaluación**			
	dL	**Lr**	**Mo**	**E**	**Corr**	**Ka**	**MAE**	**RMSE**
Mod1	N	0,3	0,2	1000	423	0.9974	0.0017	0.0106
Mod2	N	0,2	0,2	1000	423	0.9974	0.002	0.0131
Mod3	N	0,3	0,2	1500	423	0.9974	0.0013	0.0098
Mod4	N	0,2	0,2	500	423	0.9974	0.0023	0.0137

TABLA 2 Modo mapeo. Método de validación cruzada

Modelos	**Parámetros de la RNA**				**Medidas de evaluación**			
	dL	**Lr**	**Mo**	**E**	**Corr**	**Ka**	**MAE**	**RMSE**
Mod1	N	0,3	0,2	1000	421	0.9922	0.002	0.0182
Mod2	N	0,2	0,2	1000	422	0.9948	0.0023	0.0172
Mod3	N	0,3	0,2	1500	421	0.9922	0.0017	0.0179
Mod4	N	0,2	0,2	500	420	0.9897	0.0029	0.0214

Los experimentos permitieron corroborar que habría una alta probabilidad de comportamiento correcto de los modelos si se aplicaran a nuevos casos que requieren la identificación de la especie de pertenencia. Dado que la RNA, como caja negra brinda una medida global, se puede afirmar que las distintas configuraciones lograron generalizarel conocimiento del dominio.

Es decir, el modelo clasifica como positivas las instancias positivas. Lo expuesto se visualiza en el alto valor del índice Kappa, inversamente proporcional a los errores estimados por la medida RMSE

y MAE, revelando el buen desempeño del algoritmo para diferenciar las distintas especies botánicas del dominio a partir del conjunto de datos procesados.

El estudio de los valores expuestos en la TABLA 2 permite afirmar que el modelo identificado como Mod3 propone un mejor reconocimiento, dado que brinda menor valor en la medida MAE durante el proceso de entrenamiento y validación (FIGURA 2). Donde se indica como MAE-E los Errores Absolutos Medio obtenidos en los modelos en el modo entrenamiento y MAE-V los Errores Absolutos Medio de los modelos en el modo mapeo.

Un análisis de la matriz de confusión —como métrica que brinda información pormenorizada— y contrastada con el especialista del dominio permitió verificar los resultados del proceso automáticamente obtenidos en el modo mapeo, como fase previa a la implementación de este modelo de apoyo a la toma de decisiones en la comunidad de potenciales usuarios.

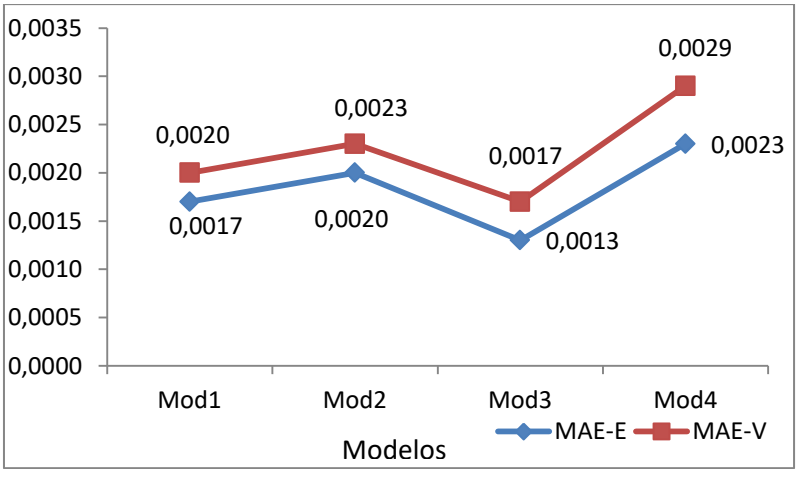

FIGURA 2. Comparativa de los errores absoluto medio obtenidos en los modelos entrenados (MAE-E) y modelos validados (MAE-V)
(Fuente: elaboración propia)

4. CONCLUSIONES

Se expusieron modelos conexionistas de Redes Neuronales Artificiales supervisadas diseñados para simular la identificación de especies de *Myrtaceae*, construidos enel modo entrenamiento y validados en el modo mapeo.

Se experimentó con los modelos y se analizaron los resultados proporcionados utilizando el algoritmo de aprendizaje automático

seleccionado. Las pruebas permitieron verificar el buen comportamiento de los modelos conexionistas diseñados para simular a los especialistas botánicos en estos procesos decisorios.

Finalmente, se eligió el mejor representante del dominio de conocimiento con miras a su transferencia a la potencial comunidad de usuarios. Las simulaciones expuestas muestran valores similares en el índice Kappa mientras que la métrica Error Absoluto Medio (MAE) presenta un menor valor en el modelo Mod3, y por ello se podría suponer un mejor comportamiento a partir de las pruebas realizadas.

Dada la relevancia de la Ciencia de los Datos en la sociedad del conocimiento, se continuarán las indagaciones en torno a la identificación de taxones mediada por otros métodos de aprendizaje automático comprendidos en la Inteligencia Artificial.

5. REFERENCIAS

CASTILLO, E.; COBO ORTEGA, A.; GUTIERREZ LLORENTE, J. M. y PRUNEDA GONZALEZ, R E. (1998): INTRODUCCIÓN A LAS REDES FUNCIONALES CON APLICACIONES. Ed. Paraninfo.

CHAPMAN, P.; CLINTON, J.; KEBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. y WIRTH, R. (2000): CRISP-DM 1.0 STEP BY STEP BIGUIDE. Editedby SPSS. Documento en línea. Disponible en: http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf

FAUSETT L. (1994): FUNDAMENTALS OF NEURAL NETWORKS. ARCHITECTURES, ALGORITHMS AND APPLICATIONS, Ed. Prentice Hall. U.S.A.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G. y MATHEUS, C. J. (1992): "KNOWLEDGE DISCOVERY IN DATABASES: AN OVERVIEW", AI Magazine 13(3): 57-70.Documento en línea. Disponible en:http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929

FREEMAN J. y SKAPURA D. (1993): "REDES NEURONALES. ALGORITMOS, APLICACIONES Y TÉCNICAS DE PROGRAMACIÓN", Addison-Wesley/Díaz de Santos, España.

GARCÍA MORATE, D. (2008): "MANUAL DE WEKA". Disponible en: http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf

HILERA J. y MARTÍNEZ V. (1995): "REDES NEURONALES ARTIFICIALES. FUNDAMENTOS, MODELOS Y APLICACIONES", Addison-Wesley Iberoamericana, México.

MARIÑO, S. I. (2001). "CONSTRUCCIÓN DE UN GENERADOR DE SISTEMAS EXPERTOS PROBABILÍSTICOS. UNA APLICACIÓN A LA IDENTIFICACIÓN DE ESPECIES VEGETALES", Tesis de Maestría en

Informática y Computación, Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste.

MARIÑO, S. I. y DEMATTEIS, M. (2014): "REVISIÓN DE SOLUCIONES DE TECNOLÓGICAS INTELIGENTES EN BIOLOGÍA", *Telematique*, Revista Electrónica de Estudios Telemáticos, 13(1): 30-50.

MARIÑO, S. I. y ALFONZO, P. L. (2016): "SIMULACIÓN DEL RAZONAMIENTO EN EL PROCESO DE IDENTIFICACIÓN BOTÁNICA BASADO EN REDES BAYESIANAS", *Investigación Operativa*, 24(39): 55-72.

MARIÑO, S. I. y ALFONZO, P. L. (2017): "UNA PROPUESTA DE INTEGRACIÓN DE INTERFACES DE USUARIO EN MÉTODOS DE MINERÍA DE DATOS", *Investigación Operativa*, 25 (41): 42-53.

MARIÑO, S. I. y TRESSENS, S. G. (2001): "ARTIFICIAL NEURAL NETWORKS APPLICATION IN THE IDENTIFICATION OF THREE SPECIES OF ROLLINIA (ANNONACEAE)", *Ann. Bot. Fennici,* 38: 215–224.

MARTINS, S.; PESADO, P. y GARCÍA-MARTÍNEZ, R. (2014): "PROPUESTA DE MODELO DE PROCESOS PARA UNA INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN: MOPROPEI". RevistaLatinoamericana de Ingeniería de Software, 2(5): 313-332,

MATHEUS J. C.; CHAN, P. K. y PIATETSKY-SHAPIRO, G. (1993): "SYSTEMS FOR KOWLEDGE DISCOVERY IN DATABASE", IEEE, TKDE, special issue on Learning& Discovery in Knowledge-Based Databases, 1-16,Documento en línea. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.169&rep=rep1&type=pdf

MOINE, J. M.; HAEDO, A. N. y GORDILLO, S. (2011a): "ESTUDIO COMPARATIVO DE METODOLOGÍAS PARA MINERÍA DE DATOS", XIII Workshop de Investigadores en Ciencias de la Computación p. 278-281

MOINE, J. M.; HAEDO, A. N. y GORDILLO, S.(2011b): "ANÁLISIS COMPARATIVO DE METODOLOGÍAS PARA LA GESTIÓN DE PROYECTOS DE MINERÍA DE DATOS", VIII Workshop Bases de Datos y Minería de Datos (WBDDM), CACIC 2011 - XVII Congreso Argentino de Ciencias de la Computación, octubre 2011 : p. 931-938

MORET BONILLO, V. (2014): "REPRESENTACIÓN DEL CONOCIMIENTO Y RAZONAMIENTO AUTOMÁTICO". España: Departamento de Computación. Facultad de Informática. Universidad de A Coruña.

RUSSELL, S. y NORVIG, P. (2004): "INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO", Ed. Prentice–Hall Hispanoamericana.

VALCÁRCEL ASENCIOS V. (2004): "DATA MINING Y EL DESCUBRIMIENTO DE CONOCIMIENTO". Revista de la Facultad de Ingeniería Industrial. (7):2: p. 83-86.

WEKA, Disponible en: http://www.cs.waikato.ac.nz/ml/weka/

Agradecimiento
Se agradece a la Lic. Sara G. Tressens, quien proporcionó los datos del dominio botánico para la elaboración del artículo y corroboró los resultados.

ANEXO

TABLA 1. Variables evidenciales y sus valores seleccionados
(Fuente: Especialista del Dominio)

Variables evidencia	Valores posibles	Identificador
Porte	1 - árbol 2 - arbusto 3 - árbol o arbusto	porte1a
Hojas	1 - lámina cartácea 2 - lámina no cartácea	lamina1a (consistencia)
	1 - lámina coriácea 2 - lámina no coriácea	lamina1b
	1 - lámina subcartácea 2 - lámina no subcartácea	lamina1c
	1 - lámina subcoriácea 2 - lámina no subcoriácea	lamina1e
	1 - lámina ovada 2 - lámina no ovada	lamina2b
	1 - lámina obovada 2 - lámina no obovada	lamina2c
	1 - lámina oblonga 2 - lámina no oblonga	lamina2d
	1 - epifilo glabro 2 - epifilo no glabro 3 - epifilo glabro o no glabro	lam_epifilo1
	1 - hipofilo glabro 2 - hipofilo no glabro 3 - hipofilo glabro o no glabro	lam_hipofilo1
	1 – ápiceagudo 2 - ápice no agudo	lam_apice1
	1 – ápiceobtuso 2 - ápice no obtuso	lam_apice2
	1 - ápice acuminado 2 - ápice no acuminado	lam_apice3
	1 - base obtusa 2 - base no obtusa	lam_base3
	1 - base cuneada 2 - base no cuneada	lam_base4

Variables evidencia	Valores posibles	Identificador
	1 - base atenuada 2 - base no atenuada	lam_base5
	1 - cáliz tetrámero 2 - cáliz pentámero 3 - cáliz tetrámero o pentámero o hexámero	caliz1
	1- cáliz siempre no persistente en el fruto 2 - cáliz siempre persistente en el fruto 3 - cáliz siempre o no siempre persistente en el fruto	caliz4
	1 - sépalos siempre reflejos después de la antesis 2 - sépalos siempre no reflejos después de la antesis 3 - sépalos siempre o no siempre reflejos después de la antesis	sepalos1
	1 - sépalos soldados en el botón floral 2 - sépalos no soldados en el botón floral	sepalos2
	1 - cara externa del sépalo siempre glabra 2 - cara externa del sépalo siempre no glabra 3 - cara externa del sépalo siempre o no siempre glabra	sepalos3
	1 - cara interna del sépalo siempre glabra 2 - cara interna del sépalo siempre no glabra 2 - cara interna del sépalo siempre o no siempre glabra	sepalos4
	1 - hipanto glabro 2 - hipanto no glabro 3 - hipanto glabro y no glabro	id-hipanto1
	1 - hipanto prolongado sobre el ovario 2 - hipanto no prolongado	id-hipanto2
	1 - inflorescencia contraída 2 - inflorescencia laxa	inflor2
	1 - inflorescencia pluriflora (más de	inflor4a

Variables evidencia	Valores posibles	Identificador
	3) 2 - inflorescencia no pluriflora	
	1 - inflorescenciatriflora 2 - inflorescencia no triflora	inflor4b
	1 - inflorescencia uniflora 2 - inflorescencia no uniflora	inflor4c
	1 - inflorescencia biflora 2 - inflorescencia no biflora	inflor4d
	1 - dicasio 2 - no dicasio	inflor7a

TABLA 2 Posibles valores de la hipótesis o variable objetivo
(Fuente: EDC en Mariño, 2001)

Hipótesis (Especies a identificar en el estudio)
Blepharocalyxsalicifolius (B. tweediei)
Calyptranthesconcinna
Campomanesiaguaviroba
Campomanesiaguazumifolia
Campomanesiaxanthocarpavar. xanthocarpa
Eugenia burkartiana
Eugenia hyemalisvar. marginata
Eugenia involucrata
Eugenia moraviana
Eugenia pitanga
Eugenia pyriformisvar. pyriformis
Eugenia pyriformisvar. uvalha
Eugenia repanda
Eugenia sp.
Eugenia uniflora
Eugenia uruguayensis
Hexachlamysedulis
Hexachlamyshumilis
Myrcialaruotteanavar. australis
Myrciaselloi (M. ramulosa)
Myrciasp.
Myrcianthescisplatensis
Myrcianthespungens

Hipótesis (Especies a identificar en el estudio)
Myrciariatenella
Myrrhiniumatropurpureumvar. octandrum
Pliniarivularis
Psidiumguajava
Psidiumguineense
Psidiumincanum
Psidiumkennedyanum
Psidiumnutans