ISSN 1870-4069

Redes neuronales artificiales y árboles de decisión para la clasificación con datos categóricos

C. Guadalupe Origel-Rivas, Eréndira Rendón-Lara,
Itzel María Abundez-Barrera, Roberto Alejo-Eleuterio

Tecnológico Nacional de México,
Instituto Tecnológico de Toluca,
México

{corigelr, erendon, iabundezb,
ralejoe}@toluca.tecnm.mx

Resumen. Las redes neuronales artificiales (RNA) y los árboles de decisión (AD) se han utilizado como clasificadores en distintas áreas, como en el sector salud, financiero, manufactura, social entre otros. Sin embargo, cada uno de estos clasificadores tiene sus ventajas dependiendo del tipo de datos (numérico, categórico o mezclados) de entrada. Por lo tanto, los árboles de decisión trabajan con valores tanto numéricos como categóricos. Teóricamente ambos clasificadores pueden trabajar con todos los tipos de datos, sin embargo, en la práctica no parecer ser así. Por esta razón en este trabajo se realiza un análisis de las ventajas y desventajas de estos clasificadores para valores categóricos. Se utilizaron bases de datos descritas en atributos de tipo categórico, que fueron codificadas, para la RNA. De acuerdo con las pruebas realizadas la eficiencia de clasificación cuando se utilizó la RNA es inferior al clasificador C4.5.

Palabras clave: Redes neuronales artificiales, árboles de decisión, codificación.

Artificial Neural Networks and Decision Trees for Classification with Categorical Data

Abstract. Artificial neural networks (ANN) and decision trees (DT) have been used as classifiers in different fields, such as in the health, financial, manufacturing, and social sectors, among others. However, each classifier has its advantages depending on the type of data (numeric, categorical or mixed) input. Nevertheless, decision trees work with both numerical and categorical values. Theoretically, both classifiers can work with all types of data, however, in practice they do not appear to be so. For this reason: in this work we analyze advantages and disadvantages of these classifiers for categorical values. Databases described in categorical attributes, which were encoded, were used for the ANN. According to the tests carried out, the classification efficiency when the ANN was used is lower than the C4.5 classifier.

Keyword: Artificial neural networks, decision trees, coding.

1. Introducción

El aprendizaje automático es un método de análisis de datos que automatiza la construcción de modelos analíticos. Mediante el uso de algoritmos que aprenden de manera iterativa de los datos, el aprendizaje automático permite a las computadoras encontrar conocimiento oculto sin necesidad de programar explícitamente dónde buscar [1].

Las Redes Neuronales Artificiales (RNA) y los Árboles de Decisión (AD) son algoritmos de aprendizaje automático, que han sido utilizados como clasificadores en diversos sectores: industrial, financiero, salud, manufactura, social entre otros, debido a la capacidad y funcionalidad de estos en el procesamiento y entendimiento de la información. En la mayoría de la literatura actual sobre métodos de clasificación de datos se enfoca principalmente a valores numéricos. Sin embargo, en muchos problemas reales la información se presenta de forma no numérica (categórica). Los datos categóricos son un desafío, debido a sus propiedades inherentes especiales de los datos que describen. Por ejemplo, el atributo categórico, "*color"* puede tener diferentes valores como rojo, verde, azul, etc. Estos valores por su naturaleza son inherentes en ciertas propiedades incertidumbre, ambigüedad y vaguedad, esto hace que la clasificación sea más complicada que cuando se trabaja con valores numéricos [2]. El buen entendimiento de los datos es importante para realizar un análisis correcto de la información y obtener los resultados interesantes.

En el campo de aprendizaje automático es muy popular el uso de la RNA, por los resultados satisfactorios que se han obtenido, sin embargo, cuando la naturaleza de los datos no es numérica, se presentan algunos problemas, es decir es necesario aplicar métodos de trasformación de valores categóricos a valores numéricos lo que no siempre resulta ser lo apropiado. Es decir, trabajar las RNA cuando se tienen datos categóricos, estos deben de ser codificados.

En la presente investigación se presenta un análisis del efecto que se tiene al realizar una transformación de los datos para trabajar con RNA y los resultados obtenidos son comparados con los obtenidos en un clasificador que no realiza trasformación de datos (Un árbol de decisión).

Este trabajo está organizado de la siguiente manera: En la sección II se presenta algunos trabajos relacionados con la utilización redes neuronales con datos categóricos y los árboles de decisión. En la sección III se proporcionan los elementos teóricos relacionados con la propuesta que se presenta. En la sección IV se detalla los pasos que se siguieron en esta investigación. En la sección V se presenta la experimentación realizada y finalmente en la Sección VI se proporcionan las conclusiones a las cuales se llegaron, así como los trabajos futuros.

2. Estado del arte

Las redes neuronales artificiales han sido implementadas como clasificadores para resolver distintos problemas por ejemplo Rajkamal Kesharwani [3] utilizó una RNA en un proceso de manufactura de herramientas de mandril en el área de calidad para clasificarlas en: desperdicio, retrabajo y utilizable.

Donde la red neuronal se programó en Matlab, diseñada para una clasificación binaria de dos fases de los datos, en la primera fase evalúa si la parte en consideración es utilizable y la segunda fase de la clasificación se centra en los datos restantes de la fase 1 (datos no utilizables). Obteniendo un 98% de exactitud en el proceso de clasificación de las herramientas. Se observa que las redes fueron implementadas para la solución de un problema de clasificación. A pesar de ello, no menciona que tipos de datos se utilizaron.

Otra investigación es la de Thomas Küfner [4], donde utilizó una red neuronal supervisada para la clasificación de los estados operativos en las plantas de producción de sistemas embebidos basados en 6 mediciones (características de los estados operativos), las cuales fueron: corriente nominal, potencia nominal, velocidad de rotación nominal, factor potencia, eficiencia y eficiencia de clase. Utilizó 250 mediciones para cada estado de operación en un perceptrón multicapa donde la capa de entrada contiene 701 neuronas, la primera capa oculta tiene 500 neuronas, la segunda capa oculta 150 neuronas y la capa de salida 6 neuronas que corresponden a los estados operativos. Siendo uno de los métodos más costo-efectivo que otros, estas redes tuvieron una clasificación del 99.82% demostrando que esta red es capaz de reconocer los estados operativos de una maquinaria de producción confiable en un ambiente de manufactura. En esta investigación los datos utilizados fueron de tipo numérico por lo que no fue necesaria ninguna codificación en los datos.

Por otro parte Harsh Patel [5] menciona que los algoritmos de clasificación de árboles de decisión obtienen mejores resultados en términos de exactitud, tiempo y precisión que otros algoritmos de clasificación. Además, realiza un estudio comparativo entre los diferentes tipos de algoritmos (ID3, C4.5 y CART) utilizando una base de datos de "*evaluación de automóviles*" para estimar cual es el mejor para este proceso. La exactitud de los resultados se encuentra entre 97.11% y 89.3%.

Sungsu Choi [6] en su investigación identifica productos defectuosos en un proceso de manufactura de equipo electrónico, para ello utilizaron el algoritmo C4.5 únicamente para tres procesos. El resultado de haber construido al árbol de decisión fue que se encontraron las causas que ocasionan que los productos sean considerados como defectuosos.

3. Aspectos teóricos

3.1. Redes neuronales artificiales

Las redes neuronales artificiales están formadas por una serie de procesadores elementales, denominados neuronas artificiales, unidades o nodos, que constituyen dispositivos simples de cálculos que, bien a partir de un vector de entrada procedente del mundo exterior, o bien a partir de estímulos recibidos de otras neuronas, proporcionan una respuesta única (salida) [13]. Hay tres tipos de unidades en las redes neuronales: de entrada, salida y ocultas. Las unidades de entrada reciben señales desde el entorno (entradas a la red), las unidades de salida envían la señal fuera del sistema (salidas de la red), las unidades ocultas son aquellas cuyas entradas y salidas se encuentran dentro del sistema, no tienen contacto con el exterior.

En términos matemáticos se puede definir una neurona *k* por medio de ec.1, ec.2:

$$u_k = \sum_{j=i}^{m} w_{kj} x_j \,, \tag{1}$$

$$y_k = \varphi(u_k + b_{k)} \,, \tag{2}$$

donde $x_1, x_2 \ldots x_m$ son los valores de las señales de entrada: $w_1, w_2 \ldots w_m$ son los pesos sinápticos de la neurona k; u_k es la suma del producto de los pesos sinápticos y los valores de la señal de entrada; b_k es el Bias; φ es la función de activación; y y_k es el valor de la salida de la neurona. La función de activación es la que determina que una neurona sea activa o pasiva de acuerdo con su valor de salida [8].

3.2. Codificación

Las variables en escala nominal no tienen valor cuantitativo, es decir, son variables puramente cualitativas. Algunos ejemplos de datos nominales son variables como el género (masculino, femenino), el estado civil (soltero, casado, divorciado) simplemente denotan categorías [9].

Existen varios métodos de codificación para convertir los datos en valores numéricos, como la Codificación Activa y Codificación Entera.

Codificación activa

Es el esquema de codificación más utilizado. Compara cada nivel de la variable categórica con un nivel de referencia fijo. Una codificación activa transforma una sola variable con *n* observaciones y *d* valores distintos, en *d* variables binarias con *n* observaciones cada una. Cada observación indica la presencia (1) o ausencia (0) de la variable dicotómica [10]

Codificación ordinal o entera

En esta codificación se asigna un número entero a cada categoría, siempre que se conozca el número de categorías existentes, no agrega ninguna columna nueva a los datos, pero implica un orden a la variable que puede no existir realmente [11].

3.3. Árboles de decisión

Los árboles de decisión (AD) son una técnica inductiva mixta que consiste en crear una estructura de árbol que clasifica un conjunto de datos (objetos o ejemplos) en un cierto número de clases, de tal manera que a posteriori pueda clasificar nuevos objetos que le sean presentados, es decir, a partir de conjuntos de objetos iniciales (Muestra de entrenamiento) se construye un árbol que permitirá clasificar o predecir a que clase pertenece a los nuevos objetos que se le van presentando. Los AD constan de dos elementos básicos: los nodos de decisión y los nodos hoja. Los nodos de decisión son aquellos que representan una característica del objeto (por ejemplo, la característica "color") de estos nodos salen tantas ramas como valores tenga la característica (rojo, azul, …) que a su vez se van a conectar a otro nodo de decisión o a un nodo hoja. Los nodos hoja son los nodos finales (último nivel del árbol) y representan la etiqueta de la clase a la que pertenece el objeto.

En la literatura se pueden encontrar diferentes algoritmos para la construcción de un árbol, uno de ellos es el algoritmo ID3 (Iterative Dichotomiser) Dichotomiser Iterativo propuesto en 1986 por Quinlan Ross [12]. El cual trabaja cuando la muestra de entrenamiento tiene dos clases. Otro algoritmo es el C4.5 es utilizado para construir un árbol de decisión también propuesto por Ross Quinlan, que es una generalización del algoritmo ID3, es decir trabaja para *n* clases. Este algoritmo utiliza como criterio de división la *ganancia de la información,* que es una propiedad estadística que mide que tan bien una determinada característica separa los objetos o ejemplos de la muestra de entrenamiento, también puede trabajar con datos numéricos, categóricos o ambos [12].

Para el cálculo de la ganancia *de la información es* necesario calcular la entropía que permite medir la impureza de la muestra de entrenamiento. Los pasos básicos del algoritmo C4.5 son sencillos, ya que se pueden aplicar recursivamente en cada nodo, así el algoritmo inicia seleccionado de entre todas las características al de mayor "Ganancia" que será seleccionado como nodo raíz, que a su vez este tendrá salidas como valores tenga la característica. De esta manera el proceso se repite de igual manera para una de las características hasta llegar a los nodos hojas que serán las clases.

4. Metodología

Los pasos que se siguieron para la realización de este trabajo se presentan en la figura 1. Seguidamente se describen detalladamente cada uno de los pasos.

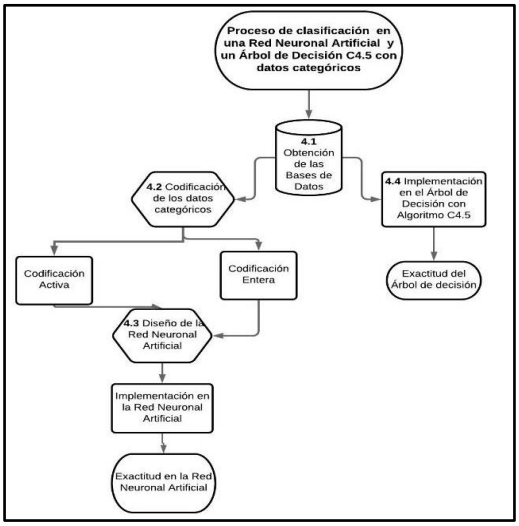


Fig. 1. Metodología.

4.1. Obtención de las bases de datos

Cuatro de las bases de datos utilizadas en este trabajo fueron adquiridas del repositorio de datos UCI "*Machine Learning Repository*" (archive.ics.uci.edu), base de datos "*Defectos en autos*" fue adquirida directamente de una planta automotriz:

a) *Evaluación de carros*: la cual está descrita en 6 características categóricas las cuales son: Compra, Mantenimiento, Puertas, Personas, Arranque y Seguridad con 4 clases (Accesible, Inaccesible, Bueno y Muy bueno) con un tamaño 1728 automóviles (objetos).

b) *Defectos en autos*: se describe en 4 características categóricas las cuales son: Modulo, Coordenada, Modelo y Tipo de defecto, con 3 clases (Ensamble, Carrocerías y Pintura) con 543 objetos.

c) *Juego de ajedrez*: se describe en 36 características categóricas, las cuales representan las posiciones en el juego, con dos clases (ganar o perder) con 3196 jugadas (objetos).

d) *Base de Juego xox:* cuenta con 9 atributos categóricos para describirlo, que son las posiciones del juego, con dos clases (juego positivo o juego negativo), con 958 juegos(objetos).

e) *Cáncer de pecho:* Se describe en 5 características de tipo categórico (Menopausia, Nodos-capas, Pecho, Irradiación) y 4 de tipo numérico (Edad, Tamaño del tumor, Nodos inv, Maligno), con dos clases (recurrente y no-recurrente) con 286 diagnósticos (objetos).

4.2. Codificación de las bases de datos

Se realizaron las codificaciones de los conjuntos de datos utilizando dos métodos, la codificación entera y la codificación activa.

Codificación entera

En las Tablas 1-5 se presenta la codificación entera, en cada una de las tablas se puede ver dos columnas por cada característica, donde la primera columna es el código que le corresponde al valor del dominio de la característica.

Tabla 1. Conjunto de datos "*Evaluación de carros*".

Codificación Entera											
Compra		**Mantenimiento**		**Puertas**		**Personas**		**Arranque**		**Seguridad**	
1	Muy A	**1**	Muy Alta	**1**	Dos	**1**	Dos	**1**	Bajo	**1**	Alta
2	Alta	**2**	Alta	**2**	Tres	**2**	Cuatro	**2**	Medio	**2**	Medio
3	Medio	**3**	Medio	**3**	Cuatro	**3**	Más	**3**	Alto	**3**	Baja
4	Baja	**4**	Baja	**4**	Más						

Tabla 2. Conjunto de datos "*Defectos en autos*".

Codificación Entera							
Módulo		**Coordenada**		**Modelo**		**Tipo de Defecto**	
1	4to D	**1**	A10	**1**	C	**1**	Abollado
2	4tO I	**2**	A11	**2**	L	**2**	Brisa
3	5ta P	**3**	A7	**3**	F	**3**	Contam
40	SI	**41**	K15			**11**	Repegón

Tabla 3. Conjunto de datos *"Juego de ajedrez".*

Codificación Entera											
Posición1		**Posición2**		**Posición3**		**Posición4**		**Posición5**		**Posición36**	
1	F	**1**	F	**1**	F	**1**	F	**1**	F	**1**	N
2	T	**2**	T	**2**	T	**2**	T	**2**	T	**2**	T

Tabla 4. Conjunto de datos *"Juego xox".*

Codificación Entera											
Posición1		**Posición2**		**Posición3**		**Posición4**		**Posición5**		**Posición9**	
1	B	**1**	B	**1**	B	**1**	B	**1**	B	**1**	B
2	O	**2**	O	**2**	O	**2**	O	**2**	O	**2**	O
3	X	**3**	X	**3**	X	**3**	X	**3**	X	**3**	X

Tabla 5. Conjunto de datos *"Cáncer de pecho".*

Codificación Entera											
Edad		**Menopausia**		**Tamaño T**		**Nodos inv**		**NodosC**		**Irradiación**	
1	20-29	**1**	Ge40	**1**	0-4	**1**	0-2	**1**	X	**1**	X
2	30-39	**2**	It4	**2**	5-9	**2**	3-5	**2**	No	**2**	No
3	40-49	**3**	Premeno	**3**	10-14	**3**	6-8	**3**	Yes	**3**	Yes
5	70-79			**11**	50-54	**7**	24-26				

Tabla 6. Conjunto de datos *"Evaluación de carros".*

Codificación Activa											
Compra		**Mantenimiento**		**Puertas**		**Personas**		**Arranque**		**Seguridad**	
0001	Muy Alta	0001	Muy Alta	0001	Dos	001	Dos	001	Bajo	001	Alta
0010	Alta	0010	Alta	0010	Tres	010	Cuatro	010	Medio	010	Medio
0100	Medio	0100	Medio	0100	Cuatro	100	Más	100	Alto	100	Baja
1000	Baja	1000	Baja	1000	Más						

Codificación activa

En las Tablas 6-10 se presenta la codificación activa, en cada una de las tablas se puede ver dos columnas por cada característica, donde la primera columna es el código que le corresponde al valor del dominio de la característica.

Tabla 7. Conjunto de datos *"Defectos en autos"*.

Codificación Activa						
Módulo o Parte	**Coordenada**	**Modelo**			**Tipo de Defecto**	
0000000000000000000000000000000001	4to D	00000000000000000000000000000001	A10	001	C	00000000001	Abollado
0000000000000000000000000000000010	4tO I	00000000000000000000000000000010	A11	010	L	00000000010	Brisa
0000000000000000000000000000000100	5ta P	00000000000000000000000000000100	A7	100	F	00000000100	Contam
1000000000000000000000000000000000	SI	10000000000000000000000000000000	K15			10000000000	Repegón

Tabla 8. Conjunto de datos "*Juego de ajedrez"*.

Codificación Activa											
Posición1		**Posición2**		**Posición3**		**Posición4**		**Posición5**		**Posición36**	
01	F	01	F	01	F	01	F	01	F	01	N
10	T	10	T	10	T	10	T	10	T	10	T

Tabla 9. Conjunto de datos *"Juego xox"*.

Codificación Activa											
Posición1		**Posición2**		**Posición3**		**Posición4**		**Posición5**		**Posición9**	
001	B	001	B	001	B	001	B	001	B	001	B
010	O	010	O	010	O	010	O	010	O	010	O
100	X	100	X	100	X	100	X	100	X	100	X

Tabla 10. Conjunto de datos *"Cáncer de pecho"*.

Codificación Activa											
Edad		**Menopausia**		**Tamaño Tumor**		**Nodos inv**		**Nodos-C**		**Irradiación**	
000001	20-29	001	Ge40	00000000001	0-4	000001	0-2	001	X	001	X
000010	30-39	010	It4	00000000010	5-9	000010	3-5	010	No	010	No
000100	40-49	100	Premeno	00000000100	10-14	000100	6-8	100	Yes	100	Yes
.	.			.	.	.	.				
100000	70-79			10000000000	50-54	100000	24-26				

4. 3. Diseño de la red neuronal artificial

La Red Neuronal Artificial utilizada se presenta en la figura 2, donde la estructura de la RNA es utilizada es la misma solo cambiando las entradas y salidas debido a las diferentes características y clases.

Todas las bases de datos se codificaron con los dos métodos anteriormente expuestos para ser implementados en la RNA en Python, la red para la clasificación mostrada en la figura 4.1, representa las neuronas de entrada dependiendo de las características que tenga la base de datos a utilizar, las capas ocultas con 4,10,10 neuronas y la capa de salida con neuronas las cuales son las clases. En la sección V se proporciona más detalle.

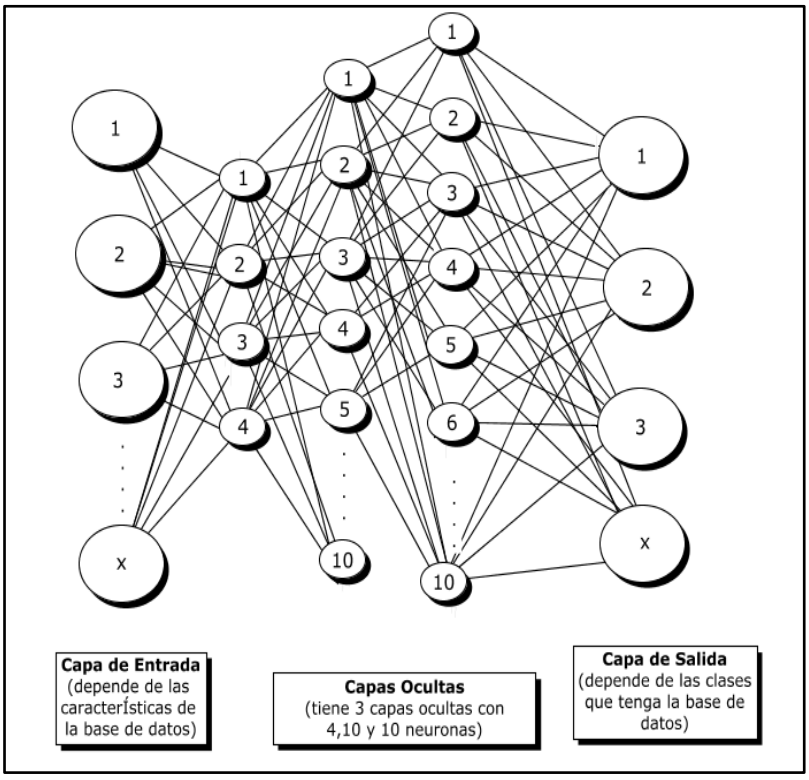


Fig. 2. Red Neuronal Artificial.

4.4. Árbol de decisión con el algoritmo C4.5

Para cada uno de los conjuntos de datos se construyó su respectivo árbol de decisión utilizando el algoritmo C4.5, para esto se trabajó con el software Weka. Por razones de espacio solo se presenta el árbol de decisión obtenido de la base de datos "*evaluación de carros*", mostrado en la figura 3, también se puede observar que el nodo que obtuvo la mayor *"ganancia de información*" fue la característica *"seguridad"*.

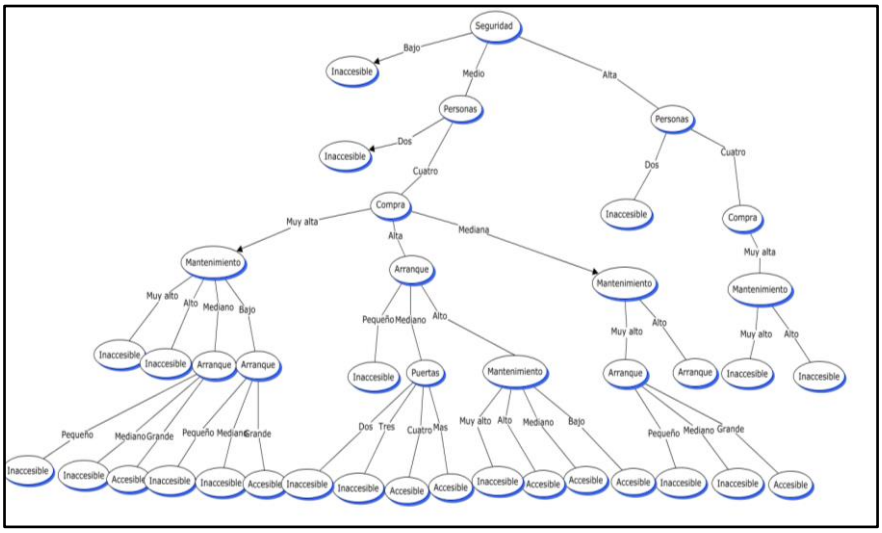


Fig. 3. Ejemplo de árbol de decisión C4.5 en la base evaluación de carros.

5. Experimentación

5. 1. Red neuronal artificial

Implementando dos redes neuronales con diferente número de neuronas en los datos categóricos con codificación entera se obtuvieron los resultados que se muestran en la figura 4. En la figura se indica la configuración de la RNA utilizada para cada una de las bases de datos, tamaño de la muestra de entrenamiento y prueba, así como la eficiencia obtenida. El número de neuronas en la capa de entrada de las redes son las características y en la capa de salida son las clases de cada una de las bases de datos. La función de activación para las capas ocultas fue RELU y para las capas de salida fue SOFTMAX en todas las bases de datos.

En el caso de la base de datos de "*Evaluación de carros*", se utilizaron 1210 objetos para entrenamiento y 518 objetos de prueba. Las redes se conformaron con 6 neuronas de entrada (características), las 3 capas ocultas generales (diferente número de neuronas) y 4 neuronas de salida que fueron las clases. Con esta codificación se obtuvo una exactitud del 68% y 74% de la clasificación.

En la base de "*Defectos en autos*" dado el tamaño, las redes neuronales se conformaron por 4 neuronas de entrada, 3 capas ocultas (con diferente número de neuronas en la red 1 y red 2) y 3 neuronas de salida, se utilizaron 380 objetos para el entrenamiento y 163 para la prueba de esta. Obteniendo una exactitud del 55% y 0.39% en el proceso de determinar la clase a la que pertenece.

Por otra parte, la base de datos de "*Juego de ajedrez*" obtuvo una exactitud del 93% y 83%, de un tamaño de 3196 objetos de los cuales se utilizaron 2237 para el entrenamiento y 959 de prueba, en una red conformada por 36 neuronas de en la capa entrada, 3 capas ocultas y 2 neuronas en la capa de salida.

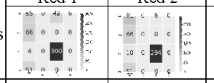
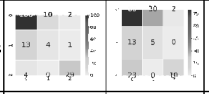
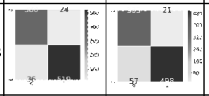
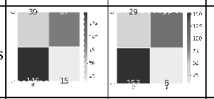
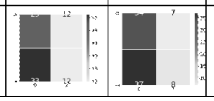
Codificación Entera								
Base de datos	**N° de neuronas en las capas oculta**		**Entrenamiento**	**Prueba**	**Matriz de confusión**		**Resultado de Exactitud**	
	Red 1	Red 2			Red 1	Red 2	Red 1	Red 2
Evaluación de Carro	4 10 10	6 4 9	1210 objetos	518 objetos			0.685	0.74
Defectos en autos	4 10 10	6 4 9	380 objetos	163 objetos			0.55	0.39
Juego de ajedrez	4 10 10	6 4 9	2237 objetos	959 objetos			0.93	0.83
Juego xox	4 10 10	6 4 9	671 objetos	287 objetos			0.16	0.1
Cáncer de Pecho	4 10 10	6 4 9	200 objetos	86 objetos			0.42	0.38

Fig. 4. Configuración de la RNA de los conjuntos de datos para "*Codificación entera*".

En la base de datos "*Juego xox*" se ocuparon 671 objetos para el entrenamiento y 287 objetos de prueba en una red compuesta por 9 neuronas de entrada, 3 capas ocultas (número de neuronas diferentes en cada red) y 2 neuronas de salida.

Teniendo una exactitud del 16% y 10% en el proceso de determinar a qué clase corresponde el objeto.

Por último, en la base de datos "*Cáncer de pecho*" se obtuvo una exactitud del 42% y 38% en una red con 9 neuronas de entrada, 3 capas ocultas y 2 neuronas de salida con 200 muestras de entrenamiento y 86 objetos de prueba. Se observa que las bases con exactitud alta entre 93% y 68% tienen más de 1000 muestras en ellas. Mientras que las otras bases tuvieron una exactitud menor de 60%.

Para la codificación activa se obtuvieron los resultados que se presentan en la figura 5. En la figura se indica la configuración de la RNA utilizada para cada una de las bases de datos, tamaño de la muestra de entrenamiento y prueba y eficiencia obtenida. El número de neuronas en la capa de entrada de las redes son las características y en la capa de salida son las clases de cada una de las bases de datos. La función de activación para las capas ocultas fue RELU y para las capas de salida fue SOFTMAX para todas las bases de datos.

En la base de datos de "*Evaluación de carros"*, las redes se utilizaron con 6 neuronas de entrada (características), las 3 capas ocultas generales (diferente número en las redes) y 4 neuronas de salida que fueron las clases. Se utilizaron 1210 objetos para entrenamiento y 518 objetos de prueba, la exactitud obtenida en ambas redes fue de 0%, debido a que no logro clasificar los datos y presento una falla en este proceso.

Para la base de datos *"Defectos en autos"* las redes neuronales utilizaron 380 objetos para el entrenamiento y 163 para la prueba en una red de 4 neuronas de entrada, 3 capas ocultas (cambia número de neuronas en las redes) y 3 neuronas de salida, presentando

un error al momento de realizar la clasificación, obteniendo una exactitud de 0% en ambas.

En la base de datos de "*Juego de ajedrez*" se obtuvo una exactitud del 90% y 92%,

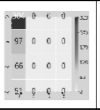
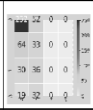
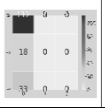
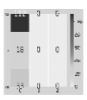
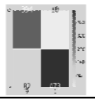
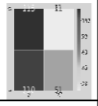
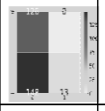
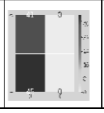
Codificación Activa								
Base de datos	**N° de neuronas en**		**Entrenamiento**	**Prueba**	**Matriz de confusión**		**Resultado de Exactitud**	
	Red 1	Red 2			Red 1	Red 2	Red 1	Red 2
Evaluación de Carro	4	6	1210 objetos	518 objetos			0	0
	10	4						
	10	9						
Defectos en autos	4	6	380 objetos	163 objetos			0	0
	10	4						
	10	9						
Juego de ajedrez	4	6	2237 objetos	959 objetos			0.9	0.92
	10	4						
	10	9						
Juego xox	4	6	671 objetos	287 objetos			0.52	0.28
	10	4						
	10	9						
Cáncer de Pecho	4	6	200 objetos	86 objetos			0	0
	10	4						
	10	9						

Fig. 5. Configuración de la RNA de los conjuntos de datos para *"Codificación Activa"*.

con 2237 objetos para el entrenamiento y 959 de prueba, en una red conformada por 36 neuronas en la capa entrada, 3 capas ocultas y 2 neuronas en la capa de salida. Probando que, de estos dos métodos de codificación, el mejor fue el de codificación Entera.

Por otra parte, la base de datos *"Juego xox"* obtuvo una exactitud de 52% y 28%, mostrando que la codificación Activa es mejor para esta base de datos. Se ocuparon 671 objetos para el entrenamiento y 287 objetos de prueba en las redes compuestas por 9 neuronas de entrada, 3 capas ocultas y 2 neuronas de salida

Por último, en la base de datos "Cáncer de pecho" se utilizaron dos redes con 9 neuronas de entrada, 3 capas ocultas (diferentes neuronas en la red 1 y red 2), 2 neuronas de salida con 200 objetos de entrenamiento y 86 muestras de prueba, los resultados obtenidos en la exactitud de este proceso para determinar a la clase que pertenece no fueron favorables y se tuvo una falla.

5. 2. Árbol de decisión con algoritmo C4.5

En la figura 6 se presentan los resultados de los árboles de decisión obtenidos de las bases de datos.

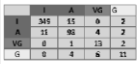
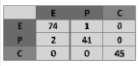
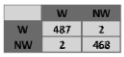
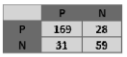
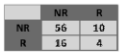
Bases de datos	Entrenamiento	Prueba	Matriz de confusión	Resultado
Evaluación de Carros	1210 objetos	518 objetos		0.9
Defectos en autos	380 objetos	163 objetos	Matriz de confusión 	0.981
Juego de ajedrez	2237 objetos	959 objetos	Matriz de confusión 	0.995
Juego xox	671 objetos	287 objetos	Matriz de confusión 	0.794
Cáncer de Pecho	200 objetos	86 objetos	Matriz de confusión 	0.697

Fig. 6. Resultado de los árboles de decisión con el algoritmo C4.5.

Base de datos	Resultados				
	Red Neuronal Artificial				Árbol de Decisión con algoritmo C4.5
	Codificación Entera		Codificación Activa		
	Red 1	Red 2	Red 1	Red 2	
Evaluación de Carros	68.50%	74%	0%	0%	90%
Defectos en autos	55%	39%	0%	0%	98%
Juego de ajedrez	93%	83%	90%	92%	99%
Juego xox	16%	10%	52%	28%	79%
Cáncer de Pecho	42%	38%	0%	0%	69%

Fig. 7. Resultados de los métodos de clasificación en las 5 *bases de datos.*

En la base de datos "*Evaluación de carros"*, se utilizaron 1210 muestras en el entrenamiento y 518 objetos para la prueba, de los cuales clasifico correctamente 470 objetos, obteniendo una exactitud del 90%.

Para la base de datos "*Defectos en los autos"* se ocuparon 380 objetos para entrenamiento y 163 de prueba, en el cual clasifico 160 objetos dando una exactitud del 98% en este proceso.

En la base de datos "*Juego de ajedrez"* se tuvo una exactitud en la clasificación del 99%, con 2237 objetos para entrenamiento y 959 de prueba.

Como se observa todas las bases de datos obtuvieron una mejor exactitud en los árboles de decisión con algoritmo C4.5 para realizar la clasificación.

6. Conclusión

De los resultados obtenidos en esta investigación podemos concluir que, el algoritmo C4.5 obtuvo mejores resultados en las 5 bases de datos utilizadas, mostrando que para datos categóricos y bases de datos con muestras reducidas se obtiene buenos resultados, mientras que los resultados obtenidos con la red neuronal fueron mejores en bases de más de 1000 muestras con codificación entera.

Las redes neuronales que utilizaron codificación activa, en la mayoría de los datos no fue viable realizarla, debido a que las exactitudes obtenidas para realizar este proceso no fueron óptimas, en la figura 7 se presenta la comparación de los resultados obtenidos por cada método y por cada base de datos.

Finalmente podríamos decir que los árboles de decisión proporcionan mejores resultados en comparación de las redes neuronales artificiales cuando los conjuntos de datos están descritos en características categóricas. Esto debido a que para utilizar una red neuronal se requiere realizar los datos categóricos a numéricos.

Referencias

1. Kedar, P., Taher S., Pardawala, C., Pai, D.: A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications 175 (2017)
2. Sarkar, J.P., Saha, I., Chakraborty, S., Maulik, U.: Machine learning integrated credibilistic semi supervised clustering for categorical data. Applied Soft Computing Journal (2020)
3. Rajkamal, K., Dagli, C., Sun, Z.: Application of neural network in shop floor quality control in a make to order business. Procedia Computer Science (2016)
4. Küfner, T., Uhlemann T., Ziegler B.: Lean data in manufacturing systems: using artificial intelligence for decentralized data reduction and information extraction. Procedia CIRP (2018)
5. Patel, H.P.: Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering (2018)
6. Choi, Sungsu: A decision tree approach for identifying defective products in the manufacturing process. International Journal of Contents, 13 (2017)
7. Hernandez-Orallo, J.: Introducción a la minería de datos. Pearson Prentice Hall (2004)
8. Haykin, S.: Neuronal networks and learning machines. Pearson Prentice Hall (1999)
9. Damodar, N.G.: Basic econometrics. The McGraw Hill (2004)
10. Brett, L.: Machine learning with R. Packt Publishing Limited (2013)
11. Von-Eye, A., Clifford, C.C.: Categorical variables in developmental research: Methods of analysis. Elsevier (1996)
12. Himani, S., Sunil, K.: A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (2013)
13. Fernández-Fernández, J.M., Flórez-López, R.: Las Redes Neuronales Artificiales. Netbiblo, (2008)