Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

COMPARACIÓN ENTRE REGRESIÓN LOGÍSTICA Y REDES NEURONALES PARA PREDECIR CÁNCER DE PIEL EN PERROS

Trabajo de investigación para optar el grado académico de bachiller en Ingeniería de Sistemas

Renato Chávez Martínez

Código 20151536

Asesor

Nadia Katherine Rodríguez Rodríguez

Lima – Perú

junio de 2019

Comparación entre regresión logística y redes neuronales para predecir cáncer de piel en perros

Renato Chávez Martínez
20151536@aloe.ulima.edu.pe
Universidad de Lima

Resumen: Determinar si un perro tiene la predisposición de desarrollar cáncer a la piel es uno de los desafíos tanto de los veterinarios como de los dueños de las mascotas. Los modelos de regresión logística y redes neuronales han sido ampliamente utilizados para realizar predicciones en el ámbito de la medicina humana, el presente estudio aborda la comparación de éstas dos técnicas para la predicción de cáncer de piel en perros. Las características que se han analizado son la edad, el sexo, raza, exposición al sol, albinismo y la aparición de dermatitis. Dichas características fueron validadas por el método de coeficiente de correlación y el análisis de componente principal. Los resultados obtenidos demostraron que la red neuronal backpropagation con validación cruzada supera al modelo de regresión logística. El valor de predicción generado por la red neuronal fue de 89.6% mientras que la regresión logística obtuvo un 84%.

Palabras Clave: Predicción, Cáncer de piel, Perros, Redes neuronales backpropagation, Regresión Lineal

Abstract: To ascertain if a dog has the predisposition to develop skin cancer is a challenge for both veterinarians and pet owners. Logistic regression models and neural networks have been used widely in the field of human medicine to make predictions; the present study approaches the comparison between these two technics to predict skin cancer in dogs. The variables we analyzed were age, sex, breed, sun exposition, albinism and, dermatitis. These variables were validated by the correlation coefficient and the principal component analysis. The obtained results showed that the backpropagation neural network technique with a cross validation is better than the logistic regression. The neural network's accuracy value was 89.6% while only 84% for the logistic regression.

Keywords: Prediction, Skin Cancer, Dog, Backpropagation Neural Network, Linear Regression

1. INTRODUCCIÓN

En la actualidad, los dueños de mascotas tienden a establecer un fuerte vínculo con sus perros, por lo que muchas veces les resulta difícil afrontar su muerte, más aún si se puede evitar al detectarse a tiempo. El cáncer de piel ha aumentado en la población de perros, en el XI Congreso Andaluz de Veterinarios, realizado en el 2015, indicaron que el 50% de los perros fallecen por algún tipo de cáncer; siendo el más frecuente el de mama con un 30%, seguido por el cáncer a la piel 20% en causas de fallecimiento (Flores, 1986).

Si el cáncer de piel es detectado demasiado tarde las probabilidades de que el perro sobreviva son casi nulas. Los dueños no suelen considerar este tipo de cáncer en sus perros debido a la creencia que su pelaje los protege (WebMD, 2018). En Lima según los estudios realizados sobre neoplásicas malignas, los perros machos entre las edades de 5 a 9 años tuvieron un porcentaje de 61.49% de desarrollar un cáncer mortal (De Vivero, 2013). Asimismo, con respecto a la raza, en los perros mestizos la incidencia es mayor (25.63%) comparada con los de raza pura y los Bóxer con 14.56%. En cuanto a la ubicación en el cuerpo, la neoplasia más se encuentra en el tronco (54.1%) seguido de las extremidades con 40.5% (Chang Huamán, 2016).

2. ESTADO DEL ARTE

2.1 Modelos de predicción de regresión logística

En la investigación realizada por Denle, Walker y Kadman (2004) comparan tres métodos de minería de datos para predecir la probabilidad de supervivencia en pacientes de cáncer de mama, en el cual utilizan un dataset de 200,000 casos para entrenar sus modelos. Los tres modelos utilizados fueron, un multilayer perceptron con backpropagation, árbol de decisiones y regresión logística. Las variables de entrada escogidas fueron la raza, estado civil, histología, comportamiento, grado, extensión, linfomas, radiación y estado del cáncer. Se obtuvo que el árbol de decisiones tuvo los mejores resultados, con un 93.6% de precisión, seguido por la RN con 91.2% y por último la regresión logística con 89.2%. Determinaron que los 3 modelos son eficientes, pero hay una ligera ventaja con los modelos de

machine learning. Además, al utilizar un análisis de sensibilidad en los modelos de redes neuronales, se pudieron obtener los factores que tenían prioridad para obtener el diagnóstico.

Por otro lado, Zhou, Liu y Wong (2004) en su trabajo de clasificación y predicción de cáncer utilizando regresión logística. Utilizando un dataset de 7129 genes de 72 pacientes con leucemia, 3226 genes de 21 pacientes con cáncer de mama y 2308 genes de 35 pacientes con tumores de célula azul, pudieron comprobar que, con una gran cantidad de registros en su dataset, el modelo llega a ser efectivo y dar resultados por encima de los benchmarks con los cuales los compararon.

2.2 Modelos de predicción utilizando redes neuronales backpropagation

Chang, C. y Chen, C. (2009) realizaron un estudio a pacientes con cáncer la piel, el cual está enfocado en seis enfermedades más frecuente a la piel, en la cual el modelo de la red neuronal backpropagation tuvo un nivel de precisión del 92.62%, por otra parte, el análisis de sensibilidad combinado con el modelo de árbol de decisión tuvo un 80.33%. Para el procesamiento de datos se contó con 366 casos de enfermedad de piel, de los cuales 244 datos se utilizaron para el entrenamiento y 122 datos para verificar la exactitud del modelo del diagnóstico clínico establecido. La tasa de error global del algoritmo del árbol de decisión en el Experimento 1 es del 9.11%, lo que indica un 90.89% de precisión en su capacidad para discriminar correctamente una enfermedad de las otras. Por otro lado, al combinar un árbol de decisión junto a la red neuronal obtuvo una precisión más baja de 86.89% y una tasa de error del 13.11%. Lo que demostró que el modelo de la red neuronal backpropagation por si solo era más eficiente.

Roffman, Hart y Girardi (2018) realizaron una investigación, pero en humanos para lo cual desarrolló y validó una red neuronal multi parametrizada utilizando la historia clínica de los pacientes. Para ello extrajeron 13 parámetros de la historia clínica. El resultado tuvo una sensibilidad de entrenamiento de 88.5% y una especificidad de 62.7%, que comparados con las estadísticas de la Sociedad Americana de Cáncer (2017) los resultados fueron aceptables. La red neuronal implementada utilizó el algoritmo de backpropagation, teniendo 12 neuronas en la primera capa, con valores normalizados entre 0 y 1. La red neuronal backpropagation fue entrenada con información del repositorio de encuesta nacional de Estados Unidos desde 1997 al 2015 y luego se comparó el resultado con las encuestas realizadas el 2016. El resultado de la investigación determina que la red neuronal backpropagation es lo suficientemente robusta para predecir si una persona tiene cáncer o no.

Por otro lado, Sánchez (2012) realizó una comparación de modelos para la predicción de cáncer de piel, comparando una red neuronal backpropagation y Random Forest (RF) utilizando 6000 registros del Hospital Universitario de Canarias, se utilizó parámetros como el tipo de sangre, alimentación e historia de cáncer en la familia, el modelo de Backpropagation obtuvo un 87% de precisión, seguido del RF con un 77%. Pero cuando se combinó el modelo de RF con un modelo de regresión logística este logro superar a la red neuronal backpropagation en un 5%, logrando 92% de precisión. Se llego a concluir que ambos métodos son efectivos bajo las circunstancias del estudio y depende mucho de las variables de entrada para poder definir cual tendrá un mejor resultado.

Lee (2011) logra demostrar que el modelo de la red neuronal backpropagation por si solo logra un 67% de certeza con un conjunto de pruebas, mientras que al aplicarle el k-fold estratificado lograba una certeza de un 85.4%, logrando así una mejora del casi 20% debido a como se estaba evaluando el modelo.

3. ANTECEDENTES

3.1 Regresión Logística

La regresión logística es un modelo estadístico usado ampliamente en el cual la probabilidad de una variable de salida dependiente está relacionada a conjunto de variables independientes, la cual tiene la siguiente formula (1).

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1\chi_1 + \ldots + \beta_i\chi_i \qquad (1)$$

Donde p es la probabilidad de la salida de interés, β_0 es un término de intercepción, β_1, …, βi son los coeficientes β asociados con cada variable. X1, …, Xi son las variables independientes. En el modelo se suele asumir que estas variables están relacionadas de una manera linear a las probabilidades del logaritmo de la variable de salida (Mehta y Patel, 1995).

En el proceso de adaptación del algoritmo se puede interpretar de manera sencilla cual es la magnitud de importancia de las distintas variables debido a sus β. Por esto la regresión logística se ha convertido en la técnica con mayor preferencia para modelos estadísticos en los cuales las variables de salida son dicotómicas (binarias) (Hosmer y Lemeshow, 2000).

3.2 Redes Neuronales

Las redes neuronales están conformadas por distintas capas, las cuales a su vez están formadas por un grupo de nodos, así como se muestra en la Figura 1. Las capas se dividen en capa de entrada que es la capa donde se ingresan las variables, la capa de salida en la cual los nodos muestran los resultados y las capas ocultas, denominados así ya que estas no tienen una relación directa con la información ingresada ni la de salida (Valencia, Yáñez, Sánchez, 2006).

Estas neuronas están vinculadas por un patrón que permite comunicación entre ellas, las cuales reciben una serie de variables que son utilizadas para predecir un resultado y aprenden al utilizar como información los procesos que se han corrido anteriormente tal como se ve en la Figura 1. Existen dos grandes métodos de aprendizaje en los cuales las redes neuronales se dividen, supervisado y no supervisado (Matich, 2001).

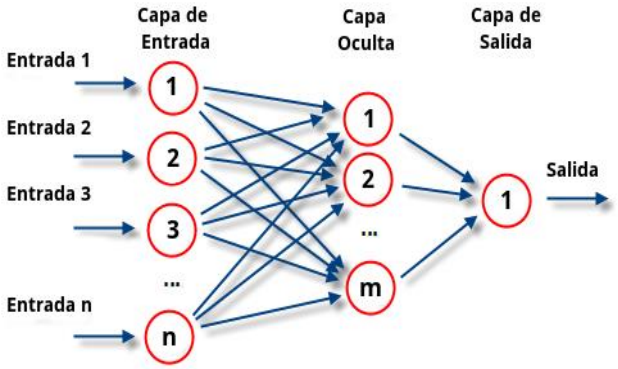

Fig. 1. Modelo de red neuronal
Fuente: Matich, D. (2001).

3.3 Modelo de entrenamiento Backpropagation

Este modelo está clasificado dentro del aprendizaje supervisado en cual el error se propaga hacia atrás desde el nodo de salida. Este tipo de modelo suele ser utilizado para el diagnóstico y predicción de problemas (Chun-Lang Chang & Chih-Hao Chen, 2008). Por cada registro que se introduce de la data de entrenamiento, el modelo recorre la red dos veces. En la ida inserta en los nodos de entrada pequeños valores arbitrarios a los pesos de las variables para obtener un valor de salida y su nivel de error comparándolo con el valor real. A la vuelta, la red determina cómo se deben cambiar los pesos para poder reducir el nivel de error.

Este es uno de los modelos más utilizados para realizar predicciones en el campo de la medicina humana, el cual sirve como una herramienta para ayudar a los doctores a analizar, modelar y entender casos clínicos complejos que tienen distintas variables (Gupta, Shreevastava, 2011).

En el siguiente caso se mostrará cómo funciona un modelo de Backpropagation.

Primero se le debe asignar un valor aleatorio a los pesos de que obtiene cada nodo (w) que se van ajustando dependiendo del valor n que es el valor del registro de la variable de entrada y también para las variables de imparcialidad (bias) b.

Una vez definido todas las variables se puede aplicar la función mostrada en la Fórmula 2 para calcular el valor de cada nodo oculto como se muestra en el caso del nodo oculto H_1.

$$\boldsymbol{H}_1 = n_1 * w_{11} + n_2 * w_{21} + n_3 * w_{31} + n_4 * w_{41} + n_5 * w_{51} + n_6 * w_{61} + b_1 \qquad (2)$$

Después, el valor calculado $\mathbf{H_n}$ se ingresa a la función de activación, que para nuestro modelo utiliza ReLu (Formula 3.1) para las capas ocultas ya que es la más efectiva (Glorot, Bordes & Bengio, 2011).

$$ReLu: \quad Output_{H_n} = \begin{cases} 0 : & H_n < 0 \\ H_n : & H_n \geq 0 \end{cases} \qquad (3.1)$$

$$\text{Sigmoidal: } Output_{H_n} = \frac{1}{1 - e^{-H_n}} \qquad (3.2)$$

Al tener todos los valores de **H**, se calcula el valor del nodo de salida aplicando la Formula 2, utilizando como parámetros de entrada los valores obtenidos para los **H.** Al resultado luego se le aplica la función de activación Sigmoidal ya que el resultado que se desea mostrar debe ser o cero o uno.

Para finalizar la primera parte, se compara el resultado de la salida con el resultado objetivo que es 1 si tiene cáncer y cero si no tiene, para ello se aplica la Formula 4.

$$E = \frac{1}{2}(Objetivo - Resutado)^2 \qquad (4)$$

Para la segunda parte, se usa este error E, para cambiar los pesos (**w**) de atrás hacia adelante, aplicando la Formula 5.

$$ErrordePeso_{w_n} = \frac{dE}{dw_n} \qquad (5)$$

Luego de simplificarlo queda como se muestra en la Formula 6.

$$ErrorPeso_{w_n} = \left(Output_{Y_1} - Objetivo\right) * \left(Output_{Y_1}\left(1 - Output_{Y_1}\right)\right) * Output_{H_n} \qquad (6)$$

Finalmente, se debe actualizar el valor de los pesos utilizando un aprendizaje (**v**). Con lo cual la fórmula para el nuevo valor de los pesos es como la mostrada en la Formula 7.

$$w_n = w_n - v * ErrordePeso_w \qquad (7)$$

ALGORITMO 1: Red de Backpropagation

"Cada entrenamiento está formado por un par $< \bar{x}, \bar{t} >$ donde $\bar{x}$ es el vector de los valores de entrada y $\bar{t}$ es el vector de los valores de salida.

η es el ratio de aprendizaje. n_{in} es el número de inputs a la red, n_{hidden} es el número de unidades en la capa oculta, y n_{out} es el número de unidades del output."

-Crear una red con n_6 inputs, n_4 nodos ocultos, n_1 nodos de salida.

-Inicializar la red con pesos pequeños aleatorios <-0.05, 0.05>

-Para cada K del K-Fold, Hacer

-Para cada $< \bar{x}, \bar{t} >$ de los registros, Hacer

Propagar el input de ida de la red

1. Input del registro de $\bar{x}$ a la red y generar la salida o_u para cada unidad de u de la red

Propagar los errores hacia atrás a través de la red

2. Para cada unidad *k* de salida de la red, calcular su error en términos de δ_k

$$\delta_k \leftarrow o_k (1 - o_k) (t_k - o_k)$$

3. Para cada unidad *h* oculta, calcular su error en termino de δ_h (2 veces porque son 2 capas ocultas).

$$\delta_h \leftarrow o_h (1 - o_h) \Sigma w_{kh} \delta_k$$

4. Actualizar los pesos w_{ji} de la red

$$w_{ji} \leftarrow w_{ji} + \eta \delta_j x_{ji}$$

Fuente: Mitchell (1997)

3.4 Overfitting

Según el principio de parsimonia, un modelo y sus procedimientos deben utilizar únicamente lo necesario para desarrollarlo y no más que eso, para así poder dar un resultado preciso (Hawkins, 2004).

En el caso del entrenamiento de una red neuronal se considera overfitting al sobreentrenar el modelo supervisado varias veces con los mismos datos en particular. Ya que, al tener el resultado, el modelo se ajustará para obtener esos resultados, resultando en un modelo poco confiable (Chicco, 2017).

3.5 K-Fold Cross Validation

El método de K-Fold es utilizado en los modelos de predicción para determinar si los resultados dados por el modelo son independientes y no hay overfitting. Para esto, los datos son divididos en k partes en donde k-1 partes son usadas para entrenar el modelo y la restante se usa para realizar las pruebas como se muestra en la figura 2. Este procedimiento luego se repite para cada una de las partes en las que se dividió. El resultado final de esta prueba se calcula realizando el promedio del porcentaje de efectividad de cada parte (Lee, 2011).

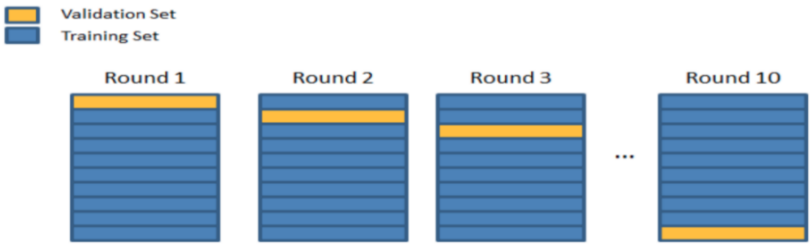

Fig. 2. Modelo de 10-Fold para validación.
Fuente: McCormick, C. (2018)

Cada resultado de los subgrupos se divide en 4, verdadero positivo (VP), cuando el resultado ha determinado correctamente que tiene cáncer, mientras que el falso negativo (FN) implica que no se tiene cáncer. Por otro lado, el verdadero negativo (VN) y el falso positivo se aplica a los casos en los cuales se ha predicho mal el resultado clasificándolo como correcto e incorrecto respectivamente cuando no los son. Con esto, el modelo se evalúa según su precisión, sensibilidad y especificidad con respecto a las siguientes formulas (Örkcü y Bal, 2011).

$$precisión = \frac{VP + VN}{VP + FN + VN + FP} \quad (8)$$

$$sensibilidad = \frac{VP}{VP + FN} \quad (9)$$

$$especificidad = \frac{VN}{VN + FP} \quad (10)$$

4. DESARROLLO DE MODELOS DE PREDICCIÓN

Para el modelo de red neuronal se decidió utilizar 6 nodos de entrada ya que se tenían 6 variables de entrada. Para los nodos de las capas ocultas y la cantidad de capas se fueron probando los resultados y se escogió el modelo más simple que tuviera a su vez resultados precisos, basándonos en el principio de parsimonia. Obteniendo así un modelo de 6 nodos de entrada, 2 capas ocultas con 4 nodos cada uno y una capa de salida con un nodo.

Para la validación del modelo de red neuronal y de regresión logística se utilizó el método de K-Fold cross validation. Se trabajó con un dataset de 2000 registros, para validar los modelos se utilizó un K de 10, lo que permite que cada partición tenga registros suficientes para entrenar y validar la data. Además, este valor es lo suficientemente alto como para obtener resultados imparciales (Kohavi, 1995).

4.1 Datos

En base a Chang G. (2016) se tomaron en cuenta las variables de edad, sexo, raza, el cual tiene como base los conceptos cubiertos por Moulton (1990) en su libro de cáncer en animales, donde determina que estas variables son las más importantes en todos los animales. Además, se consideró el tiempo de exposición al sol ya que es una de las causas más comunes de contraer cáncer la piel (WebMD, 2016). Luego, tomando en consideración el juicio experto de los veterinarios, se incluyó si el perro es albino y si suele tener dermatitis, lo cual debilita la piel.

Para poder realizar las pruebas del modelo se utilizaron registros tomados en el laboratorio de histopatología de la UNMSM, de los cuales se tomó una muestra de 2000 casos que fueron estudiados por el mismo laboratorio, en la tabla 1 se observa un resumen de los casos recompilados, en los cuales se encontraron 888 casos con dermatitis de los 2000 y 1169 casos en los cuales el perro presentaba cáncer.

Tabla 1. Resumen de razas registradas

Raza	Cantidad
Mestizo	575
Siberiano	161
Terrier	196
Schnauzer	200
Boxer	212
Shih Tzu	65
Pug	166
Labrador	155
Pastor	103
Poodle	68
Bulldog	99

Fuente: Propia (2018)

Las variables propuestas para el modelo se normalizaron en valores entre 0 y 1 según el criterio detallado en la tabla 2. Además, en la tabla 3 se puede apreciar una un grupo de registros con sus valores antes y después de ser normalizados de manera de ejemplo.

Tabla 2. Descripción de los parámetros usados en la red neuronal backpropagation

Parámetro	Tipo de entrada	Rango	Detalle
Edad	Continuo	0-1	Rango de 0 a 10 años, más de 10 años se considerará como 10
Sexo	Binario	0 o 1	0 es macho y 1 es hembra
Raza	Continuo	0-1	Se asigna el porcentaje en relación a todas las razas en la data.
Exposición al sol	Continuo	0-1	Dividido poca (25%), regular (50%), mucha (75%) y constante (75%) exposición
Tiene dermatitis	Binario	0 o 1	0 no tiene dermatitis y 1 si tiene dermatitis
Albinismo	Binario	0 o 1	0 no es albino y 1 si es albino

Fuente: Propia (2018)

Tabla 3. Normalización de registros

	Valores Recompilados						Valores Normalizados					
Registro	Edad	Sexo	Raza	Exp. Sol	Dermatitis	Albino	Edad	Sexo	Raza	Exp. Sol	Dermatitis	Albino
1	2	F	Boxer	Regular	No	No	0.2	1	=2 (boxers) /4(total de perros) = 0.50	0.5	0	0
2	5	M	Boxer	Mucha	Si	No	0.5	0	0.5	0.75	1	0
3	1	F	Bulldog	Poca	No	Si	0.1	1	0.25	0.25	0	1
4	11	F	Labrador	Constante	Si	No	1	1	0.25	1	1	0

Fuente: Propia (2018)

Para poder verificar que las variables son independientes, se utilizó el método estadístico de coeficiente de correlación, mientras más se aproxime a 1 el resultado significa que están más relacionadas entre sí. En la tabla 4 se muestran que los valores son bajos entre ellos, lo cual muestra que las variables tienen una baja dependencia entre ellos y por lo tanto no se debe descartar ni una en el modelo.

Tabla 4. Valores de factores de correlación

Edad	Sexo	Raza	Exposición al sol	Tiene dermatitis	Albinismo	Cáncer
	-0.034	0.005	-0.000	-0.015	0.083	0.269
		-0.021	-0.008	-0.034	-0.067	-0.176
			-0.210	0.008	-0.021	-0.028
				0.017	-0.000	0.012
					0.099	0.085
						0.328

Fuente: Propia (2018)

De manera complementaria se utilizó el método de análisis de componente principal (PCA) el cual pone cada grupo de datos en una dimensión distinta y a cada componente se le calcula la varianza y lo que busca es reducir la "dimensionalidad" del grupo de datos. Esto es, si ajustando los valores de los componentes se logra que los componentes que están fuertemente relacionado se puedan unir y así poder descartar una de las variables.

En la tabla 5, se pueden ver la lista de resultados el cual muestra valores altos, lo cual determina que los valores no tienen una correlación lineal entre ellos por lo que todos son considerados componentes principales.

Tabla 5. Valores de PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Eigenvalue	1.5496	1.2098	1.0306	0.9614	0.8716	0.7816	0.5955
Proportion	0.221	0.173	0.147	0.137	0.125	0.112	0.085
Parámetro	PC1	PC2	PC3	PC4	PC5	PC6	PC7

(continúa)

(continuación)

Edad	0.409	-0.063	0.568	0.208	0.554	0.086	-0.385
Sexo	-0.307	0.073	0.136	0.904	-0.078	-0.114	0.214
Raza	-0.051	-0.706	-0.06	-0.034	0.17	-0.681	0.047
Exposición al sol	0.043	0.702	-0.014	-0.108	0.213	-0.67	0.009
Tiene dermatitis	0.203	0.008	-0.781	0.258	0.506	0.156	-0.036
Albinismo	0.524	-0.009	-0.187	0.245	-0.594	-0.207	-0.485
Cancer	0.646	-0.022	0.098	0.023	-0.066	0.001	0.753

Fuente: Propia (2018)

4.2 Modelo de la red neuronal supervisada backpropagation para predicción de cáncer

Para la red neuronal, los seis nodos de entrada fueron la edad, sexo, raza, exposición al sol, dermatitis y albinismo. Debido a que la cantidad de nodos dentro de cada capa oculta y la cantidad de capas ocultas es subjetiva, se probaron distintas combinaciones hasta obtener un resultado adecuado, que fueron dos capas ocultas de 4 nodos cada una. Los valores bias de los nodos ocultos y el nodo de salida fueron definidos por una función de aleatoriedad del programa donde el modelo fue desarrollado. Los bias fueron aleatorios debido a que estos valores se van ajustando con cada iteración.

Las funciones de activación utilizadas en las capas ocultas fue ReLu, ya que es el estándar utilizado para las RN (Glorot, Bordes & Bengio, 2011), y se utilizó la función sigmoidal para la capa de salida ya que es la más adecuada cuando la información esta normalizada y los valores de salida son dicotómicos (Sibi, Allwyn & Siddarth, 2013).

A continuación, se muestra la arquitectura de la red neuronal utilizada.

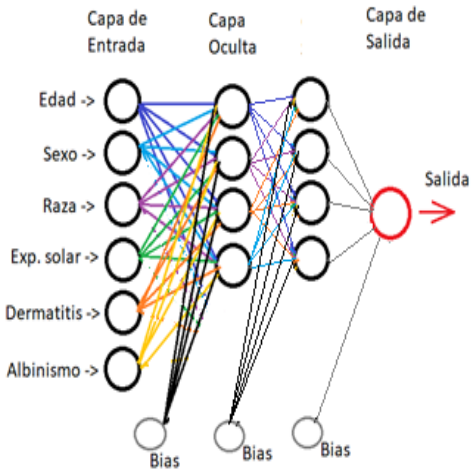

Fig. 3. Modelo de red neuronal backpropagation utilizada.
Fuente: Propia. (2018)

4.3 Modelo de predicción utilizando regresión logística para predicción de cáncer

Como el modelo debe hacer una distinción entre 2 clases, las probabilidades mayores a 50% se consideró como 1 (si desarrollara cáncer) y como 0 de ser lo contrario. Por el lado de los pesos de los coeficientes, estos se van ajustando con cada registro del dataset.

La ecuación del modelo ingresado sigue la siguiente estructura.

$$Y = \beta_0 + \beta_1 Edad + \beta_2 Sexo + \beta_3 Raza + \beta_4 Exp.Sol + \beta_5 Dermatitis + \beta_6 Albinismo$$

Después de entrenar los modelos, se probaron los resultados a través de un proceso en el cual la salida es comparada con un nuevo grupo de registros los cuales también se han ingresado al modelo. Para esto, todo el conjunto de data dividió en 10 partes las cuales se utilizaron para entrenar y evaluar a la red independientemente.

5. DISCUSIÓN Y RESULTADOS

Los modelos fueron validados realizando un 10-Fold cross validation con lo que se pudieron conseguir los resultados mostrados en la tabla 6 para los parámetros de precisión, sensibilidad y especificidad.

5.1 Resultados generales

Tabla 6. Resultados de la validación de K fold

K-Fold	VP		VN		FN		FP	
	BP	RL	BP	RL	BP	RL	BP	RL
1	121	109	56	56	15	16	8	19
2	94	83	88	91	6	6	12	20
3	135	93	57	69	3	23	5	15
4	122	96	58	56	8	27	12	21
5	122	76	52	94	11	17	15	13
6	93	97	83	80	7	11	17	12
7	98	75	81	103	9	15	12	7
8	105	111	72	66	15	13	8	10
9	76	66	107	96	7	21	10	17
10	110	107	63	57	12	20	15	16

Fuente: Propia (2018)

Como se puede observar en el cuadro comparativo, la red neuronal obtiene mejores resultados en las distintas corridas de los K-Fold, teniendo una igualdad en los verdaderos negativos. Esto demuestra que la red neuronal es capaz de discriminar mejor los casos dados.

Tabla 7. Cuadro Resumen

K-Fold	Precisión		Sensibilidad		Especificidad	
	BP	RL	BP	RL	BP	RL
1	88.50%	82.50%	89.00%	87.20%	87.50%	74.67%
2	91.00%	87.00%	94.00%	93.26%	88.00%	81.98%
3	96.00%	81.00%	97.80%	80.17%	91.90%	82.14%
4	90.00%	76.00%	93.80%	78.05%	82.90%	72.73%
5	87.00%	85.00%	91.70%	81.72%	77.60%	87.85%
6	88.00%	88.50%	93.00%	89.81%	83.00%	86.96%
7	89.50%	89.00%	91.60%	83.33%	87.10%	93.64%
8	88.50%	88.50%	87.50%	89.52%	90.00%	86.84%
9	91.50%	81.00%	91.60%	75.86%	91.50%	84.96%
10	86.50%	82.00%	90.20%	84.25%	80.80%	78.08%

Fuente: Propia (2018)

En el cuadro resumen se observa la diferencia entre los modelos de backpropagation (BP) y regresión logística (RL). En el caso de la precisión, que es la relación entre los verdaderos positivos y el total de positivos, en solo 1 de los pliegues la regresión logística fue mejor.

La red neuronal backpropagation tuvo resultados en un promedio de 5.6% mejores que los de la regresión logística, teniendo un máximo de diferencia del 15%. Estas diferencias se podrán apreciar de mejor manera con las gráficas siguientes.

5.2 Gráficos comparativos

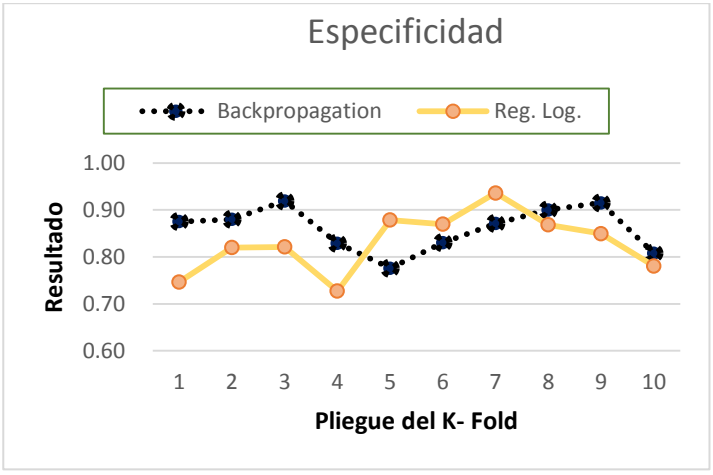

Fig. 4. Distribución de los resultados de la Especificidad.
Fuente: Propia. (2018)

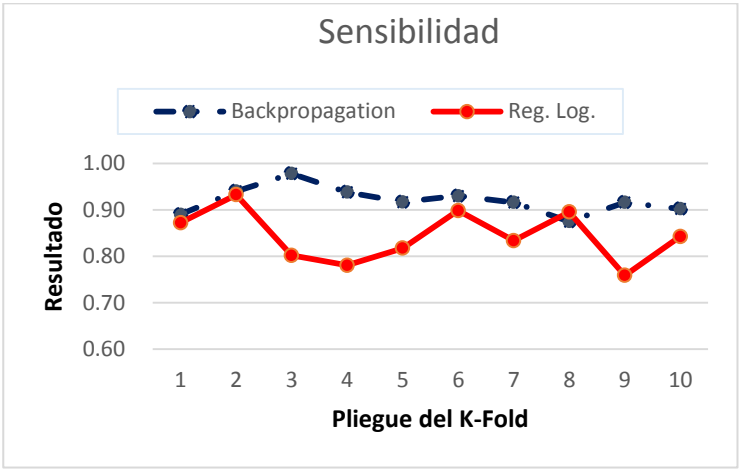

Fig. 5. Distribución de los resultados de la Sensibilidad.
Fuente: Propia. (2018)

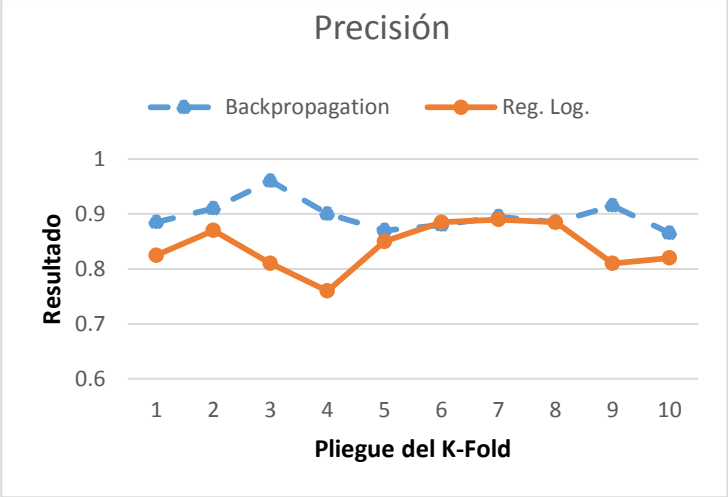

Fig. 6. Distribución de los resultados de la Precisión.
Fuente: Propia. (2018)

Se puede observar en los cuadros comparativos la diferencia entre ambos resultados en el cual claramente la ventaja la posee la red neuronal. En base a las condiciones dadas y la cantidad de registros ingresados, se puede determinar que la red neuronal backpropagation tiene una ventaja en la mayoría de resultados de la validación de K-Fold. A su vez, se observa que los resultados para el modelo de regresión presentan una dispersión mayor con respecto a los resultados que muestra, con lo que se podría deducir que le falta entrenamiento al modelo.

Considerando los resultados de los modelos mostrados en la tabla 8, se observa que el resultado de la precisión obtenido para la red neuronal la regresión logística se encuentra en rangos aceptables. El promedio de la sensibilidad de la red neuronal backpropagation a logra ser mucho mayor al de la regresión logística, lo que significa que la red neuronal es mejor detectando los casos que dan positivo a tener cáncer. Por otra parte, la especificidad de ambos fue la que más se aproximó entre ellos, mostrando que ambos modelos determinan de manera aceptable los casos negativos.

Tabla 8. Matriz de confusión

	Red Neuronal Backpropagation	Regresión Logística
Precisión	89.65%	84.05%
Sensibilidad	92.02%	84.32%
Especificidad	86.02%	82.98%

Fuente: Propia. (2018)

6. CONCLUSIONES

Se logró determinar que, con la cantidad de data de entrenamiento, la cual se considera poca para realizar análisis predictivos, se observó que la red neuronal backpropagation tuvo un mejor entrenamiento que el modelo de regresión logística y pudo obtener una mejor relación entre las variables dependientes e independientes. Esto concuerda con los resultados del trabajo de Tu (1996) en el cual concluye que las redes neuronales requieres un menor entrenamiento que los modelos estadísticos para desarrollarse. Esto se complementa con el trabajo de Dreiseitl y Ohno-Machado (2012), en el cual los casos en que se tiene una limitada cantidad de registros la red neuronal backpropagation muestra mejores resultados, pero cuando la cantidad de registros es alta, ambos modelos tienen resultados muy parecidos.

En las pruebas realizadas el modelo de red neuronal fue más propenso al overfitting en los casos que se utilizaba demasiadas veces el dataset para entrenar los modelos.

AGRADECIMIENTOS

Agradezco a mi universidad por permitirme convertirme en un profesional en lo que tanto me apasiona, a mi asesora Nadia Rodríguez por ser una guía durante todo este tiempo, por la confianza y la idea del tema. A la profesora Rosario Guzmán por apoyarme con lo que he necesitado para poder terminar este trabajo y a todos los demás profesores que estuvieron involucrados de manera directa o indirecta. A mi familia que siempre estuvo a mi lado y a mis amigos que siempre me apoyaron. Gracias.

REFERENCIAS

Álvarez Pecol, J. (2014). Perú, país perruno. Recuperado de https://www.ipsos.com/es-pe/peru-pais-perruno

Chang, C. L. y Chen, C. H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Systems with Applications. *Expert Systems with Applications, 36*(2):4035-4041.

Chang Huamán, G. S. (2016). Frecuencia de neoplasias en caninos de 0 a 5 años de edad diagnosticados histopatológicamente en el laboratorio de histología, embriología y patología veterinaria. Facultad de Medicina Veterinaria de la Universidad Mayor de San Marcos. Periodo 2003-2014. Tesis de Médico Veterinario. Lima, Perú. 43p.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1), 35-41.

Delen, D., Walker, G. y Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127.

Dreiseitl, S. y Ohno-Machado, L. (2002). Logistic regresion and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(1), 352-359.

Flowers, A. (13 de mayo de 2018). Dogs and Skin Cancer. Recuperado de https://pets.webmd.com/dogs/dogs-and-skin-cancer#1

Glorot X., Bordes, A. y Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *Proceedings of Machine Learning Research*, *15*(1), 164-172.

Gupta, A., Shreevastava, M. (2011). Medical Diagnosis using Back propagation Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 11(1), 120-132.

Hawkins, D. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1-12.

Hosmer, D. y Lemeshow S. (2000). Applied Logistic Regression Second Edition. New York: Wiley-Interscience.

Lau, H. T. y Al-Jumaily, A. (2009). Automatically early detection of skin cancer: Study based on neural network classification. Soft Computing and Pattern Recognition. *2009 International Conference of Soft Computing and Pattern Recognition*, 375-380.

Lee, T. (2011). Cross Validation Evaluation for Breast Cancer Prediction using Backpropagation Neural Network. *American J. of Engineering and Applied Sciences, 4*(4), 576-585.

Medina, I., Puicón, V. y Sandoval, N. (2017). Frecuencia de Tumores en Piel de Caninos Diagnosticados Histopatológicamente. Facultad de Medicina Veterinaria de la Universidad Mayor de San Marcos. Periodo 1999-2012. Tesis de Médico Veterinario. Lima, Perú. 32 p.

Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Universidad Tecnológica Nacional-Facultad Regional Rosario. Tesis de Ingeniería Química. Rosario, Argentina. 41 p.

Mehta, C. R. y Patel, N. R. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine*, *14*(19), 2143–2160.

Mitchell, T. (1997). Machine Learning. Portland: McGraw-Hill

Moulton, J. E. (1990). Tumors in domestic animals. (3.ª ed.). California: University of California.

Örkcü, H. H. y Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert Systems with Applications. Expert Systems with Applications, 38*(4), 3703-3709.

Sánchez, A. (2012). Análisis comparativo usando Backpropagation y Random Forest con Regresión Logística para la predicción del cáncer de mama. *Machine Learning*, *45*(1), 5-32.

Sibi, P., Allwyn, S. y Siddarth, P. (2013). Analysis of different activation functions using back propagation neural networks. *Journal of Theoretical and Applied Information Technology*, *47*(3), 1265-1268.

Timmerman, D., Van Calster, B., Testa, A. C., Guerriero, S., Fischerova, D., Lissoni, A. A., y Valentin, L. (2010). Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound in Obstetrics and Gynecology*, *36*(2), 26-34.

Torres González-Chávez, M., Peraza González, B., Fabré Rodríguez, Y., Rodríguez Aurrecochea, J. C., Calaña Seoane, L., Márquez Álvarez, M., Zamora Montalvo, Y., Rubio García, J.L., Martín Romero, J. A., y Camacho Socarrás, C. (2015). Frecuencia de presentación de neoplasias en caninos del municipio San Miguel del Padrón, La Habana, Cuba. *Revista de Salud Animal*, *37*(1), 39-46.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11),1225-1231.

Zhou, X., Liu, K.-Y., y Wong, S. T. C. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, *37*(4), 249-259.