Predicción de los parámetros nutricionales de la leche a partir de sus propiedades físico-químicas utilizando deep learning

Prediction of nutritional parameters in milk from its physicochemical properties using deep learning

Eduardo Torres Carranza[1]*, Jimy Oblitas Cruz[1]

[1]Facultad de Ciencias Agrarias, Universidad Nacional de Cajamarca, Av. Atahualpa 1070, C.P. 06003, Cajamarca, Perú

*Autor de correspondencia: etorres@unc.edu.pe

Resumen

El objetivo en el presente trabajo fue encontrar la mejor estructura de una Red Neuronal que permitan predecir los parámetros de calidad fisicoquímica de la leche, tales como la grasa, proteína, lactosa, sólidos no grasos, sólidos totales y minerales a partir de variables de fácil determinación como el tiempo de reducción de azul de metileno, densidad y pH en la empresa Nestle – Cajamarca. Se aplicó una Red Neuronal Artificial (RNA) del tipo Feedforward con los algoritmos de entrenamiento Backpropagation y de ajuste de pesos Levenberg-Marquardt, usando la topología: error meta de 10-2, tasa de aprendizaje de 0.01, coeficiente de momento de 0.5, 3 neuronas de entrada, 6 neuronas de salida y 50 etapas de entrenamiento. Se encontró que la desviación absoluta media (DAM) menor fue de 0.00715952 correspondiente a una Red Neuronal con 2 capas ocultas con 18 y 19 neuronas respectivamente y una función de tipo Tangente sigmoidal hiperbólico (Tansig) y Logaritmo sigmoidal hiperbólico (logsig) siendo su Coeficiente de regresión de 0.99837. Se comparó las predicciones con un modelo de regresión multivariable no lineal y no se encontró diferencias estadísticas ($p > 0.95$) para todas las variables de salida, excepto para la proteína.

Palabras clave: azul de metileno, calidad fisicoquímica de la leche, deep learning, densidad, pH, red neuronal

Abstract

The objective of this study was to determine the optimal structure of a Neural Network that allows for the prediction of physicochemical quality parameters of milk, such as fat, protein, lactose, non-fat solids, total solids, and minerals, based on easily determinable variables like methylene blue reduction time, density, and pH at the Nestle - Cajamarca company. A Feedforward Artificial Neural Network (ANN) was employed, using the Backpropagation training algorithm and the Levenberg-Marquardt weight adjustment algorithm, with the following topology: meta-error of 10-2, learning rate of 0.01, momentum coefficient of 0.5, 3 input neurons, 6 output neurons, and 50 training epochs. It was found that the mean absolute deviation (MAD) was minimized to 0.00715952 in a Neural Network with 2 hidden layers, consisting of 18 and 19 neurons, respectively. The activation functions used were Hyperbolic

Tangent Sigmoid (Tansig) and Hyperbolic Logarithmic Sigmoid (Logsig), resulting in a regression coefficient of 0.99837. Predictions were compared with a nonlinear multivariable regression model, and no statistical differences ($p > 0.95$) were observed for all output variables, except for protein.

Keywords: methylene blue, milk physicochemical quality, deep learning, density, pH, neural network

Introducción

En los métodos tradicionales de control de calidad es necesario hacer diversos análisis, así como utilizar grandes cantidades de muestras para ensayarlos con el consecuente deterioro de éstas (Pegolo et al., 2021). La utilización de un sistema que reduzca esta toma de muestra y el tiempo de determinación de los parámetros de calidad, podría ayudar a mejorar estos sistemas, como a la vez reducir el tiempo y costo de los análisis, dando con ello el pago rápido y oportuno a los ganaderos (Matson et al., 2021).

Un aspecto de gran importancia para los industriales es el valor monetario de los componentes de la leche, en particular de aquellos que más contribuyen a los rendimientos en los productos lácteos. Actualmente, el valor monetario de la leche cruda en la mayoría de los países es aproximadamente equivalente a U.S.$ 0.25 por litro. Se utiliza kilogramos en lugar de litros para medir la cantidad de leche, la cifra es de U.S.$ 0.24 por kilogramo de leche. Para convertir el precio, de dinero por litro a dinero por kilogramo, se divide entre la densidad de la leche, que es del orden de 1.03 kilogramos por litro (Pegolo et al., 2021).

Cajamarca es una cuenca lechera por excelencia, la cual aún no ha implementado un sistema adecuado de calidad de la leche cruda dentro de su cadena alimentaria (Boix et al., 2012), habiéndose identificado como punto crítico para la calidad de los productos lácteos. Los motivos de esta situación son el alto costo de los análisis fisicoquímicos y el desconocimiento de nuevas tecnologías que pueden permitir la estimación de estos parámetros a partir de técnicas sencillas y de bajo costo. Dentro de estas nuevas tecnologías esta la Inteligencia Artificial, la cual a través de herramientas como las Redes Neuronales podrán predecir estos parámetros, a partir de variables que normalmente se vienen evaluando como el tiempo de reducción de azul de metileno, la refractometría, densidad y el pH, con lo que se podrá estimar parámetros como los sólidos totales, por el cual el precio de la leche es estimado (Vasafi et al., 2021). Esto puede ayudar a productores y compradores a estimar estos costos de una manera rápida y sencilla.

Las llamadas Redes Neuronales o modelos conexionistas han ido progresivamente utilizándose como herramientas de predicción y clasificación (Ma et al., 2018). La presente investigación utiliza las Redes Neuronales como herramienta de deep learning para lo cual se planteó el objetivo de determinar la estructura de la Red Neuronal que permita predecir los parámetros de calidad a partir de propiedades fisicoquímicas de la leche de la microcuenca de Cajamarca y comparar la eficiencia de predicción de la red neuronal con un modelo de regresión estadístico.

Materiales y métodos

Material biológico

El material biológico constó de 252 muestras de leche de vacas de raza Holstein (40 ml / muestra) colectadas en doce centros ubicados en la campiña de Cajamarca.

Determinación de los parámetros fisicoquímicos de la leche

Se realizaron en el punto de recojo las pruebas de Tiempo de Reducción de Azul de Metileno (TRAM), densidad y pH y los demás parámetros se analizaron llegando a Planta Nestlé - Cajamarca

Sistema de cómputo

Para llevar a cabo el experimento se utilizó una computadora Harvest, multiprocesador de memoria compartida, perteneciente al Centro de Investigación Científica y Educación Superior de Ensenada, Unidad de Nayarit (CICESE-UT3), México: Procesador = 12 Intel(R) Xeon(R) CPU E5-2603 v3 @ 1.60GHz, CPUcores = 72, Memoria RAM = 64 GB. El software utilizado para la implementación de las secuencias lógicas fue Matlab versión 2015ª.

La data obtenida por cada muestra se dividió en tres valores de entrada; Densidad (Dn), Potencial de óxido reducción (Rd) y Potencial de hidrogeniones (pH). Además, se definieron seis parámetros de salida: Proteínas (Pr), Lactosa (Lc), Sólidos totales (St), Sólidos grasos (Sg), Sólidos no grasos (Sng) y Minerales (Mn). Mayores detalles se muestran en la Tabla 1 y Tabla 2.

Tabla 1. Detalles de los análisis realizados de los parámetros de entrada y salida

	Parámetro	Método	Fuente
Entrada	Densidad	Lactodensímetro (AOAC 925.22)	Scott & Helrich (1990)
	Potencial de óxido reducción	Tiempo de reacción al azul de metileno	Mayorga, Guzmán & Unchupaico (2014)
	Potencial de hidrogeniones	Potencio métrico	Urrego, Londoño, & Rosales (2014)
Salida	Proteínas	Método espectroscópico infrarrojo medio (NTP 202.130:1998)	Uria & Lucas (2003)
	Lactosa		
	Sólidos totales		
	Sólidos grasos		
	Sólidos no grasos		
	Minerales		

Tabla 2. Rangos en los parámetros estructurales

Parámetros	Rango
Neuronas capa de entrada (NE)	3
Neuronas capa de salida (NS)	6
Numero de Capas ocultas (CO)	[1-3]
Neuronas por capa por capa oculta (NCO)	[3 - 27]
Funciones de activación* (FA)	[1-3]

Modelamiento de regresión estadística

Se usó un modelo de regresión multivariable no lineal del tipo: $Y= f(x,\Phi) + \varepsilon$; basado en datos multidimensionales x,y donde, f es alguna función no lineal respecto a algunos parámetros desconocidos Φ. Como mínimo, se pretendió obtener los valores de los parámetros asociados con la mejor curva de ajuste. Se buscó la mejor relación por grupos separados. Para encontrar estas relaciones se usó el paquete estadístico DataFit. La Secuencia para creación, entrenamiento y evaluación de redes se muestra en la Figura 1.

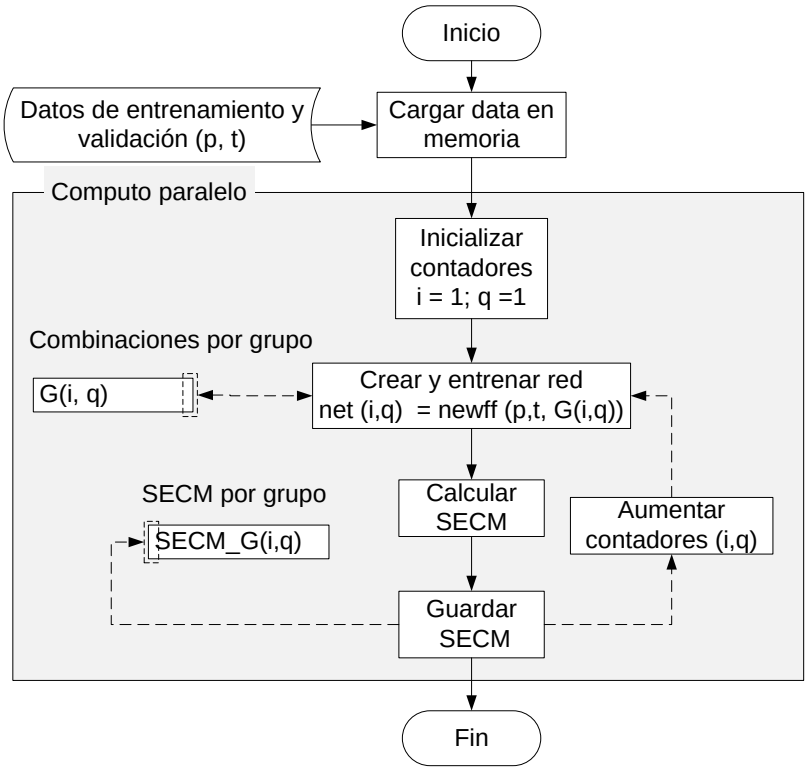


Figura 1. Secuencia para creación, entrenamiento y evaluación de redes

Resultados y discusión

De acuerdo al valor del sesgo y la curtosis estandarizada podemos observar que los valores no siguen distribuciones normales pues cuando estos valores están fuera del intervalo -2 y +2 indican un significativo

incumplimiento con la normalidad estadística (Tabla 3 y 4).

Tabla 3. Resumen estadístico de pruebas simples de calidad de la leche

	Densidad	pH	Reductasa
Recuento	252	252	252
Promedio	1.02826	6.6355	6.72751
Desviación estándar	0.00111889	0.0497268	0.653212
Coeficiente de variación	0.11%	0.75%	9.71%
Mínimo	1.0262	6.55	6
Máximo	1.03	6.79	8
Rango	0.0038	0.24	2
Sesgo estandarizado	-1.17322	2.66201	2.7901
Curtosis estandarizada	-2.65321	-0.321527	-2.28025

Tabla 4. Resumen estadístico de análisis de la leche

	Minerales	Proteína	Lactosa	S. no graso	Sólido graso	Sólidos totales
Recuento	252	252	252	252	252	252
Promedio	0.69873	2.98344	4.84148	8.5236	3.58688	12.1113
Desviación estándar	0.0151754	0.136422	0.192641	0.312609	0.181455	0.436023
Coeficiente de variación	2.17%	4.57%	3.98%	3.67%	5.06%	3.60%
Mínimo	0.5	2.69	4.31	7.73	3.15	10.89
Máximo	0.71	3.28	5.2	9.08	4.1	12.84
Rango	0.21	0.59	0.89	1.35	0.95	1.95
Sesgo estandarizado	-67.6764	-1.24462	-3.15199	-2.50502	-1.5738	-4.25536
Curtosis estandarizada	444.784	-2.01928	0.0529638	-0.624107	-0.25761	-0.216504

La regresión arrojada de la relación entre los datos reales y lo datos predichos por la red entrenada Deep learning logró un índice de correlación de R = 0.99837, lo cual indica un muy buen ajuste entre datos predichos y reales (Figura 2).

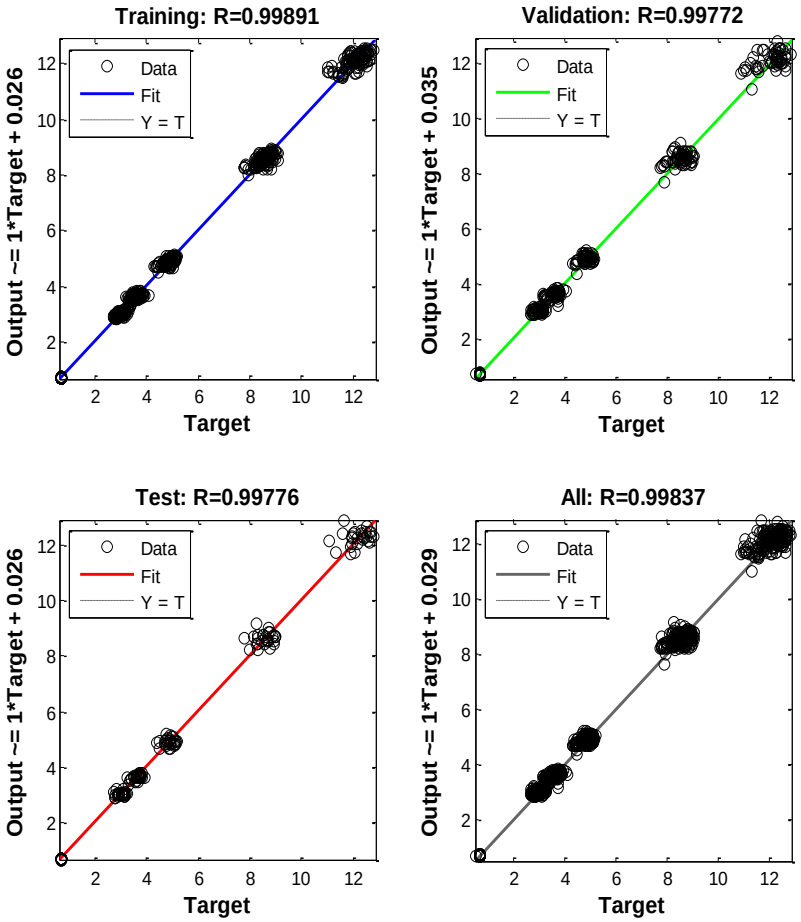
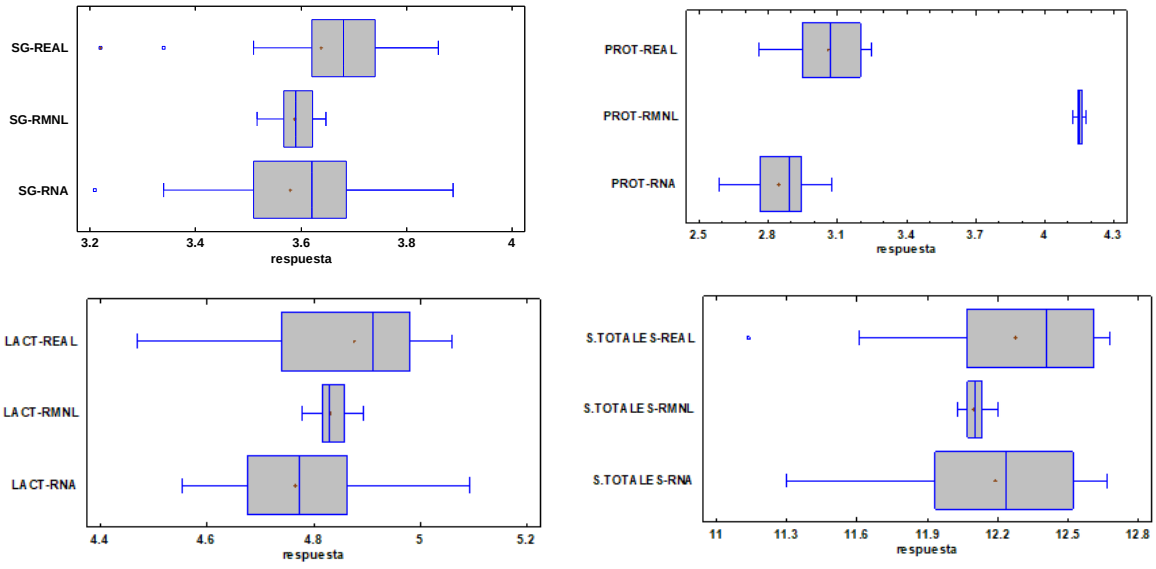


Figura 2. Comparación de datos simulados con red neuronal con todos los datos de entrenamiento reales

La eficiencia de las predicciones usando el modelo de Redes Neuronales se comparó con modelos estadísticos, siendo los modelos utilizados los de Regresión Multivariable No Lineal (RMNL), probando todos los parámetros, los cuales se muestran en la Figura 3.



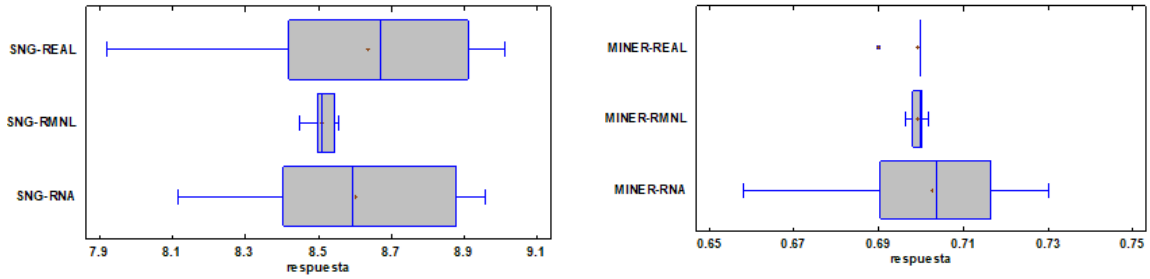


Figura 3. Gráfico de Caja y bigote para las medias de los valores y sus valores atípicos

Las correlaciones lineales para la predicción por RNAy RMNL, mostraron valores de Coeficiente de Correlación, R2, error estándar del estudio y error estándar absoluto medio (Figura 4).

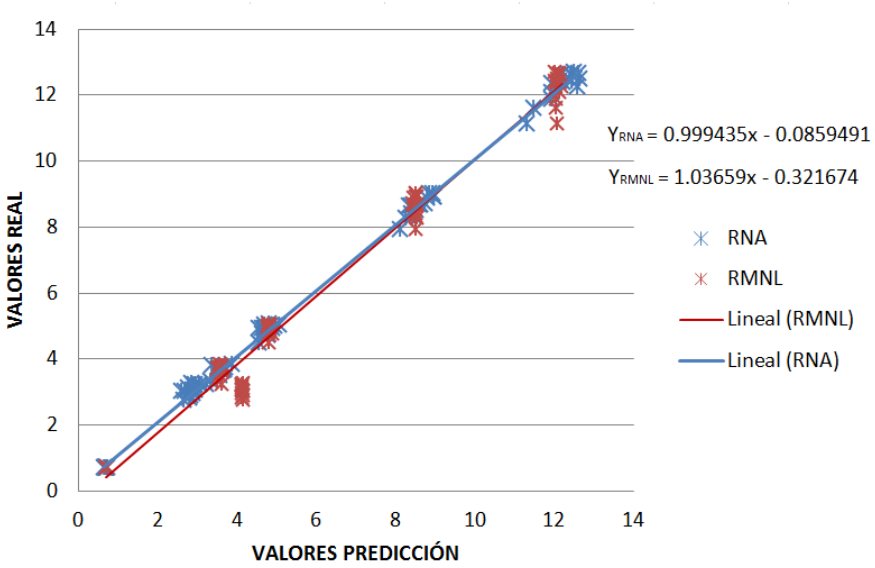


Figura 4. Ajuste lineal entre los valores esperados obtenidos por RMNL y por RNA

Y_{RMNL} = 1.03659x - 0.321674	**Y_{RNA} = 0.999435x - 0.0859491**
Coeficiente de Correlación = 0.992172	Coeficiente de Correlación = 0.999243
R-cuadrada = 98.4404 %	R-cuadrada = 99.8487 %
R-cuadrado (ajustado para g.l.) = 98.4227 %	R-cuadrado (ajustado para g.l.) = 99.847 %
Error estándar del est. = 0.486241	Error estándar del est. = 0.151462
Error absoluto medio = 0.388293	Error absoluto medio = 0.119754

Se observa que el coeficiente de correlación para la regresión con RNA es 0.999243 y su R2 = 99.85% mientras que el coeficiente de correlación para RMNL es 0.992172 y su R2 = 98.44%, comprobando con ello que la RNA presenta una mejor correlación con respecto al RMNL.

Conclusiones

De las 12 Rutas de recojo de leche de la Microcuenca de Cajamarca, mostraron que los valores del pH tienen un promedio de 6.6355, la densidad de 1.02826 g/ml, el tiempo de reducción de azul de metileno de 6.72751 horas, la grasa de 3.58688%, la proteína de 2.98344%, la lactosa de 4.84148, los sólidos no grasos 8.5336%, los sólidos

totales de 12.1113% y los minerales de 0.69873%. Estos datos están dentro de los rangos establecidos por la NTP 202.001 – 2010 para Leche Cruda.

Se determinó que la mejor estructura de la Red Neuronal que permite predecir los parámetros de calidad a partir de propiedades fisicoquímicas es una de tipo Backpropagation con una topología final de 2 capas ocultas con 18 y 19 neuronas respectivamente y con las funciones de transferencia Tangente sigmoidal hiperbólico y logarítmica sigmoidal, con un error promedio final de 0.000256%. Además, la RNA presenta mayores ventajas sobre los modelos estadísticos.

Referencias

Boix, N., Barenys, M., Llobet, J. M., Teixidó, E., Ortiz, P., & Deza, N. (2012). Developmental toxicity of triclabendazole (TCBZ) residues in milk and cheese from Cajamarca, Perú, coming from cattle with high incidence of *Fasciola hepatica*. Reproductive Toxicology, 34(2), 158. https://doi.org/10.1016/j.reprotox.2012.05.047

Ma, W., Fan, J., Li, Q., & Tang, Y. (2018). A raw milk service platform using BP Neural Network and Fuzzy Inference. Information Processing in Agriculture, 5(3), 308-319. https://doi.org/10.1016/j.inpa.2018.04.001

Matson, R. D., King, M. T. M., Duffield, T. F., Santschi, D. E., Orsel, K., Pajor, E. A., Penner, G. B., Mutsvangwa, T., & DeVries, T. J. (2021). Benchmarking of farms with automated milking systems in Canada and associations with milk production and quality. Journal of Dairy Science, 104(7), 7971-7983. https://doi.org/10.3168/jds.2020-20065

Pegolo, S., Giannuzzi, D., Bisutti, V., Tessari, R., Gelain, M. E., Gallo, L., Schiavon, S., Tagliapietra, F., Trevisi, E., Ajmone Marsan, P., Bittante, G., & Cecchinato, A. (2021). Associations between differential somatic cell count and milk yield, quality, and technological characteristics in Holstein cows. Journal of Dairy Science, 104(4), 4822-4836. https://doi.org/10.3168/jds.2020-19084

Vasafi, P. S., Paquet-Durand, O., Brettschneider, K., Hinrichs, J., & Hitzmann, B. (2021). Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. Journal of Food Engineering, 299, 110510. https://doi.org/10.1016/j.jfoodeng.2021.110510